Birmingham City University

DOCTORAL THESIS

Deep Learning-Based Decision Support for Multi-Phase Cancer Management

---

Author:
Abdullah Basaad

Supervisors:
Prof Shadi Basurra
Dr Mohammed Abdelsamea
Dr Edlira Vakaj

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Data Analytics and Artificial Intelligence
School of Computing and Digital Technology

August 20, 2025

Declaration of Authorship

I, Abdullah Basaad, declare that this thesis titled, "Deep Learning-Based Decision Support for Multi-Phase Cancer Management" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Sign:_____

Date:_____

Abstract

Early and accurate prediction of cancer is important for improving patient outcomes and making appropriate treatment decisions. This research will explore how deep learning technologies can assist clinicians in treating multi-stage cancers, specifically lung and breast cancer. The first phase is the early-stage detection of lung cancer through diagnostic imaging and generating reports. Diligent follow-up care of patients, while very important, is also a big part of ongoing patient care through routine checkups and monitoring tools such as mammograms, MRIs, and tumour markers, which fall outside of the scope of this research. The second phase concerns breast cancer recurrence; breast cancer often returns as local or regional recurrence during follow-up with the intent that it can be treated curatively. The third phase considers metastasis, which often manifests later and is a more serious event. In some cases, particularly with aggressive cancers or advanced stage cancers, metastasis may be the first sign of recurrence, but typically it comes following local recurrence.

First we consider the possibility of using convolutional neural networks (CNNs) to identify lung cancer in chest X-ray images early in the disease process to avoid treatment delays and improve the odds of survival. The CNNs are used for generating medical reports, but we also want to incorporate data from clinical patient assessments and observations. A secondary focus of this study is to develop predictive models which estimate the statistical likelihood of relapse in breast cancer to assist with treatment planning and diagnostics. Third, we want to outline a deep learning strategy for predicting the risk of metastasis in breast cancer patients, to alert clinicians to the risk of progression of the disease, which could have implications for the clinician's treatment decision. Overall our analysis suggested that deep learning has merit as a game-changer for early detection and the long-term management of our patients with complex cancers. In this thesis, three novel methods are presented to improve the decision supporting for multi-phase cancer management.

Initially, AIM-X Attention-Infused Multimodal Cross-Interaction for X-ray Clinical Report Generation is an innovative AI-powered diagnosis system which combines advances in CAD systems, including multimodal imaging analysis and radiology report generation, to enhance clinical decision-making. The system combines an attention mechanism and multiscale feature extraction with hyperparameter optimisation using a genetic algorithm, thus refining the diagnostic accuracy. It is designed to incorporate radiological images in addition to medical reports in a textual format achieving an accuracy up to 94.5%.

Secondly, GraphX-Net, a GNN based on Shapley Value, is proposed for predict-

ing cancer recurrence by modelling patients as nodes in a graph and using a set of clinical factors such as tumour cellularity and hormone therapy. With the usage of graph convolutional layers in conjunction with Shapley values, this model efficiently evaluates patient attributes and patient-to-patient relations, demonstrating best performances in recurrence prediction with 98% accuracy and 98.4% F1-score.

Third, the BG-MBC BERT-GNN approach for the prediction of metastatic breast cancer using histopathological reports, which combines LLM and GNN, is formulated for the diagnosis of MBC. This is a blend of natural language processing and graph learning paradigms, in which semantic understanding is extracted using BERT embeddings of pathology reports and understanding of patient relations is enabled using GNN guided by attention scores from the embeddings. The model achieves an accuracy of 98% and 99% accuracy in cross-validation, demonstrating strong predictive performance.

These collectively contribute to a developing field of AI-based medical insight that allows robust solutions toward disease detection in early stages of lung cancer and enhanced breast cancer recurrence and metastatic prediction.

Acknowledgment

It is dedicated to my mother, whose constant prayers and supplications smoothed the way and made my goals easier to reach. Indeed, she has always been an inspiration and a pillar of strength in my life whenever I wanted to give up. She never ceased extending her spiritual and emotional support to me.

I express my sincere gratitude to Antser Ltd and Birmingham City University for sponsoring my Ph.D. course.

I am also deeply indebted to my principal supervisor, Prof. Shadi Basurra, for encouraging me to conduct this research and for his support.

My sincere thanks go to my co-supervisor, Dr. Mohammed Abdelasamea, for guiding me through the research topic and for his valuable advice and suggestions during the research process. I also wish to thank my third supervisor, Dr. Edlira Vakaj, for her feedback and suggestions on the emerging topics we considered for study.

I would like to extend my acknowledgement to Prof. Peter Larkham, Director of Research, with whom I started the Postgraduate Certificate in Research Practice course at the start of the Ph.D. course, for his research training and academic guidance to develop my academic research skills.

Ms. Sue Witton, Doctoral Research College, for her administrative and technical assistance for this research and for being always helpful.

I sincerely thank the panel members, Dr. Issa Shuhra, Dr. Ogerta Elezaj, and the panel chair, Dr. Faisal Saeed, for their invaluable guidance, support, and contributions throughout my Ph.D. journey.

I would also like to thank the coauthors of my published research for their valuable pieces of advice, help, and constructive comments. Indeed, they contributed much to this Ph.D. advancement.

In conclusion, I would like to express my deep gratitude to my dear family for their continued support, their assistance in overcoming challenges, and their constant presence beside me without a single word of complaint.

Contents

List of Figures

List of Tables

List of Publications

This section includes a representative selection of peer-reviewed publications that advance the progress of artificial intelligence methods in medical imaging and medical diagnostics. The articles were selected based on demonstrating a graph-based deep-learning approach to improve clinical decision support (i.e., automated report-generation, cancer prediction, model interpretability) and focused on challenges within the healthcare context in the next chapter or what challenges we went through in Section 2.7.

Table 1: List of Publications with Summaries

Publication	**Summary**
Basaad, A., Basurra, S., Vakaj, E., Fu, Z., Rowlands, S., Abdelsamea, M.M. (2025). *Procedia Computer Science - Elsevier*	Proposed AIM-X: Attention-Infused Multimodal Cross-Interaction for X-ray Clinical Report Generation. The model integrates image and text features using cross-attention mechanisms to generate accurate and clinically relevant reports.
Basaad, A., Basurra, S., Vakaj, E., Aleskandarany, M., Abdelsamea, M.M. (2024). *IEEE Access*	Developed GraphX-Net: a GNN framework incorporating Shapley values to enhance interpretability in breast cancer prediction. The model outperformed standard GNN baselines in predictive accuracy and provided insights into key contributing features.
Basaad, A., Basurra, S., Vakaj, E., Eldaly, A.K., Abdelsamea, M.M. (2024). *Diagnostics, 14(13), p.1365*	Presented a BERT-GNN hybrid model for predicting metastatic breast cancer based on histopathology reports. The integration of language models with graph-based reasoning showed improved predictive performance and interpretability.

List of Abrreviations

Table 2: List of Abrreviations.

AI	**A**rtificial **I**ntelligence
MBC	**M**etastatic **B**reast **C**ancer
RBC	**R**elapse **B**reast **C**ancer
CNNs	**C**onvolutional **N**eural **N**etworks
ReLU	**Re**ctified **L**inear **U**nit
RoI	**R**egion **o**f **I**nterest
RNNs	**R**ecurrent **N**eural **N**etworks
PCA	**P**rincipal **C**omponent **A**nalysis
GAs	**G**enetic **A**lgorithms
NLP	**N**atural **L**anguage **P**rocessing
DL	**D**eep **L**earning
BERT	**B**bidirectional **E**ncoder **R**epresentations from **T**ransformers
GPT	**G**enerative **P**re-Trained **T**ransformers
XAI	E**x**plainable **A**rtificial **I**ntelligence
CAD	**C**omputer **A**ided **D**etection
NLP	**N**atural **L**anguage **P**rocessing
LIME	**L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations
MHA	**M**ulti **H**ead **A**ttention
ReLU	**Re**ctified **L**inear **U**nit
NER	**N**amed **E**ntity **R**ecognition
MSE	**M**ean **S**quare **E**rror
BLEU	**B**i**L**ngual **E**valuation **U**nderstudy Score
ROUGE	**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation
CIDEr	**C**onsensus-based **I**mage **D**escription **E**valuation
NLU	**N**atural **L**anguage **U**nderstanding
NLG	**N**atural **L**anguage **G**eneration

Continued on next page

Table 2 – *Continued: List of Abrreviations*

NSP	**N**ext **S**entence **P**rediction
BIACH & RI	IT department of **B**asavatarakam **I**ndo-**A**merican **C**ancer **H**ospital and **R**esearch **I**nstitute
METABRIC	**M**olecular **T**axonomy of **BR**east Cancer **I**nternational **C**onsortium
LSTM	**L**ong **S**hort **T**erm **M**emory
GA	**G**enetic **A**lgorithm
IU	**I**ndiana **U**niversity
XAI	**EX**plainable **A**rtificial **I**ntelligence
LIME	**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations
NCCN	**N**ational **C**omprehensive **C**ancer **N**etwork
COPD	**C**hronic **O**bstructive **P**ulmonary **D**isease

Chapter 1

Introduction

1.1 Overview

Lung and breast cancers are both common and deadly tumours worldwide as shown in Fig. 1.1, breast cancer is the most diagnosed cancer in women, and lung cancer has the most deaths from cancer among both sexes. There are published studies that suggest that these tumours, although originating in different tissues, may have shared biology and clinical relevance. For example, lung metastases are a common site for distant recurrence in breast cancer patients especially those with advanced stages (Mattiuzzi et al., 2019; Organization, 2023b). Approximately 60% of metastatic breast cancer cases involve lung or bone metastasis. Lung metastasis is the second most common site for breast cancer to spread, after bone. Specifically, triple-negative breast cancer is known to be more prone to metastasising to the lungs (Jin et al., 2018). A systematic review and meta-analysis (Zhao et al., 2022) reveals that patients with chronic obstructive pulmonary disease (COPD) have a substantially heightened risk for lung cancer as compared with patients without COPD. This study reinforces the need for heightened clinical awareness, prevention, and also addressing lung cancer risk among patients with COPD. Data were required from 13 countries, UK included. The overall pooled prevalence of lung cancer among COPD patients was 5.08% (95% Confidence Interval [CI]: 4.17%–6.00%). In Appendix 7.2, three different statistcis: (1) Analysis of the risk factors of lung cancer in COPD. (2) Subgroup analysis of the prevalence of lung cancer in COPD. (3) Basic characteristics of the included studies. This meta-analysis illustrates that patients with COPD have an increased risk of lung cancer development as compared to individuals without COPD.

Lung cancer is often asymptomatic at early stages of disease, with presented symptoms leading to a lung cancer diagnosis often being reactive and not proactive. In general, many thoracic abnormalities, such as pulmonary nodules, infiltrates,

fibrosis, and consolidation may be early indicators that a patient is developing lung cancer. If identified too late to address and identify these abnormalities based on guidelines and protocols will miss early warning signs of potentially developing lung cancer. As such, chest x-ray imaging is valuable when detecting possible thoracic abnormalities as such imaging modalities are widely available and cost-effective instrument in the clinical setting (Rajpurkar et al., 2017). While it is true that x-ray imaging has its limitations compared to more sophisticated imaging modalities, it is a relatively quicker, non-invasive imaging approach that will expose patients to lower radiation doses and is suitable for use for mass screening. According to Taiwan Health Insurance Research Database population-based performs a retrospective cohort study. The database includes more than 99% of the population of Taiwan. They established a 27,405-person pneumothorax cohort and a 109,620 person comparison cohort with data from 2000 to 2009 to evaluate the relationship between spontaneous pneumothorax and lung cancer (Chen et al., 2020a).

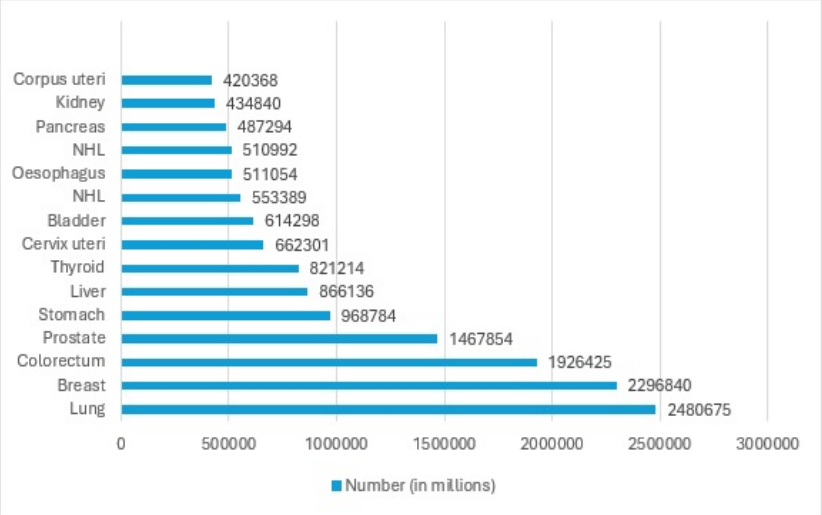


Figure 1.1: Absolute number of Cancer Incidence in both sexes in 2022 World

Chest X-ray serves as an important aspect in the early diagnosis because it can greatly limit the chance of developing the disease since it can allow for timely interventions before the cancer reaches an advanced stage. Early diagnosis can decrease the prevalence of long-term effects and increase the survival of high risk select groups especially in under-resourced settings. Additionally, early diagnosis decreases the utilisation of costly modalities like CT, MRI, or PET. These are costly to the health care system, yet they are more burdensome to radiologists and technologists in terms of increased use of time and manpower (McKee et al., 2016).

The integration of artificial intelligence (AI) early disease detecting models is not only positively poised to help save patients from potentially life-threatening medical conditions (i.e. facilitating fast-tracked treatment with many forms of lung

cancer) but providing the reduction of burden and therefore costs of consumption of resources in the healthcare system (Maleki Varnosfaderani et al., 2024). Using x-ray imaging as scalable tool combined research in deep learning and other algorithmic machine learning provides an example of a powerful strategy for managing thoracic disease before it becomes chronic or life-threatening (like lung cancer).

Breast cancer is the most common cancer in women globally, with early detection significantly improving outcomes. The survival rate in a patient with a diagnosis of breast cancer in an initial phase is around 99%. But in advanced phase breast cancer, the rate is drastically low. The influence is also found in countries with systematic breast cancer screening programmes, where the death rate has reduced quite substantially by a whopping 20-30% (Organization, 2023a). Projects like the World Health Organisation's Global Breast Cancer Initiative aim to reduce the world's mortality rate from breast cancer by 2.5% every year, which may save 2.5 million lives by 2040 (Organization, 2023a).

This research focuses on lung and breast cancers, structured around a three-phase pipeline that addresses the following key stages:

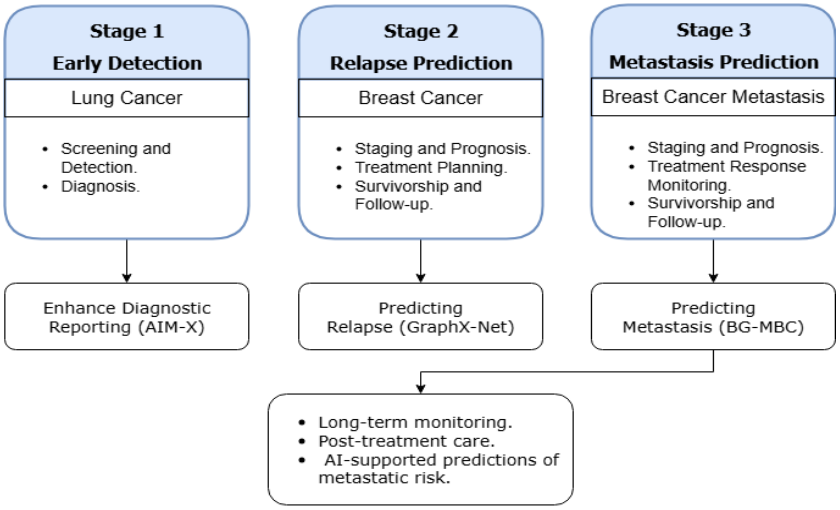


Figure 1.2: AI driven cancer management pipeline

1.1.1 Pahse 1

The journey of our research as seen in Fig. 1.2 begins with the early detection of thoracic diseases. Early detection of thoracic diseases is an important factor that improves overall mortality and long-term outcomes. In this phase time is critical. A delay to diagnosis often results in a lost opportunity for curative treatment. Lung cancer is perhaps the most notable thoracic illness in this context due to its si-

lent, rapid progression, with many patients often experiencing active small-cell or non-small cell lung cancers without noticeable symptoms early in the disease or developing otherwise subtle anomalies in perceived chest imaging (Gomez et al., 2015). Based on focusing on the earliest phase of a long list of possible pathways, we introduce the AIM-X model in Chapter 3, which is a multimodality deep learning network that generates coherent reports through the combination of visual and textual data. This architecture has innovations in multi-scale feature extraction as well as bi-directional cross-attention mechanisms to elegantly align images with texts. In contrast to feature-independence-based approaches, our method leverages a pre-trained DenseNet to enable hierarchical feature extraction of visual data. A fusion layer combines such features into one single space, while cross-attention improves accuracy through focusing on prominent inputs.

1.1.2 Pahse 2

As we moved onto stage two of our research, we turned towards those who have completed treatment for breast cancer – that is, survivors who are also at risk of recurrence or returning disease. We had a model to predict the risk of recurrence, the METABRIC data set had been used as it is a large cohort of invasive breast cancer with long follow-up and complete clinicopathological data, which includes the molecular profiling methodology in a cohort of 1980 breast cancer (Curtis et al., 2012); include clinical records, demographic data of patients, tumour characteristics, as well as treatment and follow-up data. We proposed an innovative methodology based on GNN and GraphX-Net was presented in Chapter 4, which made use of Shapley values to rank features and build an explainable graph representation by using a series of XAI techniques to create transparency and interpretability in predicting relapsed breast cancer. It used the strengths of GNNs that had the capability to deal with complex relations and non-dependencies of data structured as a graph very well and was shown to be one of the best performers for recurrence prediction with a 98% AUC score. It is possible to use this prediction model to engage in more tailored follow-up and treatment planning, so we expand on the possibility of detecting the disease before a round of other significant treatments is needed.

1.1.3 Pahse 3

In the third and final stage, we tackle one of the most severe consequences of cancer diagnosis and treatment: metastasis or spread of the disease. Metastasis from breast cancer primarily occurs in the liver, lungs, bones, and brain, and signals significant decline in patient prognosis (Yousefi et al., 2018). Our metastatic prediction BG-MCB model in Chapter 5 uses data was acquired from the IT department of

Basavatarakam Indo-American Cancer Hospital and Research Institute (BIACH & RI). It is a semi-structured Excel file comprising two main columns: patient identifiers medical record number and histopathology reports. The dataset consists of 25,652 entries in .csv format (raw text). The histopathology column includes a variety of pathological observations, such as clinical details, specimen descriptions, microscopic observations, impressions, and gross fndings (Botlagunta et al., 2023). The histopathology diagnoses pertain to metastasis, biopsy, and lymph node reports. This research focuses on detecting metastasis in patients among the different types of diagnoses. We introduce BERT Graph for Metastatic Breast Cancer (BG-MBC) to better analyse risk of distant spread of disease from breast cancer. It is an innovative integrated framework for natural language processing (NLP) and graph learning that can generate meaningful representations to support the understanding of unstructured clinical text data, for example, free-text histopathology reports, and other medical documents. It shows a promosing outcomes with a high performance accuracy 98% and cross-validation 99%.

1.2 Problem Statement

Lung and breast cancers are considerably the most common and deadly cancers around the world. Although imaging and diagnostic techniques have improved dramatically with advancements in technology, the healthcare system continues to struggle to provide optimal cancer care throughout the cancer continuum, especially early detection, timely treatment and management of the disease over the long term (Pulumati et al., 2023). Recognisably, small nodules and other subtle opacities are often overlooked as initial findings of thoracic disease in chest X-rays as they can not always be detected by the naked eye, have a non-specific appearance compared to the overwhelming number of normal cases seen every day in clinical practice. These (often) undetected findings can lead to delayed lung cancer diagnoses, and therefore the opportunity for early evidence-based treatment is lost, when most likely to have positive treatment outcomes (Aiello et al., 2021).

On the other hand, breast cancer patients are at high risk of relapse or metastatic progression after successful treatment, which can be difficult to predict with conventional methods[1], especially after primary treatment. Follow-up often did not take into account individual-level risk, which creates uncertainty for both patients and clinicians in monitoring follow-up effectively (De Rose et al., 2022).

[1]**Conventional Methods** are referring to the standard clinical approaches currently used to assess the risk of breast cancer relapse or metastasis, such as, Manual review, Basic statistical models, Follow-up schedules, etc.

These diagnostic and prognostic challenges add to challenges of care (elevated costs) based on increased care workload for clinicians, and they contribute delay and decision support in treatment in busy healthcare system like that of the NHS. We think there is a need for an integrated and intelligent system that can support clinicians at different stages along the pathway of the cancer journey starting from first suspicious findings to long-term survivorship. With these unmet needs in mind, the following research provides effective solutions towards addressing the unmet needs in cancer management by outlining a three-phase pipeline as seen in Fig. 1.2 that utilises AI to deliver the following:

1. Improve diagnostic reporting for lung cancer using multimodal models, taking advantage of chest X-rays and radiology reports.

2. Predict relapse in breast cancer through a review of the patient histories and pathology data.

3. Predict metastatic progression in breast cancer through an investigation of histopathology features.

By providing early, accurate and interpretable evidence at each critical stage, this system will aim to improve accuracy of diagnosis, reduce unnecessary intervention, enhance multidisciplinary treatment planning, and patient-centered approaches to care through optimal efficacies in the management of cancer.

1.3 Research Questions

As outlined in the problem statement Section 1.2, Fig. 1.2 illustrates the key phases of this research, which aims to exploit deep learning developments to improve digital healthcare diagnosis and prediction. The research progresses through three critical stages. At every key phase, we have a challenge that we provided in the form of a research question. Solving these questions should help achieve the research aim. This research should focus on implementing AI models to improve disease detection and prediction.

Develop models that provide transparent explanations for their predictions, helping clinicians better understand the reasoning behind the model's decisions and increasing its trustworthiness. Using different techniques to measure this objective such as SHAP and LIME to calculate fidelity and sparsity scores. Another measure that can be used is the local and global explanations technique, this approach can benefit from evaluating the consistency between the local explanations for individual predictions and global explanations (overall model behaviour) to ensure interoperability at different levels. Moreover, employing the attention mechanisms improves

the interpretability of models due to emphasis on what and how the model attends to its input, although the challenges that we have faced and explained in Chapter 2, Section 2.7. The research seeks to explore the following aspects:

1. *How might combining different sources of medical information lead to earlier and more accurate health decisions for patients?*

   To address this question, I will begin with a technical explanation, which will then help clarify the concept more effectively. Iterative refinement of attention in multimodal systems denotes a process referred to as "attention", its a method for modelling interactions between many types of data (text, images, audio, etc.) that applies to multiple steps or layers of an object state. When attention is applied repeatedly, each layer should build the object state representation progressively better cross-model representations.

   Through an iterative enhancement of attention in multimodal systems, we can improve learning of shared representations over time, and ultimately build value into multimodal communication by improving cross-modal interactions and feature alignments by optimising the correlations step by step to target possible information loss/gaps in semantics (e.g., across text, image, etc.). Where the first iterative enhancements consider grained features, later iterations refine discriminative features, while filtering out distracting signals.

   Through our experimentation, we have shown that through the development of the AIM-X model in Section 3.3. It is a framework we designed specifically to work with multimodal information, including both imagery and text, to generate accurate reports. It employs an iterative attention mechanism to progressively align and refine cross-modal relationships, enhancing inter-modal dependency learning. It introduces an advanced multimodal machine learning framework that integrates state-of-the-art deep learning methodologies to optimise predictive performance. The framework's evaluated through stringent performance metrics including data coherence, precision, recall, and F1-score. AIM-X synergises transformer architectures, CNNs, RNNs, and attention mechanisms. Tailored to address domain specific challenges, as discussed in Section 2.7, this system prioritises scalability while maximising detection accuracy. It utilised an advanced machine learning method, which is a genetic algorithm, to train and refine the model. The genetic algorithm helps boost the accuracy of detecting early signs to identify specific diseases. This can be achieved by assessing precision and recall for detecting early signs of diseases by evaluating the feature relevance using metrics such as feature importance scores and attention weights. The metrics help identify the salient features of the dataset and inform improvements in the model design and decision-

making strategies. These will be augmented by a comprehensive examination of feature importance through methods such as feature importance scores and attention weights. These tools will help to determine the most important features (medical information) in the data, for example, imaging patterns that are implicated in making precise predictions.

2. *How can smart computer systems use health information to find connections between patients and improve our understanding of healthcare?*

   Harnessing AI, with a focus on GNNs, provides a powerful approach for capturing and analysing the intricate links across various clinical data features. In our research, GNN was leveraged to capture the connections between clinical features, where edges between nodes are formed based on their corresponding significance scores. Such significance scores are obtained using machine learning models incorporating a post-hoc explainability method that calculates the contribution of clinical features. The obtained significance scores are then used to create and reinforce edges in the graph representation.

   To enhance interpretability, we investigate the effect of every input feature on individual instances and thus identify positive and negative contributions. From these contributions, we calculate the probability of each node being classified as class 0 ("No") or class 1 ("Yes"). We also measure the similarity between neighboring nodes with respect to their corresponding contributions. Here, a node is acted as a patient its grouped into a category if its contribution is close to the average contribution of the category based on a user-defined threshold parameter. This threshold parameter is a measure for controlling the grouping and connectivity of nodes. This approach that has been applied in our research, please refer to Section 4.1.2.6 for more information.

   To determine whether to add another node to a formed group, we compare the difference between the group's average contribution and the potential new node's average contribution. If the difference lies within the specified limit while maintaining the group size within its limit, the new node will be added to the group. This process ensures that node (patient) connections capture important relationships based on significance scores computed from explainability attributes.

   How can incorporating language-based insights improve a model's understanding of context in its predictions?

3. *How might understanding language help technology better interpret patient information and make better decisions?*

The importance scores extracted from language model embeddings enhance contextual awareness in predictions by capturing meaningful relationships within histopathology reports. Our approach utilises bidirectional encoder representations from LLM transformers to extract attentions and embeddings, which play distinct roles in structuring a graph representation of the data.

- Attentions define graph edges, modelling relationships between medical entities based on numerical references.
- Embeddings serve as node features, encapsulating the rich contextual information from the reports.

This graph representation is then processed by the BERT Graph Metastatic Breast Cancer (BG-MBC) model, which is specifically designed to handle graph-structured data. By leveraging both attentions and embeddings as inputs, the model effectively integrates complex dependencies between clinical attributes. This enhances the model's ability to contextually understand patient data, ultimately improving the accuracy of breast cancer metastasis prediction as illustrated in Section 5.3.

Extensive research on breast cancer from various parts of the world has also yielded varied results, will be discussed in the next chapter, and brought into focus the inherent difficulties in coming up with clear-cut remedies. Yet, despite some progress, a continuing need exists for more sophisticated approaches to improve diagnosis, treatment, and patient outcomes.

We aimed to capitalise on state-of-the-art techniques and innovations in this field, with the view to enhancing detection and prediction accuracy and offering more finer diagnostic modalities. The dire need for new contributions that can offer significant improvement in the quality of life for patients and reduction in mortality rates is our driving motivation.

Early detection is an important determinant of survival rates, we try in our research to devise a diagnostic platform that will equip the medical practitioner with the various tools necessary for identifying the disease in its most incipient stages. In so doing, we also hope to be able to make a small academic contribution towards the worldwide effort at not just saving lives but also reducing both the emotional and physical burden that the disease places on patients and their families.

1.4 Aim and Objectives

This research aims to develop deep learning models for early detection, diagnosis, and clinical decision making at various stages of cancer progression, with an emphasis on lung and breast cancer, across multisource data types, including X-ray scanned images, pathology reports, and clinical attributes. To achieve this, the following objectives have been considered:

1. To improve early-stage screening and reporting of thoracic abnormalities and potential lung cancer by developing an AI model that utilises X-ray imaging and radiology reports.

2. To design and evaluate a predictive model for breast cancer relapse, using structured and unstructured patient data to support post-treatment monitoring and timely intervention.

3. To develop a metastasis prediction framework for breast cancer patients, enabling personalised follow-up strategies and improved survivorship planning.

4. To assess the clinical impact, model generalisability, and cost-effectiveness of the proposed pipeline in terms of improving early detection, reducing unnecessary investigations, and supporting healthcare resource management.

1.5 Research Contributions

The contribution to achieving the aim outlined in the aforementioned section through the introduction of innovative approaches for addressing the main challenges while analysing patient data and understanding the contribution of different inputs in a neural network framework is the basis of this research.

The combination of visual and textual data is one of the most challenging aspects of generating medical reports using artificial intelligence. Although progress has been reported in unimodal analysis of images and texts, their combination is complex and not as well-studied. We introduce the AIM-X model in Chapter 3, which is a neural network that generates coherent reports through the combination of visual and textual data. This architecture has innovations in multi-scale feature extraction as well as cross-attention mechanisms to elegantly align images with texts. In contrast to feature-independence-based approaches, our method leverages a pre-trained DenseNet to enable hierarchical feature extraction of visual data. A fusion layer combines such features into one single space, while cross-attention improves accuracy through focusing on prominent inputs (Li et al., 2024a).

The cross-attention mechanism combines visual and textual features into a unified representation that is then processed through the LSTM decoder to produce the final report. Text embedding and positional encoding help maintain the input sequence and also enhance the output quality. This model produces diagnostic reports using images and medical histories and thus supports applications like image captioning, document summarisation, and multi-modal content generation (Guo et al., 2022).

Besides that, the thesis performs a genetic algorithm for the optimisation of hyper-parameters such as learning rates, batch size, and number of layers to ensure the very optimum performance as depicted in Fig. 1.3. These experimental results, explained in Section 3.3, prove much better improvements for three real-world CAD performances and support enhancing the accuracy of decision-making processes in clinical environments (Mathews et al., 2020).

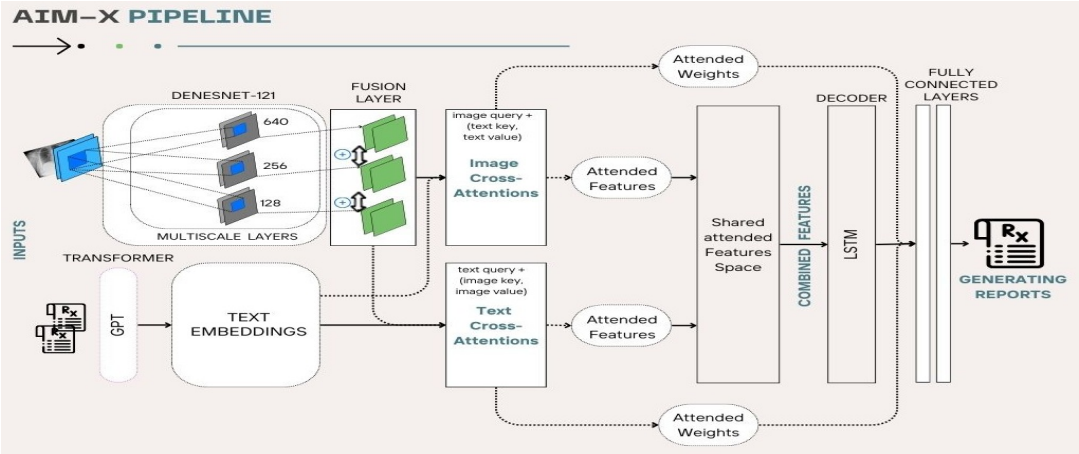


Figure 1.3: AIM-X utilises DenseNet to extract high-fidelity X-ray features and complex patterns important for diagnoses. The visual features are then blended with textual information and processed through GPT to produce rich contextual, coherent medical reports. By fusing visual and textual modalities through an attention mechanism, AIM-X generates accurate and readable diagnostic reports, effectively connecting image features with contextual information for enhanced medical diagnostics.

The integration of these sophisticated techniques and methodologies has given way to highly efficient models, which hold immense promise in the domains of early-detection and diagnose lung cancer. This model is good at the extraction of regions of interest from medical images to create correct and contextually appropriate medical summaries. This framework is based on a high-performance architecture, including multi-modal transformers, which can process multiple types of data in parallel, improving precision and diagnostic accuracy.

This research also intends to present GraphX-Net in Chapter 4, a fresh framework which proposes GNNs for enhancing interpretability and explainability in relapse prediction among breast cancer patients. Therefore, SHAP value importance in GraphX-Net enables an in-depth assessment of the contribution of various features that will facilitate an even more complete explainability of the graph representations of patient data. LIME algorithm is one more explainable approach that has been utilised to quantify the different attributes that contribute locally, for more information please refer to Section 4.2.

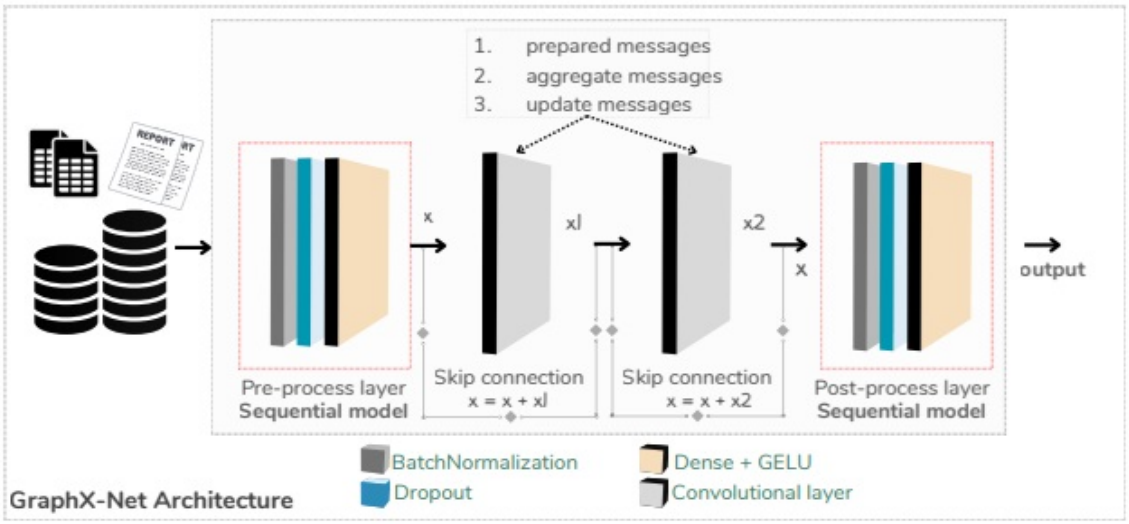


Figure 1.4: GraphX-Net Architecture

This is further supported, as the incorporation of explainable AI techniques hence allowed this model to cater not just for high accuracy in the predictions of relapse but also to give transparent insights into the decision-making process. This will be particularly important in medical contexts where understanding the reason for the predictions is really important. Adding XGBClassifier, which is interpretable, enriches the model through both local and global feature importance insights, while keeping the global explanations at the overall dataset level for personalised explanations for every patient (Delaunay, 2023). In this model, the Shapley values define the contribution of each feature in predicting the relapse. The weighted approach increases the classification for unlabelled nodes and develops a link between data points.

Leveraged LLMs, such as BERT, to extract meaningful insights from the textual data contained in histopathology reports (Alaparthi et al., 2020) as illustrated in Fig. 1.5. BG-MBC is a novel hybrid framework, illustrated in Chapter 5, that integrates GNNs and bidirectional BERT for the analysis of histopathology reports for the purpose of metastatic breast cancer (MBC) prediction. These reports are

a vital source of standardised but complex information related to tissue samples and thus are pivotal for precision in the diagnostic process. However, extracting actionable knowledge from them requires advanced machine learning mechanisms capable of extracting structural relationships within the data as well as semantic nuances within the textual information.

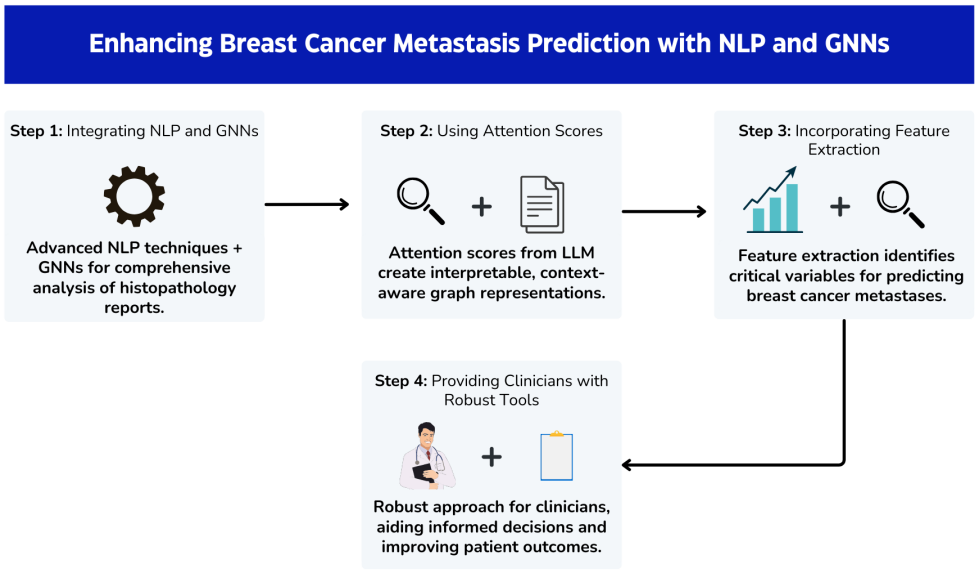


Figure 1.5: Enhancing breast cancer metastasis prediction with NLP and GNNs

The BG-MBC method addresses this issue through leveraging the complementary merits offered by GNNs and LLMs. GNNs are especially suited for modeling complex relationships and interdependencies present in structured data, as represented through the hierarchical and relational structure of histopathology reports, while BERT is capable of realising deep contextual meanings within unstructured textual content. Combining the two approaches makes BG-MBC capable of achieving better insight into the data and thus higher accuracy for the prediction of MBC.

1.6 Thesis Outline

The thesis organisation, illustrated in Fig. 1.6, comprises six chapters systematically structured to present the research methodology and findings as follows:

Chapter 1: Presents the introductory study, hence it forms the base on which the research is based. It encompasses an introduction of thoracic diseases and lung, breast cancers, its significance, its burden, and difficulties that accompany diagnosing and managing it. Further, this chapter identifies a problem statement and elaborates clearly on what is missing in the present knowledge and practices this

research shall try to cover. In addition to the problem description, aims and objectives, research questions and summary of the publications are included. Finally, it concludes by explaining the structure of this thesis.

Chapter 2: Covers the basic background and importance of major approaches and techniques that have been put to work in order to achieve the results of our undertaking. In other words, it informs in detail about varied methodologies that have been deployed with regard to how they were chosen, explained in the following chapters, and put together for smooth integration in an effort toward developing frameworks for reaching the results of the project. This chapter also considered some specific challenges faced during the whole process of the project and gave discussion to the strategies applied to cross such barriers. By highlighting these issues with their respective solutions, the chapter hopes to be a key reference for fellow researchers by offering effective counsel and offering deeper understanding for methods applied in similar contexts.

Chapter 3: Clinical Report Generation and Interpretation. Demonstrates AIM-X[2], an iterative refinement multimodal model with an added mechanism for attention, capable of generating coherent and thorough diagnostic reports from X-ray images in combination with supporting textual information. AIM-X utilises sophisticated techniques for extraction and abstraction of important medical observations, through an attention mechanism, focusing most relevant areas in the images, enhancing accuracy in diagnostics. Notable in the model is its layer for explaining its processes, providing a visualisation of attention weights, enhancing transparency in medical decision processes. Quantitative analysis of performance in model testing represents a high agreement between generated and real reports. The implications of such an observation present strong indications for the use of AIM-X in computer-aided medical report generation, positioning it as a useful tool for medical decision-support systems.

Chapter 4: GNN-Based Shapley Values Interpretation for Relapse Breast Cancer. It goes into great detail about the node classification graph in predicting recurrence among breast cancer patients, a very important aspect of cancer research. Relapse after treatment is one of the most important causes of morbidity and mortality. These clinical variables and histopathological features can be modelled as graphs of complex relationships; therefore, machine learning algorithms can

[2]Basaad, A., Basurra, S., Vakaj, E., Fu, Z., Rowlands, S. and Abdelsamea, M.M., 2025. *AIM-X: Attention-Infused Multimodal Cross-Interaction for X-ray Clinical Report Generation.* Procedia Computer Science, 264, pp.201-210.

learn patterns of these that are predictive of relapse. This chapter underlines the need for better prediction through integration between multi-omics data, radiomics information. Then, it introduces a new concept in the GNN framework, GraphX-Net[3] (Basaad et al., 2024b), which integrates Shapley values for explainable feature relevance and much improved interpretability.

Chapter 5: BERT-GNN Interpretation for Metastatic Breast Cancer. This chapter presents the approach for metastatic breast cancer (MBC) classification utilising the synergy between GNNS and LLMS. Specifically, it describes the BG-MBC model[4] (Basaad et al., 2024a), which integrates representations derived through BERT and representations derived through a graph in an attempt to boost predictive accuracy. In this model, patient medical records form nodes, and histopathology reports have important semantic information that is extracted via BERT embeddings. Several feature improvement techniques, such as univariate selection, extra trees classifier, and Shapley values, have been utilised to enhance the feature pool.

Chapter 6: Conclusion and Future Directions. This chapter sums up the thesis, in addition, it introduces a discussion of current and future works.

[3]Abdullah Basaad, Shadi Basurra, Edlira Vakaj, Mohammed A Aleskandarany, Moahmmed Abdelsamea *GraphX-Net: A Graph Neural Network-Based Shapley Values for Predicting Breast Cancer Occurrence*. IEEE Access, 2024. DOI: 10.1109/ACCESS.2024.3424526

[4]Abdullah Basaad, Shadi Basurra, Edlira Vakaj, Ahmed Karam Eldaly, Moahmmed Abdelsamea *A BERT-GNN Approach for Metastatic Breast Cancer Prediction Using Histopathology Reports*. MDPT, 2024. DOI: 10.3390/diagnostics14131365

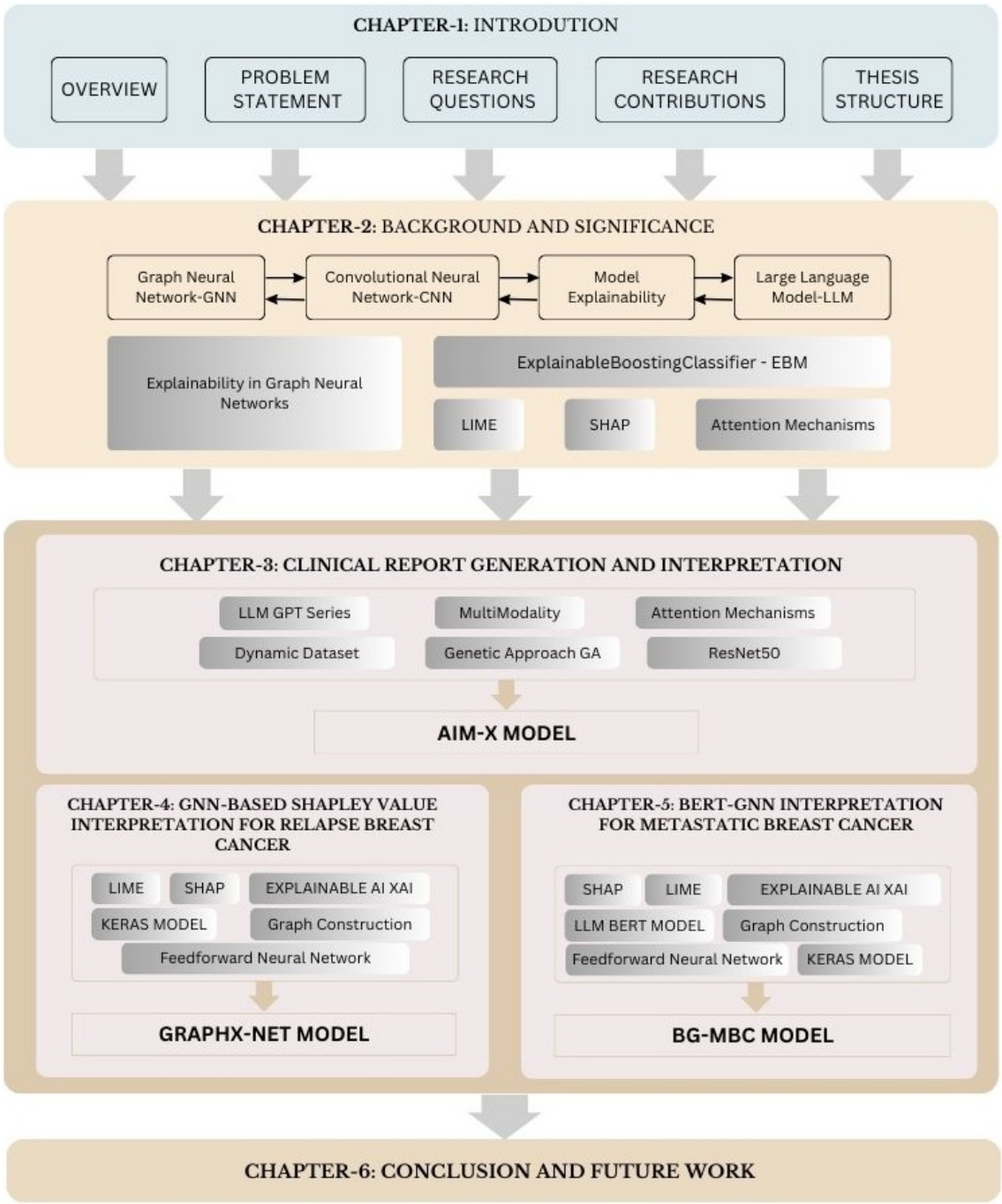


Figure 1.6: Thesis structure

Chapter 2

Background and Significance

2.1 Overview

This chapter identifies a variety of advanced technologies and new techniques to handle the problem statement defined prior in Section 1.2, such as medical image analysis, natural language processing, and explainable artificial intelligence used for prediction and detecting modelling. All these state-of-the-art methods have been combined with due care to develop a comprehensive and interpretable framework that would enhance the quality of diagnostics through improved predictions in a cancer management framework. The framework is specifically relevant for diagnosing lung and breast cancers, also it may also be useful for detecting other diseases in its role as an early stage indicator. Advanced approaches utilised will mean not only high accuracy of predictions but also explainability, which is crucial for doctors to understand and believe in recommendations provided by the system. The different approaches and methods followed in this work tackle, in a very effective way, not only the visual data represented by the medical images but also the textual and numeric data represented in the medical reports, thus providing a multimodal approach to the analysis. Each of the techniques applied will now be introduced systematically in separate sections for each approach in the following. Later, we will discuss the integration of the mentioned methodologies and review how each of them uniquely contributes to the overarching framework in enhancing predictive capability and interoperability.

In the subsequent chapters, each of them will be further detailed, where concepts lying in the background, strategies of implementation, and their contribution to constructing the final diagnostic and report generation framework will be reviewed at length. The end of this analysis gives an overview of how these work in conjunction with overcoming some of the challenges associated with modern medical diagnosis, particularly those affecting breast cancer diagnosis and the pathways

these create toward more robust and trustworthy AI-driven solutions in multi-stage cancer management.

Fig. 2.1 provides a comprehensive overview of the various themes explored in the thesis, clarifying the main areas related to the study. We have featured and highlighted the major issues and focused on their importance within the context. In addition, it also provides visual information concerning the main challenges met along the way, allowing for a more comprehensive understanding of the intricacies and complications encountered throughout the study. When considered as a combined figure, the central topics dominate most of the attention and issues associated with the entire research, thus they are organised for a better understanding of the scope, depth, and turn of developments in the thesis work.

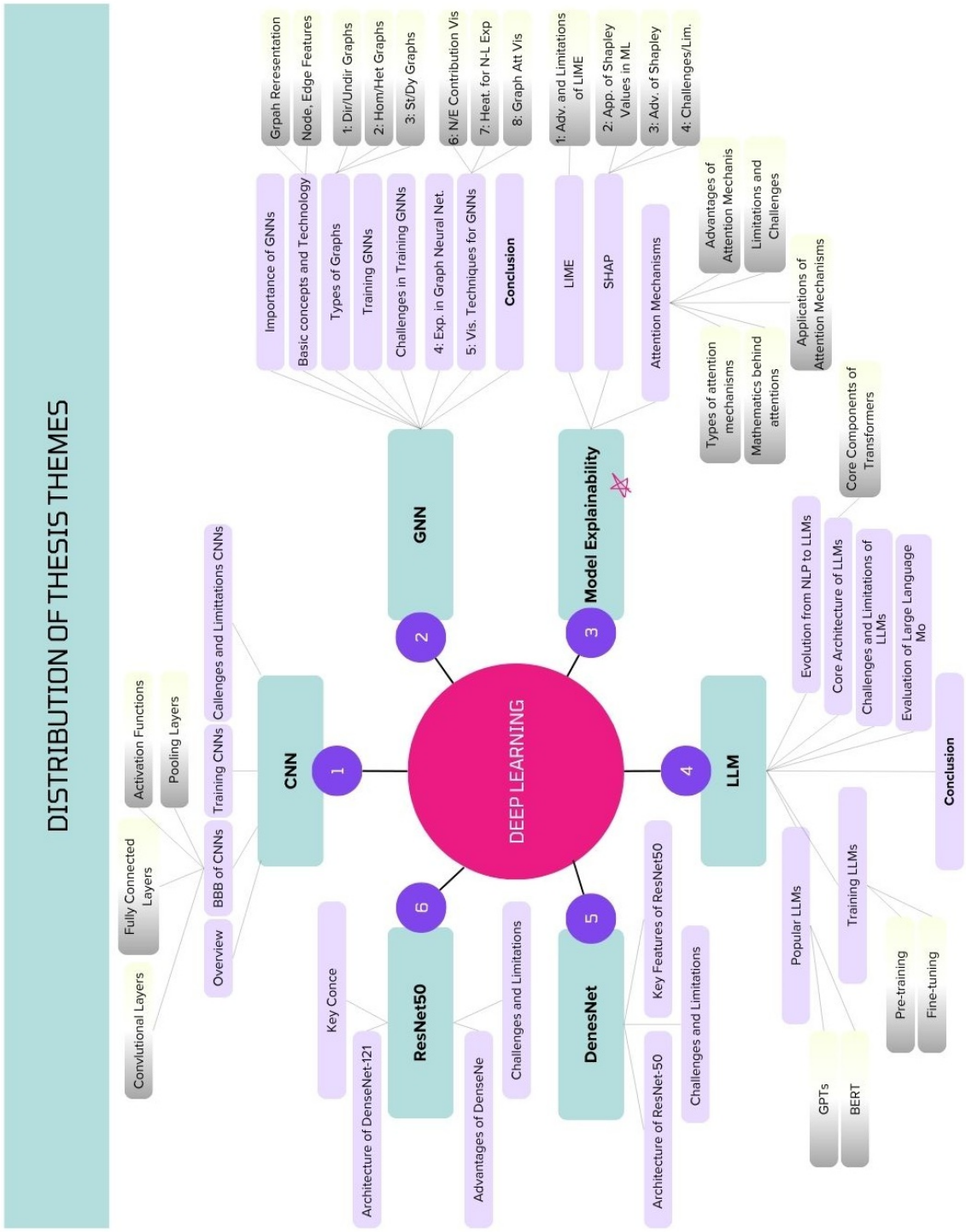


Figure 2.1: Distribution of the thesis themes

2.2 Convolutional Neural Network

CNNs are a category of deep neural networks, specially designed to handle tasks such as the analysis of structured grid data. Such a structure makes them particularly effective at tasks such as computer vision, audio processing, and some types of text data. Nowadays, CNNs form the backbone of modern machine learning image and video analysis, natural language processing, and medical imagery. With this architecture, the network can learn from the data by automatically detecting patterns concerning edges, shapes, textures, and rich spatial hierarchies, which, in many applications, are very important for meaningful feature extraction (Aloysius et al., 2017).

Their strength comes from their unique structural setup, whereas in traditional, fully connected neural networks, each neuron in one layer is connected to every neuron in its subsequent layer. CNNs rely on a mix of convolutional layers, pooling layers, and fully connected layers.

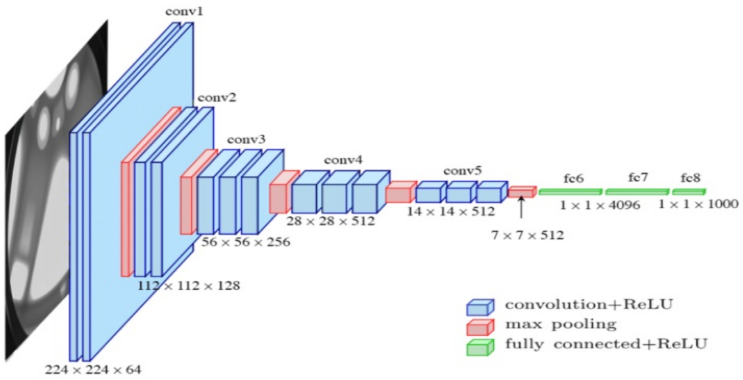


Figure 2.2: CNN Architecture

These layers cooperate in managing the data to comprehend the spatial hierarchies, focusing first on very simple features—things like edges or colours and building up to the more complex patterns that give a finer analysis. The hierarchical structure of it will, therefore, enable CNN to pick up and interpret such complex visual information as facial expression, object pose, and fine textures. These three concepts, pooling and parameter sharing, form just about the fundamentals for CNNs, each serving different purposes with the view to optimise the data processing for improving the model's learning capability. Convolutional layers apply a set of small filters on the input data. These kernels detect the local features of the data, which propagate through the network in order to obtain the feature maps (Aloysius et al., 2017; Basha et al., 2020b).

Each feature map represents one aspect of the input data that then moves to the next layer for tuning to edges or patterns. The CNN, with this capability, focuses on

just a few important features in each layer, removing the noise and other unnecessary details.

The pooling layers decrease the dimensions of the feature maps spatially; this reduces the sensitivity of the network concerning shifts in the positioning of features within the data and computationally too. Max pooling refers to a form of pooling where it takes just but the most active values within a neighbourhood in a feature map. In so doing, it will retain the most relevant information and discard the least relevant (He et al., 2015).

2.2.1 Basic Building Blocks of CNNs

The CNNs represent probably some of the most powerful tools currently being widely used in the area of machine learning and deep learning. They have become well-known owing to their outstanding performances in areas related to computer vision and everything related to images. CNNs consist of several crucial key one-dimensional building blocks, each important in helping the network learn from data as seen in Fig. 2.2. One-dimensional components include convolutional layers, pooling layers, activation functions, fully connected layers, and dropout regularisation. Each block plays its role in the model's capability to grasp, analyse, and predict such complex visual data (Aloysius et al., 2017).

2.2.1.1 Convolutional Layers

The core building block of a CNN is the convolutional layer, which has the full responsibility for feature extraction from input data. Recognising local patterns, small filters or kernels slide over the whole input image in this layer. Every filter is trained to extract specific features, such as edges, colours, or textures; the output from this operation is the feature map, representing the detected features at all spatial locations (He et al., 2015). While deeper layers capture more abstract or higher-level aspects, such as shapes or object parts, CNNs perform hierarchical feature extraction, which is considered one of the major positive points in the working of CNNs, thereby enabling them to be exceptionally good with image understanding.

2.2.1.2 Pooling Layers

Pooling layers are used after convolutional layers to downsample the feature maps, reducing their spatial dimensions. This not only reduces the computational complexity of the model but also enables CNNs to become more invariant to slight translations and rotations in the data. Several types of pooling exist; among them, max pooling is the most common (Gholamalinezhad et al., 2020). Max pooling takes the maximum value over a region of the feature map, preserving only the most

dominant feature in that neighborhood. In contrast, average pooling takes the average value over a region, which can also be useful in tasks where it makes sense to average intensities. Pooling helps the network generalise better since it retains only important features, removing less relevant details and noise (Zafar et al., 2022).

2.2.1.3 Fully Connected Layers

Fully connected (FC) layers are usually located at the end of a CNN architecture and act as the decision-making portion of the network. Unlike convolutional and pooling layers, in which neurons are connected only to a small region in the previous layer, fully connected layers connect each neuron to all neurons in the previous layer (Basha et al., 2020a). This is the connection that lets the model make high-level decisions by considering the features that it extracts from the convolutional and pooling layers. For example, in an image classification task, FC layers apply those features to predict the probabilities over every class. FC layers are essentially the classifiers synthesising the learned features into action-worthy predictions.

2.2.1.4 Activation Functions

Activation functions introduce non-linearity in the network, enabling CNNs to model richer patterns. After a convolution or pooling operation, data goes through an activation function of type rectified linear unit. By default, use ReLU because of its simplicity and efficiency, it replaces all negative values with zero and keeps positive values unchanged. It allows the network to learn non-linear relationships in data in order to improve the predictive capability. The other relatively newer variants of ReLU include leaky ReLU and ELU (Sharma et al., 2017). These variants solved some of the issues of the basic ReLU function. One such problem was that once the neurons were shut off, it hindered learning.

2.2.2 Training CNNs

CNN training is very elaborate, multistep tasks that have taken the performance of image recognition, computer vision, and even wider areas like natural language processing to the next level. CNNs are powerful and have been found to be very effective in learning from spatial hierarchies or patterns of image data (Zhang et al., 2018). Training generally consists of a number of cautious steps: data preparation, model selection together with hyperparameter selection, and iterative learning monitoring toward the best performance.

The first and foremost step in the training of the CNN is the preparation of the data through acquisition, cleaning, and organisation of the dataset (Radenović et al., 2018). Being a data-hungry model, its quality and amount decide its performance.

For an image, it should be a representative and diverse dataset of images with clear labels. Image datasets often have to go through multiple preprocessing steps in order to be prepared for training. For example, image resizing, normalisation, and data augmentation, which are standard preprocessing steps used in standard training datasets. Firstly resizing should be done, this is important to have the entire dataset resized to same dimensions and includes images of various numbers of channels. Normalising the image data usually consists of standardising the pixel values for example converting the pixel values to a range of [0, 1], or [-1, 1]. This is crucial ahead of stabilising training as it normalises feature scales improvements to gradient behaviour. Finally, data augmentation adds variance to images during training by implementing random rotations, flips, and colour shifting thus adding an artificial amount of variance that increases the number of features in training set. Data augmentation used to reduce overfitting models usually force the models to compile abstractions of images for generalisation on unseen data (Shorten et al., 2019).

Most models learnt representations of important features of images relatively invariant to orientation and lighting conditions, proving that preprocessing steps can greatly enhance dataset for improved performance models seen in the recent work on multi-modal learning as described in Section 3.1.3. Selection of the model architecture is usually driven by problem complexity and resource availability with prepared data. A quite substantial number of famous CNN architectures exist, such as AlexNet, VGG, ResNet, DenseNet and EfficientNet, all optimised for various tasks and performance needs. Architectures like ResNet have skip connections that facilitate going as deep as one wishes to avoid the vanishing gradient problem; hence, fine details may be preserved at deep representations (Zaeemzadeh et al., 2020). Most modern applications have started to adopt pre-trained models; this is now referred to as transfer learning. Transfer learning involves the adoption of an already-pretrained model, say on ImageNet, and tuning this model on your specific dataset means learning the new patterns in data while retaining general feature representations developed during its initial training.

To train the CNN on the chosen model in batches of images, optimisation through backpropagation is performed. Each batch goes through the layers of the network; neurons of the convolutional layer detect different features in an image (Basha et al., 2020b). Since this is a forward pass, it predicts a certain output and compares that with the actual label through a loss function. Cross-entropy loss is often used for classification tasks since it gives the difference between the predicted and actual probability distributions.

$$\text{Cross Entropy Loss Function} = -\sum_{x} P'(x) log P(x), \tag{2.1}$$

where $P(x)$ is minimising the cross-entropy between the "true" one-hot encoded distribution of data $P'(x)$ and the predicted class probabilities.

In this process, the backpropagation computes the gradients, followed by updates of weights through some sort of optimiser, such as stochastic gradient descent or Adam; these eventually serve to decrease the loss function in time (Azimi et al., 2018).

Other critical hyperparameters include the learning rate, batch size, and different regularisation methods used. This learning rate defines the step size at each update of weights. It is very significant to find an appropriate rate, because a too large value may cause divergence, while a too small value will significantly slack the training, and it may get converged to a local minimum (Azimi et al., 2018). The learning rate can be changed with the help of a learning rate scheduler and usually decreases during training to fine-tune the performance of the model. The batch size defines the number of samples handled at one time and controls memory usage and training speed, whereas regularisation includes dropout, batch normalisation, and weight decay to avoid overfitting for good generalisation on unseen data (Radiuk, 2017).

During training, both the training and validation losses track the learning progress of the model and hint at overfitting occurrences where the model performs well on the training data but fails to do so on the validation data. In practice, one may use early stopping to stop this training if validation performance stops improving; this preserves generalisation. Apart from accuracy, metrics such as precision, recall, and the F1 score can be used for performance evaluation, depending on the nature of the task.

After training, the generalisation capability of the CNN model is checked using a different test dataset. This is the final evaluation that will tell one how the model performs on real-world data; if necessary, one goes further with fine-tuning of hyperparameters or architectural adjustments (Zech et al., 2018). Other extra steps that might happen in a real-world application include model compression via pruning or quantisation, where one wishes to lighten the model in computations, so that it will be more suitable to run on resource-constrained devices (Sabih et al., 2020).

In all, CNN training is an art that involves reflective data preparation, the selection of model architecture, iterative optimisation, and careful regularisation toward deriving robust and generalisable models. The interaction between all these components makes CNNs exceptionally perform in extracting meaningful patterns from really complex datasets, hence being invaluable across many areas, especially in image analysis and computer vision.

2.2.3 Challenges and Limitations of CNNs

CNNs have been a revolutionary factor in computer vision and also other related fields, though they also consist of several challenges and limitations. These could affect their efficiency in the usage of certain applications and are to be overcome with care. Here's an overview of some of the major challenges and limitations associated with CNNs:

Table 2.1: Summary of CNNs Challenges and Limitations.

Challenge and Limitation
CNNs are very powerful when there are many well-labelled images available for training, so we benefit from the power of CNN over previous machine tools when analysing intractable problems like image classification and object detection. Finding large volumes of relevant data can be particularly challenging if all images require annotation (e.g., computer vision in the field of medical imaging) (Tong et al., 2022). In a data limited environment, the threat of model overfitting with a CNN is at its greatest. In relation to our training, we trained AIM-X Section 3.2 model with 7,471 images and achieved a 94.5% classification accuracy. This is positive, but it also reminds us that when analysing larger sets of images, we should balance model complexity with data quantity and variety so the model does not overfit the data and has better generalisability.
CNNs are computationally intensive, requiring significant memory, processing power, and often specialised hardware like GPUs or TPUs (Alzubaidi et al., 2021). This limits their use in resource-constrained environments unless optimised through methods like pruning or quantisation.
CNNs are often perceived as "black boxes," making it difficult to understand their decision-making process, especially in critical applications such as healthcare or autonomous driving (Das et al., 2020). Techniques like "Grad-CAM" help visualisation but don't fully resolve this limitation. Grad-CAM (Gradient-weighted Class Activation Mapping) is a popular technique for visualising what a convolutional neural network (CNN) "sees" when arriving at a decision. It provides heatmaps highlighting regions of an input image (e.g., an X-ray or MRI) that most support the model's prediction (Selvaraju et al., 2017).

(Continued on next page)

(Continued from previous page)

Challenge and Limitation

CNNs have a high potential for overfitting due to the large amount of parameters, especially when they are trained on limited, noisy, or non-representative data. Overfitting happens when the model has adapted too closely to the training data, thus retrieving noise or irrelevant patterns instead of more generalisable features to make a useful prediction. At this point, the model fits the training data well, but it will surely produce poor examples when tested against unseen data in real-world examples. Using regularisation methods such as dropout, batch normalisation, and data augmentation can help to reduce overfitting by adding variability or changing the simplicity of the model. Nonetheless, these methods may still not always be sufficient, especially in situations when the data is sparse or not diverse. In these cases, other more advanced techniques such as early stopping or transfer learning may be required to properly control overfitting and ensure that the model generalises well to new data (Santos et al., 2022).

CNN performance is sensitive to environmental factors such as lighting, noise, and background changes (Elboushaki et al., 2020; Shrestha et al., 2019). Domain adaptation is challenging, and scaling models often increases complexity, risk of overfitting, and computational demands.

2.2.4 ResNet50

ResNet50 is one of the most popular deep convolutional neural network models in The sphere of deep learning and computer vision. It is one of the members belonging to ResNet, or else said, Residual Networks (He et al., 2016). Being effective and deep, it has proved to be a strong tool in many applications like classification, object detection, and feature extraction. ResNet-50 comprises these essential building blocks: convolutional layers, batch normalisation, ReLU activation, and max pooling. Fully connected layers and identity or shortcut connections round out the assembly. The architecture can be divided into three main layers and sublayers:

1. Initial convolutional layer.
2. Residual Blocks: ResNet-50 uses residual blocks, each with three layers. Each block has a skip connection, which helps gradients flow smoothly.
3. Fully Connected (FC) Layer.

2.2.4.1 Key Features of ResNet50

The residual connections, while a distinguishing feature of ResNet architectures, allow this network to skip some layers during training. They also help in avoiding

the vanishing gradient problem and thus enable training very deep networks with minimal degradation in performance. This is an identity-mapping process that allows the input of a block to "shortcut" through layers; hence, it enables the network to focus solely on learning the residual or difference between the input and output. ResNet50 comprises 48 convolutional layers, as well as a MaxPooling layer and an AveragePooling layer. The layers are grouped into blocks. Each block consists of several convolutional layers together with residual connections (Wen et al., 2020; Mandal et al., 2021).

ResNet50 uses the architecture of convolutional layers to reduce the computation cost without compromising the representative ability and to reduce the dimensionality. Moreover, it extracts the features and restores the original dimensionality (Mukti et al., 2019; Sanmorino et al., 2024).

ResNet50 is computationally efficient for its depth, hence balancing high performance with resource utilisation reasonably well. For instance, ResNet50 models are pre-trained with large datasets like ImageNet and can easily be fine-tuned for any particular task (Sanmorino et al., 2024).

2.2.4.2 ResNet50 Challenges and Limitations

- Though ResNet50 is relatively efficient compared to deeper networks, it still requires significant computational resources and memory (Sanmorino et al., 2024).

- Training ResNet50 from scratch may require some tedious tuning of hyperparameters (ElGhany et al., 2021).

- Because of its vast number of parameters, ResNet50 might overfit the smaller dataset and cannot perform well if not using some proper regularisation methods, including dropout and data augmentation. That would be a potential source of poor generalisation on unseen data, assuming limited labelled data (ElGhany et al., 2021).

- Pre-trained ResNet50 might not generalise well for the datasets whose characteristics are very different from those of the original training dataset, for instance, ImageNet. Highly specialised tasks may require very big changes in either architecture or layer adjustments with ResNet50 (ElGhany et al., 2021).

ResNet50 is an ingredient that forms the backbone of recent successful deep learning; it's a strong foundation that high-performance models in the domain of computer vision rely on. The power of residual learning together with an effective bottleneck architecture helped move toward more sophisticated architectures and so

remained one of the favorite choices of both researchers and practitioners. With its perfect balance of depth, accuracy, and efficiency, ResNet50 remains an important tool for addressing complex image-based challenges along diverse domains (Mukti et al., 2019).

2.2.5 DenesNet-121

DenseNet-121 is one of the convolutional neural network architectures with dense connectivity between layers. It is part of the family of Densenets, proposed by (Huang et al., 2017). in their paper "Densely Connected Convolutional Networks". DenseNet-121 being one of the most widely used CNNs for tasks including image classification, object detection, and medical image analysis is because of its economy of parameters and mitigation factor to vanishing gradient problems. DenseNet-121 intends to improve further feature propagation and reuse in CNNs. While traditional architectures like ResNet use skip connections to address the vanishing gradient problem, DenseNet is built upon a more complex principle, connecting each layer to every other layer in a feed-forward fashion. This obvious feature has led to the notion that DenseNets learn additionally robust features while using far fewer number of parameters (Huang et al., 2017).

2.2.5.1 Key Concepts

In DenseNet-121, every layer gets feature maps from all of its pre-existing layers and sends its feature maps to all the other following layers. Hence, we can refer that this is a function of composite operations (e.g., convolution, batch normalisation, ReLU). Bottleneck layers are featured. In order to increase computational efficiency. Mathematically, if x_l represents the output of the $l-th$, then:

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]), \tag{2.2}$$

where x_l is the output of the l-th layer, H_l is a composite function of operations (e.g., convolution, batch normalisation, ReLU), $[x_0, x_1, \ldots, x_{l-1}]$ denotes the concatenation of feature maps from all previous layers.

DenseNet-121 integrates bottleneck layers of 1x1 convolution, which compresses the number of input feature maps before the application of 3x3 convolutions. Been decreasing the computational power of the model while still retaining its precious performance (Huang et al., 2017; Alawi et al., 2021).

Transition layers allow for the control of expanding feature maps. It consists of a 1x1 convolution layer followed by a 2x2 average pooling layer, leading to a reduced size of the feature maps in terms of width and height (Huang et al., 2017; Wang

et al., 2020).

2.2.5.2 Architecture of DenseNet-121

DenseNet-121 is a 121-layer network with four dense blocks connected by transition layers. Following the four dense blocks, there are one transition into another and one classification block at the end of the network. Therefore, there are four main sections in the entire architecture:

Initial Convolution Layer The Initial Convolution Layer represents the first phase of operation on the raw image prior to it being processed through the dense blocks. The purpose, structure, and significance of the initial convolution layer is summarised as follows: Feature extraction is the initial activity performed by the initial convolution layer, which takes an image input (in RGB pixels) and converts it directly into an initial set of features. Dimensionality reduction is another activity performed by the initial convolution layer. The results include downsampling with respect to the spatial dimensionality (height/width) to reduce computational overhead for subsequent layers (Ali et al., 2019). Channel expansion increases the number of channels (increasing from 3 RGB channels to 64, for example) to increase the richness of feature extraction.

Dense Blocks Layers Central structural elements are Dense Blocks and Transition Layers, built for maximum re-use of features and gradient flow. Beneath are succinct definitions of each of the layers, mathematical operations in operation, and implementation details. A Dense Block comprises of multiple Dense Layers (or "units"), in which each layer has direct connections with every other feed-forward layer. Also, transition Layers followed each dense block to downsample spatial dimensions and reduce channels. between Dense Blocks (Huang et al., 2017).

Classification Layer The classifier layer is the last fully connected (FC) layer that translates the extracted features into the desired output classes (for example, the 1000 classes for ImageNet) (Huang et al., 2017).

Growth Rate Growth rate is the hyperparameter defining the number of feature maps added by each layer. For example, if k=32, that layer adds 32 feature maps to the network. The growth rate functions as a "knob" to control the flow of information and model width in DenseNet, with the intention of achieving a tradeoff between performance, resource usage, and feature reuse (Huang et al., 2017; Wang et al., 2020).

2.2.5.3 Advantages of DenseNet-121

DenseNet-121 is a powerful and efficient CNN architecture that has revolutionised deep learning in computer vision tasks. Here are the main advantages of DenseNet-121 as illustrated in Table 2.2.

Table 2.2: DenseNet-121 Advantages

Advantage	Describtion
Parameter Efficiency	it reuses feature maps from all preceding layers rather than using additional parameters, compared to traditional CNNs like ResNet. This fact reduces overfitting and improves generalisation.
Feature Reuse	Dense connectivity ensures feature map reuse for the gradient flow and thus the learning of more robust features.
Mitigation of Vanishing Gradientss	This dense connectivity alleviates the problem of vanishing gradients with direct connections between involved layers, making it easier to train very deep networks.
Computational Efficiency	The bottleneck and transition layers reduce the computational cost of DenseNet-121 while keeping the accuracy level quite high.

2.2.5.4 DenseNet-121 Challenges and Limitations

Although DenseNet121 has proved its capabilities with the massive and complex problems, it has some drawbacks, as illustrated in Table 2.3.

DenseNet121, one of the model architectures based on the Densely Connected Convolutional Networks (DenseNets) architecture, is one of the most powerful deep learning models developed for image classification and other related tasks. DenseNet121 works by passing every layer's first information to every other layer in a feed-forward pass, allowing for the reuse of features. However, this also reduces the number of parameters when compared to other architectures while dealing with the problem of gradual vanishing feedback for deeper nets. Therefore, DenseNet121 becomes a good candidate when performance and computation efficiency are essential (Goyal et al., 2017).

Table 2.3: DenseNet-121 Challenges

Challenge	Describtion
Parameter Efficiency	it reuses feature maps from all preceding layers rather than using additional parameters, compared to traditional CNNs like ResNet. This fact reduces overfitting and improves generalisation.
Feature Reuse	Dense connectivity ensures feature map reuse for the gradient flow and thus the learning of more robust features.
Mitigation of Vanishing Gradientss	This dense connectivity alleviates the problem of vanishing gradients with direct connections between involved layers, making it easier to train very deep networks.
Computational Efficiency	The bottleneck and transition layers reduce the computational cost of DenseNet-121 while keeping the accuracy level quite high.

2.2.6 A review on CNN-based Thoracic Diseases and Lung Cancer Detection

The advent of CNNs has completely altered the landscape of medical imaging, and most significantly within the arena of cancer detection and diagnosis. In the following, we provide a review of the use of CNNs in the realm of early-detection lung cancer and relapse and metastasis breast cancer. The use of CNNs has elicited substantial criticism regarding data limitations, model interpretability, generalisability, and clinical reliability. A major critique of CNN-based systems for X-ray image analysis in cancer detection is the quality and the diversity of training data. CNNs require large scale datasets for supporting robust feature learning. In many cases, the datasets available for breast and lung cancer X-ray images are either limited in size, poorly labelled or extensively unbalanced. For example, the IU chest X-ray images used in this research are still limited in size and noise in labels with imbalanced diseases/classes, the same underwhelming identifications state for the ChestX-ray14 dataset (Wang et al., 2017a). Likewise most mammograms datasets like INbreast and DDSM are also relatively small and do not represent and demographic or technical diversity one might expect in real-world clinical situations (Moreira et al., 2012). Such limitations pose concerns for overfitting and poor generalisation of the model when applied outside the training domain.

The "black box" aspect of CNNs is another significant barrier to establishing the model for clinical applications. These models achieve very high accuracies, however

they are also notoriously difficult to provide clear explanations for any accuracy. The lack of interpretability to the models is especially problematic when being used in clinical practice. (Selvaraju et al., 2017) have attempted to provide some means of explainability through Grad-CAM, yet still find it challenging to interpret even the visualisation outputs, or that these features may not have clinical meanings. This inability undermines clinician trust, and prohibits the integration into clinical workflow. Another generalibility issue raised by (Zech et al., 2018) demonstrated that CNNs trained on chest X-ray images from the particular hospital performed significantly worse when those images were from outside institution.patients. This refer to differences in imaging protocols, equipment or institutional variation. For breast cancer, the same pattern is seen, for example studies that have used CNNs on the DDSM dataset report high accuracies, yet those accuracies do not carry over to other clinical datasets, or imaging standards (Ribli et al., 2018). Several studies demonstrate the upside and limitations of CNNs within the domain. (Rajpurkar et al., 2017) developed CheXNet, the CNN trained on ChestX-ray14 dataset that performed better than practicing radiologists in the detection of pneumonia, proxy for CNNs potential in lung imaging. However subsequent evaluations have questioned the validity of these results as based on label noise and overfitting concerns. (Ribli et al., 2018) developed a deep learning model that matched or outperformed radiologist performance on mammogram analyses using DDSM dataset. However, space still remains for generalisability concerns as well as challenging clinical situations. Further discussions of utilising the CNNs have been added to Section 2.6.

Table 2.4: Summary of CNN-based models in medical imaging: tasks, performance, limitations, and datasets/modalities.

Paper	**Task**	**Accuracy**	**Limitations**	**Dataset / Modality**
(Wang et al., 2017a)	Development of a large-scale chest X-ray dataset (ChestX-ray8) for weakly-supervised multi-label classification and disease localisation of 8 thoracic diseases from chest X-rays	Demonstrated the feasibility of using weakly-supervised learning for classification and localisation tasks; provided benchmarks for future research, AUC: 0.76–0.90 across diseases	Challenges in achieving high precision due to weakly labeled data; localisation performance limited by image-level annotations; no clinical validation; labels from radiology reports may contain noise; imbalanced class distribution;	ChestX-ray8 dataset comprising 108,948 frontal-view X-ray images from 32,717 unique patients, with multi-label annotations for eight thoracic diseases extracted from radiological reports using natural language processing
(Moreira et al., 2012)	Development of a full-field digital mammographic database, INbreast, with detailed annotations for breast lesions	Not applicable (database description and methodology), roposed a FFDM database with a wide variety of findings.	Limited to 115 cases; lacks large-scale diversity; no clinical validation or external dataset testing	INbreast dataset: 115 cases (410 images) from women with both breasts affected (90 cases) and mastectomy patients (25 cases); images acquired using MammoNovation Siemens full-field digital mammography with a solid-state detector of amorphous selenium; includes various lesion types (masses, calcifications, asymmetries, distortions); accurate contours provided in XML format

Continued on next page

Table 2.4 – *Continued: Summary of CNN-based models in medical imaging*

Paper	**Task**	**Accuracy**	**Limitations**	**Dataset / Modality**
(Selvaraju et al., 2017)	Provide visual explanations for deep convolutional networks using gradient-weighted class activation mapping (Grad-CAM)	Demonstrated effectiveness qualitatively in highlighting important image regions; applied to image classification and captioning tasks; quantitative performance depends on the underlying network	Grad-CAM provides coarse localisation; may not perfectly capture fine-grained features; performance depends on base CNN model	Tested on multiple datasets: ImageNet (object classification), MS COCO (image captioning), and medical imaging datasets; modality varies (RGB images, X-rays in some experiments)
(Ribli et al., 2018)	Detection and classification of breast lesions in mammograms using deep learning	Achieved AUC: 0.95 for lesion classification; outperformed traditional CAD systems	Limited by dataset size; performance may decrease on rare lesion types; model interpretability is limited	INbreast and private mammography datasets; modality: digital mammography
(Rajpurkar et al., 2017)	Detection of pneumonia in chest X-rays using deep convolutional neural networks with DenseNet-121	Achieved F1-score: 0.435, surpassing average radiologist performance on test set. Accuracy: between 0.7345 and 0.9371 across different thoracic diseases.	Limited to pneumonia only; potential dataset bias; requires large annotated dataset; interpretability of model is limited	ChestX-ray14 dataset; modality: frontal chest X-rays

2.3 Graph Neural Network GNN

GNNs represent yet another state-of-the-art development in the area of deep learning. Such models are powerful tools for processing and analysing data which naturally emerge as graphs (Zhang et al., 2021b). While classical neural networks have enjoyed tremendous success in dealing with data represented in gridlike structures, for example, images or sequences of data, GNNs extend these capabilities to the representation of non-Euclidean data structures (Chen et al., 2020c; Chami, 2021). For example, in which entities and their relationships are naturally represented as nodes and edges. That will make GNNs very effective in a wide range of applications, from social networks to molecular chemistry and knowledge graphs (Reiser et al., 2022).

It is here that GNNs come into play. GNNs design neural networks to generalise across graph-structured data in a way that models learn not only from the features of each separate node but also from the structure of the graph itself (Bianconi et al., 2009). Unlike typical neural networks, which operate on fixed-size inputs and assume a grid structure, GNNs adaptively adjust to the topology of the graph, hence being much more flexible and suitable for such tasks at hand, dealing with relational data (Cong et al., 2023). GNNs allow the aggregation of information from other nodes in a neighborhood iteratively, helping the model to effectively capture the local and global structures of the graph. This is achieved by using a message-passing mechanism where information from other adjacent nodes is passed through the network, and integrated into the current node's representation. Through multiple layers of message passing, nodes collect progressively more information from the far reaches of the graph, enabling the GNN to capture deep relational structures (Khemani et al., 2024).

A wide variety of specialised types of GNN architectures have emerged to handle diversified tasks on graphs. Among them, probably the most well-known is the Graph Convolutional Network **(GCN)**, which adapts the principle of convolution—typically employed in CNN for image data—to the graph domain. Graph Convolutional Networks aggregate and transform the node features with the help of the adjacency structure of the graph and are, therefore, really very powerful in semi-supervised learning tasks, which include node classification in citation networks (Bhatti et al., 2023). Another popular variant is the Graph Attention Network (**GAT**), which explicitly incorporates an attention mechanism that allows the model to assess the importance of neighbouring nodes differently depending on their relevance for the task under consideration. This flexibility is beneficial for applications where some relationships between nodes are more important than others (Xu et al., 2023). Graph Sample and Aggregation **(GraphSAGE)** focuses on scalabil-

ity by using node sampling so that GNNs can be more effectively trained on large graphs. Selects a fixed-size neighbourhood around a node (instead of using the entire graph) (Shao et al., 2024).

2.3.1 Importance of GNNs in Complex Data Modelling

The major advantages of GNNs concerning complex data modelling reside in the ability of learning about complicated relationships that may exist among objects. Indeed, graphs naturally represent relationships between data, such as those between users in a social network connected by friendships, atoms connected by bonds in a chemical molecule, or web pages via hyperlinks (Waikhom et al., 2023). GNNs also support an effective exchange of information among nodes using a message-passing mechanism, wherein any node can aggregate and update its feature based on the neighbouring node features. Let the GNN learn also not just the local but the global structural patterns of the graph (Munikoti et al., 2023). For instance, consider an application like social network analysis. In most cases, the behaviour of a node depends not so much on properties of the node itself but rather on some actions or interactions of neighbours. In such places, GNNs can progressively incorporate neighbours' information for predictable user behaviours, mining communities, or recommending new links (Koskinen et al., 2023). A similar explanation, in molecular chemistry, properties in a molecule are defined by the structure of the bonds between atoms. GNNs model these interactions and have thus found applications in tasks such as molecular property prediction and drug discovery. An important portion that separates GNNs from the rest of neural networks is their ability to model these kinds of relationships, which are usually captured incompletely by traditional neural networks when applied to graph-structured data (Torres et al., 2023).

2.3.2 Basic Concepts and Terminology

GNNs are powerful, deep learning architectures designed for processing graph-structured data. These are the data that, through a network, represent the interrelations among entities by means of nodes or vertices connected by edges or links. So as to understand how GNNs work, their most varied applications become relevant by understanding the basics behind some of the vocabulary about graph theory and GNNs (Khemani et al., 2024).

2.3.2.1 Graph Representation

A graph G is majorly defined, as a mathematical object $G = (V, E)$, by the following two things:

Nodes (Vertices) These represent entities or things in the graph. Every node may have its own attribute. Example: in a social network, nodes can be persons with attributes like age and name.

Edges (Links) These are the connections between the nodes. The edge can be directed or indirected and also can have attributes such as weights. A weighted edge would tell something about the closeness to each other. For example, friends in a social network (Khemani et al., 2024).

2.3.2.2 Node Features and Edge Features

In a GNN, each node is associated with a feature vector that represents the attributes of that node. These features are used as inputs to the GNN model and are updated during the learning process (Khemani et al., 2024; Yi et al., 2023). For instance, in a social network, a node's features could include age, gender, and interests (Peersman et al., 2011). Also, in a molecular graph, a node (an atom) might have features such as atomic number or electronegativity (Bergmann et al., 1996). The node features are usually denoted as h_v or x_v, where v is the node.

Considering node features, edges can also have a number of features representing properties of the connection between two nodes. These can be in the form of weights, labels, or any other measures describing the relations among nodes (Yi et al., 2023; Bilot et al., 2023). Within a transport network, for example, these edge features may capture the distance or travel time between pairs of locations. In molecular graphs, edge features may indicate bond types between atoms. Edge features are sometimes also known as rewards the property of the boundary between nodes. Edge features are often denoted as e_{uv}, representing the feature of the edge u and v.

2.3.3 Directed and Undirected graphs

A directed graph is represented by directed edges, which means the orientation does tell the direction of travel from one vertex to another. While in an undirected graph, edges are undirected and don't have any specified direction, due to that fact, nodes can interact in both ways as shown in Fig. 2.3. In the undirected graph, everything can be considered to be on the inside of two directed edges; that is, both nodes of an edge are interacting with each other (Khemani et al., 2024).

2.3.4 Homogeneous and Heterogeneous graphs

A homogeneous graph can be explained as a graph wherein the nodes and edges are categorised under one type Fig. 2.4. To be more specific, all the nodes would rep-

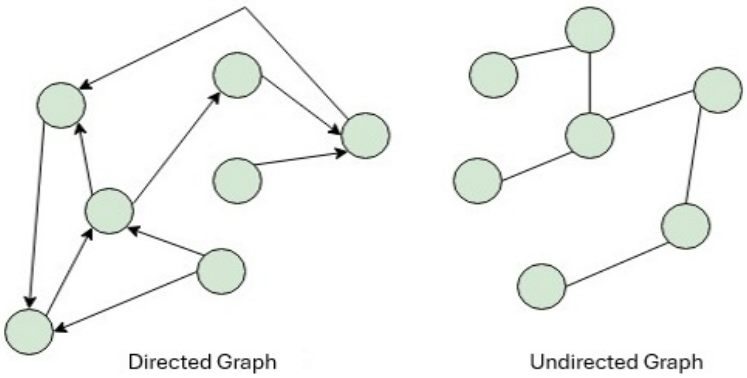


Figure 2.3: Directed/ Undirected Graphs

resent just one kind of entity, and all the edges would express the same relationship among them. Homogeneous graphs are the simplest of all graphs and are mostly utilised in those applications in which all the entities and their respective relations are of the same and similar nature. It can be applied in Social networks, citation networks, and simple computer networks (Khemani et al., 2024; Xiong et al., 2023).

Heterogeneous Graph Unlike in a homogeneous graph, in a heterogeneous graph, all node and edge types are different. That is to say, entities can be of many kinds, and relationships among them can be not only different but from other natures too. Compared with homogeneous graphs, heterogeneous graphs are much more complex but better prepared for representing a real-world system containing different entities and relationships. It can be applied in Knowledge graphs, recommendation systems, and biological networks (Khemani et al., 2024; Xiong et al., 2023).

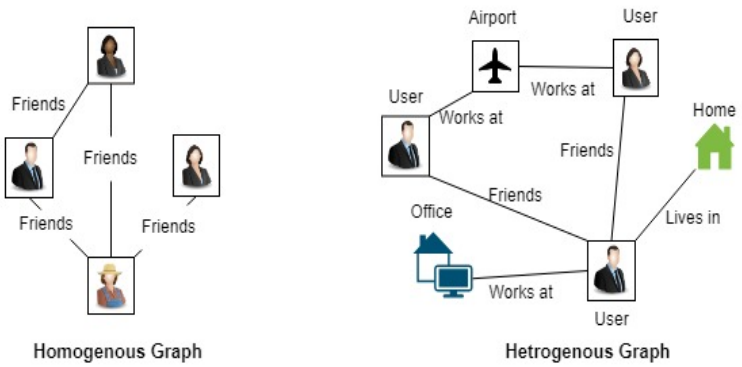


Figure 2.4: Homogeneous (Left), Heterogeneous (Right) Graph

2.3.5 Static and Dynamic graphs

A static graph is a graph whose structure, due to nodes and edges, is fixed since its definition. In the context of a static graph, the notion of time does not exist.

Consequentially, the links among objects are considered permanent and fixed. Static graphs are used for the description of systems in which there are no variations in the connections, or any change in relations is irrelevant or immaterial to the analysis as shown in Fig. 2.5.

Dynamic graph sometimes referred to as a temporal or evolving graph, models a network whose nodes, edges, or both are changing over time. In dynamic graphs, relationships between entities come and go, or evolve, with time; this surely makes them much more suitable for modeling real-world systems whose connectivities change with time (Khemani et al., 2024; Mayer et al., 1995).

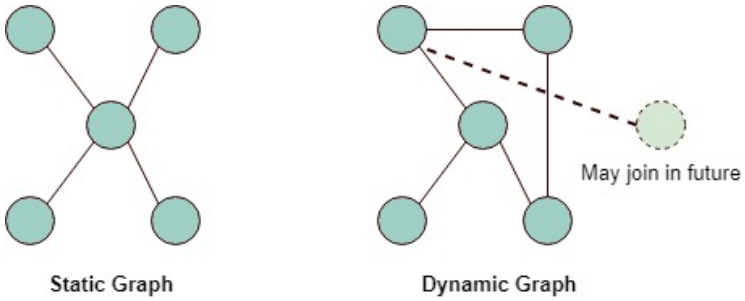


Figure 2.5: Static and Dynamic Graphs

2.3.6 Training GNNs

GNNs represent one of the most developed deep-learning architectures and are designed for graph-structured data. That fact alone makes them very powerful in a wide spectrum of areas such as social network analysis, molecular chemistry, and knowledge graph reasoning. However, GNN training inherently differs from training standard deep neural networks due to the nature of graphs and the aggregation of information from various interlinked nodes. The core points to be discussed here would be architecture, and the learning process (Munikoti et al., 2023; Hamilton, 2020).

A typical architecture of GNNs follows the message-passing framework, where every node in the graph updates its feature representation by aggregating information about its neighboring nodes as illustrated in Fig. 1.4. This process is repeated for multiple layers, at which time nodes can gather information about progressively more distant nodes in the graph. Some major moves of forward propagation at every layer in GNNs are as follows:

Aggregation or Message Passing Under the notion that neighborhoods' information is aggregated at every node, features defined as edges consider their relationship (Computer Science, n.d.).

Update It updates the node representation using the aggregated information along with its previous features.

Readout, or Pooling After some iterations of the message-passing process, the node features flow into the output layer until, ultimately, an end task is made such as node classification, link prediction, or graph classification as proved by (Computer Science, n.d.; Zhang et al., 2021a).

$$h_v^{(l+1)} = \text{Update}\left(h_v^{(l)}, \text{Aggregate}\left(\{h_u^{(l)} | u \in \mathcal{N}(v)\}\right)\right), \tag{2.3}$$

where $h_v^{(l)}$ is the feature representation of node. $N(v)$ is the set of neighbors of node v.

The Aggregate function collects information from neighboring nodes, and the Update function computes the new representation for the node. Like other deep learning models, GNNs are trained using a loss function that calculates the difference between the model's predictions and the ground truth labels, in the case of supervised learning or the expected outcome in unsupervised learning (Patrini et al., 2017). We always use Cross-entropy loss on the classification task for node, edge, or entire graph-level label predictions. Mean Squared Error (MSE) is always applied in the regression tasks where a model predicts any continuous variable (Huang et al., 2023c).

The whole process of training would be that the loss function is minimised by updating the weights of the GNN with the help of some sort of optimisation technique like stochastic gradient descent (SGD) or some derivative, such as Adam. Using backpropagation, the gradients of the loss concerning the model weights have been computed; therefore, weights have been updated with the purpose of minimising the loss. For instance, in a node classification task, the loss will be computed (Izadi et al., 2020; Bacciu et al., 2020):

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(p_i), \tag{2.4}$$

where y_i is the true label, and p_i is the predicted probability i.

2.3.7 Challenges in Training GNNs

The major benefits of using GNNs relate to the power of these models to comprehend complex relationships among entities. However, GNNs still struggle to learn node dependencies efficiently in irregularly structured graphs, where varying node connectivity complicates batch training. (Min et al., 2021). Besides, training GNNs on such diverse and complicated graph structure alone often requires some of the

most sophisticated techniques to handle such variation, explicitly with heterogeneous data, with node and edge types (Zhang et al., 2019).

Dynamic graphs also have some challenges based on changes over time. So, they need models update the structure continuously without retraining the model from scratch (Feng et al., 2024).

There is another set of difficulties, particularly in terms of computational efficiency. Graph size, memory size, and vanishing exploding gradients. In the case of the graph size, as the graph becomes larger, the computational cost of training increases dramatically. For instance, when social network graphs have millions or billions of nodes, the process of generating, aggregating, and updating the state of nodes becomes computationally expensive. This demanded a large number of memories for processing large graphs. Storing the entire graph along node features and intermediate states for backpropagation requires highly efficient memory (Zhang et al., 2019; Yang et al., 2021).

The other challenge is with the neighbor explosion in GNNs. In message passing, which is a key component in GNNs, every node will aggregate information from neighbors. With every additional layer, however, the number of nodes grows exponentially. This probably leads to the "neighbor explosion". This also increases the computational time and memory consumption, hence making it hard to scale deeply into the architecture. One major challenge with GNNs is scalability when trying to apply them to real-world large-scale datasets. To mitigate these scalability issues and reduce computational complexity, GraphSAGE is often used. It randomly selects a small, fixed number of neighbors per node (e.g., 10 neighbors per hop) instead of processing all neighbors. However, doing so may sometimes come at the cost of losing vital structural information, which might affect the overall performance of the model (Yu et al., 2022). We are still talking about the message-passing key, another challenge in this term could be the over-smoothing. So, according to the deepest architecture GNN models, after several layers of message-passing node representations tend to become indistinguishable from each other. This can distract the performance of models (Yu et al., 2022; Chen et al., 2020b).

Noisy and data incomplete also is a challenge for GNNs. Some connections in graph may be led to poor model generalisation. So, in some cases, GNNs treats all nodes equally although not all edges or connections are equally informative in GNNs training, thereby propagating noise during training. Also, training GNNs model on such incomplete graphs can lead to suboptimal performance (Huo et al., 2023).

2.3.8 Explainability in GNNs

The uprising of GNNs in most domains, from drug discovery to social network analysis, sets a growing need for these GNNs to become interpretable. Like most deep learning models, GNNs tend to behave like black boxes that take graph-structured input data and return predictions in complex representations. In most real-world applications, however, knowing why decisions have been made by the GNN is at least as important as the actual decision. It is explainability that helps in gaining confidence, ensuring accountability, and providing improvements in debugging for those models.

Why Explainability is Important in GNNs? Many sensitive deployments for which GNNs are finding applications, like health, finance, and legal, have a decision-making process that needs to be explainable to the stakeholders. As an example, in drug discovery, GNN might predict just how effective a compound would be because of molecular interactions. Without explainability, one cannot really validate such predictions or factors thereof that result in those predictions and hence reduce trust (Agarwal et al., 2023).

Knowing what has influenced a GNN to make certain predictions helps find the potential flaws or biases in the model. GNNs can exhibit multiple biases, which may stem from the data, model architecture, or training process. Structural biases may arise from topological imbalances of the graph, while sampling biases may occur when a subgraph or neighborhood is not representative (Virinchi et al., 2023). Label bias involves class labels being extremely skewed or incomplete, while temporal biases occur in dynamic graphs because models that outperform their previous state do not adapt over time (Dong et al., 2022). These types of bias can compound, resulting in unfair or inaccurate predictions which may favour high degree nodes, or push marginalised groups from a minority or majority class.

If a GNN makes its decisions with respect to insignificant parts of a graph, then this may hint at defects in the architecture of the model or at its training process and hence help the researchers tune up the model (Ghadekar et al., 2024). Much of the industry stands bound by regulations in their respective jurisdictions that enforce the interpretability of AI models. For example, the GDPR within the European Union enforces the "right to explanation" in decisions made by automated systems. Explainability in GNNs thus provides compliance with such regulations (Wu et al., 2022). This may be ensured-explainability can make sure that the model is not making partial or biased decisions using GNNs on social network analysis or recommendation systems. Hence, one would not want to discriminate due to the sensitive attribute in graph data.

2.3.9 Visualisation Techniques for GNNs

Besides, a number of different visualisation techniques reviewed that constitute a very important role in enhancing interpretability in GNNs helped us understand more about how they work. We are able to understand the inner structure of GNNs and determinants driving its prediction by node and edge contribution visualisation, generating heatmaps, and visualisation of attention. These methods make GNNs more explainable and self-trusting to the models by providing understandable explanations.

2.3.9.1 Node and Edge Contribution Visualisation

Such GNN models take as input the node attributes in addition to the structural information of the graph itself and predict the edges. Visualisation may help a user understand which part of the graph-nodes and their relationships-was more useful for a specific model task at hand, say, node classification, link prediction, or graph classification (Wang et al., 2015).

Node Contribution This would most probably explain how the GNN pools information from its neighbors, by visualising individual node contributions. The nodes should be colored based on their contribution toward the final prediction, computed over various metrics, such as attention weights, gradients, or importance scores computed using perturbation-based methods (Wang et al., 2015; Tang et al., 2021). For instance, in a social network graph, there might be a few nodes acting like highly influential nodes; highlighting those will show how much each contributes toward the influence of information.

Edge Contributions The edges in the graph encode some form of relationship between the nodes, and there are many instances when knowledge regarding which edges or even which contribute significantly toward the model's prediction. In other words, weights or importance scores are assigned to the edges in visualising these edge contributions, usually in an attention mechanism or through a gradient-based method (Han et al., 2022). One such method could be visualising the most relevant chemical bonds in a molecular graph to give reasons for some compounds' property prediction done by the GNN.

Subgraph visualisation Besides analysing all nodes and edges, often it is useful to extract and visualise the subgraph representing the most informative part of the graph to a given prediction (Yuan et al., 2021). Tools like GNNExplainer are ready to provide such subgraphs and visualise them. In such a way, the attention of the researcher is tuned to really important portions of the graph that provide an impact

on the output of the model.

2.3.9.2 Heatmaps for Node-Level Explainability

Heat maps are one of the best visualisations that could be used to represent the importance of nodes on graphs. That would underline which nodes in the graph are really influencing the GNN's decision so much more and would, therefore, help explain how the model made its predictions. Once considering the GNN framework, the heatmaps provide crucial insight into the importance of every single node. Colors here represent the importance of each independent node concerning every other dependent node on which the model is dependent for any kind of prediction. As an example, considering the classification performance of a certain node, its neighbors will have a huge contribution; in this case, such neighbor nodes can be featured with a high score, which can later be used for heatmap visualisation superimposed on graphical representation (Ying et al., 2019). In general, one can have such a diversified, important underlying node feature at the node level relevant to model prediction, including age, income, or even molecular properties. A heatmap visualises gradients or other measures of feature importance at each node, showing which node characteristics are most informative (Ying et al., 2019).

There are some applications involving GNNs where the temporal graph is able to show, through heat maps, the importance of some nodes over time (Rossi et al., 2020). Application examples are useful in financial fraud detection and the analysis of social media too, where nodes and edges have importance that varies while the graph is evolving.

2.3.9.3 Graph Attention Visualisation

Along with other attention-enhanced GNN architectures, GATs intrinsically offer an explicit method for model interpretability. The direct visualisation of the attention weights makes it possible to assess which degree of influence one node would have on another one within the graph, as this mechanism allows giving more or less importance to the neighboring nodes. In attention-based GNNs, it is possible to visualise the attention weights in order to highlight how much different neighbors are important with respect to a target node updating its representation (Knyazev et al., 2019). Think for example of a citation network, where one wants to visualise those papers or nodes that a GNN pays more attention to when predicting the category of a target paper. Color-coding, or variations of line thickness proportional to attention weights, are common in these visualisations.

Attention mechanisms can also be applied to edges in the graph and allow visualisation of which relationships are prioritised - by the edges - by the GNN. For

example, line thickness or colour intensity can be varied according to attention coefficients using a GAT to show which edges bear more weight on model predictions (Knyazev et al., 2019; Lin et al., 2020). Sometimes there's more need to visualise, instead of attention scores of individual edges or nodes; aggregated attention across multiple layers of the GNN. It gives a better holistic view of how the model processes the graph in its entirety-which parts of the graph structure have been getting higher attention on average (please refer to Section 3.1.2.4 for more details), and how that changes across different layers of the network (Zhang et al., 2022b).

2.3.10 Conclusion

GNNs have grown into very important tools that explain complicated relationships representative of graph-structured data in a wide variety of different domains: social network analysis, molecular chemistry, transport networks, and recommendation systems. The ease with which GNNs handle non-Euclidean data, where traditional deep learning methods fall short, qualifies them as a major development in machine learning. GNNs have indeed become quite improved in terms of both predictive accuracy and interpretability by embedding structural and contextual information in graph-structured data into such models for finer inferences (Jin et al., 2024). A major advantage of GNNs is their versatility. Projects are being made in which the models will be able to perform node classification, link prediction, and graph-level prediction. Various methods are hence being developed, including GCNs[1], GATs, and GRNNs[2], all with different advantages depending on the nature of the graph and the problem in hand. As is known, in homogeneous graphs in which the edges tend to link nodes that have similar attributes, GCNs often do well. On other fronts, GATs have come with the development of attention mechanisms that create a difference in the importance of nodes to target nodes, hence their better interpretability and relevance in heterogeneous or irregular graphs. Despite the excellent performance achieved, there are still many challenges that encourage or even push researchers not to give up. Scalability is still a serious problem because large-scale graph training requires super-high memory and computation resources (Serafini et

[1] **Graph Convolutional Networks (GCNs)** are a class of neural networks designed to operate directly on graph-structured data. They generalise convolutional operations from traditional grid-like data (e.g., images) to irregular graph domains by leveraging the graph's adjacency matrix and node features to learn meaningful representations (Kipf et al., 2016).

[2] **Generalised Regression Neural Networks (GRNNs)** are a specific sub-type of radial basis function (RBF) network that is relied on to conduct non-linear regression tasks. GRNNs return predictions by approximating a target function based on a weighted average of observed outputs in the training dataset where weighted assignments of the training samples are determined based on influence via the Euclidean distance from input vectors to training samples (Specht et al., 1991).

al., 2021). Such techniques as sampling-based methods and layer-wise propagation relieve such constraints at the cost of model fidelity. Moreover, explainability issues are directly related to GNNs for their wide applicability to sensitive domains like health care and finance.

These are all techniques that help improve transparency from attention mechanisms, others from subgraph extraction, and gradient-based methods on how the model derived this prediction for stakeholders. More helpful has been the inclusion of some forms of explainability and visualisation methods, including feature importance heatmaps and attention-based visualisations. Explainability is quite helpful in those domains where interpretability is as crucial as the prediction accuracy, whereby the underlying reasoning of a model for a certain decision will yield more actionable insights (Kakkad et al., 1958). In perspective, future GNNs are likely to head towards scalability, robustness, and interpretability for a wide range of application sectors. Improvement in transferability is also likely since the emergence of research in graph representation learning will sooner or later empower GNNs towards generalisation across graphs and domains. While GNNs continue to develop, their role for machine learning will only continue to grow and fundamentally change the capacity of the models on how to use relational information to solve real-world problems that could not be resolved by the standard ways of artificial intelligence. Further developments can see GNNs filling the gap in predictive and interpretative modelling and further extend the horizons of artificial intelligence and data science.

2.3.11 A Review on GNN-based Lung and Breast Cancer Prediction

GNNs are gaining traction in cancer imaging research as a compelling alternative to traditional Convolutional Neural Networks (CNNs), particularly in tasks involving structural or relational information. In the context of breast and lung cancer detection via X-ray imaging, for instance, in breast cancer mammography, (Ahmedt-Aristizabal et al., 2021) introduced a dual-graph GNN framework for mammogram analysis, modelling both spatial adjacency and feature similarity graphs simultaneously. The study demonstrated improved lesion detection and diagnostic accuracy over CNN baselines by capturing richer relationships between patches in breast X-ray images. Another GNNs based model proposed by (Du et al., 2019) a hierarchical graph neural network to detect abnormal lesions from medical images by automatically zooming into ROIs. It focusses on mammogram analysis for breast cancer diagnosis for this study. The network consists of two graph attention networks performing a node classification to predict whether to zoom into next level; and a graph classification to classify whether a mammogram is normal/benign or malignant. The

model is trained and evaluated on INbreast dataset, and we obtain comparable AUC with state-of-the-art methods. Further, (Zhang et al., 2022a) demonstrates multi-scale graph wavelet analysis in histopathology can capture tissue contextual cues more effectively than grid-like approaches. Similarly, for lung cancer, (Saihood et al., 2024) researchers have proposed Multi-side Graph Neural Network-based Attention for Local Co-occurrence Features Fusion (MS-GNN-ALCFF). The framework comprises a multi-side graph construction layer (MSGCL) that computes informative texture features and captures spatial relationships from cross-sectional and longitude orientations of the nodule, creating two sets of nodes. The LIDC-IDRI dataset, used for training and testing the proposed Multi-side Graph Neural Network-based. The LUNGx dataset used as an unseen dataset shows that our model is generalisable compared to baselines.

These studies collectively illustrate how GNNs leverage structural and scale-aware representations, offering improved interpretability and potentially greater robustness in clinical imaging contexts.

Table 2.5: Summary of graph-based models in medical imaging: tasks, performance, limitations, and datasets/modalities.

Paper	**Task**	**Accuracy**	**Limitations**	**Dataset / Modality**
(Ahmedt-Aristizabal et al., 2021)	Survey of graph neural network architectures and their applications in medical diagnosis and analysis	N/A (survey)	Scarcity of clinical GNN applications compared to CNNs; challenges in graph representation, model interpretability, scalability, dynamics, data annotation efficiency, and generalisation to new domains	Broad range of modalities: MRI, rs-fMRI, EEG; covers functional connectivity, anatomical structure, and electrical-based analyses
(Du et al., 2019)	Mammogram classification (normal/benign vs malignant) using hierarchical graph neural network that simulates radiologist zooming behavior	AUC: 0.943, comparable to state-of-the-art methods on the INbreast dataset	Evaluated only on a single dataset (INbreast) with limited cases; model complexity due to hierarchical graph design; generalisability and clinical validation untested	INbreast dataset (410 full-field digital mammograms, with lesion segmentations and BI-RADS labels)
(Zhang et al., 2022a) MS-GWNN	Histopathological image classification of breast cancer, using graph-based model capturing multi-scale tissue structure via spectral graph wavelets	Accuracy: 93.75% (BACH dataset); 99.67% (BreakHis dataset) both outperform state-of-the-art	Evaluated only on two public datasets; histopathology-only modality; model complexity and scalability not discussed; lack of clinical validation or cross-dataset generalisability analysis; may not generalise to unseen datasets	BACH and BreakHis histopathology image datasets (graph-based representation at multiple scales)

Continued on next page

Table 2.5 – *Continued: Summary of graph-based models in medical imaging.*

Paper	**Task**	**Accuracy**	**Limitations**	**Dataset / Modality**
(Saihood et al., 2024) MS-GNN-ALCFF	Lung nodule classification using a multiside graph neural network-based attention mechanism to fuse local co-occurrence features	In LIDC-IDRI 87.17 ± 0.84 %, 91.01 ± 1.14 %, 88.6 ± 1.42 %, 89.7 ± 0.87 %,95 ± 1.22 % and 76.7 ± 1.4 % in terms of Accuracy, Precision, Recall, F1-score, AUC and MCC respectively. And 69.86 ± 2.4 %, 75 ± 2.7 %, 71.4 ± 3.1 %, 73.17 ± 3.7 %, 70.2 ± 1.36 % and 71 ± 3.8 % in LUNGx	Evaluated only on the LIDC-IDRI dataset; model complexity and scalability not discussed; lack of clinical validation or cross-dataset generalisability analysis	LIDC-IDRI dataset (3D CT scans of lung nodules) LUNGx (CT).

2.4 Model Explainability

Model explainability forms one of the main pillars in AI, which deals with the understanding, interpretation, and presentation of how a machine learning model is making a certain prediction. The ability to explain a model empowers safety, fairness, and increases trust in these key verticals: healthcare, finance, and self-driving. Since AI models are normally thought of as "black boxes," especially complex ones like deep neural networks, explainability adds transparency to gain trust in users and accountability to show stakeholders why certain predictions or decisions were made.

The importance of Model Explainability is that explainable models are trusted because their predictions can be interpreted by practitioners and checked for biases and other kinds of errors (Holzinger et al., 2022). For example, in medical diagnosis, explainability allows clinicians to understand the motivation behind the diagnosis the model proposed, which may enable them to verify its reliability and align it with expert knowledge. In finance, this would mean decisions on loan approvals or fraud detection conformed to regulatory requirements, thus avoiding unfair outcomes (Amann et al., 2020).

Explainability techniques, for the most part, involve two major groups; global and local explainability.

Global Explanation This would provide insight into how generally a model makes decisions on a dataset as a whole. Global interpretability methods include feature importance analysis, that is, the identification of which features systematically contribute most to the predictions. For instance, in the case of image classification, some of the pixels or patterns weigh more in defining the objects or classes (Das et al., 2020).

Local Explanation Tries to explain individual predictions and not general model behavior; examples of techniques here include Shapley values, derived from the cooperative game theory, and LIME-standing for Local Interpretable Model-agnostic Explanations. Then, of course, comes the case of Shapley values, which take every feature regarding a certain prediction understanding in the case of why a model classifies a loan application as risky (Kakkad et al., 1958; Das et al., 2020).

There are several explanation tools, which are supposed to be model and application-dependent. Commonly used techniques and tools include:

2.4.1 LIME (Local Interpretable Model-Agnostic Explanations)

A locally adopted technique that simplifies the complex model using some simple, interpretable models around an individual prediction. The key intuition behind LIME is that an explanation of a model's behaviour locally, so around a single data point example, an image, or a sentence (Benk et al., 2020). The following discussion explores LIME's explanatory mechanisms.

Molnar describes that (Molnar, 2020), LIME creates some variations of the instance to be explained. Suppose LIME was explaining a certain image; then it would make slight changes in that image, turning the image on and off, i.e., mask segments. In the case of this being an instance of text, it could remove or replace some words. The model is generating predictions for perturbations. For each of these perturbed versions of the instance to see what the effect of that change was. In observing how the prediction changes with each perturbation, LIME can be considered to collect insight on what part of the instance is important for the model's decision. These predictions of perturbations are used to fit the interpretable model-that is, a simple model, like a linear regression or decision tree performing an approximation of the original model in the vicinity of the instance, or in other words, a simpler model than the original one but expressing the main factors which determine the prediction on a local scale.

Finally, LIME interprets the coefficients of the surrogate model as feature importance scores. This may then give visual or numerical explanations by featuring the most influential features—the image segments, words, etc. of the said prediction. For example, in image classification, LIME might highlight specific regions of the image which are most critical features for classifying an object thus helping users know what exactly the model "sees." LIME can be described as "model-agnostic" because all models are applicable without considering their internal structure. For as long as the approach may depend solely on the model's predictions for some inputs, it is allowed to explain predictions from any classifier or regressor, including those very sophisticated forms such as neural networks, random forests, and gradient-boosted trees. Such versatility in explaining such a wide range of model types, text classification, image recognition, and the analysis of tabular data is possible under this approach (Molnar, 2020; Ribeiro et al., 2016).

$$\text{explanation}(x) = \arg\min_{g \in G} \; L(f, g, \pi_x) + \Omega(g), \tag{2.5}$$

The explanation, for instance x is the model g ∈ G, such that: $g \in G \cdot L(f, g, \pi_x) + \Omega(g)$, where the loss function L: measures the difference between the explanation and the actual prediction of the model f, while model complexity is kept low by Ω. $g \in$ G

is the family of possible explanations, such as the set of all possible linear regression models. The proximity measure defines how large the neighborhood around instance x we consider for the explanation. In practice, LIME only optimises the loss part. The user needs to select the complexity, e.g. selecting the maximum number of features the linear regression model may use (Molnar, 2020).

2.4.1.1 Advantages and Limitations of LIME

Table 2.6: Summary of LIME Advantages and Limitations.

Advantages	Limitations
Interpretability LIME gives intuitive explanations on locality and how different model features arrive at a certain prediction (Ribeiro et al., 2016).	**Approximation Accuracy** It is a local approximation done with a simpler surrogate model; therefore, it might not capture all complexities of the original model, especially in highly nonlinear regions (Ribeiro et al., 2016).
Flexibility It is model-agnostic and can be applied to any model or dataset (Ribeiro et al., 2016).	**Computational Cost** LIME generates perturbed samples for every instance, making it expensive for large datasets or complex models (Shi et al., 2020).
Limit scope to local explanations Zooming in on a specific prediction, LIME provides insights into how a complex model works in individual cases, not globally. (Garreau et al., 2020; Di Cicco et al., 2019; Dieber et al., 2020).	**Stability** Explanations depend on the sampling process and may vary for the same instance (Hu et al., 2018; Burger et al., 2023).

In all, LIME is very useful to have for model interpretability, especially for applications that require an explanation of individual predictions. However, the approximation accuracy and intensive computation required make the approach more viable when the explanation scale is targeted and instance-level rather than global and large-scaled, see (Table 2.6).

2.4.2 SHAP (SHapley Additive Explanations)

Shapley values are one of the axiomatic ways in cooperative game theory to explain predictions of a model by means of the area of machine learning. Invented by Lloyd Shapley in 1953 (Roth, 1988), Shapley values were supposed to equitably split payoffs received by players in any cooperative game-one's contribution to the whole. It incorporates game theory as it directly applies Shapley values, which allocate distributions of payout (i.e., predictions) to players (i.e., input features) in a fair way, which is famously a measure from cooperative game theory (Kumar et al.,

2020). The Shapley values provide, in the context of machine learning, a method of quantifying the contribution of every feature toward some given prediction, hence helping us understand how different features individually affect model outcomes. In other words, it answers questions such as: "What is the contribution of feature X to this particular prediction for different values of the other features?

2.4.2.1 Applications of Shapley Values in Machine Learning

The shapley value for each feature assigned to an instance can show whether the impact of each feature is positive or negative toward model prediction. Therefore, Shapley values are indeed very helpful in domains like healthcare, financial analytics, and legal domains where insight into each prediction is important (Kumar et al., 2020; Giudici et al., 2021).

Global and Local Interpretability Shapley values represent both the local and global perspective. For instance, level explanation describes at the instance level how different features contribute toward one prediction; in the case of global interpretability, the aggregation of Shapley values over a dataset showcases the general pattern of important features (Kumar et al., 2020; Giudici et al., 2021).

Fairness and Bias Detection Shapley values have started to be used increasingly for the detection of bias and the assurance of fairness in machine learning models. By investigating the contributions of sensitive features-like age, gender, or race, we can better understand and therefore mitigate biases in predictions, making sure certain groups are not unduly affected (Kumar et al., 2020).

2.4.2.2 Advantages of Shapley Values

Shapley values, when applied to machine learning, are sound, theoretically valid approach to interpreting and explaining model predictions. Conceptualised from cooperative game theory, the approach has been indispensable in finding the contribution of each feature in models that have reached incredible complexity. By measuring the contribution of every feature to a particular prediction, Shapley values confer a host of advantages, more so in high-stake fields such as health, finance, and legal domains where clear and trustworthy model explanations are more than necessary.

Theoretical Grounding on Fairness and Justifiability The big advantage with Shapley values is rooted in their grounding in game theory. Shapley values have been designed in such a way as to divide up the "payout" among the players due to the

numerous different contributions of each player toward the result, hence uniquely positioned to offer fair and justifiable attribution in machine learning contexts (Heuillet et al., 2022). This stems from four major axioms underlying fairness, efficiency, symmetry, dummy, and additivity, which ensure that the contributions are consistent and reliable. By enacting these principles, Shapley values thereby offer an objective metric that can justify biases and ensure that at each prediction, the value of each feature is equitably assigned in a very transparent way (Heuillet et al., 2022). The resulting mathematical rigor makes Shapley values extremely desirable in application domains where interpretability and accountability are crucial, such as medical diagnostics or financial risk analysis.

Model-Agnostic Flexibility This means that Shapley values are model-agnostic, which could be applied to any type of machine learning model without favor over the structure of one over another. Such flexibility enables practitioners to use Shapley values with linear models, decision trees, neural networks, ensemble models, and other model types, hence allowing a consistent solution for interpretability across algorithm types. It therefore means, in practice, even very diverse machine learning ecosystems can apply one and the same standardised interpretability method across the models to improve efficiency through not having to maintain multiple interpretability frameworks (Messalas et al., 2019).

Deep Local and Global Interpretability One of the major strengths of Shapley values in explanation is the capability to offer both local and global explainability. Locally, Shapley values attribute how much each feature has contributed to the prediction on a specific instance, hence enabling the practitioner to understand why a model arrived at a certain decision for a particular case. These kinds of explanations are important in health is, knowing why a model outputs a high probability of disease for an individual patient-for clinical decision-making. Aggregating Shapley values at a global level over a dataset provides the overall feature importance, showing how the features generally drive model outputs. Indeed, such duality in interpretability could be of immense value for operational interpretation, explaining individual predictions to stakeholders, and strategic decision-making by pinpointing important drivers of the predictions at the dataset level (Messalas et al., 2019).

2.4.2.3 Challenges and Limitations

While Shapley values are indeed a powerful tool for model interpretability, there are still a few limitations that remain regarding aspects of the application in machine learning. Most of these challenges relate to the major concerns about high computation demands, interpretability, scalability, sensitivity regarding model com-

plexity, and possible challenges in nonlinear contexts. What follows is an extended overview of some of the main challenges and limitations standing in considering the application of Shapley values:

Computational-Complexity and Time Consumption One of the severe limits to applying Shapley values is related to their usual heavy computational loads. Computing Shapley values is equivalent to averaging the contribution of a feature over all possible subsets of other features. For a dataset with n features, this requires n^2 evaluations.This makes computations grow exponentially with the number of features being evaluated, which may become prohibitively expensive for high-dimensional datasets or models with a great number of features, as seen in deep learning models necessary for image or text analysis. Libraries like SHAP, for instance, allow some approximations and sampling-based methods for efforts toward the reduction of this complexity. These approximations themselves can be slightly incorrect and may be computationally expensive even for real-world applications, which demand quick interpretability (Huang et al., 2024; Kraev et al., 2024).

Interpretability in High-Dimensional Models On the one hand, for large models, several different Shapley values of individual features can be very difficult to interpret meaningfully. Confronted with hundreds of features, users cannot always make out actionable insights nor recognise the most critical factors which are contributing to a model's predictions. These limitations will reduce the applicability of Shapley values to complex models such as CNNs for image processing or transformers for language tasks with thousands of model parameters (Kraev et al., 2024). The counterfactual explanation tools, such as SHAP summary plots, might provide such insight, but these start to become heavy and difficult to decipher for high-dimensional data. Besides, summing up the individual Shapley values into global feature importance rankings loses some of the nuanced interpretability about feature interactions.

Dependence on Model Complexity and Structure The central thing to say about Shapley values is that they are model-agnostic. This means they can be calculated with any kind of model. However, their interpretability may vary based on the model's complexity. For example, in models with strong feature interactions or highly non-linear, the resulting Shapley values might not always be intuitive or easy to interpret because they average over many different types of interactions (Huang et al., 2023e).

Working Difficulty with Temporal or Sequential Data Shapley values could be even more challenging to interpret for models trained on sequential or temporal data such as time-series data (Vega García et al., 2020). In this class of data, the

importance of a feature at a one-time step is usually dependent on the values that it had or will take at previous or future time steps, respectively. Since Shapley values investigate features independently for each instance, they can hardly take into consideration temporal dependencies. This can result in interpretations that either omit or diminish the complexity with which the model identifies time-based patterns or relationships in the data.

Complexity in Non-Linear and Strongly Interactive Models Sometimes, Shapley values provide less interpretability if the model is nonlinear with strong interactions, such as neural networks. Feature interactions in such models are normally highly complex and hardly interpretable on their own. Given that Shapley values average over all subsets, they may oversimplify the importance of features within those interactions and mask the important relationships between the variables. This is particularly a problem when one tries to explain deep learning models analysing images, speech or texts: non-linear transformations abound (Kraev et al., 2024).

Shapley values themselves, as an important tool for explainability, urgently need to be tamed regarding their significant computational complexity, interpretability in high-dimensional contexts, and even for the potential to be misinterpreted. Faced with such demanding factors, advances in computational methods like the SHAP method ensure that the application of Shapley values becomes increasingly usable. By being aware of and leveraging these limitations, Shapley values can continue to play an important role in interpretable, transparent machine learning in applications where trust and accountability in an AI system's decisions or outputs are paramount. However, in practice, one has to consider the context and the application requirements, whether Shapley values are the best interpretability approach for that particular use case.

2.4.3 Attention Mechanisms

Attention mechanisms, which were first used in the field of neural machine translation (Bahdanau et al., 2015), transformed deep learning by allowing the model to determine which input element it needs to prioritise while it is predicting dynamically. By weight, we mean that attention mechanisms allow the model to assign varying weights to input elements, which are task-relevant, rather than assigning a single, prescriptive weight for every element. In a static, or traditional, neural network architecture (e.g., feedforward or recurrent), all input elements are processed according to the same weights without regard to their relevance to the task. Attention mechanisms allow the model to use context to develop weights that are dependent on the context and, therefore, to highlight important features while sup-

pressing noise. Attention allows a neural network to mimic human cognition. Like human decision-making, attention reflects actions that depend selectively on the most informative stimuli.

The core mathematics of attention mechanisms is weighted aggregation, and, in the context of attention mechanisms, the first step is assigning a query (in translation, say a word) as a point of focus and calculating compatibility scores with all keys (whose being the input elements). Next, a softmax function is applied so that the model can compute attention weights in relation to the keys. The attention weights then describe how much each input value (the value associated with each key) contributes to the output (Galassi et al., 2020). Formally, for query q, keys K,and values V:

$$\text{Attention}(q, K, V) = \sum_i \text{softmax}(q^\top K_i)V_i \tag{2.6}$$

This enables the model to "attend" meaningfully to dependencies that are spatially or temporally distant—this ability is particularly revolutionary for sequences (e.g., text data) and structured data (e.g., images) (Galassi et al., 2020).

To summarise, attention mechanisms represent a major advancement in deep learning, where individual features, which are typically processed statically, are replaced with dynamic and context-aware computations. Attention mechanisms' mathematical elegance and empirical success continue to foster important advancements across various fields.

2.4.3.1 Types of attention mechanisms

The most critical part of the transformer model is used to compute the self-attention of a word with every other word of the sequence; that is how it captures the dependency of a single sequence. In NLP tasks like translation, sentiment analysis, and question answering, maintaining sequence order and understanding context are essential—self-attention excels at both by dynamically weighting the influence of each word on others. (Zhou et al., 2024).

Global and Local Attention Global Attention computes attention weights over *all* input tokens (e.g., words in a sentence, pixels in an image). considers every token in the sequence and high computational cost $O(n^2)$ complexity for sequence length n. It is used in some applications like transformer models (e.g., BERT, GPT) (Beltagy et al., 2020). Image captioning (attending to all image regions).

Local Attention restricts attention to a *window* around the current token (e.g., nearby words or pixels). $O(n \times w)$ complexity for window size $w \ll n$. Suitable for long sequences (e.g., genomes, high-res images). It is used in some applications

like longformer (Beltagy et al., 2020) for document processing. Speech recognition (local phonetic patterns).

The multihead mechanism Is effective in letting a model attend to information in various subspaces all at once by computation of several attention scores simultaneously, each with a different set of weights. This turns out to be very helpful in modelling several aspects of the data and partially explains the success of Transformers in NLP.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(h_1, \ldots, h_i, \ldots, h_n)W_O, \tag{2.7}$$

where Q, K and V are three linearly transformed matrices and h_1 represents an output value of the ith the attention head in MHA. W_0 The output from all of these attention heads gets concatenated and transformed by a linear transformation matrix (Zhou et al., 2024).

Hierarchical Attention Hierarchical attention can be helpful for document classification and similar tasks because there is a hierarchy involved in such data; for instance, words within sentences, sentences within documents. Here, the model applies attention at every level separately to understand the context of importance (Yang et al., 2016; Ates et al., 2023; Chirasani et al., 2022).

$$p = \text{softmax}(W_c v + b_c), \tag{2.8}$$

where p is the predicted probability distribution over the document classes. W_c A weight matrix associated with the classification layer. It maps the input feature vector v to the class scores. v The feature representation of the document (e.g., from a neural network or embedding layer). b_c A bias vector added to the weighted features to account for additional adjustments.

$$L = -\sum_{d} \log p_{dj}, \tag{2.9}$$

where L is the loss function (commonly cross-entropy loss) used to optimise the classification model. $\sum_d$ Summation over all documents in the dataset. $\log p_{dj}$ The log probability of the true class j for document d. Taking the logarithm helps handle probabilities and enhances numerical stability. This involves document classification, where a framework predicts the probability of each class, while cross-entropy loss guides the training of the model towards better performance (Zhou et al., 2024; Li et al., 2017; Yang et al., 2016; Ates et al., 2023; Hassanin et al., 2024; Brauwers

et al., 2021).

Cross-Attention Usually it takes the form of encoder-decoder attention and offers the possibility that generation may attend to the outputs of some encoder. For example, in machine translation, while generating every word in the target language, the decoder can attend to all encoded words of the source language (Ates et al., 2023; Hassanin et al., 2024).

2.4.3.2 Mathematics behind attentions

The major attention mechanisms contain three components: Queries (Q), Keys (K), and Values (V). Each is a vector, which helps in estimating the importance of one word or, for that matter, one data point concerning another. The following is the basic process for computing attention:

Computing Attention Scores The attention scores of a query are usually computed against all keys. This often involves the application of some similarity measure such as the dot product. Given a query vector Q and a key vector K a score is determined by (Brauwers et al., 2021; Brownlee, 2023):

$$\text{Attention Score} = Q.K^T, \tag{2.10}$$

Apply Softmax The softmax function normalises the attention scores into probabilities that sum up to one. This allows the model to focus more on some positions compared to others. Because the "logits" converted into a probability distribution that highlights the most relevant positions while suppressing irrelevant ones (Vaswani, 2017).

Weighting Values For every value vector V is scaled by the associated attention probability for it, allowing the model to "attend" to certain values in proportion to their importance.

$$\text{Attention Weights} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right), \tag{2.11}$$

generating attention weights by applying the softmax function.$\sqrt{d_k}$ is the square root of the embedding dimension d_k (Vaswani, 2017).

Aggregating Results Weighted values are summed for scoring to provide the final output for every query, hence obtaining a context-sensitive representation of the

input (Brownlee, 2023).

2.4.3.3 Applications of Attention Mechanisms

Attention mechanisms revolutionised most tasks in NLP. This enables the model, while translating, to concentrate on relevant fragments of a sentence, and it will allow them to find specific pieces of information in question-answering tasks. Its success, with the transformer model based mostly on attention, brought powerful language models such as BERT, GPT, and T5, bringing forward the state-of-the-art for NLP. **Attention in Computer Vision** Attention mechanisms have been extended to include vision applications, which help a model focus on parts of an image that are more important. Recently, vision transformers have leveraged self-attention mechanisms by considering an image as a sequence of patches, as proposed by Dosovitskiy (Dosovitskiy et al., 2021). In image captioning, this helps in aligning sections of an image with their descriptive words, as demonstrated in the 'Show, Attend and Tell' architecture (Xu et al., 2015).

Multimodal Applications Attention mechanisms also help to put together information coming from different modes into a coherent multimodal representation. The range of such applications is not limited but includes, among others, visual question answering or multimodal sentiment analysis, where a model has to consider either the textual or visual features (Brauwers et al., 2021; Nam et al., 2017; Niu et al., 2021).

Table 2.7: Summary of Attention Mechanisms Advantages and Limitations.

Advantages	**Limitations**
Attention mechanisms are able to capture long-range dependencies in sequences quite effectively; hence, the model is able to capture context that might span very long distances in tokens or frames (Vaswani, 2017). This is a key advantage over recurrent models, which work with vanishing gradients[3] over long sequences.	Computational and memory requirement in techniques of attention mechanisms, especially self-attention, the computations increase exponentially with increased length, which may be very costly for really long sequences (Zhu et al., 2020).

Continued on next page

[3]Gradients are the partial derivatives of the loss function with respect to the model parameters in backpropagation, and they provide the information the gradient descent needs to change parameters to minimise the error (Hochreiter et al., 1997).

Table 2.7 – *Continued: Attention Mechanism Advantages and Limitations*

Advantages	**Limitations**
Attention mechanisms allow parallelisation, especially in Transformer architectures, as described in Fig. 2.6, by processing inputs independently, hence making them very efficient and scalable for large datasets. Unlike recurrent models, attention mechanisms do not require sequential processing, which significantly speeds up the training process (Vaswani, 2017).	Difficulty with local features. While attention mechanisms are good at capturing global context, sometimes they capture insufficient fine-grained local features (Wu et al., 2020). As demonstrated by (Dosovitskiy et al., 2021) such hybrid models combining attention either with CNNs or RNNs help attenuate this issue, especially in tasks requiring the recognition of local patterns, such as in image processing.
Attention weights provide some transparency, at least in terms of what parts of the input have influenced the output. That is a reason why attention-based models are more interpretable, given that it becomes possible to identify which tokens or pixels were relevant for some particular prediction.	While attention provides interpretability, correctly interpreting the weights of attention requires domain expertise and careful analysis. Moreover, an over-reliance on attention scores as the only measure of the importance of features may be misleading for complex models (Tang et al., 2022; Niu et al., 2021; Veličković et al., 2017).
Attention is a general mechanism now adapted to serve in many domains, including, but not limited to, language and vision to multimodal. The ease with which this has been done speaks volumes of its robustness and general applicability to effective contextualisation across diversified data types (Brauwers et al., 2021; Niu et al., 2021).	

Attention mechanisms are rather a paradigm shift in how deep learning models efficiently pay attention to the parts of the input chosen by themselves. Flexibility, scalability, and domain generality together have allowed the attention mechanism to establish itself as an essential building block within modern AI systems. Whether it be improving state-of-the-art NLP, bringing explainability into computer vision, or even fusion of disparate data modalities, attention continues to shape how models process and interpret rich data. Despite this challenge from computational demands and interpretive nuances, active research and efficient attention architecture developments continue to hold a promise for their extendable use and usability in a wider range of applications.

2.4.4 A Review on XAI-based Lung and Breast Cancer Interpretation

XAI has long been advocated as a means to balance the advantages of highly capable yet opaque deep learning models with the clinical need for transparency, accountability, and trust. For example, a growing number of XAI methods-LIME, SHAP, Grad-CAM, and concept bottleneck models-are presently emerging as useful for applications in breast and lung cancer. Nevertheless, there are plenty of counterarguments that identify basic challenges regarding safely and effectively applying these techniques in the clinic. (Gurmessa et al., 2024) have done a systematic review of XAI in breast cancer diagnosis from mammography and ultrasound and identified 14 primary studies-which indicated an overall lack of evidence that XAI increased either user or clinician trust in the real world. Most studies were limited by dataset-related limitations and the lack of assessments of the relationship between explanation quality and clinical interpretation. (Rafferty et al., 2025) Researchers have another perspective on evaluating XAI approaches. They write about the need for clinical experts to be involved with producing interpretable and robust deep learning systems for medical domain contexts, specifically for lung cancer detection in chest x-ray studies. In summary, in 2023, AI methods have reason to be hopeful that many recent innovations in explainable AI (XAI) techniques are helping ameliorate the opacity of black-box models by increasing transparency. However, this work shows that existing post-hoc and text-based XAI methods within the medical domain have serious limitations. These XAI methods often fail to yield clinically preferred or reliable explanations, which can severely limit their application in health care domains. Particularly in high-stakes environments such as health care, where trust and precision of use matter. Additionally, broader critiques have suggested a continuous trade-off between precision and interpretability; that is, more interpretable models yield a lower accuracy performance, and being and post-hoc methods often can't fundamentally show the internal reasoning "like" a black-box model (Ansari et al., 2025). A new interpretable deep learning process, "DeepXplainer", was developed by (Wani et al., 2024) to detect lung cancer with explanations of predictions in one integrated hybrid deep learning model. DeepXplainer is a convolutional neural network plus XGBoost. XGBoost serves as a class label prediction following the automatic input feature development phase of "DeepXplainer" with the additional convolutional layers. In order to provide explanations or explainability for predictions, DeepXplainer use an explainable artificial intelligence method "SHAP".

Table 2.8: Summary of explainable AI studies in cancer diagnosis: tasks, performance, limitations, and datasets/modalities.

Paper	**Task**	**Accuracy**	**Limitations**	**Dataset / Modality**
(Gurmessa et al., 2024)	Systematic review of explainable ML (XAI) in breast cancer diagnosis from mammography and ultrasound images	14 primary studies identified; only 1 evaluated human confidence; 92.86% flagged dataset-related issues as gaps in research	Lack of quantitative performance benchmarking; trustworthiness not demonstrated; heterogeneous XAI methods; limited human-centered evaluation; no consensus on best practices	Mammography and ultrasound imaging; heterogeneous datasets across 14 primary studies
(Rafferty et al., 2025)	Lung cancer detection using an expert-guided concept bottleneck model (ClinicXAI) that provides clinically meaningful explanations	F1-score $>$ 0.9 (surpasses baseline InceptionV3) Explanations with higher clinical relevance than LIME, SHAP, XCBs, and CXR-LLaVA Substantially increased robustness to adversarial attacks (FGSM, PGD, SimBA)	Lack of exact numeric accuracy metrics (e.g., AUC or precision/recall) Evaluations limited to publicly available datasets and annotated subsets Adversarial robustness may not translate to all real-world perturbations	Chest X-ray images from MIMIC-CXR with pathology annotations from VinDr-CXR; model incorporates expert-defined clinical concepts from radiology reports

Continued on next page

Table 2.8 – *Continued: Summary of explainable AI studies in cancer diagnosis.*

Paper	**Task**	**Accuracy**	**Limitations**	**Dataset / Modality**
(Ansari et al., 2025)	Systematic review of XAI methods enhancing ML/DL models for breast cancer diagnosis, prognosis, and treatment decision support	Comprehensive comparative analysis of XAI approaches (imaging, genomic, clinical) includes LIME, SHAP, Grad-CAM. Emphasises XAI's potential to improve clinical decision-making and patient trust, but without quantitative benchmarks.	No quantitative accuracy metrics reported; heterogeneous study designs; lack of standardised evaluation; limited clinical validation of XAI efficacy.	Diverse data modalities across reviewed studies: imaging (e.g. mammography, ultrasound), genomic data, clinical records.
(Wani et al., 2024) DeepXplainer	Detecting lung cancer using a hybrid model (CNN + XGBoost) with SHAP for interpretability	Accuracy: 97.43 % Sensitivity: 98.71 % F1-score: 98.08 %	Evaluation limited to a single open dataset; external validation pending; model's generalisability unknown; explanation quality not user-evaluated	"Survey Lung Cancer" dataset (open-source); model applied to imaging or structured features (modality unspecified)

2.5 Large Language Model LLM

LLMs represent state-of-the-art artificial intelligence engineered to understand, generate, and even process natural language professionally. Examples include the GPT series developed by OpenAI, BERT developed by Google, along with other architectures using transformers (Chang et al., 2024). These have inspired new revolutions in natural language processing-from translation and summarisation of text to sentiment analysis and even content creation. LLMS are a class of Transformer-based architectures that had their formal introduction in the seminal 2017 paper (Vaswani, 2017). Transformers use mechanisms such as self-attention to understand how words relate contextually in a sentence, independent of word position. This makes them particularly effective at understanding meaning and relationships across large corpora of text more effectively than prior models using recurrent and convolutional neural networks.

From conventionally rule-based methods to the emergence of LLMS capable of emulating human-like behaviour regarding comprehension and generation of text, the realm of NLP has indeed been one of transformation (Li et al., 2024b). It goes ahead to expound more on such trends, focusing on those milestones and novelties that have shaped language models until now.

NLP originated in the 1950s, a time when systems relied a great deal on handcrafting, rules, and symbolic logic. Those rule-based systems parsed sentences using a priori-defined grammar rules. These were pretty rigid systems, which failed at managing the variability inherent in human language (Chang et al., 2024).

The great turn of events took place in the 1980s, when statistical methods overthrew rule-based systems. Models such as n-grams, hence, introduced probabilistic methods that allowed the modelling of the next word based on its likelihood in that sequence. Consider a trigram model, given the two preceding words, it would estimate the probability of the subsequent word. Intrinsically, while statistical models were quite straightforward, they had some problems; they could not learn long-range dependencies because they considered only fixed-size contexts (Noguer i Alonso, 1950).

2.5.1 Core Architecture of LLMs

LLMs represent the pinnacle of improvement in AI, allowing machines to perform even the most complicated NLP tasks. Their architecture provides the very basis on which such feats are possible. Processing, comprehension, and generation of text that approximates human languages.

Until recently, incremental developments in the technology behind LLMs have occurred due to the transformation. Transformers replaced earlier sequential models,

like Recurrent Neural Networks and Long Short-Term Memory Networks, which couldn't parallelise and suffered from long-range dependencies (Vaswani, 2017).

2.5.1.1 Core Components of Transformers

In addition to what has been mentioned in the "Attention Mechanism's types" Section 2.4.3.1, here we have some other important components as follows:

Feed-Forward Networks FFNs are the basic building blocks of the Transformer architecture as described in Fig. 2.6, and their contribution to the processing and transformation of data between attention layers is vital. Feed-forward refers to a neural network, artificial, in which connections among nodes do not contain any cycles. This is in contrast to recurrent neural networks. The network is composed of an input layer, followed by one or more hidden layers, and then an output layer. Information moves only one way, from input through to the output end-hence the name "feedforward." (goura, 2024). Every transformer layer consists of an FFN, independently applied to each token, word, or subword of the sequence, but computed in parallel across all tokens for substantial efficiency and capturing complex relationships in data (Sonkar et al., 2023).

A typical Feedforward Neural Network inside the transformer architecture would have an architecture of two linear layers separated by a nonlinear activation. Mathematically, this would look something like (Vaswani, 2017):

$$\text{Output} = max(0, xW_1 + b_1)W_2 + b_2, \tag{2.12}$$

where x is the input vector for each token. W_1, W_2 is weight matrices for the linear transformations. b_1, b_2 are bias terms. ReLU rectified Linear Unit, a common non-linear activation function.

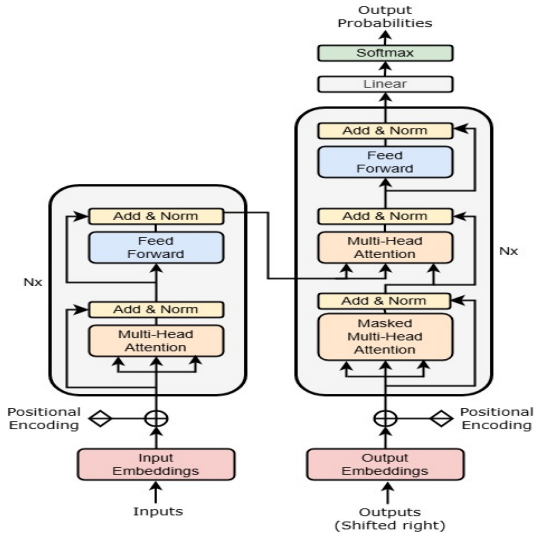


Figure 2.6: The Transformer- model architecture.

The size of a hidden layer provides its dimensionality, which is usually larger than either the input or the output. That way, a network can learn to model complex interactions among inputs before reducing the representation back to its original size. Although the linear transformations are identical in every position, they do incorporate different parameters from layer to layer. Another way of describing this is as two convolutions with kernel size 1. The dimensionality of input and output is $d_{model} = 512$, and the inner-layer has dimensionality $d_{ff} = 2048$ (Vaswani, 2017). The architecture of the feedforward neural network is divided into three main layers as follows:

Input Layer It is so named because it contains neurons that take in the input data. Each neuron in the input layer corresponds to a single feature of the input data.

Hidden Layer This model contains at least one hidden layer between the input layer and the output layer. These layers basically learn the complex pattern of data. Each neuron of a hidden layer computes the weighted sum of inputs, followed by non-linear activation.

Output Layer This is the layer that gives the output of the network. The number of neurons here depends on the number of classes involved in the classification problem or the number of outputs for the regression problem. The association of every layer in neurons has a weight that is modified in a network during training, in such a way that the error of the prediction decreases (goura, 2024; Bebis et al., 1994).

A weight is associated with each connection between neurons to adjust the loss value and minimise it while the model trains, as shown in Fig. 2.7.

Positional Encoding Transformers are different from earlier sequence-based models,

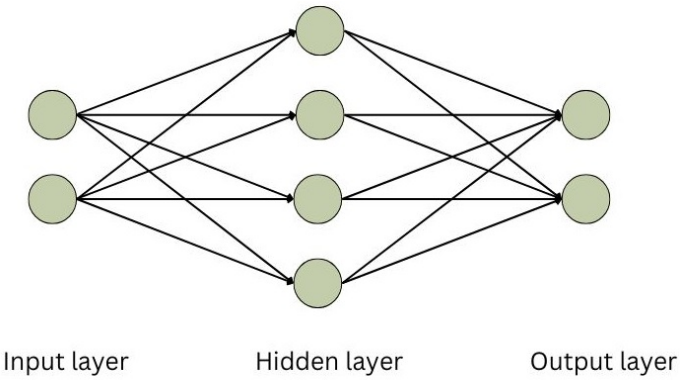


Figure 2.7: Feed Forward Neural Network

such as RNNs or LSTMs. RNNs inherently capture the order of tokens because they process the tokens one at a time. However, Transformers process all tokens simultaneously using self-attention mechanisms for efficiency, but in a position-agnostic manner. This deficiency is treated by adding some positional encoding to the input embeddings that shall explicitly provide the position of the tokens. These are fixed-size vectors, one for each token in an input sequence, which are word embeddings. In addition, positional encoding adds one more vector to this, which should indicate the position of that very token in that sequence. Then the concatenated vector gets fed into the model.

This carefully constructs the positional encoding vector to capture the information about the position in a way that is important but computationally efficient (Vaswani, 2017; agar, 2024).

$$PE(p,\ 2i) = \sin\left(\frac{p}{10000^{2i/d}}\right), \tag{2.13}$$

$$PE(p,\ 2i{+}1) = \cos\left(\frac{p}{10000^{2i/d}}\right), \tag{2.14}$$

where PE(p, 2i) Using the sine function to represent the even-indexed dimensions of the positional encoding. PE(p, 2i+1) Using the cosine function to represent the odd-indexed dimensions of the positional encoding. p is the position of the positional encoding, i is the dimension index, and d is the total dimensionality of the embedding.

These involve sine and cosine functions in order to generate wave-like patterns that oscillate over the sequence positions. The usage of sine and cosine ensures that a set of features can express the position information effectively over a wide range of sequence lengths, taking sine for even indices and cosine for odd indices (agar, 2024).

Positional encoding is a key concept in Transformer architectures; it provides

a balance between parallelisation and the need to understand order in a sequence. Transformers embed position into token representations directly, therefore giving the best performance to date on a wide range of NLP tasks. Continuous innovations, including relative and dynamic encodings, keep pushing their capabilities further with respect to longer sequences and more challenging applications.

Layer Normalisation and Residual Connections Layer Normalisation is a technique to stabilise, and therefore speed up the training of deep neural networks, by normalising the inputs to each layer over the feature dimension (as opposed to batch normalisation, which is done over the batch dimension). LayerNorm normalises the input so that the scale of the input remains consistent, and this helps to avoid both exploding and vanishing gradients. It behaves much better on generation tasks where batch size changes (Lei Ba et al., 2016; Huang et al., 2023d).

LayerNorm (Ba et al., 2016) calculates mean and variance for each input data point independently (Rodrawangpai et al., 2022); therefore, it is particularly useful in cases where:

1. Sequential Data: As it does not depend on batch statistics, LayerNorm is ideal for variable length sequences (for example, tasks in NLP or Time Series).

2. Stability: Layer norms also serve to reduce covariate shift[4], therefore retaining the same input scale per layer, even when using small batch sizes or conducting online learning.

As (Lei Ba et al., 2016) expressed in their paper, "Layer Normalisation" in deep neural networks is able to make their training model converge faster to a better result, therefore improving the performance on different tasks such as machine translation and object detection. The process of calculating layer normalisation could be expressed as a dot product "$\cdot$". Below is a formula showing the scaling and centering of the inputs.

For an input vector $x = [x_1, x_2, x_3, .., x_d]$, the normalised output y is computed as:

$$y_i = \frac{x_i - \mu}{\sigma} \cdot \gamma + \beta, \tag{2.15}$$

where x_i Mean of the input features.

[4]Covariate shift is a case of a machine learning application where the input data distribution (i.e., covariates/features) shifts between training and test time, while the conditional of the labels given the inputs $P(y \mid x)$ stays the same. The mismatch would result in decrease of performance on the model, as this model is trained from non-stationary distributions; the input data distribution at deployment will not be the same as the one it was trained on (Shimodaira, 2000; Ioffe et al., 2015)

$$\mu = \frac{1}{d}\sum_{i=1}^{d} x_i, \tag{2.16}$$

the standard deviation of the input features.

$$\sigma = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(x_i - \mu)^2}, \tag{2.17}$$

where $\gamma + \beta$ are learnable parameters (scale and shift) to allow the network to adjust the normalised output.

Residual Connections The residual connections create a skip shortcut in which the input of a specific layer can skip the operation of that layer and be directly added to its output; thus, establishing a direct route of flow and making the vanishing gradient problem minimal, while training deep networks (Xu et al., 2024a; Zhang et al., 2024).

To calculate the Residual Connection (output) of a layer with input x and output sublayer (x), the residual connection is computed as:

$$\text{Output} = x + \text{Sublayer}(x), \tag{2.18}$$

where x is the input to the sub-layer, and Sublayer x represents the output of the sub-layer (e.g., self-attention or feed-forward network)(Zhang et al., 2024). The main advantages of using Residual Connection:

1. Because the residual connection is the implementation of RL (Residual Learning) Recent times have found RL to be an effective and efficient methodology to help backward propagate the gradients computationally and hence being cost-effective. This will be primarily used for learning deep hierarchical representations and temporal dependencies (He et al., 2016).

2. Allows the flow of gradients to provide shortcuts in which gradients can flow directly backwards to earlier layers without any reduction in magnitude (He et al., 2016).

3. Enhances information propagation. The input is preserved, so that the model can use both the original representation and the transformations provided by the sublayer (He et al., 2016; Veit et al., 2016).

4. Accelerating convergence to make it easier for the model to learn identity mappings, helping layers learn faster at the time of training (Zia et al., 2020; Dolz et al., 2018; Basnet et al., 2021).

Layer normalisation and residual connections are very important to the effectiveness of the Transformer architecture. They bring stability, efficiency, and allow efficient training of deep networks. In particular, they allow Transformers to model complex sequences amenable to applications ranging from language modelling to translation, among others.

2.5.2 Popular Large Language Models

This section considers the potential influence and application of two of the most famous LLMs in the medical field. Such models have a great contribution to the development of health-related NLP tasks, which include, but are not limited to, medical report generation, patient query responses, translation of complicated medical terms, summarisation of research papers, and other related activities. Their versatility, combined with the recognition and generation of human-like text, will make them invaluable tools in the areas of efficiency and accuracy within medical communication, diagnosis, and education (Nazi et al., 2024).

2.5.2.1 GPT (Generative Pre-trained Transformer)

GPT has been another big step toward making much more advanced systems possible to understand and generate human-like texts. The GPT model follows the architecture of a transformer, and its deep neural network showcases phenomenal performance in the domains of generation, summarisation, and question-answering. It has since evolved into one of the most prominent and strong models for NLP, followed in almost all research and industries due to its robust performance and conversational aspects (Wang et al., 2019; Mars, 2022).

It is considered a pre-trained language model, which undertakes training on large textual databases, and then it is fine-tuned for specific language-related activities Fig. 2.8. Examples include the GPT series by OpenAI and BERT by Google. In particular, BERT is pre-trained on the Wikipedia and books data with over 3.3 billion tokens and is mainly applied to NLU tasks like text classification (Wang et al., 2019; Zhu et al., 2022a; Sung et al., 2019).

However, as a masked language model, it could learn only the contextual representation of words, whereas in generating coherent and organised text, it was deficient, thus limiting its further use in NLG tasks (Zhu et al., 2022a).

In contrast, GPT has emerged as the supreme NLG model due to its autoregressive properties, which allow the prediction of text in sequence and therefore text generation (Mars, 2022). This property makes GPT especially suited for creating creative content supporting complex conversational AI and automating reporting tasks and other long-form summarisations. Its architecture includes unsupervised

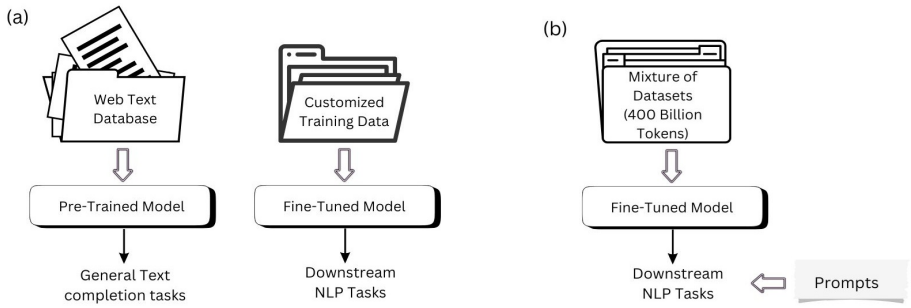


Figure 2.8: Training and re-training process of GPT-2 (a) and GPT-3 (b)

pretraining on large corpora, followed by task-specific fine-tuning, resulting in highly flexible models that can be applied to diverse tasks. The GPT model can be applied to diverse applications like healthcare, education, and customer service.

GPT-2 is a two-step training strategy: pretraining and fine-tuning. The workflow is represented in Fig. 2.8(a). In the pre-training stage, the model is trained on a text dataset compiled from millions of web pages (Radford et al., 2019). However, for any other NLP tasks, one needs to fine-tune the pre-trained model with his/her own custom-built and task-specific dataset. The fine-tuning procedure constitutes a process of training the model through iterative updating of the gradients of a big dataset consisting of examples from the particular task. It updates the weights of the pre-trained model by saving and adopting them in the target task. However, sometimes it may be impossible or complicated to have a large dataset fitting the target NLP task. So far, GPT-3 is the largest language model. It was trained on a mixture of datasets with 400 billion tokens and has up to 175 billion parameters. Compared to its predecessor, GPT-3 allows a few-shot learning: the model learns to perform any of the NLP or NLG tasks from multiple examples of that task-prompts (Brown, 2020). In the process, no update in the gradient takes place. The training process of GPT-3 is depicted in Fig. 2.8(b).

Table 2.9: Summary of GPT Advantages and Limitations.

Advantages	**Limitations**
GPT is based on the transformer architecture. A neurally designed model that controls input in a sequential manner. It uses an auto-aggressive in that, considering the past elements, it predicts the next token in sequence (Zhu et al., 2022a).	GPT learns biases from the content it was trained on, usually the internet. This can be extended even further in order to get biased or prejudicial output, nearly always on sensitive topics related to gender, race, and politics. The knowledge cutoff of GPT models is fixed by the data on which these models are trained. As such, GPT-4 has only updated knowledge up to 2023 and thus cannot perform on real-time or recent information (Brown, 2020; Yenduri et al., 2024).
The GPT models are first trained on very large-sized datasets sourced from books, websites, and other public repositories of text. This pre-training enables GPT to develop a broad understanding of structure in language, grammar, semantics, and context (Yenduri et al., 2024).	Tracing or interpreting why a particular output has occurred could be complex and prohibitive of trust in key applications like health or legal advisories. GPT is considered a "black-box" model, meaning that its decision-making process is not transparent (Yenduri et al., 2024).
It uses a unified architecture to handle diverse NLP tasks like: Translation, Summarisation, Text classification and Question answering (Mars, 2022).	It becomes difficult to explain why a particular output has been produced, thereby eroding confidence in its use for critical applications like health or legal advice. GPT models require considerable computational resources both for pre-training and fine-tuning, hence making them unreachable for smaller organisations. Large models, such as GPT-4, have the slowness inherent in the complexity (Yenduri et al., 2024).
The GPT series is continuously evolving by incorporating models of an even bigger size: GPT-2, GPT-3, and GPT-4-all have billions of parameters. While the model continues to grow in size, the understanding and generation become increasingly subtle (Yenduri et al., 2024).	Large computational resources are needed for pre-training and fine-tuning GPT models, making them unreachable for smaller organisations. Larger models, like GPT-4, are bound to be slower while generating responses (Yenduri et al., 2024; Jeong, 2024).
The domain adaptation of GPT in various domains such as healthcare, law, and education is possible with fine-tuning of domain-specific datasets that allow for: medical report generation and legal document anlysis (Brown, 2020).	While GPT does support multiple languages, it is generally better in English, since most of its training data is in English. The latter may lead to poor performance of underrepresented languages from the training dataset (Brown, 2020).

Continued on next page

Table 2.9 – *Continued: GPT Advantages and Limitations*

Advantages	**Limitations**
Can be fine-tuned GPT on task-specific datasets to adapt to specialised applications, such as: Sentiment analysis (Li et al., 2024c).	Small, fine-tuned versions of the GPT model may be too big for edge devices (IoT, mobile) and have to be distilled or pruned.

2.5.2.2 BERT (Bidirectional Encoder Representations from Transformers)

Google's BERT revolutionised language modelling and turned out to be an instant breakthrough in the area of NLU. Proposed by Kenton and Toutanova in the year 2019, it is based on the transformer architecture Fig. 2.6, which means it processes the text context bidirectionally. It processes words coming before and after the target word in a sentence, and thus captures the deep meaning of language semantics more than the previous models (Kenton et al., 2019).

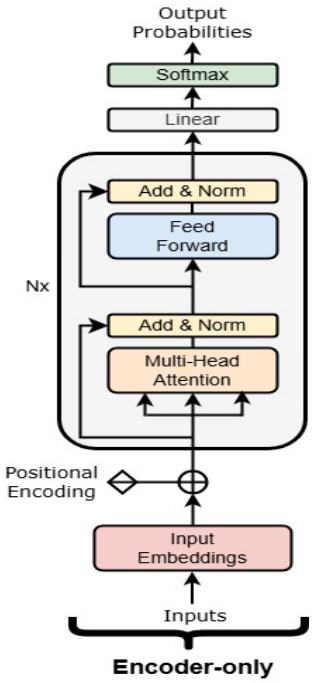


Figure 2.9: BERT Architecture

BERT is a stack of transformer encoder layers, each consisting of multi-attention "heads" (Vaswani, 2017), for each input token in a sequence; each of these heads calculates key, value, and query vectors, which are used to compute a weighted sum. All the heads for a given layer are concatenated and then passed through a fully connected feed-forward layer. This is followed by a skip connection and layer normalisation. The conventional BERT workflow consists of two parts: pre-training and fine-tuning Fig. 2.10. Pre-training involves two self-supervised tasks: MLM,

which predicts a randomly masked input token, and NSP predicts whether two input sentences are immediately next to each other. Fine-tuning for downstream tasks mostly includes adding one or more fully connected layers on top of the last encoder layer (Rogers et al., 2021).

These representations are computed through tokenisation of each word of the input into wordpieces (Wu et al., 2016); then, token, position, and segment embedding layers are all together to give one fixed vector representation. The special token [CLS] is used in classification predictions, while [SEP] separates input segments. Google and HuggingFace have come up with a number of BERT variants. They have a base and large version; some differ in head-count, layer count, and hidden state size (Wolf et al., 2020).

BERT Key Features

- BERT can capture and process information contextually in both directions, left to right and right to left (Kenton et al., 2019).
- It is designed so that during fine-tuning on most any NLP task, the architecture will be changed as little as possible, Versatile across tasks (Jeong, 2024).
- It makes use of the masked language modelling objective during pre-training, where it randomly masks some words in a sentence and trains the model to predict those tokens that were masked. In so doing, it nurtures the ability to handle context and relationships within the sentence, a process called the masked language model (MLM) (Wang et al., 2019; Wang et al., 2019).
- BERT can also predict two given sentences while training adjacent to each other in the original document. This especially made sense in tasks such as question answering or sentence similarity, where one needs to understand the level of the sentence, which is the prediction of the next sentence (Wang et al., 2019).

2.5.3 Training Large Language Models

The training of LLMs is very computationally expensive and forms the backbone of language comprehension and generation capabilities. These models leverage state-of-the-art deep learning methods and study thousands of texts to unravel complex patterns ingrained in the language. Their training is multi-staged, Fig. 2.10, starting from the pre-training stage on very large datasets, and goes all the way into fine-tuning for very specific applications (Lund et al., 2023; Shahab et al., 2024).

This is actually a very critical training process for LLMs, as it enables them to perform a wide range of activities, including text summarisation, translation,

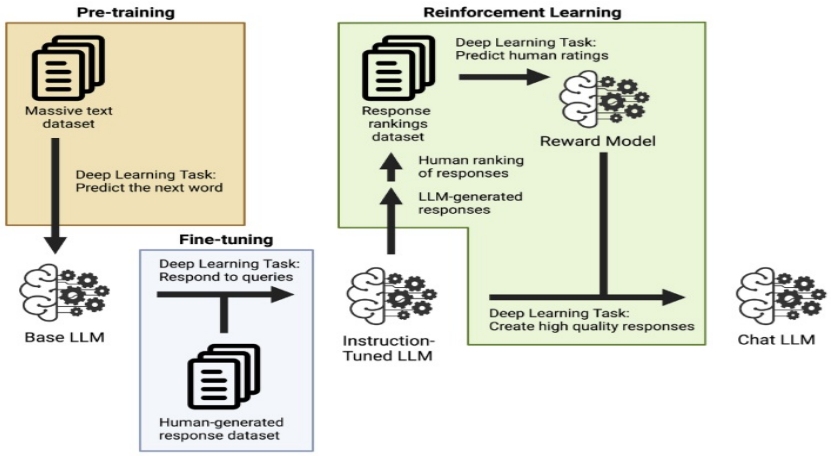


Figure 2.10: LLM training process steps. Pre-training models on enormous texts for the prediction of word sequences. Fine-tuning with paired instructions and responses for following instructions. The development of a reward model through human-ranked outputs that test the quality of responses. The reward model is combined with the fine-tuned model to yield a large language model optimised for high-quality text.

question answering, and even the generation of text. LLMs are trained on large and diversified datasets that develop an in-depth understanding of the structure of languages, including grammar, semantics, and context, which allows them to respond coherently and relevantly, across diverse domains. Furthermore, this process allows LLMs to generalise over different linguistic styles and be applied to various languages; thus, they are very handy assistants for communication in today's world and technology (Mars, 2022; Yenduri et al., 2024).

LLMs require gigantic training datasets composed of billions of words, sentences, and ideas. Such datasets, in turn, are mostly sourced from different types of sources, including books, encyclopedias, research papers, social networks, and online forums. The quality, diversity, and size of these sets will underpin the capability of the model. Quality ensures the model picks up important language patterns, not simply noise. Diversity allows the model to abstract better across different contexts, styles, and domains. It is directly proportional to the richness of the model's understanding and its capability to generate relevant outputs (Brown, 2020).

Training large language models will continue with the development of high-performance computing. This process utilises contemporary, state-of-the-art hardware specially optimised for parallel calculations in deep learning, including modern design Graphics Processing Units and Tensor Processing Units. Distributed computing uses clusters of GPUs or TPUs to divide the computation and parallelise the load across multiple machines. Using techniques such as Adam optimisers and learning rate scheduling allows a model to converge while balancing computational efficiency. In addition, training would involve sophisticated memory management

strategies along the lines of gradient checkpointing and mixed precision, since modern LLMs may contain billions of parameters (Nikolić et al., 2022).

2.5.3.1 Pre-training

The model parameters are randomly initialised. After that, the model is built, and it is trained based on a large amount of unlabeled data to learn the features from the corpus in general.

The pre-training aims to learn, from large volumes of unsupervised data in the form of text, the underlying structure and the semantic properties of the language (xinyuwei, 2023).

In unsupervised learning, the model picks up general language representations in large unannotated datasets. It would enable the model to learn statistical properties and relationships native to the syntax, grammar, and semantics of language. This part of the process represents the most demanding part in terms of resource consumption in training and, accordingly, in the cost it involves; therefore, one would typically avoid redoing it unless really necessary or when adjusting for better performance in a certain language. For most models, however, the pre-training could not be applied to NLP tasks directly, since, for example, the skill of predicting a masked word in context is rarely useful in practical situations. This is why a separate fine-tuning stage is needed after the pre-training. Fine-tuning represents a much less resource-intensive process; usually, just a few hundred examples are enough. It does often require, the use of different fine-tuning methods regarding the different tasks at hand, be it multi-label text classification, extractive question answering, or text generation (Homoki et al., 2024).

2.5.3.2 Fine-Tuning

Fine-tuning is a technique based on a pre-trained model, like a series from OpenAI called GPT. Further training is done on a smaller, domain-specific dataset. It leverages prior knowledge the model has captured and provides higher performance for select tasks with a lot less data and computation. Fine-tuning transfers the pre-trained model's learned patterns and features to new tasks; thus, it performs well with fewer training data. It is gaining popularity in the field of NLP concerning text classification, sentiment analysis, and answer (Parthasarathy et al., 2024).

Fine-truning Types

The process of fine-tuning very LLMs covers a variety of techniques to optimise pre-trained models for specific tasks (Kenton et al., 2019). All of these methods provide a way to customise LLMs to problems, while balancing performance trade-

offs, cost-per-task, or optimisation for task-specific tasks (Howard et al., 2018).

Unsupervised Fine-Tuning This approach does not require labelled dat;a hence, the LLM is simply exposed to a huge amount of unlabeled text from the target domain for its improvement in language understanding (Gururangan et al., 2020). This approach proves beneficial for newer domains like law or medicine, but gives comparatively less accuracy in certain tasks, such as classification or summarisation (Brown, 2020).

Supervised Fine-Tuning (SFT) SFT trains the LLM for the target task using labelled data. Fine-tuning an LLM on a text classification task that targets business texts would require a dataset consisting of text snippets annotated with their corresponding class labels (Kenton et al., 2019). While this approach works well, it does depend on substantial amounts of labelled data, which can be costly and sometimes hard to obtain Howard et al., 2018.

Instructional Fine-tuning via prompt engineering It relies on providing the LLM with instructions in natural language, which is particularly helpful in developing specialised assistants. While it limits the requirement of lots of labelled data, the approach heavily relies on the quality of the prompts (Parthasarathy et al., 2024; Ye et al., 2023). Prompt engineering thus tries to find the textual prompt p^* that gives the best performance on a given dataset D when using a given LLM M_{task} as the task model. More precisely, we assume that all datasets are able to be formatted into text input-output pairs, i.e., let $D = (x, y)$. In this regard, we present a prompt development set D_{dev}.

$$p^* = \arg\max_{p} \sum_{(x,y)\in D_{\text{dev}}} f(M_{\text{task}}(x; p), y), \tag{2.19}$$

where $f(M_{\text{task}}(x; p), y)$ is the output from the task model given prompt p and f is a per-example evaluation function (Ye et al., 2023).

2.5.4 Evaluation Techniques

When assessing LLMs, all facets of performance, robustness, and safety must be evaluated, and many different ways to assess these facets are available. The regular ways are to benchmark on tasks with standardised benchmarks (e.g GLUE or MMLU)[5].

[5](GLUE) General Language Understanding Evaluation, to evaluate general NLP capabilities. (MMLU) Massive Multitask Language Understanding tests multitask knowledge in 57 subjects (Wang et al., 2018a; Hendrycks et al., 2020). Using qualitative metrics such as coherence and relevance, or adversarial tests to reveal biases or vulnerabilities. New methods, including self-

Table 2.10: A comparison between pre-training and fine-tuning, emphasising their distinct features and processes.

Aspect	**Pre-training**	**Fine-tuning**
Definition	Training on a vast amount of unlabelled text data	Adapting a pre-trained model to specific tasks
Data Requirement	Extensive and diverse unlabelled text data	Smaller, task-specific labelled data
Objective	Build general linguistic knowledge	Specialise model for specific tasks
Process	Data collection, training on large dataset, predict next word/sequence	Task-specific data collection, modify last layer for task, train on new dataset, generate output based on tasks
Model Modification	Entire model trained	Last layers adapted for new task
Computational Cost	High (large dataset, complex model)	Lower (smaller dataset, finetuning layers)
Training Duration	Weeks to months	Days to weeks
Purpose	General language understanding	Task-specific performance improvement
Examples	GPT, LLaMA 3	Fine-tuning LLaMA 3 for summarisation

2.5.4.1 Quantitative Metrics

Quantitative metrics are numerical indicators used to objectively assess the functionality, efficiency, and quality of LLM models. The most commonly used metrics that have been used in LLM models are as follows:

BLEU (Bilingual Evaluation Understudy) Evaluates text generation by comparing model output with a reference text in respect to their n-gram overlap. (Huang et al., 2023b).

ROUGE, Recall-Oriented Understudy for Gisting Evaluation It calculates the quality of summarisation by checking for overlap with human-written summaries. It describes different automated scoring methods that measure the similarity of

assessment (e.g. truthfulness probes) and situation-based evaluation, can sample the complexities of actual deployment (Ribeiro et al., 2020; Lin et al., 2021)

summaries (Lin, 2004).

F1 Score It measures the precision and recall, especially in tasks like classification and named entity recognition ER.

Perplexity It characterises how well a language model is predicting a sample, and lower perplexity means better performance (Parthasarathy et al., 2024).

2.5.4.2 Human valuation

Quality assessment of the generated text, whether from a language model or authored by a human, is one of the most historic daunting tasks, which has always attracted wide attention (Chan et al., 2023). Human judges assess outputs for qualities like relevance, coherence, creativity, and appropriateness. Often used in subjective tasks like storytelling or open-ended question answering.

2.5.5 Challenges and Limitations of LLMs

Despite such tremendous possibilities for transformation, LLMs also raise many challenges and limitations on various levels, be it technical, ethical, or social, which require reflective deliberation in order to ensure full benefit while limiting injury.
Computation Cost LLM training requires massive computation resources such as GPU and TPU, which are really expensive and out of budget for many organisations (Zhu et al., 2022a; Sun et al., 2024). AIM-X experiments were conducted in Google Colab and also high-performance virtual machines (VMs) at the university, Section 3.2. The training alone took five days to train, and this was across 3,247 data points, including X-Ray images with their relevant radiology reports. Time was also spent on hyperparameter tuning with a genetic algorithm. The training/tuning process highlights the amount of computational power needed to optimally utilise large language models (LLMs) and get the best performance.

Energy Consumption This takes a lot of energy to train, which has major environmental impacts and increases carbon footprints (Sun et al., 2024). In the last job I mentioned in the first point, I was required to continuously run the process for training the model. The server was running for an extended period, more than 5 days, for training and tuning the model.

Training Data Biases and Data Quality Normally, LLMs are trained on uncurated large datasets that can be biased because of the data itself, for example, gender, racial, and cultural biases can be present. Results may therefore be discriminatory

or even hurtful. Noise, redundancy, and inconsistencies in training data can make models perform differently (Parthasarathy et al., 2024).

Domain knowledge LLMs may lack domain-specific knowledge, for instance, in legal and medical fields, unless specifically fine-tuned with domain-specific data (Brown, 2020; Parthasarathy et al., 2024).

Overfitting and memorisation could happen with LLMs having them overfit the training data by memorising specific details, not generalising well, may lead to inappropriate responses or irrelevant ones (Prechelt, 2002).

Lack of Explainability The black-box nature of LLMs is not interpretable; hence, it is difficult to comprehend such decisions, making debugging and building trust for critical applications complicated (Hsieh et al., 2024).

Regulatory Oversight LLMs started to hold greater power across manifold industries; their oversight by way of regulation became highly critical. This will ensure that these kinds of powerful technologies will be developed, deployed, and used responsibly with ethical principles and conception of societal values. If left unwatched, misuse, bias, and other unintended consequences could definitely outweigh their benefits (Parthasarathy et al., 2024; Meskó et al., 2023).

Energy Efficiency Energy efficiency is key in the development and utilisation of LLMs. LLMs do magic while executing their tasks; however, they act as the major consumers of energy. Training and running such models, especially those sophisticated ones like GPT and BERT, requires heavy computation; hence, they consume very high levels of energy. This becomes a concern for many in light of the environmental repercussions it causes, especially now when most of the world is striving to fight climate change and head towards sustainability (Parthasarathy et al., 2024; Stojkovic et al., 2024).

Under the specified hardware configuration (a Dell PowerEdge R750xa server with an NVIDIA A100 80GB set up), on a continuous operating basis (168 hours/7 days), the system would consume approximately 134.4kWh of electricity. Therefore, a carbon footprint would be approximately 28.2 kg CO_2 (National Grid ESO, 2024), or the equivalent of: Driving a petrol vehicle 224 km in the UK (based on average UK emissions of 0.126 kg CO_2/km), or powering a typical UK home for 5.6 days (based on average daily usage of 24kWh). The estimates were based on 800W (0.8kW) average power draw when using the system under conditions typical of a normal AI workload (Dell Technologies, 2023). This calculation is based on:

$$0.8\,\text{kW} \times 168\,\text{hours} = 134.4\,\text{kWh}, \tag{2.20}$$

where (0.8kw) is the Avg power draw under a typical AI workload.

$$134.4\,\text{kWh} \times 0.21\,\text{kg}\,CO_2/\text{kWh} = 28.2\,\text{kg}\,CO_2, \tag{2.21}$$

where carbon intensity is 0.21 kg CO_2/kWh (National Grid ESO, 2024).

Bias Mitigation The development and deployment of LLMs are primarily about how to mitigate their biases. LLMs are powerful yet inherit the biases within their training data, which could further propagate or magnify stereotyping and inequities in society. That's why there is a need for bias mitigation to ensure that LLMs are equitable, inclusive, and ethical, especially for key domains of applications such as health, jurisprudence, hiring, and education (Huang et al., 2023a). Our research indicates that both of our models, the AIM-X, which is described in Section 3.2.1 and BG-MBC, which is described in Section 5.1.7, showed biases, most notably confirmation bias and clinician cognitive bias. This bias stemmed from potential subjective suggestions in the reports generated from known radiology and histopathology. For example, some of the suggestions potentially provided for things that were not evidence-based (i.e., inappropriately suggesting a disproportionate number of biopsies in some demographics) may skew the LLMs learning associations to bias the embeddings one way or the other, as the LLMs association for embeddings would have formed based on what would have been relatively equal representation among the learning set of models. Moreover, as the LLMs would try to differentiate between factual observations, there was likely to be interference of noise from clinician suggestions, which would have obscured the representation of the pure medical concept. We contend that the systematic removal of suggestions starting at the training data phase would improve the extraction of the embeddings considerably and improve their performance.

2.5.6 Evaluation of Large Language Models

Efficiency, accuracy, robustness, and ethical consequences of LLMs shall be evaluated in real-life applications. This means it needs to expose models to various test aspects, linguistics, tasks that adapt, and limitations that belong to themselves. Evaluations of such complex LLMs therefore also consist of several dimensions and methodologies (Ray, 2023). The generation performance and generalisation capability of LLMs can be considered only if the model checks whether the generated text is grammatically correct and fluent. The model is also supposed to track and understand the contextual relationship between sentences or paragraphs while handling multiple languages and their translations effectively. Evaluation might also be about

how well the model generalises to new, unseen tasks with small amounts of retraining involved, and how transferability of knowledge would be measured across domains. Evaluation extent to which it can provide information on whether the model exhibits its biases in responses, especially concerning sensitive topics such as gender, race, or culture. Testing models and ensuring that tendency to generate harmful, offensive, or inappropriate content would improve the evaluation. Testing the evaluation of this model is necessary to ensure its robustness and reliability. Testing shall be carried out with fabricated, ambiguous, or misleading prompts, not to mention the addition of noise or error in the input data. Evaluating includes checking the efficiency and scalability by checking how much it costs to train and run the model. The response times and memory usage at inference time shall be evaluated. Besides, the scalability shall be modelled. It further extends to cover users' feedback about the facility of interaction provided, relevance, and satisfaction, especially in chat-based LLMs (Ye et al., 2023; Sun et al., 2024; Ray, 2023).

Most importantly, transparency in developing and using LLMs is of great importance. It helps build trust, hence being kept accountable. Transparency may mean the creation of models understandable by people and-most importantly, be investigated by people so as to keep everybody on the same page concerning how and why a model comes up with a particular output. This becomes crucially important in domains touching on human life, health and other fields (Parthasarathy et al., 2024). LLMs represent the most revolutionary change in natural language processing, unlocking unprecedented levels of understanding, generation, and interaction with human language. LLMs can be applied to extend accessibility through automated tools in all sorts of highly specialised domains, from health and education to research. Ethical issues, computing resource consumptions, and biased or/and unjust outcomes are some major obstacles in using such mighty models. As LLMs continue to improve, these limitations would need to be discussed with a view to stimulating collaboration, innovation, and responsible use. A proper balance between technological progress and ethical issues will ensure that the LLM is a transformative tool for a more inclusive and informed world.

Finally, the DenseNet121 model was used to develop and train our AIM-X model, please refer to Section 3.1.2. The model performed quite well with 94.5% accuracy on the test dataset. Learning rate warmup, gradient clipping, and dropout rate embolisation are worthy mentions among the diverse techniques that made learning stable and generalised. Furthermore, with cross-entropy loss and teacher forcing, great learning and modelling features have been successfully attained for generative sequence tasks.

2.5.7 A Review on LLM-based Lung and Breast Cancer Utilisation

Increasingly, large language models (LLMs), including GPT, BERT variants, and domain-specific models like CancerLLM, are being researched for their applicability to breast and lung cancer care, including clinical text extraction, treatment recommendations, and prognosis modelling. While their linguistic capabilities provide engaging support for information retrieval and question-answering from guidelines, numerous serious critiques challenge their appropriateness for clinical use, given the consequences associated with clinical decision-making. A systematic review paper entitled "Utilising large language models in breast cancer management" from (Sorin et al., 2024) found evidence that there is a large gap in performance across tasks for LLMs in breast cancer care. While models demonstrated high accuracy (88–98%) on menial tasks such as information extraction from patient notes or structured information extraction from patient notes, their performance and accuracy supporting clinical decisions for complex aspects like treatment planning fell off dramatically (50–70%). Models were particularly sensitive to changes in the prompt, and could vary dramatically depending on the clinical scenario, raising concerns about their reliability in the clinical context.

(Chen et al., 2023) reported that LLMs were known to produce confident, but incorrect, or invented responses. They presented results evaluating ChatGPT 3.5 responses to the following question type: if a doctor asked ChatGPT to provide a treatment recommendation for an established cancer, what percentage of the recommendations were concordant with guidelines. They found that about one-third of recommendations were non-concordant with established guidelines (NCCN), and 12.5% of novelties included hallucinations—invented therapies that did not exist in the oncology protocols. These errors are an immediate patient safety issue if employed in a health system without human oversight, especially for lung and breast cancer care, where precision is crucial. In another evaluation (Griewing et al., 2024), researchers had GPT 3.5, Llama2, Bard, and GPT 4, complete a simulated breast cancer management protocol. Even though GPT 4 had a higher level of concordance with recommendations from a multidisciplinary tumour board, surprisingly, GPT 3.5 demonstrated daily clinical and habitat changes, which were characteristics to be concerned about within models with respect to model drift and the potential for misleading recommendations to clinical practice accuracy. Such evaluations are preliminary, and their fully simulated design limits any conclusions pertaining to real-world safety or efficacy. Studies such as (Cheligeer et al., 2024) validated LLMs in the context of announcing the detection of pathologic complete response within breast cancer pathology reports. Although impressive, this study demonstrated a

narrow focus and kept research tasks specific and without integrating imaging, radiology, or patient histories, which are typically triaged in multidisciplinary oncology care. The studies' lack of multimodal input or external validation in patients across diverse populations limits their generalisability and utility.

Table 2.11: Summary of LLM-based studies in cancer care: tasks, performance, limitations, and datasets/modalities.

Paper	**Task**	**Accuracy**	**Limitations**	**Dataset / Modality**
(Sorin et al., 2024)	Systematic review of LLM applications in breast cancer management	Accuracy varied across studies: 50–98%; structured tasks reached up to 98%	High variability; prompt dependency; missing critical clinical info; mixed real vs. internet data	Multiple studies clinical notes, guideline QA, patient management (breast cancer text-based tasks)
(Chen et al., 2023)	Utility of ChatGPT for cancer treatment information	Not explicitly reported as numeric accuracy; findings indicate partial alignment with clinical guidelines, but with notable misinformation generation	Reliability issues: ChatGPT often generated partially or completely inaccurate recommendations; prompt phrasing significantly impacted responses; risk of misinformation amplification	Various cancer treatment questions (breast, prostate, lung) submitted to ChatGPT

Continued on next page

Table 2.11 – *Continued: Summary of LLM-based studies in cancer care.*

Paper	**Task**	**Accuracy**	**Limitations**	**Dataset / Modality**
(Griewing et al., 2024)	Comparing treatment recommendations by publicly available LLMs (ChatGPT-4, ChatGPT-3.5 (Sept 21, Jan 22), Llama2, Bard) with those of a multidisciplinary tumour board for complex breast cancer profiles	"Invasive breast cancer": GPT-4 70.6%; GPT-3.5 (Sept 2021) 58.8%; GPT-3.5 (Jan 2022) 41.2%; Llama2 35.3%; Bard 23.5% "Including DCIS": GPT-4 60.0%; GPT-3.5 (Sept 2021) 50.0%; GPT-3.5 (Jan 2022) 35.0%; Llama2 30.0%; Bard 20.0% "Radiotherapy recommendations": GPT-4 100% corpus concordance "Genetic testing": GPT-4 up to 85% concordance; GPT-3.5 (Jan 22) at 55%	- Limited to 20 synthetic patient profiles, not real patient data - Early feasibility: prompt engineering and generalisability not tested - Only the consensus of a single tumour board is used as the gold standard - No real-world validation or multicenter data	20 complex synthetic breast cancer profiles; modalities include surgical, endocrine/systemic therapy, radiotherapy, genetic testing recommendations; the "dataset" is a set of clinical vignettes compared against MTB recommendations

Continued on next page

Table 2.11 – *Continued: Summary of LLM-based studies in cancer care.*

Paper	**Task**	**Accuracy**	**Limitations**	**Dataset / Modality**
(Cheligeer et al., 2024)	Detecting pathologic complete response (pCR) in narrative pathology reports using LLMs and traditional ML classifiers	Sensitivity: 95.3 % (95 % CI: 84.0–100.0%) PPV: 90.9 % (95 % CI: 76.5–100.0%) F1-score: 93.0 % (95 % CI: 83.7–100.0%)	Limited to a single regional cohort; no external validation; LLMs tested were limited to transformer embeddings and GPT-2 fine-tuning; privacy concerns prevented use of models like ChatGPT	351 pathology reports post-neoadjuvant chemotherapy from four tertiary care hospitals in Calgary, Canada; modality: narrative surgical pathology reports (text)

2.6 Deep Learning in Lung and Breast Cancer: A Systematic Review of Models' Evaluation and Dataset Selection Rationales

Deep learning has recently emerged as a disruptive technology in digital health, allowing for advanced analytics of complex multiple data types and high-dimensional data. Traditional machine learning models require a manual step of engineering features to describe the data. DL models automatically extract features from the data, be it image data (e.g., medical images), free text (e.g., clinical notes), structured data like electronic health records (EHRs), or genomic sequences. This is especially valuable for tasks that require detection of diseases, diagnosis, prognosis, and treatment planning (Miotto et al., 2018). Specific neural networks, such as CNNs, have differentiated themselves in imaging tasks, such as classifying chest X-rays and tumour detection on mammograms (Basha et al., 2020b). Recurrent neural networks (RNNs) and models based on the Transformer architecture have become popular for new healthcare tasks, such as analysing clinical text and examining sequential records of patient data (Specht et al., 1991). GNNs also allow for relational data modelling, such as graphs of patient similarities or molecular shapes. The hope is that as deep learning becomes integrated into healthcare, diagnostic accuracy will improve, clinician time and burden will decrease, and medicine will shift towards a data-driven culture (Reiser et al., 2022).

We organise the next part of the literature review to describe a number of deep learning architectures by their structure, their performance, and their appropriateness for different health-related applications. This section reflects on our method and compares it to existing methods, and evaluates the objectives and datasets used in each study. We thoughtfully examine the advantages and shortcomings of these works in order to have a comprehensive and balanced evaluation. In addition, we analyse each model from a variety of viewpoints, so that we can offer an entire discussion of its major enhancements.

2.6.1 AIM-X Multimodality Detection Model

Table 2.12 compares four multimodal models for chest X-ray report generation and summarises important distinctions with respect to their architecture, fusion approaches, and clinical applications. While (Alqahtani et al., 2024) used a transformer architecture and (Liu et al., 2019) employed a BERT-enhanced architecture, (Singh et al., 2019) focused on fusion-based strategies with a simpler architecture. Our AIM-X model (Basaad et al., 2025) outperformed these models and achieved

the best state-of-the-art scores (BLEU-1: 0.727, ROUGE-L: 0.727, accuracy: 94.5%) with a multiscale CNN to extract the image features, a text encoded using GPT, and implementing bidirectional cross-attention for reasoning. However, AIM-X requires the most computational resources. Each model has trade-offs found in (Alqahtani et al., 2024) excel in structured reporting, but are limited in sparse vocabulary, and (Liu et al., 2019) focuses on clinical fluency but requires a comprehensive range of annotations. The IU Chest X-ray dataset is a common denominator amongst every model, which establishes the known value of utilising the IU Chest X-ray dataset for medical-AI research.

Table 2.12: Comparison of four state-of-the-art multimodal models for chest X-ray understanding and report generation. The models differ in architecture, fusion techniques, datasets used, and clinical applicability.

Aspect	**(Alqahtani et al., 2024)**	**(Liu et al., 2019)**	**(Singh et al., 2019)**	**(Basaad et al., 2025) (Ours)**
Domain	Medical imaging	Medical imaging and NLP (chest X-rays)	Medical imaging and NLP	Medical imaging and NLP
Task	Chest X-ray report generation	Radiology report generation	Radiology report generation	Radiology report generation
Modality	X-ray + text	Chest X-ray images + Reports	X-ray + Radiology report	X-ray + Radiology report
Attention Mechanism	Transformer-based encoder-decoder with multi-head attention	CNN features + text decoder (LSTM); explores early vs. late fusion	Attention via CNN-LSTM, fusion strategies explored	CNN features + Multiscale feature levels + text encoder (GPT) + Cross-attentions fusion + text decoder (LSTM)
Innovation	Combines DenseNet201 CNN for image features and Transformer decoder for text; uses cross-modal fusion with attention	Uses BERT embeddings + attention to align with clinical correctness	First to compare fusion strategies (early, late, joint) in medical report generation	iterative attention refinements + Bidirectional Cross-attentions
Interpretability	Attention heatmaps and section-level control	High: aligned with clinical terms and sections	Moderate, uses attention visualisation for image-text fusion	Moderate, uses attention visualisation for image-text fusion
Clinical Utility	High: generates coherent reports with structured sections	Moderate	(targets real-world radiology setting)	High: generates coherent reports
Dataset Used	IU Chest X-ray dataset	MIMIC-CXR + clinical notes	IU Chest X-ray Collection	IU Chest X-ray dataset

Continued on next page

Table 2.12 – *Continued: Comparison of four state-of-the-art multimodal models for chest X-ray understanding and report generation.*

Aspect	**(Alqahtani et al., 2024)**	**(Liu et al., 2019)**	**(Singh et al., 2019)**	**(Basaad et al., 2025) (Ours)**
Limitations	Limited dataset size; domain-specific vocabulary can be sparse	Requires large annotated clinical corpora	Simpler architecture; fusion strategy selection not always optimal	Requires high machine resources for training and customising hyperparameter uses genetic algorithm.
Results	BLEU-1: 0.45, BLEU-4: 0.23; ROUGE-L: 0.41; METEOR: 0.26	Better CheXpert compliance; strong clinical fluency	BLEU-1 0.26; late fusion slightly outperforms early fusion in coherence and accuracy	BLEU-1: 0.727, BLEU-4: 0.469; ROUGE-L: 0.727; Accuracy up to: 94.5% — outperforming existing methods

2.6.1.1 The Rationale for IU Chest X-ray Dataset Selection

The *IU Chest X-ray* dataset provides a compelling resource for multimodal early detection of thoracic diseases as it contains pairs of contextualised diagnostic quality radiographs alongside radiology reports, permitting the simultaneous analysis of visual and textual data (OpenI, 2024). The radiographs provide morphological evidence towards diagnoses (e.g., lung opacities, nodules), while the radiology reports provide textual explanations of findings (e.g., "suspicious mass," "bilateral effusions") which contextualise the imaging modality and reduce uncertainty for early-stage interpretation. The final advantage of this structure is that, as it represents a dual-modality, one can use advanced fusion methods, such as cross-modal attention (e.g., aligning image segments to their textual descriptions) or a contrastive learning approach that prompts the model to explicate the relevance of an image and its textual associated text, potentially isolating an important clinical finding. The dataset enumerates very pertinent pathologies (e.g., pneumonia, pneumothorax, lung cancer precursors) where early detection is critical, while the public dataset affords opportunities for benchmarking themselves against state-of-the-art approaches. Given the potential for a multimodal model with this dataset to outperform not only unimodal models, but furthermore for even the worst performing multimodal AI tool trained on this dataset to outperform a unimodal AI tool, particularly within edge cases pertaining to differentiating between benign calcification vs malignancies, offers compelling promise. In conclusion, considering the exceptional annotations' variety, the clinical relevance of the dataset, and the modalities' respective relevancy, the IU dataset is a primary candidate dataset for the design of effective AI tools designed for early detection of thoracic disease.

2.6.2 GraphX-Net Relapse Breast Cancer Model

This analysis in Table 2.13 illustrates the variety of methods used in previous research of breast cancer relapse prediction, based on the METABRIC data and hospital data, with regard to the more specific context of deep learning and graph-based methodology. The model proposed by (Basaad et al., 2024b), GraphX-Net, illustrates the inclusion of Graph Neural Networks (GNNs) with Shapley values with high interpretability, and a novel graph-based representation of clinical data. In contrast, the study by (Rupavath et al., 2024) takes advantage of the capability for temporal modelling with recursive neural networks on multi-omics data to show a comprehensive view of metastasis. (Gupta, 2022) maintains a more traditional machine learning pipeline, using models SVMs and Random Forests on clinical data, which obtains moderate performance, while providing better explainability compared to some of the previous deep learning work. Lastly, the work by (Kumari

et al., 2025) included an ensemble of both CNN and ANN architectures to obtain the highest predicted accuracy, while sacrificing explainability. In conclusion, while ensemble and deep models may project better predictability, graph-based and more interpretable models, such as GraphX-Net, provide potential for higher accuracy along with clinical explainability - particularly singularly important for real-world healthcare settings.

Table 2.13: Comparison of Deep Learning and Machine Learning Models for Predicting Breast Cancer Recurrence.

Aspect	**(Basaad et al., 2024b)** ***GraphX-Net***	**(Rupavath et al., 2024)** ***Multi-omics + RNN***	**(Gupta, 2022)** ***Classical ML models***	**(Kumari et al., 2025)** ***DL Ensemble***
Approach	Graph Neural Network + Shapley explainability	Recursive Neural Network on multi-omics data	SVM, Random Forest, Logistic Regression	Ensemble of ANN and CNN models
Objective	Predict recurrence with interpretable predictions	Metastasis prediction through multimodal inputs	Estimate tumour recurrence time	Forecast recurrence based on mixed clinical data
Data Modalities	Clinical tabular data modelled as graphs	Genomic, proteomic, transcriptomic data	Clinical + histopathology features	Clinical + demographic + pathologic data
Explainability	High — interpretable via Shapley values	Moderate — omics feature saliency	Moderate — decision trees and feature ranking	Low — lacks interpretability mechanisms
Performance	AUC: 0.98, F1-score: 0.98, Accuracy: 98.9%	AUC: 0.92, F1-score: 0.88	AUC: 0.91, Accuracy: 85%	AUC: 0.95, Accuracy: 93.5%
Innovations	Graph-based architecture + feature attribution	Temporal modelling of multiple omics sources	Early-stage prediction via classical modelling	High-accuracy ensemble learning framework
Dataset Source	METABRIC (processed into graphs)	METABRIC dataset	METABRIC dataset	Hospital-based anonymised cohort
Limitations	Requires graph construction; high computation cost	Needs high-dimensional omics input; generalisation limited	May underperform on complex nonlinear data	Low interpretability; lacks external validation

2.6.2.1 The Rationale for METABRIC Dataset Selection

The METABRIC dataset is an ideal option for GNN-driven relapse prediction because it has a unique multi-omic depth (genomic characteristics, transcriptomic patterns, and clinical data) and relapse-related annotations (e.g., relapse-free survival), allowing researchers to create biological networks of molecular interactions. Having a large cohort of 2,500 patients of the breadth of subtypes of breast cancer can help alleviate the understanding of a model's propensity for overfitting, as well as allow for a better search for recurrent outcomes (Curtis et al., 2012). The dataset is well-structured to take advantage of GNN's representation, where patients are modelled as nodes and edges are regulatory or clinical relationships. Lastly, by integrating the network-ready relationships found in METABRIC and GNNs, the model will be able to capture more complex and non-linear drivers of relapse that prior methods may have missed.

2.6.3 BG-MBC Metastatic Breast Cancer Model

In the following Table 2.14, we provide the efficiency of combining two potent DL techniques, LLMs and GNNs, to predict metastatic breast cancer using data from BIACH and RI providers. Our proposed BERT-GNN model provides a high opportunity for performance in the detection of metastases. Compared to publications that used other architectures and the same datasets, our work displays a promising opportunity to enhance disease monitoring ability. (Basaad et al., 2024a) reported a BERT-GNN that integrated a graph neural network and natural language processing of histopathology reports, providing contextual and relational learning with strong performance, but only moderate interpretability. A baseline model was developed to show the power of infusing LLM and GNN to understand the nuanced relationships of medical data attributes. The baseline model is a Keras sequential model with two feedforward networks. (Botlagunta et al., 2023) reported a set of traditional machine learning models, such as XGBoost and SVM, using structured clinical data, which had adequate predictive accuracy while providing greater interpretability from model features. (Maini et al., 2025) optimised deep learning models using hyper-parameter tuning, producing the most accurate predictions, but considerably less interpretability. Deep models were determined to produce more accurate predictions than graph and machine learning-based models collectively in this systematic synthesis; however, in a clinical context, interpreting models is preferred over blindly recommending the models. While deep models are ubiquitous and more accurate than graph and machine learning models, models based on graph learning and machine learning are interpretable and use significantly less computational resources to conduct inference, making them a preferable model in a clinical

context, where resources (time and computer) and interpretability is preferred.

Table 2.14: Comparison of recent approaches using deep learning and machine learning for metastatic breast cancer prediction.

Aspect	**(Basaad et al., 2024a)** ***BERT-GNN for Histopathology Reports***	**Baseline** *Baseline*	**(Botlagunta et al., 2023)** *ML on Clinical Data*	**(Maini et al., 2025)** *DL with Hyperparameter Tuning*
Approach	Graph Neural Networks combined with BERT for textual histopathology reports	Two feedforward networks + Keras model	Supervised machine learning algorithms (SVM, RF, XGBoost) applied to structured clinical datasets	Deep learning models with optimised hyperparameters (e.g., CNN, DNN)
Data Type	Unstructured text from pathology reports	Pathology reports	Structured clinical and demographic data	Imaging and structured diagnostic features
Data Source	BIACH and RI	BIACH and RI	Structured clinical and demographic data	Imaging and structured diagnostic features
Model Type	BERT for feature extraction + GNN for learning graph structure among entities	BERT for feature extraction + Keras sequential model	Traditional classifiers (SVM, Decision Trees, XGBoost, etc.)	Custom deep learning architectures with fine-tuned parameters
Objective	Metastasis prediction using contextual information from medical reports	Metastasis prediction using contextual information from medical reports	Classification of metastatic status based on routine clinical parameters	Classification of breast tumours (including metastatic cases) with high accuracy

Continued on next page

Table 2.14 – *Continued: Comparison of Deep Learning and Machine Learning Models for Predicting Breast Cancer Recurrence.*

Aspect	**(Basaad et al., 2024a)** ***BERT-GNN for Histopathology Reports***	**Baseline** *Baseline*	**(Botlagunta et al., 2023)** *ML on Clinical Data*	**(Maini et al., 2025)** *DL with Hyperparameter Tuning*
Explainability	Moderate to high — uses attention weights from BERT and GNN node importance + Shapley values	Moderate uses Shapley values	Moderate — feature importance from ML algorithms	Low — DL models not inherently interpretable unless paired with explainers
Performance	Accuracy: 98%, AUC: 0.98	Accuracy: 91.44%	Accuracy: 89.6%, AUC: 0.91 (XGBoost highest)	Accuracy: 94.1%, F1-score: 0.92
Innovation	Integrating advanced natural language processing (NLP) techniques with graph neural networks (GNNs) and proposing a novel approach for comprehensive analysis of histopathology reports.	-	Demonstrates comparative performance of multiple ML algorithms on real-world clinical datasets	Focus on training efficiency and accuracy via systematic hyperparameter tuning
Limitations	Depends on availability of annotated pathology text; graph construction may introduce complexity	-	Limited generalisability to imaging or text data; model choice impacts performance significantly	Limited interpretability; dataset origin and size not extensively validated externally

2.6.3.1 The Rationale for BIACH and RI Dataset Selection

The BIACH and RI dataset provides real histopathology reports documenting metastatic breast cancer, with clinically validated annotations (e.g., biopsy-confirmed lymph node involvement, distant metastasis) in a resource-limited healthcare setting. Unlike research-only datasets (e.g., METABRIC), these reports highlight diagnostic challenges intrinsic to late-stage presentations and the local patient demographic (e.g., Indian cohorts and higher BRCA incidence), which likely enhances model generalisability and robustness (Botlagunta et al., 2023). This enables us to employ LLM to extract structured features (e.g., tumour grade, nodal status) for downstream analysis. GNN has been applied to model relationships between entities (patient IDs) to predict metastatic progression using text-derived embeddings.

2.7 How Our Research Addressed Challenges

As we integrated the key subjects in deep learning into our research, we have fused the challenges of key components together (e.g., GNNs and LLMs). As for some of these challenges we have surpassed, and the others we have been working hard to mitigate the impacts as to formulate frameworks embedded in a pipeline workflow, to accomplish the research's objective and to aid/enhance the diagnosis and process of exposing these diseases at early intervention stages. The aforementioned challenges and limitations were, in any case, very general the following: all of the challenges we ran into and how we treated them:

- Initial Challenges in Graph Construction for GNN Implementation. The primary issue we faced when working with Graph Neural Networks (GNNs) was the building and construction of the graph structure through the software. As we highlighted, the shape value is fundamental in overcoming many of the complexities of the data, and helps us to specify distinctive relationships between data elements, which are managed through a specific component of the GNN, the Graph Constructor Module.

  We made sure to establish additional parameters to complete the graph construction and modifications to:

  - Restrict structural properties of the graph which preserve meaning through connectivity
  - Manage computational complexity through the graph generation and feature propagation processes.

  In conclusion, these modifications we made were necessary to ensure were

delivering an approach through the GNN that was both interpretable as well as scalable. I highlighted this issue in detail in the following Section 4.1.2.2.

- Challenges in AI-Driven Frameworks: Data Acquisition and Preparation. A main challenge of our AI-based research was to identify proper datasets as the basis for our project. It was extremely important to find a high-quality paired dataset of medical images (X-rays) and diagnostic radiology reports that would align with our goal of building a multimodal framework. After extensive exploration, we located a dataset at Indiana University Hospital that consisted of chest X-ray images and radiology reports, as discussed in Sections 2.6.1.1, 2.6.2.1, 2.6.3.1.

  However, this dataset required preprocessing. Specifically, we removed images and reports for two reasons related to:

  – Non-relevant material related to reports that could unintentionally introduce bias.

  – Poor image quality, which would impede reliable feature extraction.

  The cleaning step was essential to establish an integrity solid dataset that we could utilise towards our multimodal *AIM-X* model.

- Among the significant challenges I had experienced was the visualisation of the region of interest (ROI) as explained in Chapter 3 through heatmaps. This problem usually happens because of the class imbalance in the dataset, where normal chest X-rays outweigh images with disease. This is intended from my argument as the model will focus on the patterns of what is being seen normally (pathology of importance). The model's ability to produce sensible reports for the common disease was adequate; however, patterns of attention would miss infected areas in rare circumstances. The only way to reduce this problem, which I believe existed in the model, is to have domain experts curate and balance the dataset and have a structured labelling mechanism, which may allow the model to improve robustness and generalisability.

- Ethical Approval Process and Data Usage Considerations: Acquiring ethical approval to use this data set involved a lengthy and rigorous approval process. Even though the data was made publicly available and stated explicitly for research purposes, we structured our ethical application so that we were fully compliant with proper ethics policies and reduced potential future liabilities. We provided a detailed description of the data development, usage process and complied with established ethical procedures for publicly available medical data.

- Addressing Explainability and Clinical Transparency: To enhance interpretability of our model and clinical applicability, we implemented multiple explainable artificial intelligence (XAI) strategies to add to our framework's architecture:

  1. Feature Importance Analysis: We utilised Shapley values (SHAP) to analyse the most contributing features towards predictions, giving us specific make sense explanations of the model's decision making locally or globally as discussed in Section 4.1.2.6.
  2. LIME Explainability: LIME techniques have been employed to show the feature importance score towards the decisions. It works locally with each individual prediction, not overall model behaviour. As shown in Sections 4.1.2.6, 5.2.1.
  3. Visual Explainability: We incorporated attention mechanisms with attention heatmaps on anatomical regions in medical images. This configuration made clinical reasoning concrete and assisted our model with additional processing.

  Our strategies provide clinical trustworthiness, where black box decisions can be audited. Actionable insight through associating explanations back to clinical knowledge (linking heatmaps back to radiological markers).

- Data Annotation and Labelling Process: Annotation and labelling are an important step in our data preparation workflow. A simple cleaning process that attempted to:

  – Strip out any obvious irrelevant text markers (e.g. 'XX')
  – Remove unnecessary whitespace and formatting bolts
  – Standardise, where possible, the report structure

  Yet we view this approach as transitional. A better data cleaning approach would fully involve:

  – Workshops with medical domain experts
  – Workshops with data scientists
  – Prototypes of quality assurance checks on the annotation process
  – Templates standardising how to label reports

  Using an expert-guided process would enable better data preparation for advanced analytical methods and allow proper clinical context for the data.

- Computational Requirements for Real-Time Processing: Developing and evaluating our real-time processing systems required a considerable computational load to manage computationally demanding training regimes. Because of this, we took the following multi-faceted approach focused on our compute usage:

  For computational resource uses:

  - Google Colab Pro for initial model prototyping and smaller-scale experimentation.
  - HPC resources within our institution, for large-scale training runs and final model evaluation with the following specs:
    * CPU: Intel(R) Xion(R) W-2245 CPU @ 3.90GHz.
    * Memory: 96.0 GB.
    * Windows 10 Version 22H2.
    * GPU: NVIDIA RTX A5000. Dedicated GPU memory: 24.0 GB.

  For computing challenges:

  - Optimised training pipelines where possible.
  - Parallel Processing methods to speed up convergence.

  Taking a hybrid approach helped to engage the required computational resources necessary to investigate deep learning architectures, while accommodating the limitations of access to institutional-level resources.

- Throughout the model development process, we faced a number of critical technical issues. Applied data augmentation methods (geometric transformations and synthetic data creation) to alleviate the limited training data available.

  Some algorithms have been used to detect bias. To deal with the overfitting, we used early stopping through validation monitoring and k-folds for performance evaluation. Also added batch normalisation to the input layer to reduce internal covariate shift.

  All of these methods resulted in a model that could generalise better, and still performed well on our test datasets.

2.8 Conclusion

In this chapter, we have considered the fundamental architectures and frameworks which have given rise to contemporary AI-based diagnostics. We began with CNNs, and we examined how they extract a series of hierarchical visual features - an ability

that is vital to the analysis of medical imaging data (X-rays). Following this, we explored GNNs and how relational data and structured dependencies can be modelled; we began to investigate how to incorporate multimodal data from medical records. Our examination of the importance of model explainability illuminated the need for transparent clinical AI, where it is important that any clinical AI would yield diagnostic predictions that are clear to the medical professionals involved. Finally, the advancements in LLMs indicated that diagnostic accuracy could also be augmented and revised by utilising unstructured text data (i.e., pathology reports).

Further discussions were conducted around the key themes of our research, highlighting how each element contributes to the central focus on lung and breast cancers. The following section presents a detailed comparison between our proposed models and existing published studies. In addition to the challenges and obstacles that we faced along the journey.

In Chapter 3, we are building on these themes as we introduce AIM-X, a unified model that is distant from imaging and radiology report diagnostics. AIM-X uses DenseNet-121 (CNN) as a base model for visual feature extraction, and provides cross-attention in both directions to the X-ray images and radiology reports for co-reference, made possible by the inherent attention mechanisms that clarify the challenges and opportunities.

In short, the chapter sets the stage for a description of two additional novel models:

- GraphX-Net Chapter 4, which invokes Shapley values to form patient-specific graphs for relapse on predictions of breast cancer, demonstrates that explainability.

- BG-MBC Chapter 5, which uses a GNN that first extracts embeddings using pathology reports integrating LLMs, deploys feature selection algorithms before employing a GNN to predict metastatic breast cancer.

Chapter 3

Clinical Report Generation and Interpretation

As outlined in the previous chapters particularly in the introduction and Section 1.1.1, we provided an in-depth discussion of thoracic diseases and their potential progression to lung cancer. Early screening, detection of abnormalities, and accurate diagnosis are critical for managing the various phases of cancer addressed in this research. Our strategy starts with the detection and diagnosis of abnormalities in the chest, and systems developed will help establish a strong framework for identifying disease at an early stage for patient survival. Following the various deep learning approaches and themes showcased in Chapter 2, we are introducing AIM-X, which is explained in detail in this chapter.

This research represents a step further in the expansion of AI-assisted medical diagnostic systems, and the merging of multimodal data. Traditionally, radiologists and clinicians relied on making observations in isolation either visual data (e.g., X-Ray, CT) or textual data (e.g., patient history, radiology reports). With recent developments, for example, multimodal report generation models (Xu et al., 2024b; Lipkova et al., 2022), it is now possible to analyse both image and textual data, improving the diagnostic process's accuracy and thoroughness. Advances in AI architectures (e.g., DenesNet-based convolutional neural networks) will further enhance this system, underlining AIM-X's potential to create a broad, data-driven approach to healthcare. This modality fusion will enhance the accuracy and completeness of diagnostic reports, advancing digitalised healthcare.

Sophisticated designs like DenseNet of CNNs become part of most multi-modal disease detection systems. Since it has a dense layers learning structure, DenseNet has superior performance on complicated and high-dimensional medical image data. As such, since it captures all the minute pattern visual features indicative of disease, it turns out to be very useful in the domain of medicine, especially in radiology. This capability is hence of utmost importance in scenarios such as early diagnosis

in cancer, where the detection of minor irregularities within images could improve a patient's prognosis by leaps and bounds (Ashraf et al., 2020; Yu et al., 2021; Yang et al., 2022b).

The integration of multimodal data into AI-driven medical diagnostics represents a substantial advancement in healthcare technology (Xu et al., 2024b). Historically, medical diagnostics have relied on visual data, such as X-rays and CT scans, or textual data, such as patient histories and clinical notes, which were independently reviewed by radiologists and clinicians. Nonetheless, the advent of multimodal report generation models has transformed this process by enabling the concurrent analysis of both image and text data (Xu et al., 2024b; Lipkova et al., 2022). This fusion of modalities enhances the precision and comprehensiveness of diagnostic reports, offering a more holistic approach to patient care.

Convolutional Neural Networks (CNNs), especially sophisticated designs like DensNet, play a pivotal role in multi-modal disease detection systems (Li et al., 2021). DenseNet leverages the network's capabilities by reusing the network's features, thereby producing models that are concise, easy to train, with high parameter efficiency. A concatenation of feature maps learned by different layers would increase variation in the input of subsequent layers and improve efficiency. That's where DenseNets really differ from ResNets. This capability is especially vital in contexts like early cancer detection, where spotting minor irregularities in images can significantly improve patient outcomes (Huang et al., 2017; Ashraf et al., 2020; Yang et al., 2022b).

Recent versions of the (GPT 2,3,4) series have been employed to process text data for comprehending and generating human language (Yenduri et al., 2024). When used alongside RNNs and transformers, these models are particularly effective in visual analyses. These transformer-based methods are capable of contextual analysis across lengthy text sequences, which is advantageous for generating coherent and contextually accurate medical reports (Rasmy et al., 2021; Tunstall et al., 2022). A pivotal part of multi-modal report generation models is the attention mechanism, depicted in Fig. 3.1, which enhances focus and interpretability. This mechanism weights the significant sections of inputs, both textual and visual, then emphasises the most pertinent parts (Serrano et al., 2019). This approach not only boosts the accuracy of the model's predictions but also improves interpretability by explaining why certain data points were highlighted. For instance, in medical imaging, attention maps can identify sections of an X-ray that are most indicative of a condition, thus making the AI's decision-making process clearer and more reliable for healthcare professionals (Wang et al., 2017b; Tutek et al., 2022). The cooperation between visual and textual data in multi-modal models leads to more accurate diagnostics and explainability (Wang et al., 2017b). For instance, in the detection

of thoracic diseases from chest X-rays, the integration of clinical notes with image analysis can provide a richer context, improving the model's ability to correctly identify conditions (Li et al., 2023). Furthermore, the use of attention mechanisms helps in creating more explainable AI systems, which is essential in the medical field where understanding the rationale behind a diagnosis can be as important as the diagnosis itself (Wiegreffe et al., 2019; Salvi et al., 2024). In addition, utilising the genetic approach enhanced the model performance and outcomes. While the primary application of these models has been in radiology, the disciplinarity and the application principles can be extended to other areas of medicine and beyond.

The major contributions of our work are as follows:

1. AIM-X introduces a dual cross-attention architecture that fuses visual features from X-ray images with textual data from radiology reports. This is by alternately assigning query roles to both modalities, where the model learns richer, more discriminative representations and generates highly accurate and more patient-specific clinical reports.

2. Our dual cross-attention mechanism integrates multiple cross-attention strategies with multiscale feature extraction, allowing AIM-X to progressively focus on the most relevant image regions and align them effectively with textual context.

3. AIM-X also incorporates genetic algorithm-based optimisation to tune hyperparameters based on the structure of medical reports, improving the model's understanding of complex clinical content and improving the quality of report generation.

3.1 Material and Methodology

3.1.1 Dataset

The data we used in this work was supplied by the Indiana University Hospital Network and is divided into two parts (OpenI, 2024):

X-rays Chest X-ray images with two types of view, including frontal and lateral views. The dataset includes a total of 7,472 x-ray images. These images are representative of chest radiographs from different individuals.

Reports Different reports, in XML, for patients, where each XML document represents one patient and includes clinical information associated with that patient. The XML tag <parent Images id="image-id"> is used for the specific images that relate to the report. The id attribute within it contains the name of the image,

which aligns with the PNG images. A report or one XML document may pertain to a series of images.

The XML files consist of four key features:

Comparison This section provides information about the methods for serial follow up, for instance, if no changes since the last X-ray.

Indication Includes patient-relevant clinical information, typically when the radiologist is writing the report.

Findings This attribute represents information related to the x-ray image, whether there are abnormal findings in specific regions like the right or left lung.

Impression This can be inferred from the indicators and findings; this represents the results of the medical report, whether the lungs are clear or if there are pulmonary conditions.

The AIM-X architecture is designed specifically to work with multimodal information, including both imagery and textual, in generating accurate reports. An integral part of AIM-X involves feature extraction both from image and text sources. In specific, in our case, convolutional neural networks (CNNs), such as DenseNet, have been utilised to extract important features out of chest X-ray images. For the textual part, a model in the form of a Transformer (e.g., GPT-2, GPT-3) has been used to generate textual embeddings. This model converts each word in the text into a high-dimensional vector space, in turn, representing semantic meaning of the text.

The features or embeddings produced in this stage serve as representations of both textual and graphical information for future analysis. After producing textual and graphical embeddings, the AIM-X attention mechanism arranges attention values in terms of weighted representations through computation of importance values, such as normalised values for attention, via application of the Softmax function, in relation to features and embeddings in an input sequence. In specific, we first designed a fully connected layer to map the input embeddings onto a shared, uniform, or constant coordinate frame in space. By converting all representations to a uniform, normalised form, comparative analysis is both efficient and effective. Next, normalised representations serve as input for a decoder RNN layer, such as LSTM networks, to generate a sequential ordering of words that forms a report.

Concurrently, a mechanism for attention was developed through the use of the technique of projecting representations of both visual and textual information into a shared feature space, allowing for selective prioritisation of the most relevant inform-

ation in generating reports. That mechanism assigns an attention value to each such representation in terms of its contribution towards model output. The computation of attention values is performed through an evaluation of each embedding in terms of its comparative importance when measured with regard to others in the mapped feature space, through a computation between a query (Q) representation—embodied in the model's current focus and in relation to state in the decoder when generating a word—and a number of key (K) representations, including representations learned for both image and previously generated textual items mathematically:

$$\text{query} = W_q \cdot \text{input\_features} + b_q, \tag{3.1}$$

where W_q is the weight matrix of the linear layer. b_q is the biase and input_features are the original features (e.g., image or text embeddings).

Comparison is performed through a function of similarity (e.g., a dot-product function) between each key with regard to a query mathematically:

$$\text{dot}(q_i, k_j) = q_i \cdot k_j = \sum_{d=1}^{D} q_{i,d} \cdot k_{j,d}, \tag{3.2}$$

where D is the dimensionality of the embeddings (image embeddings dim). The dot product between a query q_i (a single row in Q) and a key k_j (a single row in K). Repeating for all pairs of queries and keys, resulting in a matrix of attention scores (shape: batch size, image seq len, text seq len). To avoid the danger of dot products becoming too large and potentially causing the gradients of training to become unbalanced, the scores are normalised through division by the square root of the embedding dimension:

$$\text{scaled\_scores} = \frac{Q \cdot K^T}{\sqrt{D}}, \tag{3.3}$$

where D is the dimensionality of the embeddings. This process adapts the model refocusing, and the mechanism enables the model to dynamically refocus its attention onto a variety of aspects of an input, whether regarding specific parts of an image or previously generated words, in an effective and coherent creation of reports regarding medical cases.

The attention scores go through processing through the Softmax function, normalising them to a cumulative value of one, effectively transforming these scores into attention weights. Next, these attention weights are utilised in calculating attention values, understood in terms of weighted representations or weighted embeddings. Next, the attention values are arranged in a manner such that larger-magnitude ones are emphasised, and these represent specific features or embeddings. In an at-

tempt to boost the model's ability to work with increased intensity, attention weights go through iterative refinement during training. By iterative updating, the model can go through a systematic examination of the input and realign its focal point at each stage, enhancing its accuracy in its mechanism of attention with increased training. Techniques utilised in such a technique include:

- **Feedforward Network**: Transformed embeddings are fed into the decoder recurrent neural network (RNN) layer, for example, LSTM networks, to generate a sequence of words in report format. In each time step t, x_t (the transformed embedding) is processed together with the current hidden state h_{t-1} and the memory of the previous time step (cell state) m_{t-1}, with a new hidden state generated through the following equation:

$$h_t = \text{LSTM}\left(x_t, h_{t-1}, m_{t-1}\right), \tag{3.4}$$

  where h_t is the hidden state at time step t, x_t is the input at time step t, and h_{t-1}, m_{t-1} represent the hidden state and memory from the previous time step (Lu et al., 2017). The LSTM generates the report sequentially, using updated attention weights at each time step.

- **Loss Computation:** After generating the report, the predicted sequence of words is compared to the ground truth using a loss function (CrossEntropyLoss). The loss is computed as follows:

$$\text{Loss} = -\sum_{t=1}^{T}\sum_{c=1}^{C} y_{t,c} \cdot \log(p_{t,c}), \tag{3.5}$$

  where $y_{t,c}$ represents the ground truth probability distribution (or one-hot encoding) for the target word or token at time step t, and $p_{t,c}$ is the predicted probability for that word or token at time step t. The loss provides feedback for iteratively refining the attention mechanism, as the model learns to focus on more relevant parts of the data during each step of training.

- **Backpropagation:** The gradients of the loss with respect to the model parameters (including the attention weights and LSTM weights) are computed. These gradients are used to update the model parameters via gradient descent. The update rule for the attention weights is as follows:

$$W \leftarrow W - \eta \cdot \frac{\partial \text{Loss}}{\partial W}, \tag{3.6}$$

  where η is the learning rate, and $\frac{\partial \text{Loss}}{\partial W}$ is the gradient of the loss with respect to the attention weights. This iterative process ensures that attention weights

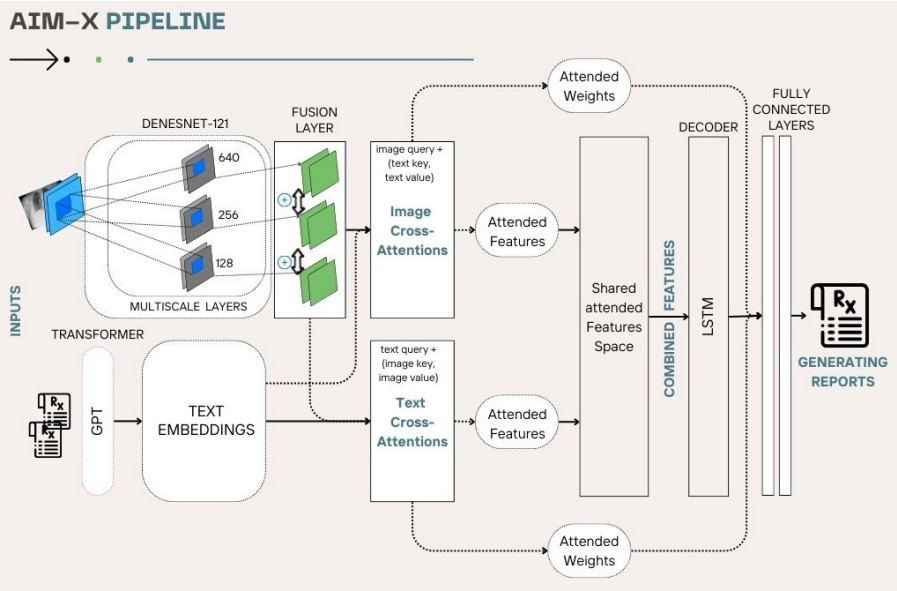


Figure 3.1: AIM-X pipeline high-fidelity X-ray features and complex patterns important for diagnoses. The visual features are then blended with textual information and processed through GPT to produce rich contextual, coherent medical reports. By fusing visual and textual modalities through an attention mechanism, AIM-X generates accurate and readable diagnostic reports, effectively connecting image features with contextual information for enhanced medical diagnostics.

are progressively refined at each step of training, improving the model's ability to focus on the most relevant features. Additionally, the (`clip_grad_norm_`) function is applied to prevent exploding gradients, ensuring stable learning during the update process.

3.1.2 Model Components

The framework is designed to combine features extracted from images and text. After preprocessing, the X-ray images are then aligned to their respective impressions using a custom-designed dataset class in the framework. The system generates a merged feature representation wherein every image is aligned with its textual features. This enables the model to be trained from images and text concurrently. Fig. 3.1 illustrates the full architecture of the proposed model, incorporating various advanced components. DenseNet121 functions as the core for extracting image features, with the derived visual features being processed through subsequent layers to further refine and distil the information.

3.1.2.1 Image Feature Extraction (DenseNet121)

- The model uses a DenseNet-121 pretrained on ImageNet to extract features from input images.

- DenseNet is one of the CNN architecture models that is noted for being highly connected, allowing for feature reuse as well as better propagation of gradients.
- The multiscale feature extraction layer is part of the model responsible for extracting features at various scales from the images.

3.1.2.2 Feature Projection Layers

It basically means that the model architecture makes predictions by processing multimodal data, which involves images and text. The features derived from DenseNet are grouped into three different levels:

- Foundational features: Early layers (128 channels) that capture basic textures and contours.
- Mid-level features: Middle layers (256 channels) capturing more complex patterns.
- Higher-level features: Following layers (640 channels) covering semantic information

Every feature level is projected into the standardised embedding dimension (image_embedding_dim) via the use of different linear layers (low_proj, mid_proj, high_proj) (Dalmaz et al., 2022; Yang et al., 2020).

3.1.2.3 Feature Fusion

The multi-scale features (low, mid, and high-level) are concatenated and passed through a fusion layer (feature fusion) to combine them into a single representation.

3.1.2.4 Cross-Attention Mechanisms

The attention mechanism is adopted here to combine both modalities of image and text for enhancing the accuracy in prediction.

- The model uses two cross-attention mechanisms to align and combine information from images and text:
    1. **Visual Cross-Attention**: Attends to image features based on text embeddings.
    2. **Textual Cross-Attention**: Attends to text embeddings based on image features.
- Combined the attended image and text features into a unified representation. These attended features allow the model to focus on the most relevant parts of the input data when generating the report.

3.1.2.5 Decoder (LSTM)

- The integrated features are passed into the LSTM network, which is a special case of recurrent neural networks (RNNs) that are often used for sequence generation work such as generating text.
- The LSTM network processes the combined features and generates a sequence of hidden states.

3.1.2.6 Fully Connected Layers

- The output from the LSTM is then passed through two fully connected layers, to map the hidden states to the vocabulary space (vocab size).
- Dropout is applied after each fully connected layer to prevent overfitting.

3.1.2.7 Positional Encoding

Positional encoding is implemented in the embeddings to encode information and keep the order of the words in the input text. This is important since the model needs to understand the order of words to enable the model to produce well-formed reports.

3.1.3 Model Pipeline

1. **Image Processing**:
    - The input images are reshaped before being passed through the DenseNet architecture for multi-scale feature extraction.
    - The features are divided into low, mid-level, and high-level features, each of which corresponds to a standardised embedding dimension.
    - The multi-scale features are incorporated and merged into an integrated representation.
    - Data augmentation was applied to enhance dataset variability, encouraging the model to learn robust, invariant features instead of memorising the training data. By simulating real-world variations such as rotations, crops, and noise, the model improves its ability to generalise to unseen data.
2. **Text Processing**:
    - Positional encoding is added to the text embeddings to capture the order of words.

3. **Cross-Attention**:
    - Visual cross-attention and textual cross-attention are used to enable the alignment of features of images and textual content.
    - The cross-attention mechanism employed across modalities combines the attended features into one single representation.
4. **Sequence Generation**:
    - The integrated features are handled through the LSTM and produce a sequence of hidden states.
    - The hidden states are converted into vocabulary space using fully connected layers with the aim of eventually generating the final output.
5. **Output**: The model returns the generated report (output), along with the attention weights for visualisation or analysis.

3.1.4 hyperparameter settings

Here, we used a Genetic Algorithm (GA) to fine-tune hyperparameters and enhance model performance. Genetic algorithms utilise principles from natural selection and evolution, where a set of candidate solutions progresses through generations to increase their fitness (Lambora et al., 2019; Kanada, 2016). This evolutionary approach is particularly effective in navigating complex hyperparameter spaces where traditional optimisation methods may struggle. GA is used to modify the following essential hyperparameters:

- Learning Rate (lr): The learning rate determines the magnitude by which the weights are adjusted in the model during training. A genetic algorithm can try multiple values to come up with some decent learning rate that will result in good convergence without overshooting the values (Kanada, 2016).
- Batch Size: It defines the number of training samples in a single batch that would require processing before the weights get updated. Through GAs, it's feasible to optimise batch size to create a trade-off between training speed against model performance as larger the batch size stabilises gradients but requires more memory (Kandel et al., 2020).
- Dropout Rate: Dropout is a form of regularisation technique that prevents overfitting by randomly turning off some proportion of neurons during training. Considering various dropouts, genetic algorithms may find the optimal trade-off between regularisation and model's learning capability (Tian et al., 2020).

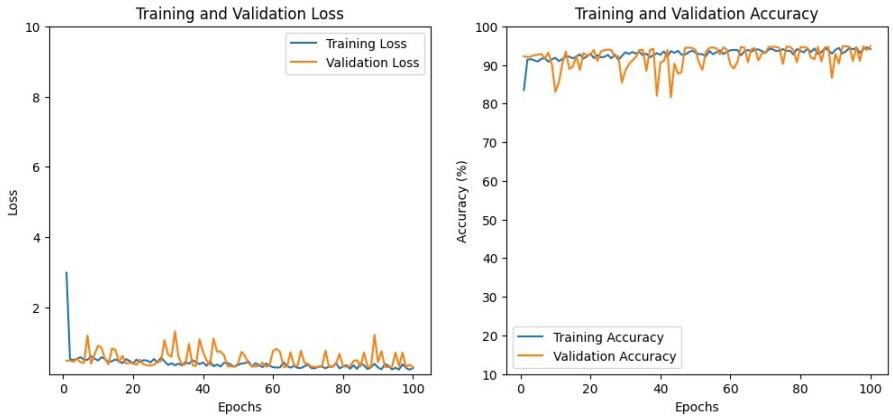


Figure 3.2: **Training and Validation Curves of the AIM-X Model**. The plot demonstrates the model's high performance, with validation accuracy converging between 92-94.5%, indicating robust generalisation. Simultaneously, the loss function (e.g., CrossEntropyLoss) steadily decreases, confirming effective optimisation and minimised prediction errors. These results highlight the model's stability and suitability.

- Number of Layers: The model's capability to capture the data's inherent complexity may be estimated based on GAs can evolve neural network architectures to determine the optimal number of layers, finally improving model performance (Deng, 2023).

The six steps that have been taken to implement GA:

1. Initialisation: Initialise a population of hyper parameter sets. Each individual represents a different set of hyperparameters –'lr', 'batch size', 'dropout', 'num layers'.

2. Fitness function : The training of the model with each set of hyperparameters evaluates the fitness - loss. Calculate how the model is performing good for each hyperparameter setting.

3. Selection : Select the best performing hyperparameters set will for reproduction

4. Crossover: Take two parents from the individuals of a different hyperparameter setting and allowed offspring to improve the performance more

5. Mutation One or more hyperparameter were changed to prevent converging to a local optima.

6. Evolution The process was applied for multiple generations in order to reach best individual - hyperparameter settings- by its fitness score (Lambora et al., 2019; Mirjalili et al., 2019).

3.2 Experiments

Essentially, medical imaging reports form one of the most critical aspects of diagnosis. Report generation from medical multimodal data involves imaging and text reports, which have been a very potential area of clinical assistance in diagnostic tasks. This model exploits an attention mechanism in pointing out the most critical areas within the images for the healthcare expert to make the most precise and effective diagnosis. This model further improves coherence and relevance in the generated reports by embedding the model with an explainability layer, allowing the visualisation of the attention weights over images. Further, genetic methodology is applied to better the values of the hyperparameters of this model for its best performance and results.

3.2.1 Experimental Setup

Data provided by the Indiana University hospital network and are divided into two parts (OpenI, 2024):

1. X-rays: IU data has provided chest X-ray images; hence, two types of views are shown: frontal and lateral views. The dataset consists of 7,472 x-ray images in total. These images represent chest X-rays of different patients.

2. Reports: The second part of this dataset contains various reports in XML, for the patients. Each XML document refers to one patient and contains the report about the x-ray images referred to that patient. The id attribute includes the name of the image, which corresponds to the PNG images. One report or one XML document may deal with a sequence of images.

The "impressions" feature will be used as a medical report. It is the most essential element of insights that the medical report will convey. Some partial examples are shown in Fig. 3.3. Image IDs will also be extracted to relate a report to an X-ray image.

1. Text Preprocessing : For text processing, several packages such as NLTK were utilised, which extracted key features from XML files and prepared them for further processing. In the dataset, we ignored the data points that did not carry any associated images or impressions. To clean the text, several steps were taken:

    - Deleting extra spaces.
    - Removing words like "XX" and "XXX".

Table 3.1: Examples of refining the dataset we came to split and integrate into the framework

Images	**Impression**
CXR1_1_IM-0001-3001.jpg; CXR1_1_IM-0001-4001.jpg	Normal chest
CXR1000_IM-0003-1001.jpg; CXR1000_IM-0003-3001.jpg; CXR1000_IM-0003-2001.jpg	One increased opacity right upper lobe associated atelectasis may represent focal consolidation mass lesion atelectasis recommend chest CT evaluation. Two opacity overlying left fifth rib may represent focal airspace disease
CXR1008_IM-0009-2001.jpg; CXR1008_IM-0009-4004.jpg	No acute cardiopulmonary abnormality
CXR1015_IM-0001-2001.jpg; CXR1015_IM-0001-1001.jpg; CXR1015_IM-0013-1001.jpg; CXR1015_IM-0013-2001.jpg	Bibasilar opacities right greater than left, features suggest combination consolidation and atelectasis

- Removing stop words, and numbers were transformed into words (e.g., 8 to “eight”, & to “and”).
- All special characters, parentheses, and brackets were removed.
- All text was changed to lowercase and contractions like “isn’t” and “doesn’t” expanded.
- Words like “dt, “rt, have also been removed.

After processing, the final structure of our data framework looks like the following:

We have 3,820 entries in the dataset. There are altogether 7,814 X-ray images both from the frontal and lateral view. The impression variable summarises data that often are used to provide diagnostic insight to medical data. The CDF plot of Fig. 3.4a shows how the cumulative probability of the values of Impression varies across the dataset giving insight about the general distribution and concentration of the values. Fig. 3.4b provides more details with respect to the distribution of the Impression values by emphasising the density and frequency of such values across intervals. This detailed analysis will allow for insight into the frequency of certain Impression values occurring and how they cluster in the whole dataset, integral components in the conveyance of the consistency and variation within the medical impressions (Tuerlinckx, 2004).

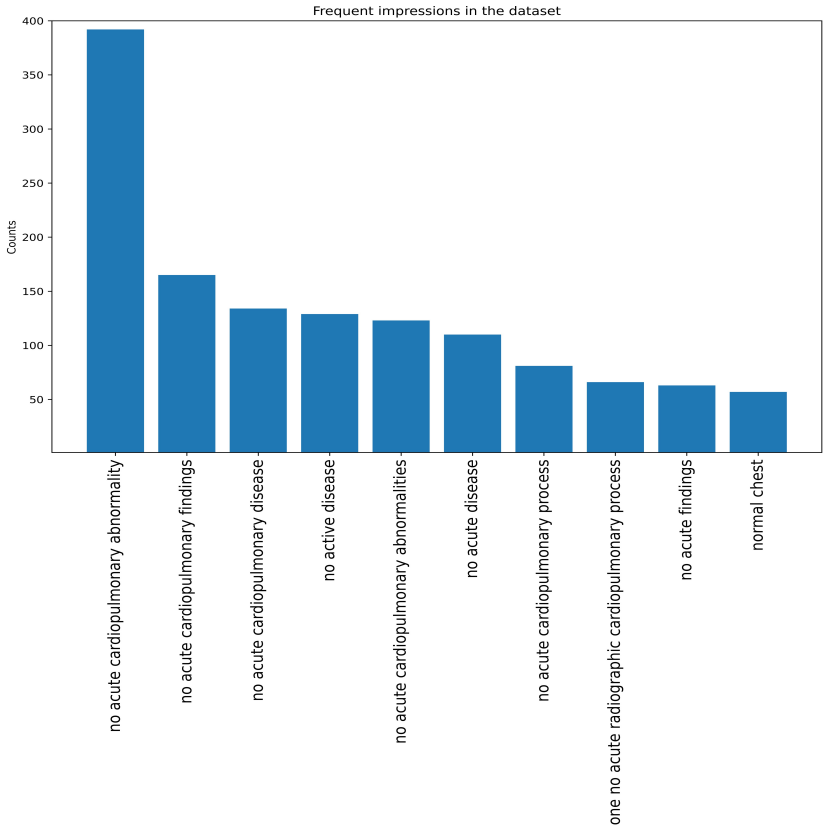
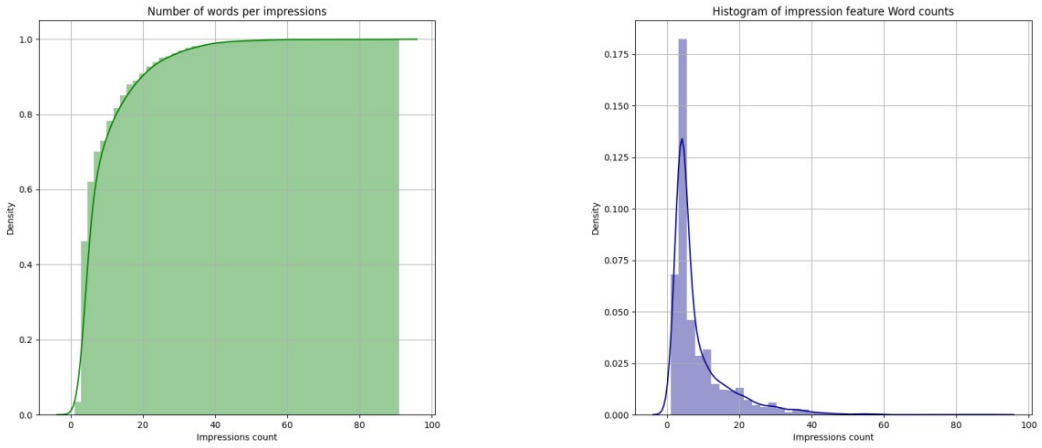


Figure 3.3: A frequency analysis was performed on the impressions to identify the most commonly observed entries in the dataset. This process involved counting the occurrences of each impression and creating a visualisation of the top 10 most frequent ones

2. Image Preprocessing: Original X-ray images are resized into regular shapes of 224x224 pixels to provide standardised input to the machine learning model. Normalisation of pixel values is performed in the range of 0 to 1 for fast convergence during training.



(a) (b)

Figure 3.4: The Impression feature's distribution is visualised through (a) the Cumulative Distribution Function (CDF) and (b) the Probability Density Function (PDF). The CDF plot shows the cumulative probability of the Impression feature values, while the PDF illustrates the density and frequency of these values across the dataset. Together, these plots provide a comprehensive overview of the distribution characteristics of the Impression feature.

3.3 Results and Discussion

We adopted a few measures to evaluate the model performance Based on Bilingual Evaluation Understudy Score (BLUE), and Recall-Oriented Understudy for Gisting Evaluation (ROUGE). They were utilised to evaluate the generated report/text against the ground truth reports. Consequently, all evaluation measures were applied exclusively to the "impression" portion of each data point to ensure a focused and precise assessment of the model's performance in generating contextually relevant and clinically accurate outputs. The AIM-X model Fig. 3.1 demonstrates strong performance in medical report generation, achieving high accuracy up to 94.5% on validation data while maintaining low loss values, indicating robust learning and generalisation. The training curves show stable convergence, with no signs of overfitting, suggesting effective regularisation.

We perform a comparison with several models that were proposed for the generation of medical reports and the captioning of images. These include the state-of-the-art methodologies including a benchmark method for image captioning, we compared our model with attention-based mechanisms such as Attend and Tell AdaTT) (Lu et al., 2017). TieNet (Wang et al., 2018b) was also used in the comparison. Finally, a reinforcement learning-based model, called the CCR model, was included, which is optimised for accuracy and inherently boosts precision (Liu et al., 2019).

All methods were applied to the Open-I dataset using the same set of evaluation metrics for consistency. The resulting scores for each model across the respective metrics are reported in Table 3.2 providing a comprehensive assessment of their performance.

Table 3.2: Performance of various models in terms of NLG metrics. BL-n represents the BLEU score using up to 4-grams, while RG-L stands for ROUGE-L. AIM-X shows the improvement across all NLG metrics relative to the Baseline versions (1,2). TieNet (Wang et al., 2018b), AdaTT (Lu et al., 2017), CCR (Liu et al., 2019)

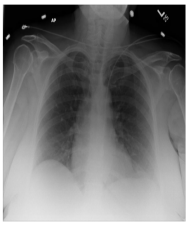
Model	**NLG Metrics**				
	BL_1	**BL_2**	**BL_3**	**BL_4**	**RG_L**
TieNet	0.330	0.194	0.124	0.081	0.311
AdaTT	0.220	0.127	0.089	0.069	0.308
CCR	0.162	0.084	0.055	0.036	0.244
Baseline V1	0.083	0.035	0.026	0.023	0.082
Baseline V2	0.207	0.105	0.069	0.058	0.199
AIM-X	**0.727**	**0.6595**	**0.601**	**0.469**	**0.727**

In contrast, our baseline models follow a conventional design:

- The first baseline model (v1) uses the image modality's query to interact with the text modality.
- The second baseline model (v2) employs the text modality's query to interact with the visual modality.

Additionally, both baseline models combine self-attention and cross-attention mechanisms within their respective modalities—visual for the first model and text for the second. As shown in the evaluation results Table 3.2, this comparison highlights the superior performance of AIM-X model over these baselines.

Table 3.3: Analysis of the alignment between ground truth reports and AIM-X generated outputs for sample data points

Image Sample				
Ground Truth	no active disease	no acute cardiopulmonary findings	hyperinflated lungs no acute cardiopulmonary abnormalities	no radiographic evidence acute cardiopulmonary disease
Generated Report	no active disease	no acute cardiopulmonary abnormalities	lowype volumes lungs lungs no acute cardiopmonary abnormalities abnormalities	no radiographic evidence acute cardulmonary disease disease

3.4 Conclusions

A novel architecture, AIM-X, designed for generating diagnostic reports from multimodal medical data. AIM-X focuses on the integration of X-ray images with their corresponding textual reports, leveraging an enhanced attention mechanism as its core innovation. This mechanism enables the model to selectively prioritise critical features in both the visual and textual data, ensuring that the generated diagnostic reports are both contextually relevant and clinically accurate. By effectively aligning visual and textual data, the attention mechanism enhances the coherence and relevance of AIM-X's outputs.

AIM-X framework uses a CNN-based architecture for the extraction of robust image features through DenseNet and a Transformer-based model for text processing. The integration of these features, via attention mechanisms, enables AIM-X to achieve highly accurate and rich-in-context reports. This fusion not only improves diagnostic accuracy but also enhances the interpretability of AI-driven decisions, making AIM-X a valuable tool for medical diagnostics.

AIM-X's demonstrates potential as a proof-of-concept, showcasing its great capabilities in enhancing diagnostics for medical diagnosis with explainable, contex-

tually appropriate reports. Moreover, the inclusion of attention mechanisms for multimodal data integration positions AIM-X as a versatile platform that could extend beyond radiology diagnostics into other medical domains.

Despite its promising potential, AIM-X has some limitations. A major challenge lies with class bias inherent in medical datasets. The training datasets are often imbalanced in their representation of frequently occurring conditions (e.g., pulmonary or cardiovascular conditions), yet underrepresent normal results (i.e., reporting of "Normal chest X-ray"). This inherent bias distorts the model's output, leading to reasonable yet inaccurate conclusions about under-presented or minor conditions. For instance, analyses using data from the IU hospital system demonstrated high accuracy by the model when it was used to diagnose common cases, but performed very weakly with infrequent conditions—particularly in longer reports where subtle details were critical.

As illustrated in Fig. 3.3, this disparity reflects the model's bias towards more common diagnoses. Follow-on studies should quantify this effect by comparing systematic accuracy measures among high-frequency and low-frequency conditions.

In a nutshell, bidirectional refining attention mechanism is the central contribution of AIM-X, which enables the effective fusion of multimodal data to improve both the accuracy and interpretability of medical report generation. Future work will focus on refining this attention mechanism, extending its applicability to other medical imaging modalities, and incorporating feedback from clinical professionals to further enhance its real-world impact.

Chapter 4

GNN-Based Shapley Values Interpretation for Relapse Breast Cancer

The present chapter can be viewed as an extension of the disease detection framework as described in Section 1.1.2. In the previous chapter, we had dealt with detecting the existence of primary lung tumours and thoracic abnormalities through AIM-X model in Chapter 3. We are now moving toward a more sophisticated and clinically relevant goal, the prediction of breast cancer recurrence. This transition from detection to prediction is a reflection of not only an advanced approach, moving from detection to something more sophisticated, but a reactive response to a clinically significant question in oncology; being able to understand the outcome following treatment and the likelihood of its relapse. Recurrence prediction is both a novel and distinct undertaking with respect to diagnostic tasks, as it requires understanding the temporal and complex dynamic aspect of patient-specific considerations over time, where the task at hand is to model the big picture risk factors that are associated with relapse.

To achieve this, we need to extend our methodological toolbox. Effective methods that we used in the prior chapter where convolutional neural network and attention mechanisms were used to detect the existence of disease, we will employ core node classification methodology in graph-based models. The future of deep learning in medically-based research is set to be woven through the growing application of GNNs to facilitate modelling of the ties between different, heterogeneous clinical, pathological and imaging data. Whether we bind patients together through their clinical characteristics as nodes is better suited to modelling the risk factors associated with not only local and distant relapse but in predicting outcomes.

Relapse or recurrence of breast cancer after treatment continues to be a ma-

jor concern for both patients and healthcare providers. It has long been reported that morbidity and mortality observed in patients with breast cancer primarily result from disease progression through local-regional or systemic dissemination, as well as the occurrence of recurrence post-treatment, or a combination of these factors (Rabbani et al., 2007).Upon development of metastatic deposits outside the primary location, the prognosis of patients is dramatically worsened, and cure would be unlikely (Nicolini et al., 2006). The ability to accurately predict the probability of relapse can greatly help in treatment planning, monitoring, and ultimately improving patient outcomes (Riggio et al., 2021; Hutchinson et al., 2010).

Researchers have achieved notable advancements in the development of predictive models of relapse status in breast cancer through the effective use of node classification graphs (Chen et al., 2011). One particularly interesting research study was explored the predictive potential of K-Banhatti and Zagreb type degree-based topological indices in quantitative structure–property relationship(QSPR) analysis for medications used to treat type-I and type-II diabetes. These indices were computed for 14 anti-diabetes drug molecules using edge and vertex partitioning techniques. By utilising these topological indices, researchers developed QSPR regression models to predict the physicochemical properties of the drugs under study (Ullah et al., 2024).

Another interesting study discussed the use of topological indices (TIs) to predict the physical and biological properties of drugs used to treat Alzheimer's disease. Degree-based It were generated using edge partitioning for drugs like Tacrine, Donepezil, and Rivastigmine. A QSPR model was developed using linear regression to predict characteristics such as boiling point, flash point, molar volume, molecular weight, complexity, and polarisability. The findings suggest that TIs can be valuable tools in drug discovery and design for Alzheimer's disease treatment (Ahmed et al., 2024).This study focuses on a node classification graph representing the interconnectivity between nodes, where each node corresponds to an individual patient. This representation encapsulates the intricate relationships and interdependencies among various factors that contribute to relapse, encompassing clinical variables, histopathological features, and treatment regimens. Through the utilisation of machine learning algorithms on these graphs, researchers can discern patterns and features that possess predictive value for relapse occurrence.

A notable advantage of employing node classification graphs lies in their capacity to accommodate the heterogeneous nature of breast cancer data (Song et al., 2014). Breast cancer is a multifaceted ailment influenced by diverse variables that interact with one another. By representing these variables as features of the nodes and their relationships as edges, node classification graphs offer a comprehensive framework for modelling and analysing these intricate connections' recent years, several studies

have been dedicated to improving the precision of predicting re-lapse recurrence in breast cancer through the use of various techniques (Testa et al., 2020).

Using medical imaging techniques such as mammography, magnetic resonance imaging, or PET-CT scans, researchers can gain pertinent information on tumour properties (Luo et al., 2019). Radiomic [1] models, which involve the extraction of a multitude of features from the imaging data, enable the capture of subtle patterns and textures that are indicative of relapse. These extracted features are sub-sequent employed to train machine learning models, thereby enhancing the accuracy of re-lapse prediction. Furthermore, the inclusion of clinical and pathological information (Acharya et al., 2008) and the analysis of longitudinal data are instrumental in elucidating the dynamic changes associated with relapse (Zhu et al., 2022b). These additional factors contribute valuable information to advance our overall understanding of relapse recurrence in breast cancer.

This chapter aims to tackle the challenges associated with understanding the relational nature of patient data and determining the contribution of inputs within a neural network framework. To address these challenges, we propose a novel approach based on Graph Neural Networks (GNNs), we called GraphX-Net Fig. 4.1, that harnesses Shapley values to evaluate the significance of features and generate an explainable graph representation. This is achieved by employing XAI techniques to enhance transparency and interpretability in predicting relapse. To gain insight at both local and global levels, we incorporate Shapley values as a means of quantifying the contribution of each feature to relapse prediction for individual patients, as well as the entire dataset. To classify unlabelled nodes and establish edges between nodes, we adopt a weighted approach that considers the contributions of features to measure probabilities. Furthermore, we initialised two graph convolutional layers as message passing layers to aggregate, trans-form, and update node representations, incorporating the information learned from the underlying graph structure.

Key Contributions of GraphX-Net:

1. GraphX-Net leverages the power of GNNs that handle complex relationships and dependencies of graph-structured data very effectively. Patients and their relevant features are treated as nodes of a graph, and edges depict the relationship between patients on the basis of similarity.

2. GraphX-Net is able to capture both local and global information of patients and features, thus giving more accurate predictions compared to traditional

[1] The term radiomics is referring to the high throughput extraction of quantitative features from medical images (CT, MRI, PET) in order to characterise tumours, predict treatment response, and/or assess disease progression. In effect, it is transforming images into data which can be mined using complex computational analysis—which can include machine learning or deep learning analysis (Kumar et al., 2012).

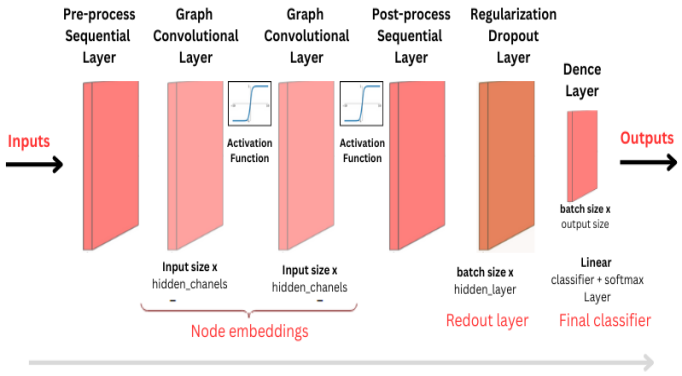


Figure 4.1: GraphX-Net model architecture. The model applies preprocessing using feed-forward network to the node features to generate initial node representations. Also applies two graph convolutional layers, with skip connections, to the node representation to produce node embeddings. And finally feed the node embeddings in a Softmax layer to predict the node class. Each graph convolutional layer added captures information from a further level of neighbors.

methods.

3. GraphX-Net uses Shapley values for the interpretability and explainability of predictions: it can identify the importance of different features in predicting a relapse, hence giving a ground for clinicians and researchers to understand the underlying factors driving the prediction.

4. Infusing GNNs and Shapley values within GraphX-Net empowers the process of predicting relapsed breast cancer. In fact, the architecture allows the risk of relapse to be precisely predicted and interpreted at the level of underlying factors that drive such predictions.

4.1 Materials and Methodology

4.1.1 Dataset

This study utilised a large cohort of invasive breast cancer with long-term clinical follow-up and complete clinicopathological data. This dataset consists of 1980 breast cancer cases within the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort. The METABRIC study protocol, detailing the molecular profiling methodology in a cohort of 1980 breast cancer samples is described by (Curtis et al., 2012). Patient demographics are summarised in Appendix 7.1. Data was downloaded from the cBioportal data repository and it include of clinical records, patient demographics, tumour characteristics, treatment details, and follow-up data sourced from trusted repositories (Gao et al., 2013).

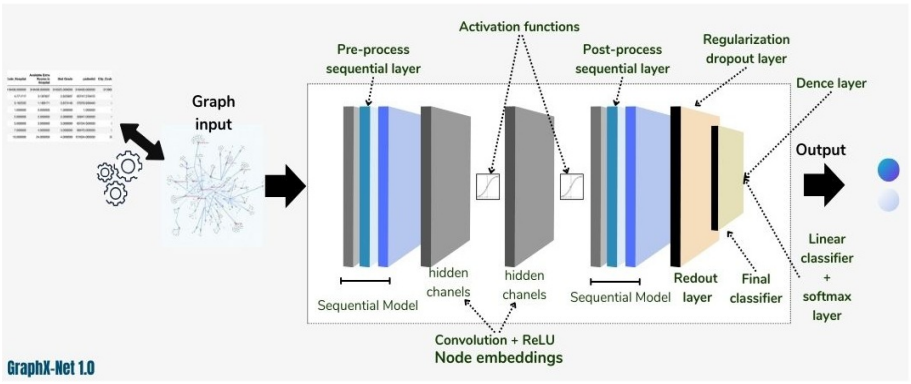


Figure 4.2: GrpahX-Net Architecture

4.1.2 Model Components

The GraphX-Net model incorporates a fusion of machine learning algorithms and deep learning algorithms. We first transformed the data into a graph representation, where each patient is represented as a node within the graph, while the edges symbolise the interconnections between patients. The determination of node edges is based on the assessment of input contributions. By analysing the positive or negative contributions of each input, the corresponding node labels can be identified.

The utilisation of Feature Shapley Values Fig. 4.5 enables quantification of each feature's contribution to the predictions made by GraphX-Net model. This computation involves iteratively training the model, while systematically excluding distinct features during each iteration. Through a comparison of prediction differences between the complete model and models with excluded features, Shapley values were attributed to individual features. These values accurately depict the incremental importance of each feature in the prediction process.

GNN is used to train the GraphX-Net model for node classification. More precisely, we investigate the model architecture that includes both pre-process and post-process layers to improve the model performance.

4.1.2.1 Pre-process Layer

It is to preprocess the input data to set the stage for effective learning. It has been built using three different layers: BatchNormalisation, Dropout, and Dense with Gelu activation.

- BatchNormalisation: Normalising the node features during training reduces the internal covariate shift. This layer helps in stabilising the learning process, hence fastening convergence to improve the overall performance of the model.

- Dropout: The Dropout layer will apply regularisation by randomly dropping

out a fraction of nodes during every epoch of training. This prevents overfitting of the network. The network will learn generalisable representations.

- Dense with Gelu activation: Apply linear transformation followed by a non-linear activation function called Gelu. Gelu stands for Gaussian Error Linear Unit type of activation function, which works very well on most deep learning models. Its smooth gradient during training contributes greatly to stability (Lee, 2023).

4.1.2.2 Graph Construction

The beginning of the graph construction process involved the application of a graph construction algorithm or technique, specially designed to establish edges between nodes, following predefined criteria. This key task is of paramount importance, as it forms the basis for the creation of pertinent connections between patients within the network. In this endeavor, we utilised the collective contributions of all input attributes encapsulated within each node to discern the intricate relationships that govern the adjacency's between the focal node and its counterparts in the graph.

$$f(N_c) = \sum_{i=0}^{n} (l_1 sv, l_2 sv, l_3 sv, ..., l_n sv) \tag{4.1}$$

This equation calculates the Shapley value scores for specific features associated with each node i from 0 to n. Here, $l_i sv$ represents the Shapley value for feature i, and the sum aggregates these values to quantify the contribution of all relevant features across the nodes.

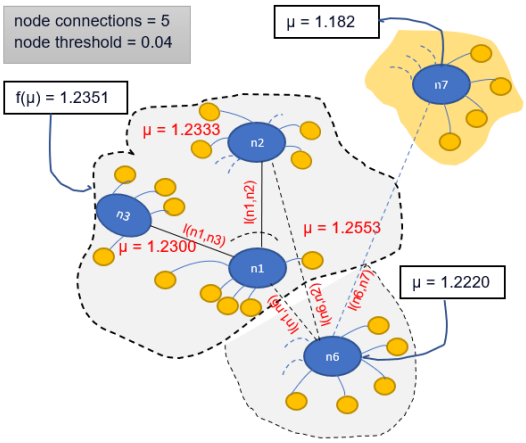


Figure 4.3: This diagram depicts the process of connecting a node by considering the information from its neighboring nodes. The figure highlights that the information of clusters is influenced by both the node threshold value and the requirement that the number of nodes in each cluster not exceed the node connections parameter.

Through the computation of contributions for all inputs within each observation, accounting for both positive and negative contributions, we can determine the probabilities assigned to each node's classification as either 0 (indicating "No") or 1 (indicating "Yes"). Moreover, by considering the average contribution of adjacent nodes, we can assess their respective adjacencies.

By performing calculations on the contributions of all inputs for each observation, encompassing both positive and negative contributions, we derived the probabilities associated with the classification of each node. Additionally, we evaluated the adjacencies of neighboring nodes based on their respective contributions. For example, it suggests that the node should be connected to a group where the average contribution is either less than or equal to a predefined threshold. This threshold serves as an integer parameter that is incorporated into the grouping of nodes. To determine the connectivity of a new node to a group, the difference between the average contributions of the group and the contribution of the new node was assessed against the threshold. If this difference was less than or equal to the threshold, and the total number of nodes in the group remains within the permitted group size, the new node should be connected to the group as depicted in Fig. 4.3. The figure uses different colours to indicate clusters of nodes. Nodes within the gray area all belong to a single cluster, where the mean contributions meet the node threshold value. In contrast, the node in the yellow area, although connected to the gray cluster, belongs to a different cluster. This is because its mean contribution does not align with the overall mean of the gray cluster.

4.1.2.3 Feature Extraction

Here we used XGBoost method as the basis for calculating the relative contribution of each feature to the model. This is based on the gain metric to evaluate the importance of a feature by considering its contribution between individual trees within the model (Ben Jabeur et al., 2023). A higher gain value, relative to other features, signifies greater significance in generating predictions. Consequently, the gain metric served as a pivotal attribute for interpreting the relative importance of each feature. By calculating the gain values, we can effectively pinpoint the most influential features within the model as shown in Table 4.1.

Patient attributes and clinical data are of significant importance in the prediction of relapse in breast cancer cases. Various features, including patient demographics, medical history, genetic information, and treatment records, are carefully extracted and associated with each patient node. These features were obtained using the gain metric, which allows for the identification of relevant attributes to predict relapse. These features are incorporated as node features in the graph representation. This

Algorithm 1 Calculate the total contributions of each node

Input: lst_shap_values ▷ List of all Shapley values
Input: df ▷ Data frame
Output: df ▷ Data frame updated

for each $i, \text{row} \in \text{df}$ **do**
    pos_cont $\Leftarrow 0$
    neg_cont $\Leftarrow 0$
    for each $c \in$ df.cols **do**
        f_cont $\Leftarrow$ lst_shap_values$[i]$
        if f_cont > 0 **then**
            pos_cont $\Leftarrow$ f_cont + pos_cont
            pos_cont $\Leftarrow$ pos_cont
        end if
        if f_cont < 0 **then**
            neg_cont $\Leftarrow$ f_cont + neg_cont
            neg_cont $\Leftarrow$ neg_cont
        end if
    end for
end for
return df

Algorithm 2 Create Graph's Adjacency Matrix

Input: df ▷ Data frame
Output: adjacency Matrix

for each `row` **in** df **do**
    neg_cont $\Leftarrow$ row.neg_cont
    pos_cont $\Leftarrow$ row.pos_cont
    classi $\Leftarrow$ row.relapse_free_status
    patient_id $\Leftarrow$ row.patient_id
    node_ad $\Leftarrow$ get_node_adjacents(neg_cont, pos_cont, classi, patient_id)
    adjacency matrix $\Leftarrow$ node_ad
end for
return adjacency Matrix

Algorithm 3 Get Node Adjacents

Input: node_connections ▷ Cluster nodes = 5
Input: node_threshold ▷ Node's threshold parameter = 0.04
Input: neg_cont ▷ Node negative contributions
Input: pos_cont ▷ Node positive contributions
Input: classi ▷ Node classification
Output: edge_indexes

```
for each row in df do
    if (classi = 0 and row.relapse_free_status = 0) then
        Add node, negative contributions into an array
    end if
    if (classi = 1 and row.relapse_free_status = 1) then
        Add node, positive contributions into an array
    end if
end for
if (classi == 0) then
    Sorting negative contributions
    for each i, v in neg.dic do
        if (neg_cont < v and i ≠ patient_id) then
            if (adjacents ≥ node_connections) or (count > 1 and (v -
(sum/count)) > node_threshold) then
                Exit Loop
            end if
        end if
        adjacentsArray ⇐ node
        count ⇐ count + 1
        sum ⇐ sum + v
    end for
end if
if (classi == 1) then
    Sorting positive contributions
    for each i, v in pos.dic do
        if (pos_cont < v and i ≠ patient_id) then
            if (adjacents ≥ node_connections) or (count > 1 and (v -
(sum/count)) > node_threshold) then
                Exit Loop
            end if
        end if
        adjacentsArray ⇐ node
        count ⇐ count + 1
        sum ⇐ sum + v
    end for
end if
return adjacentsArray
```

#	**Feature**
1.	Patient's Vital Status
2.	AGE_AT_DIAGNOSIS
3.	Relapse Free Status (Months)
4.	Overall Survival (Months)
5.	LYMPH_NODES_EXAMINED_POSITIVE
6.	Overall Survival Status
7.	TUMOUR CELLULARITY
8.	Nottingham Prognostic Index
9.	Integrative Cluster

Table 4.1: The most significant 9 input features in the Relapse breast cancer dataset.

integration enabled a thorough analysis of the breast cancer relapse graph by leveraging the insights gained from the XGBoost model. The presence of these features as attributes associated with each node offers valuable information for this node.

By incorporating the insights obtained from the XGBoost model within the graph context, a deeper understanding of the underlying patterns and relationships within the breast cancer relapse graph can be achieved. Each node within the graph corresponds to a patient, with the colour assigned to each node indicating the occurrence of relapse for that particular patient.

4.1.2.4 GraphConvLayers Module

The core of our node classification model involves the two graph convolutional layers. Graph convolutional layers have been specially designed to handle graph-structured data and can capture local and global dependencies among nodes. Such a layer aggregates information from neighboring nodes to update node embeddings using the adjacency matrix of a graph for weighing the importance of each neighboring node to capture structural dependencies of the graph. Each layer refines the node representation so that the model can, in a progressive manner, learn to understand the underlying topology of the graph.

4.1.2.5 Post-process Layer

This post-processing layer was the last piece of our GNN-based node classification model, which served to polish the learned representations and generate the final prediction of nodes. Also, similar to the pre-processing layer, the post-process layer included three layers: BatchNormalisation, Dropout, and Dense with Gelu activation. The output of the model is a probability distribution over all classes for each node, reflecting the likelihood of each node to belong to each class.

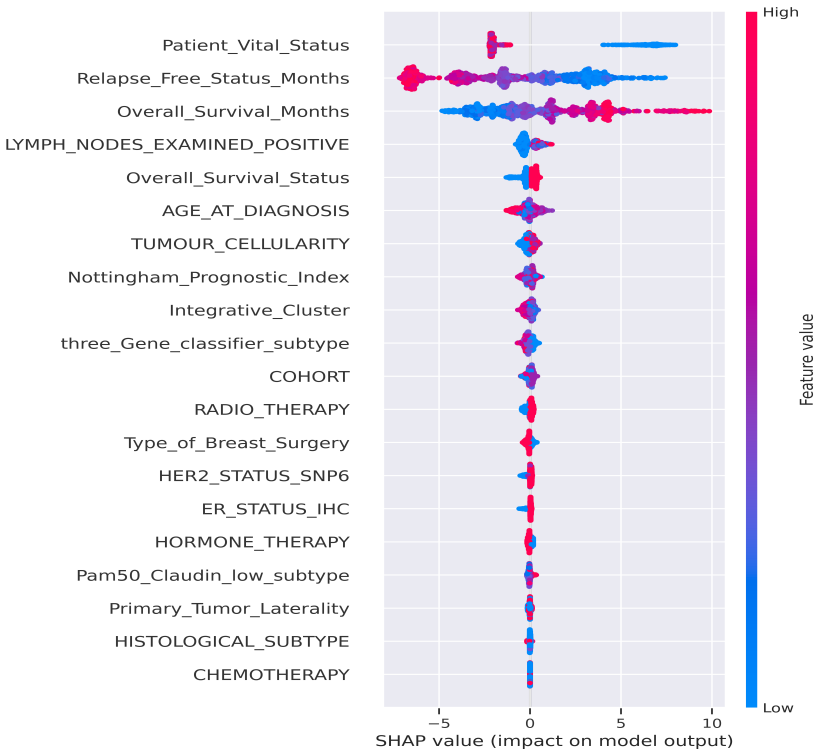


Figure 4.4: Features are ranked according to their Shapley values, which are highly useful in making sense of and interpreting the feature effects on the output of a machine learning model

Thus, our model architecture for training a GNN using the GraphConvLayer module for node classification was carefully designed. Both pre-process and post-process layers enhance this model's capability of capturing features and structural dependencies of the graph by incorporating BatchNormalisation, Dropout, and Dense with Gelu activation. Such a powerful zero-shot node classification approach allows achieving very good results of node classification on many applications and exploits all the potentials of GNNs to solve real-world problems in the most effective way.

4.1.2.6 GrpahX-Net Features

The features are mapped to a graph representation, where every patient was a node of the graph; such edges of the graph connect patients to each other. Determining the edges of a node depends on an evaluation of the contributions by the input. Node labels associated with such nodes can be inferred by looking at which contributions of the inputs are positive and which are negative. Fig. 4.4 provides the contribution of each feature toward the prediction from the GraphX-Net model. Calculations would be done in iteration; in each iteration, features of interest are left out. Comparison Shapley values were assigned to each of the features based on the differences in the

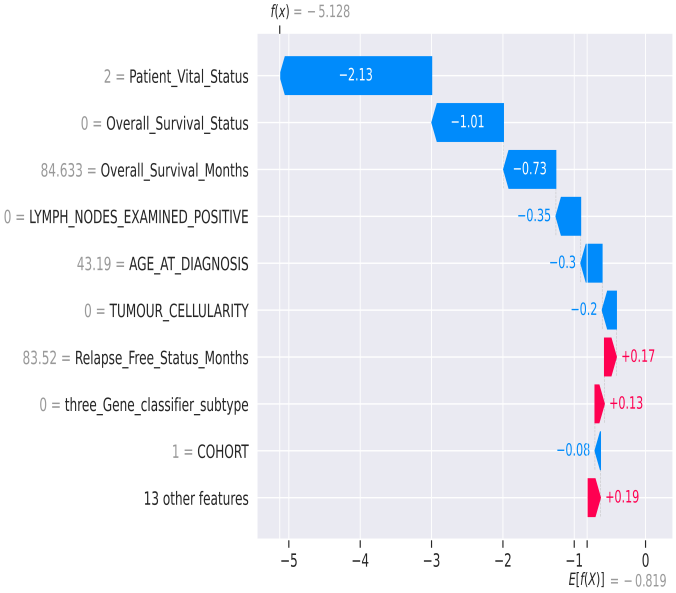


Figure 4.5: SHAP Values Waterfall plot represents how the SHAP values impact the individual predictions in GraphX-Net model

predictions between the complete model and the models where those features had been excluded. That gives the right description of how each feature is incrementally important in making the prediction. The SHAP summary plot really helped identify which were the most prominent features and the range of effects in that feature through the whole data. When the Feature Shapley Values were evaluated against the GraphX-Net model, rich and valuable insight and interpretations would have been yielded. It also allows the finding of influential features and explain their impact on GNN-derived predictions.

Fig. 4.5 Shapley Values analysis provides further visualisation and insight into a deeper understanding of the relationships and dependencies involved within the graph. In addition, this procedure helped to pinpoint possible biases or constraints inherent within the model.

Fig. 4.6 LIME local explanation displays the raw feature values included in the model, such as Relapse_Free_Status_Months (0.29) and Integrative_cluster (0.01). Each of these feature values depicts a patient-specific input used to inform the model's prediction. It shows the amount of weight assigned to each feature towards the predicted outcome (e.g., Patient_Vital_Status received the largest positive weight (0.55) and Overall_Survival_Months has positive weight (0.27), indicating are strongly back the prediction, while CHEMOTHERAPY and AGE_AT_DIAGNOSIS have negligible contributions (0.00)). The values shown under each feature expresses how much they either pushed the prediction towards class 1 (e.g., relapsed) or class 0 (e.g., not relapsed).

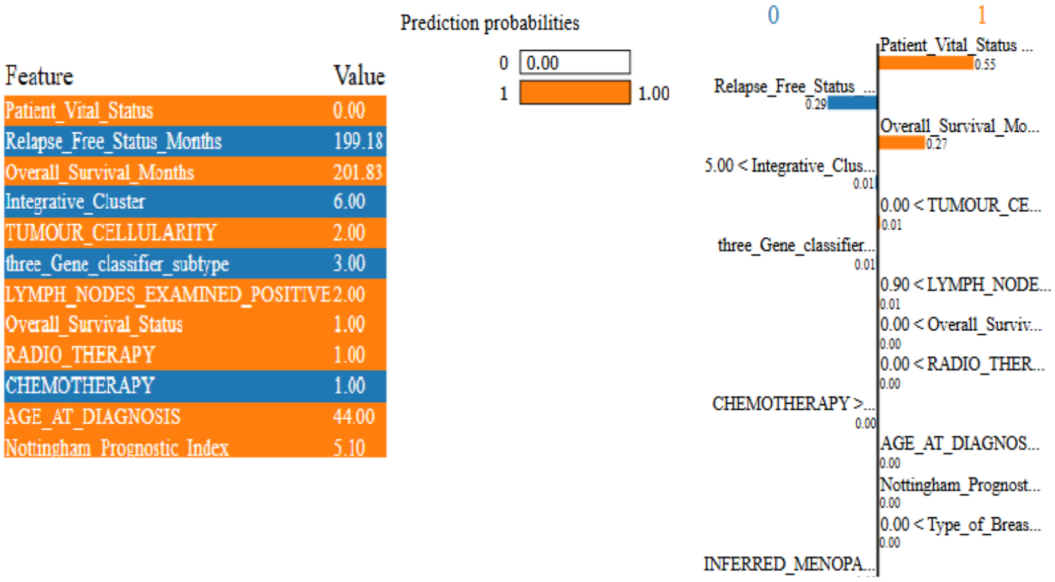


Figure 4.6: The LIME explanation of a clinical prediction model illustrates the contributions (weights) to the predicted outcome defined by the model.

For the measurement of the effectiveness of the GraphX-Net model, its performance was ranked as an estimate of how well it generalises to unseen data. There are a few techniques that work for such purposes. The major techniques for evaluation that were considered for GraphX-Net involve:

Splitting into Train-Validation-Test Data is split into trains 0.07 and 0.15 for validation and test data, respectively.

Tuning the Learning Rate Lambda Callback package has been used in detecting the best learning rate value with callback actions; all loss values have been recorded. The best learning rate may correspond to the learning rate where the loss drops off the fastest before it begins divergence or oscillation. In this case, LR between 1.2 x 10^{-2} and 1.3 x 10^{-2}, as shown in Fig. 4.7.

CrossValidation kfold cross-validation was performed in order to get more robust performance estimates with k-fold = 3

Task-Specific Metrics As shown in Table 4.2.

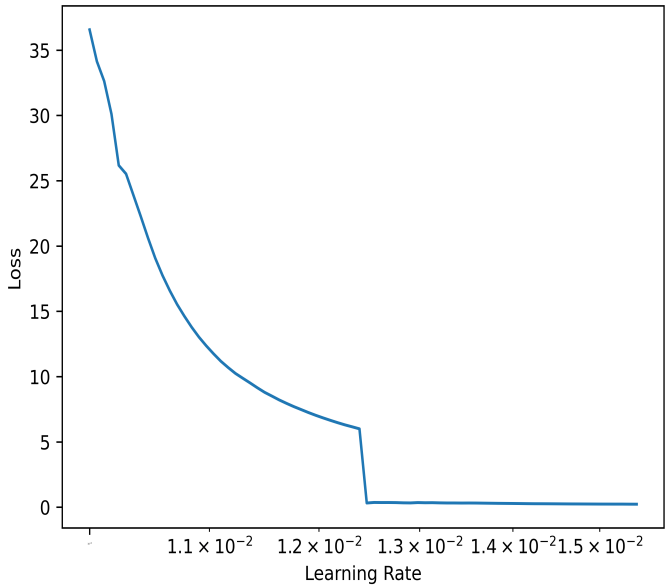


Figure 4.7: Exploring the Impact of Learning Rate on Model Training: An In-Depth Analysis of Learning Rate vs. Loss

4.2 Experiments

We describe in this section the different experiments done to evaluate the performance of the proposed method, including explanations of the values adopted during parameter tuning. Finally we introduce the default versions and the refined versions of the evaluation measures used to assess performance of the methodologies.

In conclusion, we discuss experimental results with statistical significance testing and prove merits of GraphX-Net in comparison to other methods. Using a large cohort of invasive breast cancer with long-term follow-up of clinical characteristics in addition to complete clinicopathological data. This dataset, therefore, comprises 1980 cases of breast cancer. AppendixA 7.1 summarizes patient demographics, characteristics of the tumours, the details of the treatment.

The follow-up data sourced from trusted repositories (Gao et al., 2013).

4.2.1 Experimental Setup

Table 4.2 contains raw data, already performing through different preliminary preparation processing on which allows the suitability of analysis performed using its clean variant. In the first sight, data is provided on a tabular format needed several data wrangling in order to process the problem:

Firstly, it contains missing values that are to be uniformly imputed or assigned throughout the data. Here, this data contained were removed for ensuring compatibility at these columns. Furthermore, certain parameters are categorised to make

them more useful for further analysis. Inconsistencies were checked and corrected to ensure data integrity. Then, we cleaned the data, and after this, some pre-processing was done such as feature selection and normalisation.

The key steps involved in processing the METABRIC dataset to make it ready for efficient analysis and modeling are as follows: The first and foremost important task involves the selection of the target variable, say, 'Relapse free status'. Further, feature selection techniques identify the most relevant attributes related to the discussed clinical outcome. In this regard, SMOTE was performed in order to handle the class imbalance problem by generating artificial samples corresponding to the minority class. Similarly, the ExtractTreeClassifier class can deliver informative features from the dataset, extending beyond raw attributes. The data set was further divided into training, validation, and test sets subsequent to feature selection and extraction. Feature importance was identified using, but not limited to, the XGBoost classifier and Shapley values, thus providing a clear idea of the contribution of each feature. Then, traditional machine learning algorithms, Boosting, Random forest, and Stacking algorithms were developed and analysed for their predictive performance. A graph-based approach is also taken to explore feature relationships, whereby features are treated as nodes and their relationship as edges. Graph neural networks for node classification on this feature graph gave a more holistic representation of the features, feature interaction, and their impact on the target variable. The approach presented here is designed to be comprehensive in improving our understanding of breast cancer prognosis and treatment decisions.

Processing of the METABRIC data included target selection, feature engineering, SMOTE for class balancing, advanced feature analysis using XGBoost and Shapley values, classical ML modeling, and GNN-based feature interaction explorations. Together, these have allowed strong predictions and further insight into the outcomes of relapse breast cancer patients.

Feature selection helped to choose most informative variables that can predict relapse. It reduces the dimensionality and enhances the performance of the model as shown in Table 4.2. Normalisation ensures that features are on a common scale to avoid bias due to different magnitudes. These cleaning and preprocessing steps formed a vital backbone for our later analysis and modeling task because they guaranteed consistent, and hence meaningful results.

Table 4.2: Experiment findings on feature selection and other metrics

Illustration	**METABRIC cross-sectional data**
Size of the training set	1063
Size of the validation & test set	456
Total class count	2
Total number of features in the dataset	24
Number of trainings data features after correlation	10
Number of features in the optimal set	9
Learning rate	1.2 x 10^{-2}
Hidden units	[32,32]
Dropout rate	0.5
Epochs	300
Batch size	128

4.3 Results and Discussion

Through the applications of GNNs and Shapley values, the prediction for relapse in breast cancer patients distinctly improves regarding accuracy and explainability. The intricate relationships between clinical variables, treatment history, and genetic markers in patient data are modeled efficiently by GraphX-Net. The inclusion of Shapley values improves this method as a mathematically grounded mechanism that assesses the relative importance of individual features and interactions present in a patient's treatment graph. The synergy of these elements assures predictive superiority along with a framework in which clinicians to determine the basic Driving factors for relapse. Like what has been depicted in Fig. 4.8, GraphX-Net illustrates how it achieves a much better result than the predictive and explainable approaches of classical machine learning models. Although traditional methods appear as a black box that provides little information back to the decision process, ours allows for pinpointing the focal points of relapse in patients. The utmost gap in breast cancer prognosis that this study focuses on is the absence of an exact and explainable predictive model. Current models offer limited means of prediction or insufficientness to provide an intelligible explanation of output, and thus become less useful for clinicians' purposes. In extreme contrast to that, GraphX-Net bridges this gap by building state-of-the-art deep learning systematisation together with those principles of explainability lying in cooperative game theory. As a result, for the very first time, clinicians and researchers will gain profound insights into the mechanisms of observations related to relapse. The rigorous experiments conducted across other methods further strengthen the claims of superiority of GraphX-Net regarding performance and interpretability. Overall, the comparative analysis shows that this is

capable of producing higher accuracy than applying the traditional approaches.

Predictive Performance and Interpretability As illustrated in Fig. 4.8, GraphX-Net, owing to the aforementioned specific merits, outperforms conventional machine learning models not merely in explanation but in predictive quality. Unlike traditional black-box models, which seem to offer little insight into how they got there.

Addressing a Critical Gap in Breast Cancer Prognosis This study provides breast cancer prognosis with the contribution of long-awaited resolved explanation-based predictive models. Traditional statistical methods are usually dependent on linearity in their assumptions while ignoring the complex, highly non-linear interactions typical of cancer progression. On the other hand, machine learning models, which are presumed to be highly effective, often lose the inverse of interpretation, making them untrustworthy and imprudent in terms of application in high-stakes clinical decision-making. GraphX-Net creates an important bridge by providing a GNN-based architecture, strengthened by explainability principles in cooperative game theory and LIME techniques. More details about the gaps and limitations discussed in the following Section 2.6.

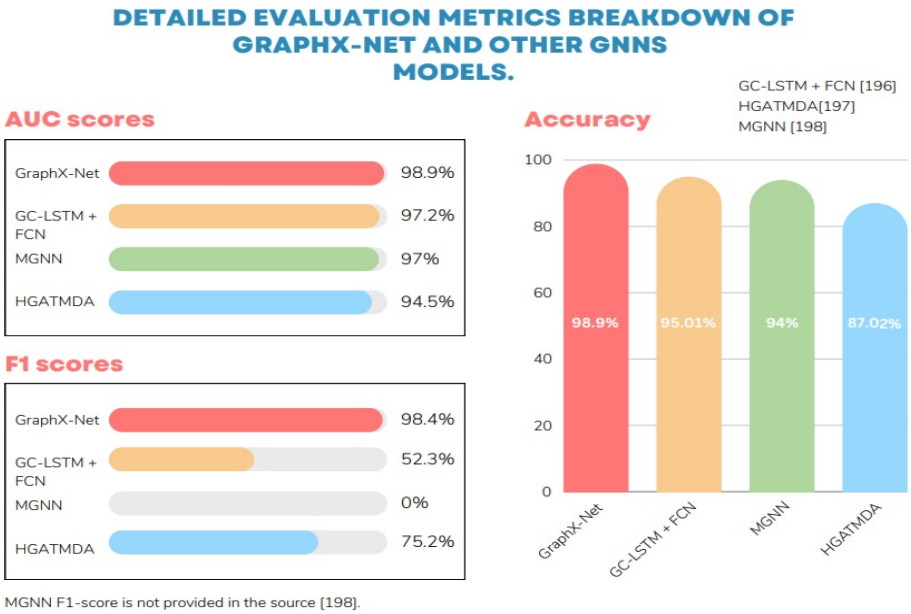


Figure 4.8: A comprehensive breakdown of evaluation metrics for GraphX-Net and other GNN models. (Ji et al., 2021; Gao et al., 2020; Gupta, 2022)

The results show that cautioning against the employment of traditional models such as Random Forest, Support Vector Machines, and classic deep learning architecture puts GraphX-Net ahead in aspects of prediction accuracy, at least in terms of:

- Lower false positive and false negative rates (prediction accuracy).

- With regard to Feature importance attribution, being more interpretable in that they shed clearer light on a few relapse-related factors.
- With regard to Model generalisability, being more resistant to overfitting than are comparable traditional models, across different cohorts.

4.4 Conclusions

The proposed model of GraphX-Net has shows good performance. Computationally, by using GNNs, structured information is dealt with effectively, capturing richer dependencies that may evade classic models. Enhanced prediction capabilities thus assure more accurate predictions. The results to be presented in this work will go beyond just advancing the current state of bioinformatics but will also point to which direction future works should proceed in order to adopt GNNs for a variety of predictive modeling tasks in digital healthcare and beyond.

One other future research direction could be the inclusion of longitudinal data of the patients to further enhance the predictive performance of the model. Such information, when put into use with temporal features such as how gene expression has changed during the period, will capture dynamic patterns and make even better predictions about recurrence. Another fruitful direction would be to study to what extent GraphX-Net generalises to other types of cancer or diseases, thus further establishing the utility of this technique and its extension to a large number of other disorders.

Chapter 5

BERT-GNN Interpretation for Metastatic Breast Cancer

In line with the framework for the detection and diagnosis of chest abnormalities and the predictive model for breast cancer recurrence presented in Chapters 3 and 4, the current chapter represents the next step in our analytical trajectory, the analysis of metastatic breast cancer (MBC) as illustrated in Section 1.1.3. MBC represents the origin of a pivotal moment in the natural history of metastatic breast cancer, as we can integrate timely and accurate prognostic indicators into treatment decisions, and ultimately improved patient outcomes. Chapter 3 educated readers on a foundation of strategies for disease identification through imaging, and Chapter 4 expanded these initial strategies used to incorporate risk prediction using graph-based modelling of recurrence residue to a hopefully level of insight for prediction. This chapter elaborates on the trajectory by transitioning the scope to a more sophisticated and semantically rich reality, to predict metastatic progression using unstructured medical narratives.

We present here BERT Graph for Metastatic Breast Cancer (BG-MBC) a new integrated framework for natural language processing (NLP) and graph learning capable of creating meaningful representations from free-text histopathology reports, and other medical documents. This often underutilises unstructured text.

The advanced metrics MBC illustrate the complexity of acting on data defined by volume instead of quality and structural modelling, as well as episodic data that are animated and predicted by instinctual actions, as opposed to "pre-structured" or "defined", to observe structural elements. This transition not only illustrates a higher methodological transition from structure-defined, label-driven data analysis to context-driven learning based and impacting evidence from text, but to show how GNNs can increasingly be added as diverse interpretations of data to further predictive insights as an alternate approach. As such, this chapter demonstrates an extension to the analytic arc formed in previous chapters and an extension of

the growing importance of integrating more profound language-based understanding with relational modelling for medical AI in contemporary health systems.

MBC is a considerable challenge to oncology and has been considered the principal cause of death among patients with breast cancer. It contributes significantly to the tumourigenic causes of mortality among women worldwide (Redig et al., 2013). Despite the improvement in mode of detection and treatment, metastatic dissemination of breast cancer remains a highly important clinical problem for which new strategies must be developed for the purpose of early diagnosis and intervention (Bertucci et al., 2019). A critical step in the progress of tumour spread, metastasis involves the spread of tumour cells from the site of its origin to a distant organ or tissue. Statistics reflect the serious nature of MBC disease, with close to 30% of all early-stage breast cancers progressing to a late stage (Siegel et al., 2021). In addition, most of the deaths caused by this disease are related to MBC; therefore, there is an urgent need to develop superior diagnostic and therapeutic methods. Cancer is a complex disease process that imposes enormous challenges concerning diagnosis and management. Precise diagnosis is the cornerstone toward the creation of appropriate treatment strategies, which would improve prognosis. Conventionally, histopathology was the backbone of diagnosis in cancer, offering a great deal of useful information about cellular and tissue-level changes in malignancies (Hoda et al., 2017; Aziz, 2007). The advent of molecular and genomic techniques for cancer diagnosis and management did not reduce the importance of histopathology; rather, it became complementary information that added to our in-depth knowledge of the disease (Gurcan et al., 2009). The inclusion of histopathological information from the molecular profiles and imaging modalities allows for a holistic approach toward cancer prognosis and thus provides for personalised treatment according to characteristic (Nowak-Sliwinska et al., 2019). Histopathology involves the microscopic examination of tissue specimens so that the assessment of morphological characteristics may be pursued: cell morphology, tissue architecture, and any pathological abnormalities, nowak2019drug, rosai2011rosai including tumour formation (Rosai, 2011). Regarding this, therefore, histopathology allows the pathologist not only to identify the kind of cancer involved but also to assess the grade or aggressiveness of such lesions, its staging, and most importantly, a certain prognosis, all important in treatment planning (Elmore et al., 2017). Histopathological techniques have been evolving over the years, improving various staining techniques, advanced imaging, and molecular diagnostics, which have enhanced sensitivity and specificity in cancer diagnosis (Colling et al., 2019).

The histopathological reports are an important way of prognosis and management of cases, including metastatic breast cancer. Histopathology reports give the histological nature of tumour specimens, which highlights critical information like

the type of tumour, grading, hormone receptor profile, and possible metastasis. Detailed histopathological analysis, therefore, lies at the heart of guiding treatment options and assessing prognosis in cases of metastatic breast cancer (Hicks et al., 2005). Large language models, such as BERT, open up newer avenues in mining useful knowledge from histopathology reports. Such LLMs with advanced NLP may be utilised to give detailed descriptions of complex textual data by extracting vital features and their context, which are indispensable in cancer prognosis (Kenton et al., 2019). The integration of LLMs into the prognostic pipeline therefore has the potential to further enhance the detection of metastatic breast cancers.

Application of ML models has lately gained significant momentum in oncology, especially for metastasis prediction and improvement of patient outcomes. Various studies have been conducted which use the technique of machine learning in order to overcome the problems associated with metastatic disease. Esteva et al (Esteva et al., 2017), used deep learning algorithms in analysing histopathology images for the prediction of the presence of metastatic breast cancer in the lymph nodes with high accuracy. The research illustrated the potentials of CNN in detecting metastatic disease from digitised tissue slides, hence promising a bright perspective for automated histopathological analysis. Liu et al. (Liu et al., 2017) has proposed an automated framework using a CNN for detecting and localising tumours in gigapixel microscopy images. On Camelyon16, the method detects (92.4%) of tumours at eight false positives per image, surpassing the previous best (82.7%) and human pathologists (73.2%). They succeeded in attaining AUC scores above (97.0%) and found mislabeling in the dataset, after which the false negative rate could be reduced to a minimum. Other than the study presented on histopathological analysis, several authors have studied the use of clinical data coupled with molecular markers to predict metastatic outcomes. Li et al. (Hamamoto et al., 2022), for example, proposed a prediction model on the possibility of distant metastasis in patients suffering from breast cancer by integrating clinical features with gene expression profiles. Here, the proposed approach could effectively make proper prognostic predictions by integrating machine learning algorithms with gene expression data to identify high-risk patients in the early stages. Another paper (Gupta, 2022) compares several predictive models to predict the time for tumour recurrence in breast cancer patients and provides a prediction for the time of tumour recurrence with accuracy up to one year. The research analysed 198 patients and came to notice that (40.0%) of the patients were predicted to develop tumour recurrence within the first year of diagnosis. The study used several classification models like spectral clustering, DBSCAN, and k-means, along with prediction models like support vector machines, decision trees, and random forests. These were proof that these machine learning techniques were effective, with SVMs at their pinnacle, accurately predicting the time of recurrence

of tumours or recovery of a patient at (78.7%). Authors in (Yang et al., 2022a) have proposed a new multimodal deep learning method for estimating the risk of relapse and metastasis in HER2-positive breast cancer patients by integrating whole-slide H&E images with clinical data. Images of 512 × 512 pixels were analysed for feature extraction using a deep CNN and then combined with clinical data. This model yielded an AUC of (76,0%) in two-fold cross-validation and (72,0%) on independent testing with TCGA data, hence showing its potential in prognosis prediction despite demographic and experimental differences.

Most of the previous works suffer from remarkable limitations and deficiencies in their approaches. For instance, one of the works analysed data from 198 cases of breast cancer patient (Gupta, 2022) and applied several machine learning methods such as spectral clustering, DBSCAN, k-means, support vector machines, decision trees, and random forests. The highest accuracy obtained with support vector machines was (78.7%) for predicting tumour recurrence within one year. This work suffers from a relatively small number of samples and moderate accuracy of predictions, which means the necessity of more reliable and scalable methods.

This work integrates LLMs with GNNs and provides a robust framework for cancer prediction tasks using these embeddings, which encode words and phrases from histopathological reports semantically. That is, words with similar meanings or context will be mapped to similar embeddings even if they occur in different positions of the report, and GNNs are able to uncover latent patterns and dependencies that are important to predict disease states with high accuracy. Application of GNNs on top of LLMs fulfills a complementary approach toward detecting metastatic breast cancer, by leveraging textual and structural knowledge towards better prognostic precision. In any case, although the efficiency of the proposed system, further validation and optimisation should be performed in order to determine the performance and reliability of the developed system. As depicted in Fig. 1.5, the major contribution of this work can be summarised.

1. This submission integrates the latest techniques of NLP coupled with the technology of GNNs for proposing a state-of-the-art comprehensive analysis in histopathology reports.

2. Use the attention scores from LLMs to construct interpretable, context-aware graph representations.

3. Hence, the addition of feature extraction methods to determine vital variables for metastases prediction in breast cancer enhances its predictive accuracy and interoperability.

4. A robust methodology, which it's hoped will give the clinicians much more insight into the underlying data, informing their decisions and helping improve patient outcomes.

5.1 Materials and Methodology

5.1.1 Dataset

Data was acquired from the IT department of Basavatarakam Indo-American Cancer Hospital and Research Institute (BIACH & RI), Hyderabad, Telangana, India in an Excel format. It is a semi-structured Excel file comprising two main columns: patient identifiers (medical record number—MR No) and histopathology reports (Hist_report). The dataset consists of 25,652 entries in .csv format (raw text). The Hist_report column includes a variety of pathological observations, such as clinical details, specimen descriptions, microscopic observations, impressions, and gross findings (Botlagunta et al., 2023). The histopathology diagnoses pertain to metastasis, biopsy, and lymph node reports. This research focuses on detecting metastasis in patients among the different types of diagnoses.

Analysis of histopathology reports with BERT typically involves tokenising these into individual words or subword units using the BERT tokeniser. First, the BERT tokeniser breaks up this text into a set of tokens, mapping each token into an index in the BERT vocabulary. The tokenised sequences have to pass through the BERT model to get embeddings for each token, which represents all contextual features of the report. In many downstream applications, the embedding is mostly used from the last hidden state layer of BERT because this layer captures more contextual information than its lower layers. Attention scores can also be extracted and will help analyse the relation between tokens. Mathematically, calculate the mean embeddings of all tokens in the report.

$$\text{Mean Embedding} = \frac{1}{N}\sum_{i=1}^{N} E_i, \tag{5.1}$$

where N is the total number of tokens in the report and E_i represents the embedding for token i (Kenton et al., 2019; Vaswani, 2017).

The Lambda Callback package was used to monitor the loss values that occur during training in determining the optimum learning rate. From the resultant graph, the optimum learning rate was where the loss was decreasing fastest before divergence or oscillation-that is, between 1.2×10^{-2} and 1.3×10^{-2}. The literature suggests that, on dropout rate, the starting rate should be between 0.2 and 0.5. Empirical tests within this range have yielded results to show that the dropout rate of

Table 5.1: BG-MBC Model Hyperparameters

Model Hyperparameters	**Value**
hidden units	[34,34]
learning rate	0.01
dropout rate	0.2
num epochs	300
batch size	128
k-fold	3

0.2 offers the best balance between model complexity and performance. Early stopping was also done to prevent overfitting. This works by stopping the training of the models at a point where the model performance on the validation sets stops improving. In so doing, it avoids overtraining on the training data, thereby retaining its ability to generalise on new data. Last but not least, it had two unique parameters, node threshold and node connections, both of which handled the relationship among nodes. These two parameters were subjected to several trial-and-error combinations to come up with an optimum setup. Performances in the BG-MBC were finally evaluated by some criteria: ROC graphs between true positive rate and false positive rate, AUC, F1 scores, accuracy, and cross-validation metrics. Generally speaking, such a multi-model approach for the selection of parameters by cross-validation and empirical testing has ascertained a high degree of accuracy and generalisability in predicting metastatic breast cancer by the BG-MBC model.

5.1.2 Building and Refining Attentions per Node in the Dataset

Padding and truncation parameters represent variable-length histopathology reports as fixed-size feature vectors during the preprocessing stage. Padding is the process of matching a shorter sequence to the length of the longest sequence by adding special tokens, whereas truncation involves cutting longer sequences to a pre-specified length. This normalises all sequences to a fixed length (Tao et al., 2023). Then, each report goes through a feature extraction wherein the embeddings from the last hidden state layer of BERT are utilised, thereby capturing rich contextual information (Kenton et al., 2019; Vaswani, 2017). All the tokens in each report are combined or aggregated into a resultant feature vector of 767 dimensions. Mathematically, the feature vector F_i can be represented for token i as:

$$E_i = [E_{i1}, E_{i2}, \ldots, E_{iN}], \tag{5.2}$$

where E_{ij} represents the (j-th) dimension of the embedding for token i, and N is the total number of dimensions in the embedding space (e.g., 768 for BERT-base).

Attention in BERT typically involves multiple heads to capture various aspects of the contextual relationship among tokens. Scores from each attention head are combined into a single score per token by averaging along the head. Repeat such for all tokens in a report to obtain a matrix of attention scores of shape number of tokens $\times 767$. Mathematically speaking, the attention score A_{ij} for token i and dimension j can be computed as:

$$A_{ij} = \frac{1}{H} \sum_{h=1}^{H} A_{ij}^{(h)}, \tag{5.3}$$

where H is the total number of attention heads, and $A_{ij}^{(h)}$ is the attention score for token i and the dimension j from attention head h.

5.1.3 Graph Building

The first step in graph construction is to compute the edges linking nodes according to the criteria predefined by the physician based on which the establishment of meaningful relationships among medical records in the network should stand. It includes the attributes with their attentions the model has the capacity to selectively focus on crucial parts of the input text while predicting the results, hence achieving a good relationship among them. It contains the attention mechanism; through the mechanism, each word or token within the input sequence receives different weights such that it allows the model to concentrate on the most critical information and disregard irrelevant parts, hence attesting to why BERT did exceptionally well in tasks concerning natural language processing (Clark, 2019) which lie in every node and explore the depth of connections defining adjacencies between nodes within a graph. Computing for all the inputs within each observation, Algorithm 1 is performed, thereby resulting in assigning ascertained probabilities to each node in its classification as either 0 no metastasis or 1 metastasis.

Algorithm 4 Report outputs calculations

Input: reports ▷ List of all Histopathology reports
Output: attentions, embeddings

1: **for** each **report do**
2: Calculate tokens, attentions and embeddings for each report
3: **end for**
4: **return** attentions, embeddings

The probabilities associated have been obtained by classifying each node as 0-

no metastasis or 1-metastasis by computing the attention scores for all features in each node using Algorithm 4. Moreover, it estimates the linkage between neighboring nodes according to their respective attention scores using Algorithm 5. For example, it prescribes that a node should be connected to a cluster whose average attention score is less than or equal to some predefined threshold. The threshold was obtained experimentally by systematically changing its value and assessing model performance for different thresholds, which means choosing a few threshold values and then training and testing the model with each to see how it would affect its performance metrics, such as accuracy, precision, recall, or F1 score. Finally, by judging against all these criteria, the adjacency metric building comes into place in Algorithm 6.

Algorithm 5 Attentions per Node Calculations

Input: attentions ▷ attentions
Output: attention scores per node

1: **for** each **attention matrix do**
2: Stacked all the layers per node
3: Find the average attention per node
4: **end for**
5: **return** attention scores per node

Algorithm 6 Adjacency Matrix calculation

Input: attention ▷ Node attention
Input: cls ▷ Node class
Input: node threshold = 200
Input: node connections = 5
Output: Adjacency Matrix

1: **for** each **row do**
2: add attention to the zeroes dictionary
3: add attention to the ones dictionary
4: **end for**
5: Sorting zeroes dictionary
6: dic ⟸ zeroes dictionary
7: Sorting ones dictionary
8: dic ⟸ ones dictionary
9: **for** each *adjacentvalue* in dic **do**
10: Exit where number of nodes = node connections parameter.
11: add the node to the Adjacency matrix
12: increment node counter
13: adjust the mean of the cluster
14: **end for**
15: **return** Adjacency Matrix

Based on these evaluations, an optimum threshold can be selected based either

on the general performance or on any desired trade-off between the various metrics. This threshold is applied in connecting entities in case of their similar entity, helping thereby in grouping entities based on their attention level. Now, this threshold, being an integer parameter, forms part of the node grouping process. To connect a new node with the already available group, one needs to compare to find the difference between the current group's average attention score with the new node's attention score against its threshold. If the difference comes out to be less than or equal to the threshold set, coupled with the cumulative count of nodes within the particular group not crossing the permissible limit of the group size, the new node is connected to the ongoing group.

5.1.4 Feature Extraction

Results from BERT have impressively proved to be a game-changer concerning the analysis of textual data in various domains, even to histopathological reports concerning diagnosis of diseases. It works by tokenising the raw text of the histopathology reports into an understandable format with which machine learning algorithms work. BERT is especially unmatched in state-of-the-art context with regard to the capture of contextual information and semantic meaning of text (Kenton et al., 2019). After tokenising the histopathology reports using BERT, an embedding is generated for every token that encodes its semantic context. These embeddings act as input features to downstream tasks such as classification. In nature, histopathology reports are such that when embedded, they transform the text data into a high-dimensional space, capturing complex relationships and patterns underlying the data. These embeddings are very useful in predicting BCM. However, there is the creation of a large number of embeddings-in this case, 767 columns were created-and hence, one should seek to look into the feature selection methods to identify the most informative features that will ensure that accurate predictions can be made. That is where feature selection techniques play a crucial role. Feature selection techniques include univariate selection, extra trees classifier, and Shapely values that have been considered the most impactful on the model. Though several other methods can also be applied, such as recursive feature elimination 25, these algorithms go through the BERT-generated embeddings and select the most relevant features toward the model, which introduces a high value to the predictive performance of the model. This process aims to reduce the dimensionality of the feature space while keeping the most discriminate information, and the selected features are further employed for building the BG-MBC model. Each patient number here acting as an entity in the graph will be considered as a node in the graph, having selected features act as node attributes. Moreover, attentions obtained from BERT will be utilised

in building the adjacency matrix, capturing relationships among different medical records based on semantic similarity and contextual dependency, explained in the above-mentioned algorithms. This idea utilises the power of BERT embeddings and attention mechanisms in extracting meaningful information from histopathology reports in developing a predictive model for cancer metastasis. Through the GNN, it allows capturing complex relationships and dependencies that are there in the data, which leads to better results in prediction, as in (Kenton et al., 2019).

5.1.5 BG-MBC Model and Interpretability

The BG-MBC model architecture is a neural network designed to work on graph-structured data. It applies techniques from GNN to process and analyse dependencies and relations that exist in the data.

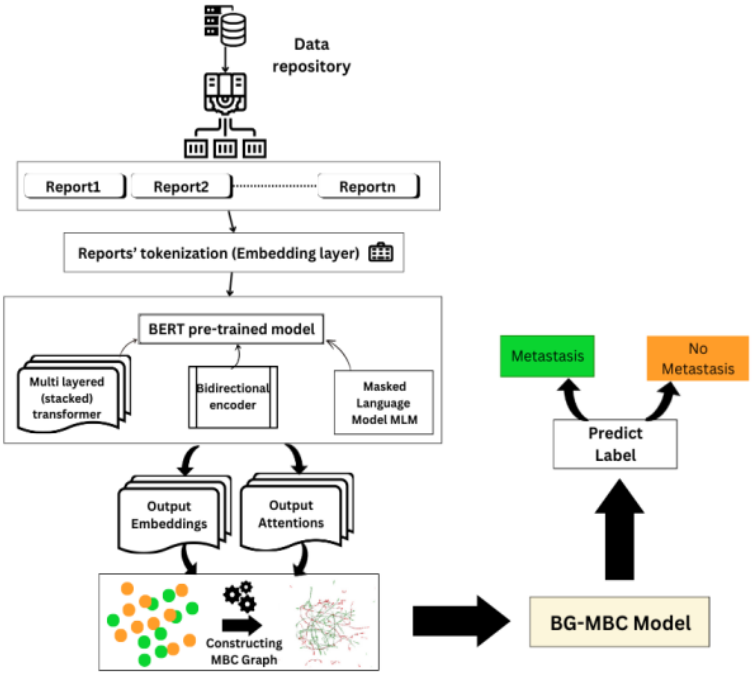


Figure 5.1: The above diagram displays how a BERT model can be leveraged in extracting embeddings and attentions out of histopathology reports. These embeddings and attentions are utilised to provide a graphical representation of data. This is fed to a BG-MBC model, whose framework incorporates GNN for the prediction of BCM based on the structured data arising out of the BERT model outputs

One of the most important advantages of the BG-MBC model architecture is its interpretability. The BG-MBC model enhances the interpretability of its predictions through the usage of Shapley values that allow the model to underline which features are important for the prediction of metastasis. This helps clinicians and researchers alike in understanding what the most relevant drivers for such predictions are, offering deep insight into the metastatic BC underlying drivers. Therefore,

analysis of the learned parameters and how information propagates throughout the graph enables one to find the most contributive features and relationships regarding the model's predictions. This interpretable level is very critical in metastasis prediction; it secures clinical acceptance and trust due to the evident reason behind each decision.

5.1.6 Training Procedure

First, the BG-MBC model is trained in a supervised way to learn how to predict the correct labels of the nodes in a graph given some input features and graph structure. Training usually is performed by optimising some loss mean square error (MSE), sparse categorical cross-entropy loss through an Adam optimiser.

5.1.7 BG-MBC and LLM (BERT) Models Integration

Here, it concerns the integration of the LLM, BERT, and BG-MBC into one framework for both histopathology report analysis and metastatic breast cancer prediction as shown in Fig. 5.1. The last hidden layer of neural network models, especially in architectures like BERT, is of much importance in most aspects related to natural language processing. In the analysis of histopathology reports for the prediction of metastatic breast cancer, understanding the contribution of the last hidden layer in deriving meaningful insights becomes very important (Kenton et al., 2019; Tenney, 2019). More importantly, the last hidden layer also captures substantial semantic information regarding input text. It is capable of encoding contextualised representations of words and phrases so that important nuances and intricacies that are critical in medical reports are conveyed. Such semantic representation forms the very foundation for the derivation of relevant fkenton2019berteatures and the spotting of patterns linked with metastatic breast cancer (Lu et al., 2019). On the other hand, one may want to analyse the activations and embeddings in the last hidden layer to allow for the extraction of important features that have contributed to the classification of histopathology reports (Puccetti et al., 2021). Particular keywords, phrases, or context features of cancer metastasis can be included in such features. In this process, the model is interested in only the most appropriate aspects of the input data (Tang et al., 2019). This can be done in the following way:

$$F = \text{ExtractFeatures}(E), \tag{5.4}$$

where E denotes the embeddings generated by the BERT model for a histopathology report.

Attention mechanisms are also built into BERT and other models, letting it

dynamically compute the relative importance of words and context tokens in the input sequence. The attention scores obtained at the last hidden layer give an idea about the salient aspects in the input text that guide the model's predictions. In this context, the attention pattern provides insight into the relevant elements in the histopathology report that it uses to make a prediction regarding metastatic breast cancer (Vaswani, 2017). The average score for the attention scores for layer heads is given by:

$$\bar{A} = \frac{1}{H} \sum_{i=1}^{H} A_i, \tag{5.5}$$

where A denotes the attention scores obtained from the BERT model and H the number of layer heads. For the computation of attention means for each node considered, the following formula can be followed:

$$\text{Attention}_{\text{node}} = \frac{1}{H} \sum_{i=1}^{N} \text{Attention}_{\text{layer}_i}, \tag{5.6}$$

where $Attention_{node}$ means attention mean score of the particular node and H is the number of layers, $Attention_{laye}$ denotes the attention score for the $i-th$ layer. Besides, in most situations, there is either fine-tuning or transfer learning of the pre-trained models. For instance, BERT: during such processes, the last hidden layer would be the most useful since it carries most of the subtle representations that were learned on huge pretraining from large texts of histopathology reports. Fine tuning will let the model re-adapt its parameters for the present task to predict cancer metastasis knowledge since it is encoded there (Howard et al., 2018; Clark, 2020). We use the BERT model to generate the embeddings from the histopathology reports and draw out relevant features that give insight into the contents of the report. This shall include identifications of important variables or dimensions from within the embeddings that indicate specific metastatic breast cancer. Furthermore, the extraction of important attention is carried out by focusing on the last hidden state layer from the BERT model and extracting the important attention scores from each report. Taking the mean across all the layer heads for each embedding variable produces an overall measure of attention for every report. This shall be used in order to quantify the relevance of different aspects from the histopathology reports.

5.2 Experiments

Our proposed approach uses the BERT model to extract attentions and embeddings from histopathology reports, as shown in Fig. 5.2. These attentions define

the edges in graphs, which represent the relations between various entities acted by medical numbers. Meanwhile, the embeddings extracted in reports are treated as node features in this graph. These attentions incorporate with the embeddings to construct a graph representation of histopathology. After that, the graph representation was fed into the proposed model named BERT Graph Metastatic Breast Cancer (BG-MBC). The BG-MBC model takes graph-structured data as input and can incorporate complicated relationships expressed by the graph edges and node features powerfully. It is expected to improve metastasis prediction performance for breast cancer by mining the rich contexts provided in histopathology reports using the attentions and embeddings that BERT outputs as the input to BG-MBC.

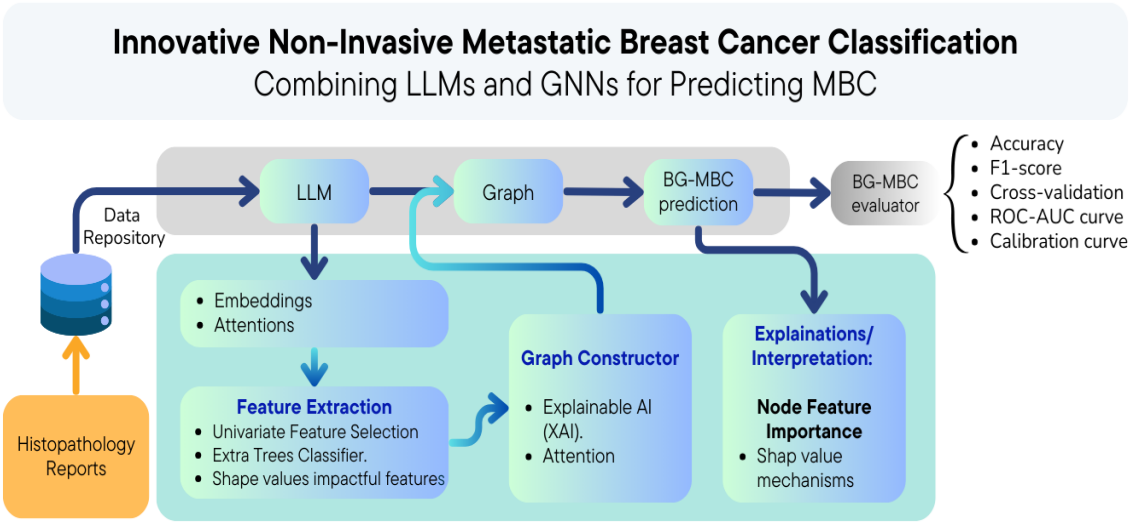


Figure 5.2: It uses natural language processing in order to pick out relevant features within the histopathology reports. BG-MBC forms a graph of these embeddings, with a graph neural network applied on top. Shapley values were used to interpret and explain its predictions.

5.2.1 Experimental Setup

Data were gathered from BIACH and RI in a semi-structured Excel file containing two key columns, namely, patient identifiers—medical records and histopathology reports. The dataset consists of 25,652 entries in CSV format, raw text. The Reports contain all kinds of pathological observations, such as clinical details, specimen descriptions, microscopic observations, impressions, and gross findings (Botlagunta et al., 2023). The histopathology diagnoses are related to metastasis, biopsy, and lymph node reports. Among the various kinds of diagnoses, this addresses the prediction of metastasis in patients.

Various datasets containing medical records, histopathology reports, and diagnostic results were combined into one file for analysis. The original data was in

several unstructured CSV files and was preprocessed in Google Colab, structured in tabular form, and combined into one dataset. Cleaning the data involved feature selection and normalisation to ensure that the model performed well. Feature selection identified key variables that best predicted metastasis, reducing dimensions, while normalisation scaled features to a standard range to avoid features with large magnitudes dominating the model.

Feature selection was done using three algorithms: univariate selection, extra trees classifier, and SHAP values. Univariate selection identified the most statistically significant features for this dataset. An extra trees classifier gave the importance of features, ranking them in order of their influence on the predictions of the model. This ensured a sound dataset for predictive modeling. Below, this summary outlines that:

- Integration of various datasets into a single structured format: Data preprocessing involved cleaning, feature selection using several methods—univariate selection, extra trees, SHAP values—and normalisation.

- Objective: An attempt has been made to enhance the performances of models with a reduction in biases during the analysis.

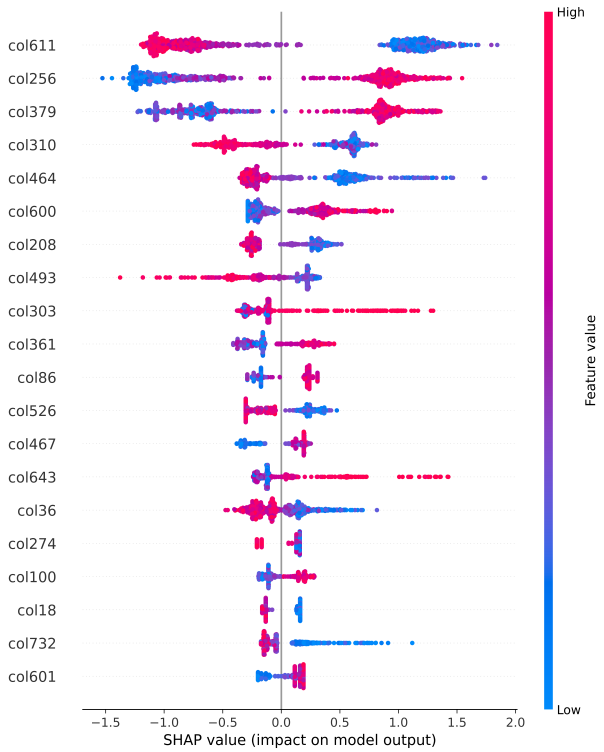


Figure 5.3: Shapley additive explanations (SHAP): the SHAP values highlight the features with the greatest impact on the model

Shapley additive explanations (SHAP) values were employed to enhance the feature selection process by providing a detailed understanding of feature importance in machine learning models. SHAP values reveal the contribution of each feature to individual predictions, offering a more nuanced perspective than traditional methods. This approach improves interpretability and trust in model predictions, aiding decision-making (Lundberg, 2017; Munn et al., 2022). Fig. 5.3 illustrates the influence of each feature on predictions, underscoring their role in result interpretation.

5.2.1.1 Bidirectional Encoder Representations from Transformers

BERT is a transformer-based model proposed by Google AI for creating deep, bidirectional representations from input text. Its architecture, basically an encoder, processes the input sequence both in the forward direction and in the reverse direction to capture effective contextual information (Kenton et al., 2019). BERT has two key outputs: embeddings and attention scores. Regarding embeddings, BERT will compute the contextualised word embedding for each token in the input sentence. These are representations of the semantic meaning of the token in context-informative about its role or function within the meaning of the whole input sentence. BERT also produces an attention score, representing the relationship of each token to every other token in the sequence input, represented as a matrix A. The entry A_{ij} indicates means the attention weight of token i during the processing of token j

5.3 Results and Discussion

Table 5.2 provides the performance comparison of the proposed approach and the existing methods for metastatic breast cancer prediction using histopathology reports. Specifically, one may observe that the proposed approach, namely BG-MBC, outperforms the existing methods. More specifically, BG-MBC achieves an accuracy score of 0.98. It outperforms the existing methods because of the incorporation of state-of-the-art graphical techniques, namely GNN and large language models, while constructing the graph. Moreover, it has a good record while capturing the relationship and dependencies between data. Thus, its higher performance indicates that it can model complex patterns and hence predict more compared to the existing methods. In contrast, the baseline model achieves an accuracy score of 0.94. While it still performs decently, it falls short when compared with BG-MBC, indicating further that there is still room for improvement in modeling the underlying data structure and relationship. CNNI_BCC (Ting et al., 2019) achieves an accuracy score of 0.94, which is a bit lower than BG-MBC but still competitive. Thwin et al. give in (Thwin et al., 2024) a model that achieves an accuracy score of 0.90,

indicating thereby that most likely it is not as good at capturing the underlying pattern as either BG-MBC or CNNI_BCC does. The deep convolutional neural network model proposed in (Mullooly et al., 2019) reaches an accuracy of 0.93, which is just a little bit lower than BG-MBC, but still is relatively good.

Table 5.2: Detailed evaluation metrics breakdown of BG-MBC and existing methods. Best results are highlighted in bold, and second-best results are underlined. All scores given as '-' are not provided in their sources. CNNI_BCC (Ting et al., 2019), DL-CNN (Thwin et al., 2024), DCNN (Mullooly et al., 2019).

Models	**AUC Score**	**F1 Score**	**Balanced Accuracy**	**Cross-Validation**
CNNI_BCC	0.90	-	0.90	-
DL-CNN	0.91	-	0.90	0.89
DCNN	<u>0.96</u>	-	0.93	-
Baseline model	0.90	0.90	<u>0.94</u>	<u>0.92</u>
BG-MBC	**0.98**	**0.98**	**0.98**	**0.99**

Fig. 5.4 shows the calibration curve of BG-MBC. The calibration curve plots the fraction of positive examples predicted by a model against the true fraction of positive examples, for a range of thresholds on the prediction probability. It gives information regarding the reliability of the predicted probabilities of a classification model, which in essence means that if a model is well-calibrated, the estimated probabilities need to be close to the real probability of a case really belonging to a class. That makes the calibration curve helpful for tasks such as medical diagnostics or risk analysis (Hastie et al., 2005). Thus, the plot depicts that the probability estimate from the model is reliable and hence can be confidently used in making decisions for metastatic breast cancer diagnosis. This good alignment can be said to show that the predicted probabilities by the model are well-calibrated. If these probabilities are going to be used to decide on thresholds or thresholds to make decisions, then accurate and representative model predictions of the true underlying probabilities would be expected.

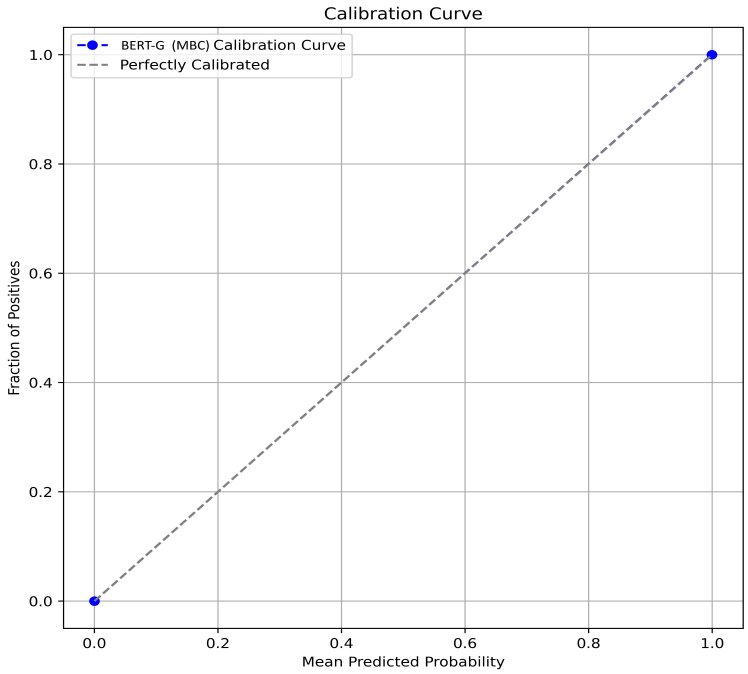


Figure 5.4: The above diagram depicts how the BERT model can be employed for the extraction of embeddings and attentions from histopathology reports. The obtained embeddings and attentions are utilised in the visualisation of the data in graphical form. Obtained graphical data is used as input to the BG-MBC model. The overall framework involves utilising GNN for the prediction of BCM from the structured data obtained due to the BERT model outputs.

5.4 Conclusions

A diagnosis of MBC from histopathology reports is advantageous on many grounds: the reports are very detailed and follow a certain standard, so one gets good value for the diagnosis; secondly, because of advanced machine-learning algorithms like GNN and BERT, one could get high-accuracy performances for this prediction task. In the current paper, we propose a new hybrid mode using GNN and BERT for MBC prediction-namely, BG-MBC. Our method not only furthers our understanding of the diseases but also contributes to developing more powerful strategies for diagnosis and therapy within the medical area. To that end, this work proposes a novel hybrid model named BG-MBC to analyse histopathology reports based on integrated GNNs and BERT for the task of MBC prediction. The BG-MBC model leverages both GNN and BERT with their strengths to further understand metastatic breast cancer and, ultimately, develop diagnostic and treatment strategies. Below are the summary statements of the novelty and practical impact of the work.

Combination of GNN-BERT Until now, no such combination of GNN and BERT methods has been done for analysing medical data. Although BERT captures semantic relationships among text data with unparalleled performance, GNN meth-

ods are designed to capture the structural relationships within the graph-structured data. This way, BG-MBC can leverage both the linguistic context from the histopathology reports and the intricate connections among different medical features.

By capturing the semantic meaning of the text using BERT and modeling the relationships and dependencies of data points using GNN, BG-MBC will be able to reason holistically. The final result will yield higher prediction accuracy due to this dual approach and will thus act as a strong tool for diagnosing MBC.

Comprehensive data representation The model uses BERT to form embeddings that can capture the semantic richness in the report text; these embeddings, being the node features of the graph constructed based on the attention scores, hence reflect the important connections within this text. This full representation allows the model to gain an enhanced view of the complicated interaction of various medical features.

Better understanding of the disease BG-MBC may provide much-needed insight into the latent patterns and factors underlining metastatic development of breast cancer due to the precise modeling of relationships and dependencies that they may exhibit. Such understanding guides the clinicians to make informed decisions on diagnosis and treatment.

In conclusion, the BG-MBC model is a milestone in medical diagnosis, prepared by harvesting complementary strengths brought by GNN and BERT. Novelty improves the accuracy of predictions but also sheds useful light on diseases that will contribute to better diagnosis and treatment. It is only with further research that more potential can be unleashed: considering other LLM architectures, improving graph construction, and testing the efficiency on larger and more diverse datasets.

Chapter 6

Conclusion and Future Work

This research has presented a comprehensive multi-phase model for cancer progression management that represents key transition points in the patient-care pathway from detection to monitoring a progression of cancer. In phase one, we applied a multimodality deep learning model to chronic obstructive pulmonary disease (COPD) as a high-risk factor for lung cancer, specifically for the purpose of early screening and detection as explained in Chapter 1. This stage allowed for the earliest identification of an abnormality or lesion identified in the thorax, that triggers further investigational diagnosis and referral.

In phase two, we introduced a predictive model for breast cancer relapse in an incomplete, progressing state of the disease from a clinical perspective, taking into consideration that staging and prognosis can be useful indicators for directing clinical decisions around cancer treatment. The second phase was conducted not only to assess the risk of recurrence but aim to assist in the delineation of individualised treatment plans and organised follow-up, resulting in delayed at best or improved at least survival and a better quality of life for patients.

The third phase included a discussion of predicting cancer metastases, where determining staging and diagnosis can still aid in determining the best course of therapy. The model development for phase three, provides a model for determining the actual treatment response, validating continued adaptation of the treatment plan, while also monitoring the patient on their potential path.

Each of these phases represents connected, integrated context for cancer management that is between early diagnosis for early detection, relapse prediction, and monitoring for metastasis, illustrating our use of advanced deep learning methods at different stages of development of disease. This article represents initial work toward developing a robust, AI integrated clinical informed decision support system to guide early detection through the cancer journey, while considering enhanced diagnostic accuracy, personalised interventions aimed at improving patient outcomes for lung cancer and breast cancer. To operationalise these three

phases, three dedicated models were developed for integration into the cancer management framework pipeline.

AIM-X in Chapter 3 is a diagnostic framework that merges X-ray images with their textual reports to offer comprehensive diagnostic reports. This architecture leverages deep CNNs, such as DensNet, for image feature extraction, while the GPT transformer-based model process the texts to underline the strengths of the multiple data modalities combined. Attention mechanisms zoom in on the most critical image areas, while genetic algorithms further optimise the involved hyperparameters, raising both the performance and relevance of the model. It incorporates an explainability module for increased transparency through visualisation of attention weights, an issue has been raised with this feature explained earlier in Section 2.7.

GraphX-Net see Chapter 4 represents the state of the art in the applications of GNN, gives an improvement in the prediction accuracy because GNNs can handle structured information and represent complex dependencies among variables, enabling the model to be a great tool for risk assessment, early intervention, and personalised treatment strategies.

Another significant contribution is the BG-MBC model Chapter 5, which combines GNN and LLM to further improve performance in predicting MBC. This hybrid approach presents novelty in methodological contributions to medical data analysis via the combination of structural relationships captured by GNNs with the semantic context of LLM. Predictive performance is therefore improved by embedding the linguistic subtlety of histopathology reports together with the complex relations among medical features. The completeness of the data representation the model could offer, as it embeds semantic richness with structural relationships, provides useful insights into the latent pattern underlying the development of MBC. These shall go a long way in aiding clinicians to make clinically informed decisions.

6.1 Overview

This research aims to develop deep learning models for early detection, diagnosis, and clinical decision making at various stages of cancer progression, with an emphasis on lung and breast cancer, across multisource data types, including X-ray scanned images, pathology reports, and clinical attributes. To achieve the aim, four general objectives were discussed in Chapter 1. More detail on the fulfilment of each objective is discussed below.

We have achieved all of the aims below all the objectives outlined in Section 1.4 by addressing key stages of cancer management through constructing a multi-stage

AI pipeline for lung and breast cancer, whereby each stage has been noted as a specific research aim.

- The first objective: To improve early-stage screening and reporting of thoracic abnormalities and possible lung cancer with a multimodality framework AIM-X Chapter 3, Section 3.1.2 was realised through the development of a deep learning model that uses X-ray images and radiology reports. The model produced by this project exhibited good performance when it comes to the detection of abnormalities and lesions in the thoracic region. Thus, the project provides a powerful tool for early diagnosis and for reducing the time frame for clinical decision-making.

- The second objective: To create and evaluate a predictive model for relapse in breast cancer patients using structured and unstructured patient data was achieved in the second phase. GraphX-Net Chapter 4, Section 4.1.2.6, a model that produces a good performance for predicting the risk of relapse, and thus will enable further investigation to help accurately stage cancers in terms of size, grade and evidence of metastasis, and could help support personalised treatment planning. The model provides a mechanism to not only identify patients with a risk of recurrence promptly but also improve the post-treatment plan and follow-up accordingly.

- The third objective: To develop a framework for predicting metastasis in breast cancer patients. BG-MBC Chapter 5, Section 5.1.5 is a model addressed in the final phase. The model developed here thus effectively identifies patients at high risk of metastasis, enabling clinicians to monitor therapy responses carefully and make informed decisions about follow-up treatments.

- Ultimately, the fourth aim to examine the clinical impact of the proposed pipeline, generalisability of the models, and cost-effectiveness was approached by accepting that clinical validation was not possible because the models had not been deployed in a real-world setting. However, the models were trained and evaluated on datasets that were gathered through multiple cohorts. This meant it was possible to robustly explore the generalisability of models to multiple patient populations. The major contribution here will be to lessen unnecessary investigations and so use healthcare resources more effectively, and the early detection phase as part of the pipeline will be vital in creating these efficiencies.

The proposed cancer management pipeline integrates early detection, prediction of relapse and prediction of metastases in one AI-based clinical decision support

tool. This work represents a meaningful contribution towards an improved patient journey and patient outcomes, a more efficient cancer care pathway, and the better configuration and use of resources in healthcare systems by leveraging and harnessing next-generation deep learning modalities in consideration of clear clinical outcomes.

The achievement of this research has marked a significant advancement in machine learning, particularly in predictive and diagnostic modelling. GNNs have advanced not only the state of artificial intelligence but also provided scalable, robust, interpretable, and applicable tools that can be applied to real-world challenges in many disciplines. To conclude all, here we have some tables to brief the outcomes of the research.

Table 6.1: Include various models or techniques performance metrics used in the research.

Model	**Task**	**BLEU1**	**ROUGE**	**Accuracy**	**Explainability**
AIM-X (Basaad et al., 2025)	Detection and Generation Reports	0.73%	0.73%	94.5%	-
GraphX-Net (Basaad et al., 2024b)	RBC Prediction	-	-	0.98%	High
BG-MBC (Basaad et al., 2024a)	MBC Prediction	-	-	0.98%	High

Table 6.2: Include in the table, which summarises information on the dataset, sample distribution, feature, and class distribution used in the research.

Dataset Attribute	**Details**
X-Ray images and reports	7,472
BCP Total Patients	1,500
Metastatic Breast Cancer Cases	2,500
Features Used	Histopathology Reports, Gene Expression, X-Ray Images, Histopathological Features
Types of Data	Numeric, Textual, Images
Temporal Data Included	Yes

Table 6.3: Show: What are the practical implications of this research for clinical medicine?

Question / Use Case	**Impact**
How could these findings be translated into real-world medical diagnostics?	Utilizing AIM-X, GraphX-Net, and BG-MBC for early diagnosis and predictions may improve multi-stage cancer management reasons. The evaluation of performance should include a broader base of patient cohorts that can use the GraphX-Net relapse predictions alongside medical notes while explaining critical predictors (tumour grade, lymph nodes status, etc.) through Shapley values and LIME. In addition, using the predictive feature of BG-MBC's metastatic breast cancer, baseline clinical records, alongside the histopathology reports should facilitate and support more informed decisions for medical technicians.
What potential impact might this research have on clinical decision-making or patient outcomes?	• Personalised Treatment Planning. • Patient Care. • Better targeting of therapies.
Multimodal Analysis	Holistic patient understanding
What barriers and opportunities exist for adopting this methodology in healthcare settings?	While regulatory, financial, and cultural barriers exist, there are many obstacles related to the application and model validation will be as follows: • Collecting and refining various data to promote the scalability and model robustness. • Availability of high machine resources to train and validate the model. • Testing the model. • validating outcomes based on subjective opinion from the rigor of experts. • Ethical concent. • Automation (e.g., AI for administrative tasks, diagnostics) can reduce clinician burnout and streamline workflows.

Continued on next page

Table 6.3 – *Continued: Practical implications*

Question / Use Case	**Impact**

6.2 Future Work

This research presents some contributions from new perspectives synthesis of GNNs and CNNs, combined with LLMs, with different techniques and approaches in deep learning, which we hope they help in diagnosing, detecting thoracic diseases, and automating the prediction of relapse and metastatic breast cancer in a management cancer multi-stages pipeline. Despite all these advances, challenges remain. First, scalability is the big challenge; it is computationally very expensive to train large-scale graphs. Explainability is another challenge, particularly when models are being deployed in sensitive domains like healthcare. Some open issues that must be addressed regarding wider acceptance or real-world effectiveness concern the same. The following, thus, might be some of the ways that this work can go forward:

6.2.1 Improve the Interpretability

As shown in Table 6.1, the GNN models were developed by exploiting relationships derived from the contributions of relevant medical attributes. This design choice enhanced the transparency of the models' decision-making process, enabling clinicians to better understand and derive insights for each individual case. For the multimodality model, certain limitations outlined in Section 2.7 remain and could be addressed in future work to further improve interpretability. A more professional and comprehensive explanation of predictions could be achieved by integrating intuitive interpretability techniques beyond those already explored, such as attention mechanisms, Shapley values, LIME, and the combination of multiple methods (e.g., visualising node importance or generating feature heatmaps). Such enhancements would strengthen clinicians' confidence in AI-generated insights and promote trustworthiness by making the decision-making process more transparent.

6.2.2 Incorporation of Longitudinal Data

Future research intends to merge longitudinal data would substantially strengthen our multi-phasee pipeline (AIM-X, GraphX-Net, BG-MBC) Chapters 3, 4, 5 respectively, as doing so would promote their predictive accuracy, clinical relevance and ultimately, disease management. Time sequencing would precisely determine

disease progression rather than relying on snapshots at a single point-in-time. Integration of serial imaging, follow-up reports, clinical data and treatment timelines can incorporate temporal fusion models and joint longitudinal–survival models, to generate patient-specific time-stamped risk patterns and trajectories. This would further refine early detection, reduce forecasting inaccuracies related to relapse and metastasis otherwise based on traditional singular imaging and promote treatment by tracking risk patterns over time with structured treatment regimens. It would support focusing on the patient, trustworthy approach, to better support patient-specific cancer management decisions.

6.2.3 Mitigate Bias and Improve Fairness

As described in sections 2.6.1.1, 2.6.2.1, and 2.6.3.1, this research used three different datasets to develop an algorithmic cancer management pipeline involving multiple phases. In the first phase, the Indiana University dataset of chest X-rays was used for the AIM-X model in Chapter 3. The IU dataset included multiple cohorts and multiple diseases, but it was very imbalanced in terms of disease representation. The diversity of classes was substantially uneven, and many of the diseases were extremely over-represented, creating an inherent dataset bias that caused the models behaviour to be negligibly affected.

In practice, when algorithms are trained on imbalanced datasets, the algorithm will tend to favour the majority classes. These issues have recognised effects with AIM-X in our operational environment: model sensitivity for minority classes and recall values decreased, resulting in missed detections with lower presented diseases.

Mitigating these issues requires targeted strategies. At the modelling stage. With respect to the modelling phase, we may use methods to balance the classes of the majority via sampling, (e.g. over-sampling the minority classes or under-sampling the majority classes), but again relying on the expertise of a domain expert is necessary due to the very sensitive nature of medical data, and likely cause differing shows of clinically-reasonable balancing (prior to re-blancing). Using other approaches such as class-weighted loss function, or, focal loss can also limit predominated major class. Data augmentation, continental to synthetic sample generation (e.g. GAN based augmentation), to increase the effective use of their minority classes, are other avenues one may pursue.

6.2.4 Further Experimental Studies

Future research could continue the experimentation, and further enhance the generalizability of the models, by also exposing them to larger and more heterogeneous patient samples from multiple organizations. A true cross-center evaluation would

provide an opportunity to assess overall domain shifts, as a consequence of different imaging equipment, document styles for clinical notes and the post-collection preparation for histopathology. More experiments could develop longitudinal records for the patients that could allow models like GraphX-Net to not only refine predictions of the relapse over time, but also better represent disease progression patterns. We could take this one step further by integrating the relapse predictions of BG-MBC with extracted image embeddings from histopathology and the structured clinical variables. Predictive models describing each of those components in isolation would enhance our multi-modal performance. Additionally, comparative evaluations of the newer explainability approaches like SHAP, LIME and counterfactual explanations could help provide a quantification of the extent of feature importance as it fluctuates across the patient cohorts, both within a singular disease state and across domains. In conclusion, if the ultimate objective is to provide high-quality, reliable, and trustworthy decision support to clinicians and/or technicians, additional studies should be considered in areas like explainability and generalizability in the era of decision support.

Table 6.4: Summarise potential areas for further exploration based on our outcomes.

Future Direction	**Goal**
Longitudinal Data Integration	Capture temporal patterns for progression.
Multi-Modal Data Fusion	Combine imaging, text, radiomics, proteomics and genomics data (Li et al., 2022).
Improved Explainability Tools	Build trust and ensure clinical usability.
Disease Generalisation	Expand applicability to other diseases, for instance, extend *AIM-X* model to apply different modalities like CT-scan, MRI, etc.
Enhancing Data Quality and Accessibility	• Implement data augmentation techniques for dealing with shortages in labeled data. • Foster collaborative efforts among institutions to develop larger, more diverse datasets. • Develop federated learning approaches to allow the training of models on distributed data without the need to share sensitive information.

Continued on next page

Table 6.4 – *Continued: Summarise potential areas*

Future Direction	**Goal**
Improving Model Robustness and Interpretability	• Use XAI methods to facilitate the transparency of model predictions. • Perform regular external validation of models to check for generalisability. • Foster the incorporation of domain knowledge, such as the insights of radiologists, into model design.
Fostering Fairness and Mitigating Bias	• Work with balanced datasets representing diverse populations. • Apply fairness-aware algorithms to mitigate prediction bias.
Facilitating Clinical Integration	• Involve clinicians in the development and evaluation of AI tools to ensure usability and relevance. • Perform clinical trials to provide evidence that AI models are safe and effective.
Leveraging Advanced Techniques	• Use multi-modal learning to leverage imaging, genomic, and clinical data. • Explore the use of GNNs to model relationships between patients, diseases, and treatments.

Our analysis revealed significant opportunities to develop AI in medical imaging, as shown in Table 6.4. For instance, longitudinal data integration will deal with the time aspect of modelling disease progression, and Multi-Modal data fusion will bring together imaging, radiomics and genomics to provide a holistic view of the health of patients. Enhancing explainability tools and model robustness, particularly XAI methods and external validity, can improve clinical adoption. Generalising disease (e.g., generalisation of AIM-X from lung CT to liver MRI) and increasing data quality (e.g., through augmentation) will help to support scalability. Ethical concerns should consider mitigating bias with fair algorithms and balanced cohorts. Finally, advanced techniques (e.g., GNNs as relays of patient-disease network) can also provide solutions to the multimodal learning aspect of disease aetiology. Overall, the next steps in all of these opportunities are to balance the advancement of

technology with the application of clinical knowledge and practice.

Bibliography

Acharya, Chaitanya R et al. (2008). 'Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer'. In: *Jama* 299.13, pp. 1574–1587.

agar (May 2024). *Positional Encoding in Transformers*. URL: https://www.geeksforgeeks.org/positional-encoding-in-transformers/?ref=header_outind.

Agarwal, Chirag et al. (2023). 'Evaluating explainability for graph neural networks'. In: *Scientific Data* 10.1, p. 144.

Ahmed, Wakeel et al. (2024). 'Molecular insights into anti-Alzheimer's drugs through predictive modeling using linear regression and QSPR analysis'. In: *Modern Physics Letters B* 38.27, p. 2450260.

Ahmedt-Aristizabal, David et al. (2021). 'Graph-based deep learning for medical diagnosis and analysis: past, present and future'. In: *Sensors* 21.14, p. 4758.

Aiello, Marco et al. (Nov. 2021). 'How does DICOM support big data management? Investigating its use in medical imaging community'. In: *Insights into Imaging* 12.1, p. 164.

Alaparthi, Shivaji et al. (2020). 'Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey'. In: *arXiv preprint arXiv:2007.01127*.

Alawi, Abdulfattah E Ba et al. (2021). 'Performance Analysis of Deep Dense Neural Networks on Traffic Signs Recognition'. In: *2021 International Conference of Technology, Science and Administration (ICTSA)*. IEEE, pp. 1–4.

Ali, Redha et al. (2019). 'Fused deep convolutional neural networks based on voting approach for efficient object classification'. In: *2019 IEEE National Aerospace and Electronics Conference (NAECON)*. IEEE, pp. 335–339.

Aloysius, Neena et al. (2017). 'A review on deep convolutional neural networks'. In: *2017 international conference on communication and signal processing (ICCSP)*. IEEE, pp. 0588–0592.

Alqahtani, Fawaz F et al. (2024). 'Cnx-b2: a novel cnn-transformer approach for chest x-ray medical report generation'. In: *Ieee Access* 12, pp. 26626–26635.

Alzubaidi, Laith et al. (2021). 'Review of deep learning: concepts, CNN architectures, challenges, applications, future directions'. In: *Journal of big Data* 8, pp. 1–74.

Amann, Julia et al. (2020). 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective'. In: *BMC medical informatics and decision making* 20, pp. 1–9.

Ansari, Zulfikar Ali et al. (2025). 'The role of explainable AI in enhancing breast cancer diagnosis using machine learning and deep learning models'. In: *Discover Artificial Intelligence* 5.1, p. 75.

Ashraf, Rehan et al. (2020). 'Deep convolution neural network for big data medical image classification'. In: *IEEE Access* 8, pp. 105659–105670.

Ates, Gorkem Can et al. (2023). 'Dual cross-attention for medical image segmentation'. In: *Engineering Applications of Artificial Intelligence* 126, p. 107139.

Azimi, Seyed Majid et al. (2018). 'Advanced steel microstructural classification by deep learning methods'. In: *Scientific reports* 8.1, p. 2128.

Aziz, Noreen M (2007). 'Cancer survivorship research: state of knowledge, challenges and opportunities'. In: *Acta oncologica* 46.4, pp. 417–432.

Ba, Jimmy Lei et al. (2016). 'Layer Normalization'. In: *arXiv:1607.06450*.

Bacciu, Davide et al. (2020). 'A gentle introduction to deep learning for graphs'. In: *Neural Networks* 129, pp. 203–221. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2020.06.006. URL: https://www.sciencedirect.com/science/article/pii/S0893608020302197.

Bahdanau, Dzmitry et al. (2015). 'Neural machine translation by jointly learning to align and translate'. In: *ICLR*. URL: https://arxiv.org/abs/1409.0473.

Basaad, Abdullah et al. (2024a). 'A BERT-GNN approach for metastatic breast cancer prediction using histopathology reports'. In: *Diagnostics* 14.13, p. 1365.

Basaad, Abdullah et al. (2024b). 'GraphX-Net: A Graph Neural Network-Based Shapley Values for Predicting Breast Cancer Occurrence'. In: *IEEE Access* 12, pp. 93993–94007.

Basaad, Abdullah et al. (2025). 'AIM-X: Attention-Infused Multimodal Cross-Interaction for X-ray Clinical Report Generation'. In: *Procedia Computer Science* 264, pp. 201–210.

Basha, S.H. Shabbeer et al. (2020a). 'Impact of fully connected layers on performance of convolutional neural networks for image classification'. In: *Neurocomputing* 378, pp. 112–119. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2019.10.008. URL: https://www.sciencedirect.com/science/article/pii/S0925231219313803.

— (2020b). 'Impact of fully connected layers on performance of convolutional neural networks for image classification'. In: *Neurocomputing* 378, pp. 112–119.

Basnet, Ramesh et al. (2021). 'A deep dense residual network with reduced parameters for volumetric brain tissue segmentation from MR images'. In: *Biomedical Signal Processing and Control* 70, p. 103063.

Bebis, George et al. (1994). 'Feed-forward neural networks'. In: *Ieee Potentials* 13.4, pp. 27–31.

Beltagy, Iz et al. (2020). 'Longformer: The long-document transformer'. In: *arXiv preprint arXiv:2004.05150*.

Ben Jabeur, Sami et al. (2023). 'Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering'. In: *Computational Economics* 61.2, pp. 715–741.

Benk, Michaela et al. (2020). 'Explaining interpretable machine learning: Theory, methods and applications'. In: *Methods and Applications (December 11, 2020)*.

Bergmann, D et al. (1996). 'Electronegativity and molecular properties'. In: *Angewandte Chemie International Edition in English* 35.2, pp. 150–163.

Bertucci, François et al. (2019). 'Genomic characterization of metastatic breast cancers'. In: *Nature* 569.7757, pp. 560–564.

Bhatti, Uzair Aslam et al. (2023). 'Deep learning with graph convolutional networks: An overview and latest applications in computational intelligence'. In: *International Journal of Intelligent Systems* 2023.1, p. 8342104.

Bianconi, Ginestra et al. (2009). 'Assessing the relevance of node features for network structure'. In: *Proceedings of the National Academy of Sciences* 106.28, pp. 11433–11438.

Bilot, Tristan et al. (2023). 'Graph neural networks for intrusion detection: A survey'. In: *IEEE Access* 11, pp. 49114–49139.

Botlagunta, Mahendran et al. (2023). 'Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms'. In: *Scientific Reports* 13.1, p. 485.

Brauwers, Gianni et al. (2021). 'A general survey on attention mechanisms in deep learning'. In: *IEEE Transactions on Knowledge and Data Engineering* 35.4, pp. 3279–3298.

Brown, Tom B (2020). 'Language models are few-shot learners'. In: *arXiv preprint arXiv:2005.14165*.

Brownlee, Jason (Jan. 2023). *The Attention Mechanism from Scratch*. URL: https://machinelearningmastery.com/the-attention-mechanism-from-scratch/.

Burger, Christopher et al. (2023). 'Are Your Explanations Reliable? Investigating the Stability of LIME in Explaining Text Classifiers by Marrying XAI and Adversarial Attack'. In: *arXiv preprint arXiv:2305.12351*.

Chami, Ines (2021). *Representation Learning and Algorithms in Hyperbolic Spaces*. Stanford University.

Chan, Chi-Min et al. (2023). 'Chateval: Towards better llm-based evaluators through multi-agent debate'. In: *arXiv preprint arXiv:2308.07201*.

Chang, Yupeng et al. (2024). 'A survey on evaluation of large language models'. In: *ACM Transactions on Intelligent Systems and Technology* 15.3, pp. 1–45.

Cheligeer, Ken et al. (2024). 'Validation of large language models for detecting pathologic complete response in breast cancer using population-based pathology reports'. In: *BMC Medical Informatics and Decision Making* 24.1, p. 283.

Chen, Chia-Kuei et al. (2020a). 'Patients with Spontaneous Pneumothorax Have a Higher Risk of Developing Lung Cancer: A STROBE-Compliant Article'. In: *Medicine* 99.30, e21411. DOI: 10.1097/MD.0000000000021411. URL: https://doi.org/10.1097/MD.0000000000021411.

Chen, Deli et al. (2020b). 'Measuring and relieving the over-smoothing problem for graph neural networks from the topological view'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 04, pp. 3438–3445.

Chen, Fenxiao et al. (2020c). 'Graph representation learning: a survey'. In: *APSIPA Transactions on Signal and Information Processing* 9, e15.

Chen, L et al. (2011). 'P5-13-24: A Predictive Model of Early Systemic Disease Relapse after Standard Adjuvant Therapy for Breast Cancer.' In: *Cancer Research* 71.24_Supplement, P5–13.

Chen, Shan et al. (2023). 'The utility of ChatGPT for cancer treatment information'. In: *MedRxiv*, pp. 2023–03.

Chirasani, Sateesh Kumar Reddy et al. (2022). 'A deep neural network for the classification of epileptic seizures using hierarchical attention mechanism'. In: *Soft Computing* 26.11, pp. 5389–5397.

Clark, K (2020). 'Electra: Pre-training text encoders as discriminators rather than generators'. In: *arXiv preprint arXiv:2003.10555*.

Clark, Kevin (2019). 'What Does Bert Look At? An Analysis of Bert's Attention'. In: *arXiv preprint arXiv:1906.04341*.

Colling, Richard et al. (2019). 'Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice'. In: *The Journal of pathology* 249.2, pp. 143–150.

Computer Science, McGill University School of (n.d.). *The graph neural network model*. Accessed: 13-10-2024.

Cong, Shuang et al. (2023). 'A review of convolutional neural network architectures and their optimizations'. In: *Artificial Intelligence Review* 56.3, pp. 1905–1969.

Curtis, Christina et al. (2012). 'The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups'. In: *Nature* 486.7403, pp. 346–352.

Dalmaz, Onat et al. (2022). 'ResViT: residual vision transformers for multimodal medical image synthesis'. In: *IEEE Transactions on Medical Imaging* 41.10, pp. 2598–2614.

Das, Arun et al. (2020). 'Opportunities and challenges in explainable artificial intelligence (xai): A survey'. In: *arXiv preprint arXiv:2006.11371*.

De Rose, Francesca et al. (Sept. 2022). 'Rethinking breast cancer follow-up based on individual risk and recurrence management'. In: *Cancer Treatment Reviews* 109. Epub 2022 Jul 1, p. 102434. ISSN: 0305-7372. DOI: 10.1016/j.ctrv.2022.102434.

Delaunay, Julien (Dec. 2023). 'Explainability for machine learning models : from data adaptability to user perception'. Theses. Université de Rennes. URL: https://theses.hal.science/tel-04496068.

Dell Technologies (2023). *New Frontiers: Dell EMC PowerEdge R750xa Server with NVIDIA A100 GPUs*. Technical overview of the PowerEdge R750xa server with NVIDIA A100 GPU integration. Dell Technologies. URL: https://infohub.delltechnologies.com/en-us/p/new-frontiers-dell-emc-poweredge-r750xa-server-with-nvidia-a100-gpus/ (visited on 15/03/2024).

Deng, Tiancheng (2023). 'Effect of the Number of Hidden Layer Neurons on the Accuracy of the Back Propagation Neural Network'. In: *Highlights in Science, Engineering and Technology* 74, pp. 462–468.

Di Cicco, Vincenzo et al. (2019). 'Interpreting deep learning models for entity resolution: an experience report using LIME'. In: *Proceedings of the second international workshop on exploiting artificial intelligence techniques for data management*, pp. 1–4.

Dieber, Jürgen et al. (2020). 'Why model why? Assessing the strengths and limitations of LIME'. In: *arXiv preprint arXiv:2012.00093*.

Dolz, Jose et al. (2018). 'HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation'. In: *IEEE transactions on medical imaging* 38.5, pp. 1116–1126.

Dong, Yushun et al. (2022). 'Edits: Modeling and mitigating data bias for graph neural networks'. In: *Proceedings of the ACM web conference 2022*, pp. 1259–1269.

Dosovitskiy, Alexey et al. (2021). 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale'. In: *ICLR*.

Du, Hao et al. (2019). 'Zoom in to where it matters: a hierarchical graph based model for mammogram analysis'. In: *arXiv preprint arXiv:1912.07517*.

Elboushaki, Abdessamad et al. (2020). 'MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences'. In: *Expert Systems with Applications* 139, p. 112829.

ElGhany, Sameh Abd et al. (2021). 'Diagnosis of Various Skin Cancer Lesions Based on Fine-Tuned ResNet50 Deep Network.' In: *Computers, Materials & Continua* 68.1.

Elmore, Joann G et al. (2017). 'A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis'. In: *Journal of pathology informatics* 8.1, p. 12.

Esteva, Andre et al. (2017). 'Dermatologist-level classification of skin cancer with deep neural networks'. In: *nature* 542.7639, pp. 115–118.

Feng, ZhengZhao et al. (2024). 'A Comprehensive Survey of Dynamic Graph Neural Networks: Models, Frameworks, Benchmarks, Experiments and Challenges'. In: *arXiv preprint arXiv:2405.00476*.

Galassi, Andrea et al. (2020). 'Attention in natural language processing'. In: *IEEE transactions on neural networks and learning systems* 32.10, pp. 4291–4308.

Gao, Jianjiong et al. (2013). 'Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal'. In: *Science signaling* 6.269, pl1–pl1.

Gao, Jianliang et al. (2020). 'MGNN: A multimodal graph neural network for predicting the survival of cancer patients'. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1697–1700.

Garreau, Damien et al. (2020). 'Explaining the explainer: A first theoretical analysis of LIME'. In: *International conference on artificial intelligence and statistics*. PMLR, pp. 1287–1296.

Ghadekar, Premanand et al. (2024). 'Multi-Class Malware Detection using modified GNN and Explainable AI'. In: *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)*. IEEE, pp. 1–8.

Gholamalinezhad, Hossein et al. (2020). 'Pooling methods in deep neural networks, a review'. In: *arXiv preprint arXiv:2009.07485*.

Giudici, Paolo et al. (2021). 'Shapley-Lorenz eXplainable artificial intelligence'. In: *Expert systems with applications* 167, p. 114104.

Gomez, Daniel R et al. (2015). 'Time to treatment as a quality metric in lung cancer: staging studies, time to treatment, and patient survival'. In: *Radiotherapy and Oncology* 115.2, pp. 257–263.

goura (June 2024). *Feedforward neural network*. URL: https://www.geeksforgeeks.org/feedforward-neural-network/.

Goyal, Priya et al. (2017). 'Accurate, large minibatch sgd: Training imagenet in 1 hour'. In: *arXiv preprint arXiv:1706.02677*.

Griewing, Sebastian et al. (2024). 'Evolution of publicly available large language models for complex decision-making in breast cancer care'. In: *Archives of Gynecology and Obstetrics* 310.1, pp. 537–550.

Guo, Meng-Hao et al. (2022). 'Attention mechanisms in computer vision: A survey'. In: *Computational visual media* 8.3, pp. 331–368.

Gupta, Siddharth Raj (2022). 'Prediction time of breast cancer tumor recurrence using Machine Learning'. In: *Cancer Treatment and Research Communications* 32, p. 100602.

Gurcan, Metin N et al. (2009). 'Histopathological image analysis: A review'. In: *IEEE reviews in biomedical engineering* 2, pp. 147–171.

Gurmessa, Daraje kaba et al. (2024). 'Explainable machine learning for breast cancer diagnosis from mammography and ultrasound images: a systematic review'. In: *BMJ Health & Care Informatics* 31.1, e100954.

Gururangan, Suchin et al. (2020). 'Don't stop pretraining: Adapt language models to domains and tasks'. In: *arXiv preprint arXiv:2004.10964*.

Hamamoto, Ryuji et al. (2022). 'Application of non-negative matrix factorization in oncology: one approach for establishing precision medicine'. In: *Briefings in bioinformatics* 23.4, bbac246.

Hamilton, William L (2020). *Graph representation learning*. Morgan & Claypool Publishers.

Han, Kai et al. (2022). 'Vision gnn: An image is worth graph of nodes'. In: *Advances in neural information processing systems* 35, pp. 8291–8303.

Hassanin, Mohammed et al. (2024). 'Visual attention methods in deep learning: An in-depth survey'. In: *Information Fusion* 108, p. 102417.

Hastie, Trevor et al. (2005). 'The elements of statistical learning: data mining, inference and prediction'. In: *The Mathematical Intelligencer* 27.2, pp. 83–85.

He, Kaiming et al. (2015). 'Spatial pyramid pooling in deep convolutional networks for visual recognition'. In: *IEEE transactions on pattern analysis and machine intelligence* 37.9, pp. 1904–1916.

— (2016). 'Deep residual learning for image recognition'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hendrycks, Dan et al. (2020). 'Measuring massive multitask language understanding'. In: *arXiv preprint arXiv:2009.03300*.

Heuillet, Alexandre et al. (2022). 'Collective explainable AI: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with shapley values'. In: *IEEE Computational Intelligence Magazine* 17.1, pp. 59–71.

Hicks, David G et al. (2005). 'Assessment of the HER2 status in breast cancer by fluorescence in situ hybridization: a technical review with interpretive guidelines'. In: *Human pathology* 36.3, pp. 250–261.

Hochreiter, Sepp et al. (1997). 'Long Short-Term Memory'. In: *Neural Computation* 9.8, pp. 1735–1780.

Hoda, Syed A et al. (2017). *Robbins basic pathology*.

Holzinger, Andreas et al. (2022). 'Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence'. In: *Information Fusion* 79, pp. 263–278.

Homoki, Péter et al. (2024). 'Large language models and their possible uses in law'. In: *Hungarian Journal of Legal Studies*.

Howard, Jeremy et al. (2018). 'Universal language model fine-tuning for text classification'. In: *arXiv preprint arXiv:1801.06146*.

Hsieh, Weiche et al. (2024). 'A Comprehensive Guide to Explainable AI: From Classical Models to LLMs'. In: *arXiv preprint arXiv:2412.00800*.

Hu, Linwei et al. (2018). 'Locally interpretable models and effects based on supervised partitioning (LIME-SUP)'. In: *arXiv preprint arXiv:1806.00663*.

Huang, Dong et al. (2023a). 'Bias assessment and mitigation in llm-based code generation'. In: *arXiv preprint arXiv:2309.14345*.

Huang, Gao et al. (2017). 'Densely connected convolutional networks'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Huang, Hui et al. (2023b). 'Towards making the most of llm for translation quality estimation'. In: *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, pp. 375–386.

Huang, Kexin et al. (2023c). 'Uncertainty Quantification over Graph with Conformalized Graph Neural Networks'. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., pp. 26699–26721. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/54a1495b06c4ee2f07184afb9a37abda-Paper-Conference.pdf.

Huang, Lei et al. (2023d). 'Normalization techniques in training dnns: Methodology, analysis and application'. In: *IEEE transactions on pattern analysis and machine intelligence* 45.8, pp. 10173–10196.

Huang, Xuanxiang et al. (2023e). 'The inadequacy of Shapley values for explainability'. In: *arXiv preprint arXiv:2302.08160*.

— (2024). 'On the failings of Shapley values for explainability'. In: *International Journal of Approximate Reasoning* 171. Synergies between Machine Learning and Reasoning, p. 109112. ISSN: 0888-613X. DOI: https://doi.org/10.1016/j.ijar.2023.109112. URL: https://www.sciencedirect.com/science/article/pii/S0888613X23002438.

Huo, Cuiying et al. (2023). 'T2-gnn: Graph neural networks for graphs with incomplete features and structure via teacher-student distillation'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 4, pp. 4339–4346.

Hutchinson, Lisa et al. (2010). 'Focus issue on biomarkers'. In: *Nature Reviews Clinical Oncology* 7.6, pp. 295–295.

Ioffe, Sergey et al. (2015). 'Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift'. In: *International Conference on Machine Learning (ICML)*. PMLR, pp. 448–456.

Izadi, Mohammad Rasool et al. (2020). 'Optimization of graph neural networks with natural gradient descent'. In: *2020 IEEE international conference on big data (big data)*. IEEE, pp. 171–179.

Jeong, Cheonsu (2024). 'Fine-tuning and utilization methods of domain-specific llms'. In: *arXiv preprint arXiv:2401.02981*.

Ji, Cunmei et al. (2021). 'Predicting miRNA-disease associations based on heterogeneous graph attention networks'. In: *Frontiers in genetics* 12, p. 727744.

Jin, Bowen et al. (2024). 'Large language models on graphs: A comprehensive survey'. In: *IEEE Transactions on Knowledge and Data Engineering*.

Jin, Li et al. (2018). 'Breast cancer lung metastasis: Molecular biology and therapeutic implications'. In: *Molecular and Clinical Oncology* 9.2, pp. 123–134. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6300341/.

Kakkad, J et al. (1958). *A Survey on Explainability of Graph Neural Networks (2023)*.

Kanada, Yasusi (2016). 'Optimizing neural-network learning rate by using a genetic algorithm with per-epoch mutations'. In: *2016 International joint conference on neural networks (IJCNN)*. IEEE, pp. 1472–1479.

Kandel, Ibrahem et al. (2020). 'The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset'. In: *ICT express* 6.4, pp. 312–315.

Kenton, Jacob Devlin Ming-Wei Chang et al. (2019). 'Bert: Pre-training of deep bidirectional transformers for language understanding'. In: *Proceedings of naacL-HLT*. Vol. 1, p. 2.

Khemani, Bharti et al. (2024). 'A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions'. In: *Journal of Big Data* 11.1, p. 18.

Kipf, Thomas N et al. (2016). 'Semi-supervised classification with graph convolutional networks'. In: *arXiv preprint arXiv:1609.02907*.

Knyazev, Boris et al. (2019). 'Understanding attention and generalization in graph neural networks'. In: *Advances in neural information processing systems* 32.

Koskinen, Johan et al. (2023). 'Multilevel longitudinal analysis of social networks'. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 186.3, pp. 376–400.

Kraev, Egor et al. (2024). 'Shap-Select: Lightweight Feature Selection Using SHAP Values and Regression'. In: *arXiv preprint arXiv:2410.06815*.

Kumar, I Elizabeth et al. (2020). 'Problems with Shapley-value-based explanations as feature importance measures'. In: *International conference on machine learning*. PMLR, pp. 5491–5500.

Kumar, Virendra et al. (2012). 'Radiomics: the process and the challenges'. In: *Magnetic resonance imaging* 30.9, pp. 1234–1248.

Kumari, Deepa et al. (2025). 'Predicting breast cancer recurrence using deep learning'. In: *Discover Applied Sciences* 7.2, p. 113.

Lambora, Annu et al. (2019). 'Genetic algorithm-A literature review'. In: *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*. IEEE, pp. 380–384.

Lee, Minhyeok (2023). 'Gelu activation function in deep learning: a comprehensive mathematical analysis and performance'. In: *arXiv preprint arXiv:2305.12073*.

Lei Ba, Jimmy et al. (2016). 'Layer normalization'. In: *ArXiv e-prints*, arXiv–1607.

Li, Chun et al. (2021). 'Transfer learning for establishment of recognition of COVID-19 on CT imaging using small-sized training datasets'. In: *Knowledge-based systems* 218, p. 106849.

Li, Linghui et al. (2017). 'Image caption with global-local attention'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1.

Li, Lingyao et al. (2024a). 'A scoping review of using Large Language Models (LLMs) to investigate Electronic Health Records (EHRs)'. In: *arXiv preprint arXiv:2405.03066*.

Li, Mengfang et al. (2023). 'Medical image analysis using deep learning algorithms'. In: *Frontiers in Public Health* 11, p. 1273253.

Li, Ming et al. (2024b). 'Ruler: Improving llm controllability by rule-based data recycling'. In: *arXiv preprint arXiv:2406.15938*.

Li, Peng et al. (2024c). 'Table-GPT: Table Fine-tuned GPT for Diverse Table Tasks'. In: *Proceedings of the ACM on Management of Data* 2.3, pp. 1–28.

Li, Simin et al. (2022). 'A review of radiomics and genomics applications in cancers: the way towards precision medicine'. In: *Radiation Oncology* 17.1, p. 217.

Lin, Chin-Yew (2004). 'ROUGE: A Package for Automatic Evaluation of Summaries'. In: *Annual Meeting of the Association for Computational Linguistics*. URL: https://api.semanticscholar.org/CorpusID:964287.

Lin, Lu et al. (2020). 'Graph attention networks over edge content-based channels'. In: *proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1819–1827.

Lin, Stephanie et al. (2021). 'Truthfulqa: Measuring how models mimic human falsehoods'. In: *arXiv preprint arXiv:2109.07958*.

Lipkova, Jana et al. (2022). 'Artificial intelligence for multimodal data integration in oncology'. In: *Cancer cell* 40.10, pp. 1095–1110.

Liu, Guanxiong et al. (2019). 'Clinically accurate chest x-ray report generation'. In: *Machine Learning for Healthcare Conference*. PMLR, pp. 249–269.

Liu, Yun et al. (2017). 'Detecting cancer metastases on gigapixel pathology images'. In: *arXiv preprint arXiv:1703.02442*.

Lu, Jiasen et al. (2017). 'Knowing when to look: Adaptive attention via a visual sentinel for image captioning'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 375–383.

Lu, Yiping et al. (2019). 'Understanding and improving transformer from a multi-particle dynamic system point of view'. In: *arXiv preprint arXiv:1906.02762*.

Lund, Brady D et al. (2023). 'ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing'. In: *Journal of the Association for Information Science and Technology* 74.5, pp. 570–581.

Lundberg, Scott (2017). 'A unified approach to interpreting model predictions'. In: *arXiv preprint arXiv:1705.07874*.

Luo, Wei-quan et al. (2019). 'Predicting breast cancer in breast imaging reporting and data system (BI-RADS) ultrasound category 4 or 5 lesions: a nomogram combining radiomics and BI-RADS'. In: *Scientific reports* 9.1, p. 11921.

Maini, Yash et al. (2025). 'Breast Tumor Classification with Fine-Tuned Hyperparameter Training using Deep Learning Models'. In: *2025 AI-Driven Smart Healthcare for Society 5.0*. IEEE, pp. 54–59.

Maleki Varnosfaderani, Shiva et al. (2024). 'The role of AI in hospitals and clinics: transforming healthcare in the 21st century'. In: *Bioengineering* 11.4, p. 337.

Mandal, Bishwas et al. (2021). 'Masked face recognition using resnet-50'. In: *arXiv preprint arXiv:2104.08997*.

Mars, Mourad (2022). 'From word embeddings to pre-trained language models: A state-of-the-art walkthrough'. In: *Applied Sciences* 12.17, p. 8805.

Mathews, Roshan P et al. (2020). 'CAD Applications and Emerging Research Potential in Medical Imaging'. In: *arXiv preprint arXiv:2009.14657*.

Mattiuzzi, Camilla et al. (2019). 'Current cancer epidemiology'. In: *Journal of epidemiology and global health* 9.4, pp. 217–222.

Mayer, Alain et al. (1995). 'Local computations on static and dynamic graphs'. In: *Proceedings Third Israel Symposium on the Theory of Computing and Systems*. IEEE, pp. 268–278.

McKee, Brady J et al. (2016). 'Performance of ACR Lung-RADS in a clinical CT lung screening program'. In: *Journal of the American College of Radiology* 13.2, R25–R29.

Meskó, Bertalan et al. (2023). 'The imperative for regulatory oversight of large language models (or generative AI) in healthcare'. In: *NPJ digital medicine* 6.1, p. 120.

Messalas, Andreas et al. (2019). 'Model-agnostic interpretability with shapley values'. In: *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, pp. 1–7.

Min, Seung Won et al. (2021). 'Pytorch-direct: Enabling gpu centric data access for very large graph neural network training with irregular accesses'. In: *arXiv preprint arXiv:2101.07956*.

Miotto, Riccardo et al. (2018). 'Deep learning for healthcare: review, opportunities and challenges'. In: *Briefings in bioinformatics* 19.6, pp. 1236–1246.

Mirjalili, Seyedali et al. (2019). 'Genetic algorithm'. In: *Evolutionary algorithms and neural networks: theory and applications*, pp. 43–55.

Molnar, Christoph (2020). *Interpretable machine learning*. Lulu. com.

Moreira, Inês C et al. (2012). 'Inbreast: toward a full-field digital mammographic database'. In: *Academic radiology* 19.2, pp. 236–248.

Mukti, Ishrat Zahan et al. (2019). 'Transfer learning based plant diseases detection using ResNet50'. In: *2019 4th International conference on electrical information and communication technology (EICT)*. IEEE, pp. 1–6.

Mullooly, Maeve et al. (2019). 'Application of convolutional neural networks to breast biopsies to delineate tissue correlates of mammographic breast density'. In: *NPJ breast cancer* 5.1, p. 43.

Munikoti, Sai et al. (2023). 'Challenges and opportunities in deep reinforcement learning with graph neural networks: A comprehensive review of algorithms and applications'. In: *IEEE transactions on neural networks and learning systems*.

Munn, Michael et al. (2022). *Explainable AI for practitioners*. " O'Reilly Media, Inc."

Nam, Hyeonseob et al. (2017). 'Dual attention networks for multimodal reasoning and matching'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 299–307.

National Grid ESO (2024). *UK Carbon Intensity Dashboard*. Real-time and historical data for UK electricity carbon intensity (0.21 kg CO_2/kWh as of 2024). National Grid Electricity System Operator (ESO). URL: https://carbonintensity.org.uk (visited on 15/03/2024).

Nazi, Zabir Al et al. (2024). 'Large language models in healthcare and medical domain: A review'. In: *Informatics*. Vol. 11. 3. MDPI, p. 57.

Nicolini, Antonio et al. (2006). 'Metastatic breast cancer: an updating'. In: *Biomedicine & pharmacotherapy* 60.9, pp. 548–556.

Nikolić, Goran S et al. (2022). 'A survey of three types of processing units: CPU, GPU and TPU'. In: *2022 57th international scientific conference on information, communication and energy systems and technologies (ICEST)*. IEEE, pp. 1–6.

Niu, Zhaoyang et al. (2021). 'A review on the attention mechanism of deep learning'. In: *Neurocomputing* 452, pp. 48–62.

Noguer i Alonso, Miquel (1950). 'Key Milestones in Natural Language Processing (NLP) 1950-2024'. In: *Key Milestones in Natural Language Processing (NLP)* 2024.

Nowak-Sliwinska, Patrycja et al. (2019). 'Drug repurposing in oncology: Compounds, pathways, phenotypes and computational approaches for colorectal cancer'. In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1871.2, pp. 434–454.

OpenI (2024). 'Indiana University - Chest X-Rays (PNG Images)'. In: *Indiana University - Chest X-Rays*. URL: https://openi.nlm.nih.gov/faq.php.

Organization, World Health (2023a). *Breast Cancer*. Fact sheet. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/breast-cancer (visited on 25/02/2025).

— (2023b). *Cancer*. Fact sheet. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/cancer (visited on 22/02/2025).

Parthasarathy, Venkatesh Balavadhani et al. (2024). 'The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities'. In: *arXiv preprint arXiv:2408.13296*.

Patrini, Giorgio et al. (2017). 'Making deep neural networks robust to label noise: A loss correction approach'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952.

Peersman, Claudia et al. (2011). 'Predicting age and gender in online social networks'. In: *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pp. 37–44.

Prechelt, Lutz (2002). 'Early stopping-but when?' In: *Neural Networks: Tricks of the trade*. Springer, pp. 55–69.

Puccetti, Giovanni et al. (2021). 'How Do BERT Embeddings Organize Linguistic Knowledge?' In: *Proceedings of deep learning inside out (DeeLIO): the 2nd workshop on knowledge extraction and integration for deep learning architectures*, pp. 48–57.

Pulumati, Akhil et al. (2023). 'Technological advancements in cancer diagnostics: Improvements and limitations'. In: *Cancer Reports* 6.2, e1764.

Rabbani, Shafaat A et al. (2007). 'Evaluating distant metastases in breast cancer: from biology to outcomes'. In: *Cancer and Metastasis Reviews* 26, pp. 663–674.

Radenović, Filip et al. (2018). 'Fine-tuning CNN image retrieval with no human annotation'. In: *IEEE transactions on pattern analysis and machine intelligence* 41.7, pp. 1655–1668.

Radford, Alec et al. (2019). 'Language models are unsupervised multitask learners'. In: *OpenAI blog* 1.8, p. 9.

Radiuk, Pavlo M (2017). 'Impact of training set batch size on the performance of convolutional neural networks for diverse datasets'. In.

Rafferty, Amy et al. (2025). 'Leveraging Expert Input for Robust and Explainable AI-Assisted Lung Cancer Detection in Chest X-rays'. In: *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)*. IEEE, pp. 576–587.

Rajpurkar, Pranav et al. (2017). 'Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning'. In: *arXiv preprint arXiv:1711.05225*.

Rasmy, Laila et al. (2021). 'Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction'. In: *NPJ digital medicine* 4.1, p. 86.

Ray, Partha Pratim (2023). 'ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope'. In: *Internet of Things and Cyber-Physical Systems* 3, pp. 121–154.

Redig, Amanda J et al. (2013). 'Breast cancer as a systemic disease: a view of metastasis'. In: *Journal of internal medicine* 274.2, pp. 113–126.

Reiser, Patrick et al. (2022). 'Graph neural networks for materials science and chemistry'. In: *Communications Materials* 3.1, p. 93.

Ribeiro, Marco Tulio et al. (2016). '" Why should i trust you?" Explaining the predictions of any classifier'. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Ribeiro, Marco Tulio et al. (2020). 'Beyond accuracy: Behavioral testing of NLP models with CheckList'. In: *arXiv preprint arXiv:2005.04118*.

Ribli, Dezső et al. (2018). 'Detecting and classifying lesions in mammograms with deep learning'. In: *Scientific reports* 8.1, p. 4165.

Riggio, Alessandra I et al. (2021). 'The lingering mysteries of metastatic recurrence in breast cancer'. In: *British journal of cancer* 124.1, pp. 13–26.

Rodrawangpai, Ben et al. (2022). 'Improving text classification with transformers and layer normalization'. In: *Machine Learning with Applications* 10, p. 100403. ISSN: 2666-8270. DOI: https://doi.org/10.1016/j.mlwa.2022.100403. URL: https://www.sciencedirect.com/science/article/pii/S2666827022000792.

Rogers, Anna et al. (2021). 'A primer in BERTology: What we know about how BERT works'. In: *Transactions of the Association for Computational Linguistics* 8, pp. 842–866.

Rosai, Juan (2011). *Rosai and Ackerman's surgical pathology e-book*. Elsevier Health Sciences.

Rossi, Emanuele et al. (2020). 'Temporal graph networks for deep learning on dynamic graphs'. In: *arXiv preprint arXiv:2006.10637*.

Roth, Alvin E (1988). 'Introduction to the Shapley value'. In: *The Shapley value* 1.

Rupavath, Rana Veer Samara Sihman Brahmatej et al. (2024). 'A Deep Learning Based Metastasis Prediction System Using Multi Omics Data and Recursive Neural Network'. In: *2024 4th International Conference on Mobile Networks and Wireless Communications (ICMNWC)*. IEEE, pp. 1–5.

Sabih, Muhammad et al. (2020). 'Utilizing explainable AI for quantization and pruning of deep neural networks'. In: *arXiv preprint arXiv:2008.09072*.

Saihood, Ahmed Ali et al. (2024). 'Multiside graph neural network-based attention for local co-occurrence features fusion in lung nodule classification'. In: *Expert Systems with Applications* 252, p. 124149.

Salvi, Massimo et al. (2024). 'Multi-modality approaches for medical support systems: A systematic review of the last decade'. In: *Information Fusion* 103, p. 102134.

Sanmorino, Ahmad et al. (2024). 'Fine-tuning a pre-trained ResNet50 model to detect distributed denial of service attack'. In: *Bulletin of Electrical Engineering and Informatics* 13.2, pp. 1362–1370.

Santos, Claudio Filipi Gonçalves Dos et al. (2022). 'Avoiding overfitting: A survey on regularization methods for convolutional neural networks'. In: *ACM Computing Surveys (CSUR)* 54.10s, pp. 1–25.

Selvaraju, Ramprasaath R et al. (2017). 'Grad-cam: Visual explanations from deep networks via gradient-based localization'. In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Serafini, Marco et al. (2021). 'Scalable graph neural network training: The case for sampling'. In: *ACM SIGOPS Operating Systems Review* 55.1, pp. 68–76.

Serrano, Sofia et al. (2019). 'Is attention interpretable?' In: *arXiv preprint arXiv:1906.03731*.

Shahab, Omer et al. (2024). 'Large language models: a primer and gastroenterology applications'. In: *Therapeutic Advances in Gastroenterology* 17, p. 17562848241227031.

Shao, Yingxia et al. (2024). 'Distributed graph neural network training: A survey'. In: *ACM Computing Surveys* 56.8, pp. 1–39.

Sharma, Sagar et al. (2017). 'Activation functions in neural networks'. In: *Towards Data Sci* 6.12, pp. 310–316.

Shi, Sheng et al. (2020). 'A modified perturbed sampling method for local interpretable model-agnostic explanation'. In: *arXiv preprint arXiv:2002.07434*.

Shimodaira, Hidetoshi (2000). 'Improving predictive inference under covariate shift by weighting the log-likelihood function'. In: *Journal of Statistical Planning and Inference* 90.2, pp. 227–244.

Shorten, Connor et al. (2019). 'A survey on image data augmentation for deep learning'. In: *Journal of big data* 6.1, pp. 1–48.

Shrestha, Ajay et al. (2019). 'Review of deep learning algorithms and architectures'. In: *IEEE access* 7, pp. 53040–53065.

Siegel, Rebecca L et al. (2021). 'Cancer statistics, 2021'. In: *CA: a cancer journal for clinicians* 71.1, pp. 7–33.

Singh, Sonit et al. (2019). 'From chest x-rays to radiology reports: a multimodal machine learning approach'. In: *2019 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, pp. 1–8.

Song, Won-min et al. (2014). 'Planar filtered gene regulatory networks in breast cancer'. In: *Cancer Research* 74.19_Supplement, pp. 363–363.

Sonkar, Shashank et al. (2023). 'Investigating the Role of Feed-Forward Networks in Transformers Using Parallel Attention and Feed-Forward Net Design'. In: *arXiv preprint arXiv:2305.13297*.

Sorin, Vera et al. (2024). 'Utilizing large language models in breast cancer management: systematic review'. In: *Journal of cancer research and clinical oncology* 150.3, p. 140.

Specht, Donald F et al. (1991). 'A general regression neural network'. In: *IEEE transactions on neural networks* 2.6, pp. 568–576.

Stojkovic, Jovan et al. (2024). 'Towards Greener LLMs: Bringing Energy-Efficiency to the Forefront of LLM Inference'. In: *arXiv preprint arXiv:2403.20306*.

Sun, Wenbo et al. (2024). 'Cebench: A benchmarking toolkit for the cost-effectiveness of llm pipelines'. In: *arXiv preprint arXiv:2407.12797*.

Sung, Chul et al. (2019). 'Pre-training BERT on domain resources for short answer grading'. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6071–6075.

Tang, Hao et al. (2022). 'Learning attention-guided pyramidal features for few-shot fine-grained recognition'. In: *Pattern Recognition* 130, p. 108792.

Tang, Haoteng et al. (2021). 'Commpool: An interpretable graph pooling framework for hierarchical graph representation learning'. In: *Neural Networks* 143, pp. 669–677.

Tang, Raphael et al. (2019). 'Distilling task-specific knowledge from bert into simple neural networks'. In: *arXiv preprint arXiv:1903.12136*.

Tao, Mingxu et al. (2023). 'A frustratingly easy improvement for position embeddings via random padding'. In: *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, pp. 301–313.

Tenney, I (2019). 'BERT rediscovers the classical NLP pipeline'. In: *arXiv preprint arXiv:1905.05950*.

Testa, Ugo et al. (2020). 'Breast cancer: a molecularly heterogenous disease needing subtype-specific treatments'. In: *Medical Sciences* 8.1, p. 18.

Thwin, Su Myat et al. (2024). 'Attention-Based Ensemble Network for Effective Breast Cancer Classification over Benchmarks'. In: *Technologies* 12.2, p. 16.

Tian, Yang et al. (2020). 'An unsupervised regularization and dropout based deep neural network and its application for thermal error prediction'. In: *Applied Sciences* 10.8, p. 2870.

Ting, Fung Fung et al. (2019). 'Convolutional neural network improvement for breast cancer classification'. In: *Expert Systems with Applications* 120, pp. 103–115.

Tong, Yin-Syuen et al. (2022). 'Deep Learning for Image-Based Plant Growth Monitoring: A Review.' In: *International Journal of Engineering & Technology Innovation* 12.3.

Torres, Luis HM et al. (2023). 'Few-shot learning with transformers via graph embeddings for molecular property prediction'. In: *Expert Systems with Applications* 225, p. 120005.

Tuerlinckx, Francis (2004). 'The efficient computation of the cumulative distribution and probability density functions in the diffusion model'. In: *Behavior Research Methods, Instruments, & Computers* 36.4, pp. 702–716.

Tunstall, Lewis et al. (2022). *Natural language processing with transformers*. " O'Reilly Media, Inc."

Tutek, Martin et al. (2022). 'Toward practical usage of the attention mechanism as a tool for interpretability'. In: *IEEE access* 10, pp. 47011–47030.

Ullah, Asad et al. (2024). 'Predictive potential of K-Banhatti and Zagreb type molecular descriptors in structure–property relationship analysis of some novel drug molecules'. In: *Journal of the Chinese Chemical Society* 71.3, pp. 250–276.

Vaswani, A (2017). 'Attention is all you need'. In: *Advances in Neural Information Processing Systems*.

Vega García, María et al. (2020). 'Shapley additive explanations for NO2 forecasting'. In: *Ecological Informatics* 56, p. 101039. ISSN: 1574-9541. DOI: https://doi.org/10.1016/j.ecoinf.2019.101039. URL: https://www.sciencedirect.com/science/article/pii/S1574954119303498.

Veit, Andreas et al. (2016). 'Residual Networks Behave Like Ensembles of Relatively Shallow Networks'. In: *Advances in Neural Information Processing Systems (NeurIPS)* 29.

Veličković, Petar et al. (2017). 'Graph attention networks'. In: *arXiv preprint arXiv:1710.10903*.

Virinchi, Srinivas et al. (2023). 'BLADE: Biased neighborhood sampling based graph neural network for directed graphs'. In: *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 42–50.

Waikhom, Lilapati et al. (2023). 'A survey of graph neural networks in various learning paradigms: methods, applications, and challenges'. In: *Artificial Intelligence Review* 56.7, pp. 6295–6364.

Wang, Alex et al. (2018a). 'GLUE: A multi-task benchmark and analysis platform for natural language understanding'. In: *arXiv preprint arXiv:1804.07461*.

Wang, Chenguang et al. (2019). 'Language models with transformers'. In: *arXiv preprint arXiv:1904.09408*.

Wang, Shui-Hua et al. (2020). 'DenseNet-201-based deep neural network with composite learning factor and precomputation for multiple sclerosis classification'. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16.2s, pp. 1–19.

Wang, Xiaosong et al. (2017a). 'Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106.

Wang, Xiaosong et al. (2017b). 'Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases'. In: *IEEE CVPR*. Vol. 7. sn, p. 46.

Wang, Xiaosong et al. (2018b). 'Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9049–9058.

Wang, Yong et al. (2015). 'Ambiguityvis: Visualization of ambiguity in graph layouts'. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1, pp. 359–368.

Wani, Niyaz Ahmad et al. (2024). 'DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence'. In: *Computer Methods and Programs in Biomedicine* 243, p. 107879.

Wen, Long et al. (2020). 'A transfer convolutional neural network for fault diagnosis based on ResNet-50'. In: *Neural Computing and Applications* 32.10, pp. 6111–6124.

Wiegreffe, Sarah et al. (2019). 'Attention is not not explanation'. In: *arXiv preprint arXiv:1908.04626*.

Wolf, Thomas et al. (2020). 'Transformers: State-of-the-art natural language processing'. In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45.

Wu, Bingzhe et al. (2022). 'A survey of trustworthy graph learning: Reliability, explainability, and privacy protection'. In: *arXiv preprint arXiv:2205.10014*.

Wu, Haiping et al. (2020). 'Visual Transformers: Token-based Image Representation and Processing for Computer Vision'. In: *arXiv:2006.03677*.

Wu, Yonghui et al. (2016). 'Google's neural machine translation system: Bridging the gap between human and machine translation'. In: *arXiv preprint arXiv:1609.08144*.

xinyuwei (Jan. 2023). *Differences between Pre-Training and Supervised Fine-Tuning (SFT)*. URL: https://techcommunity.microsoft.com/blog/machinelearningblog/differences-between-pre-training-and-supervised-fine-tuning-sft/4220673.

Xiong, Xinye et al. (2023). 'Graph Representation Learning for Interactive Biomolecule Systems'. In: *arXiv preprint arXiv:2304.02656*.

Xu, Chenghao et al. (2023). 'Intelligent task allocation for mobile crowdsensing with graph attention network and deep reinforcement learning'. In: *IEEE Transactions on Network Science and Engineering* 10.2, pp. 1032–1048.

Xu, Guoping et al. (2024a). 'Development of Skip Connection in Deep Neural Networks for Computer Vision and Medical Image Analysis: A Survey'. In: *arXiv preprint arXiv:2405.01725*.

Xu, Kelvin et al. (2015). 'Show, Attend and Tell: Neural Image Caption Generation with Visual Attention'. In: *ICML*.

Xu, Xi et al. (2024b). 'A Comprehensive Review on Synergy of Multi-Modal Data and AI Technologies in Medical Diagnosis'. In: *Bioengineering* 11.3, p. 219.

Yang, Han et al. (2021). 'Rethinking graph regularization for graph neural networks'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 5, pp. 4573–4581.

Yang, Jialiang et al. (2022a). 'Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning'. In: *Computational and structural biotechnology journal* 20, pp. 333–342.

Yang, Shaokang et al. (2020). 'Automatic medical image report generation with multi-view and multi-modal attention mechanism'. In: *International Conference on Algorithms and Architectures for Parallel Processing*. Springer, pp. 687–699.

Yang, Xiaoli et al. (2022b). 'PD-ResNet for classification of Parkinson's disease from gait'. In: *IEEE Journal of Translational Engineering in Health and Medicine* 10, pp. 1–11.

Yang, Zichao et al. (2016). 'Hierarchical attention networks for document classification'. In: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480–1489.

Ye, Qinyuan et al. (2023). 'Prompt engineering a prompt engineer'. In: *arXiv preprint arXiv:2311.05661*.

Yenduri, Gokul et al. (2024). 'Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions'. In: *IEEE Access*.

Yi, Yang et al. (2023). 'Graph classification via discriminative edge feature learning'. In: *Pattern Recognition* 143, p. 109799. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2023.109799. URL: https://www.sciencedirect.com/science/article/pii/S0031320323004971.

Ying, Zhitao et al. (2019). 'Gnnexplainer: Generating explanations for graph neural networks'. In: *Advances in neural information processing systems* 32.

Yousefi, Meysam et al. (2018). 'Organ-specific metastasis of breast cancer: molecular and cellular mechanisms underlying lung metastasis'. In: *Cellular Oncology* 41.2, pp. 123–140.

Yu, Haiyang et al. (2022). 'GraphFM: Improving large-scale GNN training via feature momentum'. In: *International Conference on Machine Learning*. PMLR, pp. 25684–25701.

Yu, Hang et al. (2021). 'Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives'. In: *Neurocomputing* 444, pp. 92–110.

Yuan, Hao et al. (2021). 'On explainability of graph neural networks via subgraph explorations'. In: *International conference on machine learning*. PMLR, pp. 12241–12252.

Zaeemzadeh, Alireza et al. (2020). 'Norm-preservation: Why residual networks can become extremely deep?' In: *IEEE transactions on pattern analysis and machine intelligence* 43.11, pp. 3980–3990.

Zafar, Afia et al. (2022). 'A comparison of pooling methods for convolutional neural networks'. In: *Applied Sciences* 12.17, p. 8643.

Zech, John R et al. (2018). 'Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study'. In: *PLoS medicine* 15.11, e1002683.

Zhang, Aston et al. (2024). *Dive into Deep Learning*. Cambridge University Press, p. 574.

Zhang, Chuxu et al. (2019). 'Heterogeneous graph neural network'. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 793–803.

Zhang, Guangyun et al. (2018). 'Hierarchical spatial features learning with deep CNNs for very high-resolution remote sensing image classification'. In: *International Journal of Remote Sensing* 39.18, pp. 5978–5996.

Zhang, Mo et al. (2022a). 'MS-GWNN: multi-scale graph wavelet neural network for breast cancer diagnosis'. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 1–5.

Zhang, Muhan et al. (2021a). 'Nested Graph Neural Networks'. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 15734–15747. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/8462a7c229aea03dde69da754c3bbcc4-Paper.pdf.

Zhang, Wentao et al. (2022b). 'Graph attention multi-layer perceptron'. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4560–4570.

Zhang, Xiao-Meng et al. (2021b). 'Graph neural networks and their current applications in bioinformatics'. In: *Frontiers in genetics* 12, p. 690049.

Zhao, Guixiang et al. (2022). 'Prevalence of lung cancer in chronic obstructive pulmonary disease: A systematic review and meta-analysis'. In: *Frontiers in Oncology* 12, p. 947981.

Zhou, Zijie et al. (2024). 'Towards Understanding Neural Machine Translation with Attention Heads' Importance'. In: *Applied Sciences* 14.7, p. 2798.

Zhu, Hu et al. (2020). 'Multimodal Fusion Method Based on Self-Attention Mechanism'. In: *Wireless Communications and Mobile Computing* 2020.1, p. 8843186.

Zhu, Q. et al. (2022a). 'Generative Pre-Trained Transformer for Design Concept Generation: An Exploration'. In: *Proceedings of the Design Society* 2, pp. 1825–1834. DOI: 10.1017/pds.2022.185.

Zhu, Yajing et al. (2022b). 'Expression patterns and prognostic implications of tumor-infiltrating lymphocytes dynamics in early breast cancer patients receiving neoadjuvant therapy: A systematic review and meta-analysis'. In: *Frontiers in Oncology* 12, p. 999843.

Zia, Tehseen et al. (2020). 'Learning deep hierarchical and temporal recurrent neural networks with residual learning'. In: *International Journal of Machine Learning and Cybernetics* 11, pp. 873–882.

Chapter 7

Appendices

7.1 Appendix A

Variables	**N (%)**
Age at diagnosis[Median (range)]	61.8 (21.93-96.29)
Tumour size [Median (range)]	23 (1, 182)
NPI [Median (95% CI)]	4.46 (4.41-4.51)
Survival [Median (Months, 95% Cl)]	149 (141-159)
Axillary lymph nodes status	
0	1035
1	337
2	171
3	114
>3	314
Axillary lymph nodes status	
Positive	1497
Negative	438
Null	42
PAM50 subtype	
Basal	330
HER2	238
Luminal A	715
Luminal B	489
Normal-like	199
Not classified	6
Adjuvant systemic therapy (AT)	
No AT	305
Hormone therapy (HT)	1216
Chemotherapy	416
Hormone + chemotherapy	192

Table 7.1: Clinicopathological characteristics in the METABRIC cohort.

7.2 Appendix B

Risk factors	Studies	Model	OR	Heterogeneity		$P_{difference}$
	n		(95% CI)	I^2	*P*	
Gender						
Male	4	Random	0.48 (0.09, 2.66)	99.50%	0	0.398
Female	2	Random	0.13 (0.00, 4.86)	99.70%	0	0.268
COPD severity						
Mild	3	Fixed	1.79(1.23, 2.60)	21.90%	0.278	0.002
Moderate	3	Fixed	2.14(1.44, 3.18)	0	0.931	0.000
Severe	2	Fixed	1.36(0.80, 2.31)	0	0.419	0.251
Very severe	1	Fixed	0.60(0.18, 1.98)	0	0.569	0.404
Smoking status						
Never smoking	3	Fixed	2.94(2.38, 3.64)	31.40%	0.233	0.000
Former smoking	4	Random	3.17(1.30, 7.74)	91.10%	0	0.011
Current smoking	5	Random	3.94(1.28, 12.12)	95.10%	0	0.017

Figure 7.1: Analysis of the risk factors of lung cancer in COPD.

Subgroups	Studies	Total	Events	Model	ES	Heterogeneity		$P_{difference}$
	n				(95%CI)	I^2	*P*	
Gender								
Male	9	62627	3472	random	5.09% (3.48%, 6.70%)	98.80%	0	0.000
Female	8	45620	1724	random	2.52% (1.57%, 4.05%)	99.90%	0	0.000
Smoking status								
Never smoking	4	52863	744	random	0.68% (0.10%, 4.65%)	100%	0	0.000
Former smoking	4	20812	323	random	3.42% (1.51%, 5.32%)	97.600%	0	0.000
Current smoking	5	9879	731	random	8.98% (4.61%, 13.35%)	98.40%	0	0.000
COPD severity								
Mild	6	5311	151	random	3.89% (2.14%, 7.06%)	99.40%	0	0.000
Moderate	3	1986	141	random	6.67% (3.20%, 10.14%)	87.00%	0	0.000
Severe	2	835	70	random	5.57% (1.89%, 16.39%)	94.70%	0	0.000
Cancer type								
Small cell carcinoma	3	8213	35	random	0.78% (0.78%, 1.77%)	99.70%	0	0.000
Adenocarcinoma	3	8213	68	random	1.59% (0.23%, 2.94%)	90.90%	0	0.022
Squamous cell carcinoma	3	8213	75	random	1.35% (0.57%, 3.23%)	99.70%	0	0.000
Region								
European	15	531191	18711	random	3.21% (2.36%, 4.06%)	99.6%	0	0.000
Western Pacific region	12	287245	17558	random	7.78% (5.06%, 10.5%)	99.9%	0	0.000
Americas	2	6888	180	random	3.25% (0.88%, 5.61%)	94.40%	0	0.007

CH, Cohort study; CS, Cross-sectional study; CC, Case-control study.

Figure 7.2: Subgroup analysis of the prevalence of lung cancer in COPD.

Reference	Country	Study design	COPD				Lung cancer			Duration or range of follow-up, years
			Diagnosis	Sample size	Age (years)	M/F	Diagnosis	Sample size	M/F	
Sandelin et al.2018 (22)	Swedish	Retrospective cohort	ICD-10-CM code J44	19894	–	9452/ 110442	ICD-10 code C34	594	291/303	1999.1.1-2009.12.31
Ahn et al., 2020 (23)	Korean	Retrospective cohort	ICD-10 codes J43-J44	11551	–	6172/ 5379	ICD-10 codes C33-C34	1136	–	2004.1.1-2015.12.31
Husebøet al., 2019 (24)	Norway	Prospective cohort	Clinical and Spirometry confirmed	433	63.5 ± 6.9	258/175	Norwegian Cancer Registry	28	–	9
Park et al., 2020 (25)	Korean	Retrospective cohort	ICD-10 codes J43-J44	58972	–	–	ICD-10 code C33 or C34	290	–	2002.1.1-2013.12.31
Machida et al., 2021 (26)	Japan	Prospective cohort	Spirometry confirmed	224	70.4 ± 8.4	214/10	CT	19	19	2014.1-2020.4
Sakai et al., 2020 (27)	Japan	Retrospective cohort	Spirometry confirmed	198	69.7 ± 8.0	184/14	–	43	–	2011.4.1-2015.7.16
Montserrat et al., 2021 (28)	Spain	Retrospective cross-sectional	Spirometry confirmed	24135	72 ± 11	18612/ 5523	ICD-10	552	–	2012.1.1-2017.12.31
Jurevičienė et al., 2022 (29)	Lithuanian	Retrospective cross-sectional	ICD-10-AMD J44.8	4834	67.2 ± 8.4	3338/ 1496	ICD 10 code C33, C34	186	–	2012.1.1-2014.6.30
Thomsen et al., 2012 (34)	Denmark	Prospective cohort	ICD8: 490–492; ICD10: J44	8656	65 (57, 74)	47%/ 53%	ICD10 code C34	93	–	5
Chubachi et al., 2016 (35)	Japan	Prospective cohort	Spirometry confirmed	311	72.3 ± 8.2	278/33	clinical history and medical records	13	–	2
Divo et al., 2012 (13)	USA + Spain	Prospective cohort	Spirometry confirmed	1659	66 ± 9	1477/ 182	medical record and direct questioning	151	–	1997.11-2010.3
Westerik et al., 2017 (36)	Dutch	Retrospective cohort	ICPC code R95 in the electronic medical record	14603	66.5 ± 11.5	7749/ 6854	ICPC code R84	317	–	2012–2013.12.31
Lin et al.2013 (37)	China	Retrospective case-control	ICD-9-CM code 496	2630	–	2096/ 534	cytologically or histologically confirmed	181	–	2006.1.1-2011.12.31
de Torres et al., 2007 (38)	Spain	Prospective cohort	Spirometry confirmed	1166	54 ± 8	74% vs 26%	CT and Biopsy	23	–	2000.9-2005.12
Purdue et al., 2007 (39)	Swedish	Retrospective cohort	Spirometry confirmed	6849	–	6849	ICD-7 codes 162, 163	175	175	1971-2001
Wilson et al., 2008 (40)	USA	Prospective cohort	Spirometry confirmed	1486	–	–	medical records and pathology reports	67	–	3.26
Rodríguez et al., 2010 (41)	UK	Prospective cohort	Oxford Medical Information System [OXMIS] and Read codes	1924	–	–	Oxford Medical Information System [OXMIS] and Read codes	48	–	1996.1.31-2001
de Torres et al., 2011 (42)	USA + Spain	Prospective cohort	Spirometry confirmed	2507	65 ± 9	2307/ 200	medical records and pathology reports	215	205/10	1997.1-2009.12
Kornum et al., 2012 (43)	Danish	Prospective cohort	ICD-8 codes:491-492; ICD-10 codes: J41-J44	236494	–	129344/ 107150	medical records and pathology reports	10118	–	1980-2008
Shen et al., 2014 (44)	China	Retrospective cohort	ICD-9-CM 491, 492, and 496	20730	70	13291/ 7439	ICD-9-CM 162	729	575/154	1998-2011
Hasegawa et al., 2014 (45)	Japan	Retrospective cohort	ICD-10 codes: J41, J42, J43, J44	172707	–	136632/ 36075	ICD-10 codes C34	13930	–	2010.7.1-2013.3.31
Roberts et al.,	UK	Prospective	ICD10 code J44	9716	73 ± 10	4906/	Medical records	180	–	2008.3-2008.8

Figure 7.3: Basic characteristics of the included studies - part1

2011 (46)		cohort	and J45/46 (asthma) later confirmed as COPD			4810	confirmed by physician			
Ställberg et al., 2018 (47)	Swedish	Retrospective cohort	ICD-10 code: J44	17479	–	–	ICD-10 code: C34	1091	–	2000-2014
Mannino et al., 2003 (48)	USA	Prospective cohort	Spirometry confirmed	5402	–	2473/ 2929	ICD-9 code: 162	113	–	1971-1992
Schneider et al., 2010 (49)	UK	Retrospective case-control	OXMIS codes	35772	–	18351/ 17421	OXMIS codes	2585	1526/ 1059	1995.1.1-2005.12.31
Greulich et al., 2017 (50)	Germany	Retrospective case-control	ICD-10: J41, J43, J44	146141	67.2 ± 12.41	51%/ 49%	ICD-10 code not provided	2663	–	2013.1.1-2014.12.31
Jo et al., 2015 (51)	Korean	Retrospective cross-sectional	ICD-10 code: J44	744	65.0 ± 9.40		ICD-10 code: C34	97	–	2010-2012
Deniz et al., 2016 (52)	Turkey	Retrospective cross-sectional	Spirometry confirmed	3095	71.9 ± 10.5	2434/ 661	Medical records	58	–	2014.1.1-2014.12.31
Jung et al., 2018 (53)	Korean	Retrospective cross-sectional	ICD 10 code J44	15949	69 (60, 76)	9039/ 6910	ICD 10 code C34	753	590/163	2011.1-2011.12
Masuda et al., 2017 (54)	Japan	Retrospective cohort	Spirometry confirmed	920	–	651/269	self-reported and confirmed by a physician	13	10/3	2009.4-2010.3
Nishida et al., 2017 (55)	Japan	Retrospective cross-sectional	Spirometry confirmed	2309	69.06 ± 10.53	1549/ 760	ICD-10 code C34	354	–	2005.9-2008.12

COPD, chronic obstructive pulmonary disease; F: female; M: male; ICD, International Classification of Diseases; -: No mentioned.

Figure 7.4: Basic characteristics of the included studies - part2

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2025) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

International Neural Network Society Workshop on Deep Learning Innovations and Applications

AIM-X: Attention-Infused Multimodal Cross-Interaction for X-ray Clinical Report Generation

Abdullah Basaad[a], Shadi Basurra[a], Edlira Vakaj[a], Zeyu Fu[b], Sareh Rowlands[b], Mohammed M. Abdelsamea[b,*]

[a]*School of Computing and Digital Technology, Birmingham City University, 15 Bartholomew Row, Birmingham, UK*
[b]*Department of Computer Science, University of Exeter, North Park Road, Exeter, UK*

Abstract

Medical imaging reports are a crucial part of the medical diagnosis process. The generation of reports from multimodal medical data includes images and text reports, which have been a potential area of clinical assistance in diagnostic tasks. In this work, we introduce a novel architecture, we called attention-infused multimodal cross-interaction (AIM-X), designed to extract and summarize information from X-ray images and their corresponding textual reports into coherent and comprehensive diagnostic reports. AIM-X model introduces a dual-attention mechanism to emphasize the most significant areas within the images, enabling precise diagnoses by healthcare professionals. AIM-X further reinforces the coherence and relevance of the generated reports, embedding the model with an explainability layer that enables visualization of the attention weights. The generated reports have a high proportion of extracting significant medical observations out of the ground truth accurately. Quantitatively, the performance of AIM-X was validated through BLEU measures of 0.7272 (BLEU1), 0.6595 (BLEU2), 0.6012 (BLEU3), 0.4686 (BLEU4), and a ROUGE value of 0.7272, representing a high level of similarity between generated reports and target diagnoses, achieving an accuracy of 94.5%.

Keywords: Medical report generation; Pre-trained Vision-Language Models; Large Language Models; Attention mechanism and LSTM.

1. Introduction

The integration of multimodal data into AI-driven medical diagnostics represents a substantial advancement in healthcare technology [1]. Historically, medical diagnostics have relied on visual data, such as X-rays and CT scans, or textual data, such as patient histories and clinical notes, which were independently reviewed by radiologists and clinicians. Nonetheless, the advent of multimodal report generation models has transformed this process by enabling

* Corresponding author.
E-mail address: m.abdelsamea@exeter.ac.uk

the concurrent analysis of both image and text data [1, 2]. This fusion of modalities enhances the precision and comprehensiveness of diagnostic reports, offering a more holistic approach to patient care.

Convolutional Neural Networks (CNNs), especially sophisticated designs like DensNet, play a pivotal role in multi-modal disease detection systems [3]. DenseNet leverages the network's capabilities by reusing the network's features, thereby producing models that are concise, easy to train, with high parameter efficiency. A concatenation of feature maps learned by different layers would increase variation in the input of subsequent layers and improve efficiency. This capability is especially vital in contexts like early cancer detection, where spotting minor irregularities in images can significantly improve patient outcomes [4, 5, 6].

Recent versions of the Generative Pre-trained Transformers (such as GPT-2,3,4) series have been employed to process text data for comprehending and generating human language [7]. When used alongside CNNs, these models are particularly effective in visual analyses, and ideal for medical report interpretation. These transformer-based methods are capable of performing deep contextual analysis across lengthy text sequences, which is advantageous for generating coherent and contextually accurate medical reports [8, 9]. A pivotal part of multi-modal report generation models is the attention mechanism, which enhances the model's focus and interpretability. This mechanism weights the significant regions of the inputs, both textual and visual, then emphasizes the most pertinent parts [10]. This approach not only boosts the accuracy of the model's predictions but also improves interpretability by elucidating why certain data points were highlighted. For instance, in medical imaging, attention maps can identify regions of an X-ray that are most indicative of a condition, thus making the AI's decision-making process clearer and more reliable for healthcare professionals [11, 12]. The cooperation between visual and textual data in multi-modal models leads to more accurate diagnostics and explainability [11]. For instance, in the detection of thoracic diseases from chest X-rays, the integration of clinical notes with image analysis can provide a richer context, improving the model's ability to correctly identify conditions [13]. Furthermore, the use of attention mechanisms helps in creating more explainable AI systems, which is essential in the medical field where understanding the rationale behind a diagnosis can be as important as the diagnosis itself [14].

The major contributions of our work are summarized as follows.

1. AIM-X introduces a dual cross-attention architecture that fuses visual features from X-ray images with textual data from radiology reports. This is by alternately assigning query roles to both modalities, where the model learns richer, more discriminative representations and generates highly accurate and more patient-specific clinical reports.
2. Our dual cross-attention mechanism integrates multiple cross-attention strategies with multiscale feature extraction, allowing AIM-X to effectively focus on the most relevant image regions and align them effectively with textual context.
3. AIM-X also incorporates genetic algorithm-based optimization to tune its hyperparameters based on the structure of medical reports, improving the quality of report generation.

2. Related Work

The generation of reports from multi-modal medical data, encompassing both imaging and text, has emerged as a promising area for clinical support in diagnostic tasks. Nakaura et al. (2024) carried out a retrospective study at a single center to explore the potential of GPT-series models in generating radiology reports. The study was approved by the institutional review board, with informed consent waived due to its retrospective design. Reports generated by the models were compared to those written by board-certified radiologists, focusing on 28 common imaging findings and their differential diagnoses based on the Radiology Review Manual. The dataset included cases from January 2020 to February 2023. To assess performance, the researchers evaluated mean average precision (mAP) and specifically compared the "image findings," "impression," and "differential diagnosis" sections of the reports. The findings suggest that GPT-based models can approximate the structure and content of radiologist-authored reports with promising accuracy [15].

Li et al. (2023) introduced MedEPT, a parameter-efficient approach for medical report generation that uses image-only datasets, which had previously been underexplored in the field. Their method applied prompt-based efficient tuning (PET) to adapt vision-language pretraining (VLP) models, aiming to address key inefficiencies related to fine-

tuning time and computational demands. To enhance dataset adaptability, they incorporated the MRGPID framework to generate relevant textual descriptions for text-sparse image datasets. MedEPT was systematically evaluated and achieved competitive results, outperforming state-of-the-art methods while using fewer than 10% of trainable parameters and less than 30% of the typical training time [16].

Initial attempts showed approaches that primarily utilized pre-trained models like BERT for generating text and CNNs for extracting image features, achieving limited success in generating structured reports [17]. Nevertheless, these models often treated image and text modalities separately. A Multi-modal Memory Transformer Network that integrates images, textual reports, and medical terminologies to enhance medical report generation. The MMTN encoder aligns and retains the relationships between visual and terminological features, while the multi-modal fusion layer assesses the contributions from both visual and linguistic data. Comprehensive experimental results on three real-world datasets have shown that the MMTN outperforms current leading methods [18].

Raminedi et al. (2024) proposed a multimodal transformer-based framework for automated medical report generation that integrates advanced image feature extractors such as ViT, BEiT, and DEiT—with a language model based on GPT-2. The architecture uses a cross-attention mechanism to align visual features with corresponding textual components in the report, facilitating more accurate and context-aware descriptions of medical conditions. The study demonstrated that this approach outperforms traditional recurrent neural network (RNN) models in generating clinically relevant and structurally coherent medical reports [19].

Unlike prior work that primarily utilizes a single cross-attention stream with either image or text as the query source, our solution introduces a dual cross-attention mechanism that alternately assigns query roles to both visual and textual modalities. This bidirectional querying solution allows each modality to contextualize its features based on the complementary modality, thus yielding a richer and more discriminative shared representation. Our design leads to consistent performance improvements across clinical report generation benchmarks. To our knowledge, this dual-asymmetric configuration remains underexplored in medical report generation literature.

3. AIM-X

Here, we introduce AIM-X, a multimodal model that produces a coherent medical reports. AIM-X combines DenseNet for image feature extraction and GPT for text generation. Enhanced with a novel dual cross-attention mechanisms, AIM-X produces actionable and meaningful diagnostic reports.

Figure 1 demonstrates the architecture of our proposed AIM-X model, which integrates advanced components for joint visual–textual medical report generation. At its core is a DenseNet-121 module (pretrained on ImageNet), which is selected for its dense connectivity and efficient gradient flow, and served as the primary image feature extractor. Chest X-ray images are first preprocessed and aligned with their corresponding textual impressions using a custom dataset class. The extracted visual features are then captured at three hierarchical levels: low-level (128 channels), mid-level (256 channels), and high-level (640 channels), corresponding to textures, structural patterns, and high-level semantics, respectively. These multi-scale features are then projected into a shared embedding space via linear transformation layers to facilitate downstream fusion with textual information, ultimately enabling the generation of clinically relevant reports.

To integrate information across modalities, AIM-X employs a multi-stage fusion strategy. Projected features are concatenated and passed through a fusion layer to create a joint image representation. This is combined with the textual embedding via our proposed dual cross-attention mechanism. Our dual cross-attention mechanism is composed of visual cross-attention (image features guided by text), textual cross-attention (text features guided by image), and shared-representation layer (to combine representations). This configuration enables reciprocal contextualization and enhances the model's ability to generate coherent reports. Here, the attention scores used in the dual cross-attention mechanism are calculated by taking the relative importance of each embedding compared to the others in that transformed space, done by contrasting a query (Q) representation (which is the current focus of the model and represented by the hidden state of the decoder during the generation of a word) with multiple key (K) representations, including representations learned for both image and previously generated textual items mathematically:

$$\text{query} = W_q \cdot \text{input\_features} + b_q, \tag{1}$$

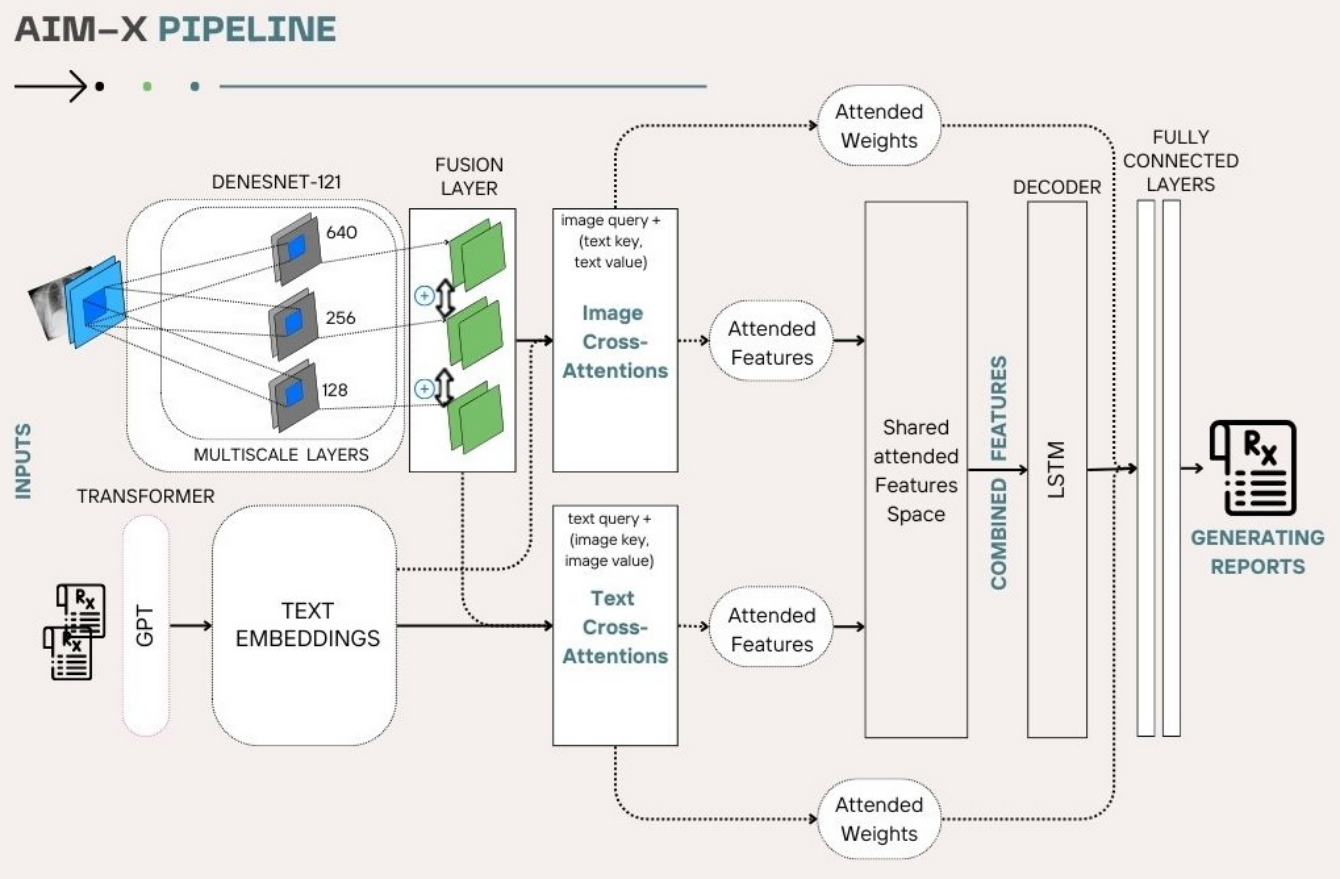


Fig. 1: AIM-X utilizes DenseNet to extract high-fidelity X-ray features and complex patterns important for diagnoses. The visual features are then blended with textual information and processed through GPT to produce rich contextual, coherent medical reports. By fusing visual and textual modalities through an attention mechanism, AIM-X generates accurate and readable diagnostic reports, effectively connecting image features with contextual information for enhanced medical diagnostics.

where W_q is the weight matrix of the linear layer. b_q is the biase and input_features are the original features (e.g., image or text embeddings).

Comparison is performed through a function of similarity (e.g., a dot-product function) between each key with regard to a query mathematically:

$$\text{dot}(q_i, k_j) = q_i \cdot k_j = \sum_{d=1}^{D} q_{i,d} \cdot k_{j,d}, \tag{2}$$

where D is the dimensionality of the embeddings (image_embeddings_dim). The dot product between a query q_i (a single row in Q) and a key k_j (a single row in K). Repeating for all pairs of queries and keys, resulting in a matrix of attention scores (shape: [batch_size, image_seq_len, text_seq_len]). To avoid the danger of dot products becoming too large and potentially causing the gradients of training to become unbalanced, the scores are normalized through division by the square root of the embedding dimension:

$$\text{scaled\_scores} = \frac{Q \cdot K^T}{\sqrt{D}} \tag{3}$$

In this way, our dual cross-attention configuration enhances multimodal understanding by enabling each modality to contextualize its features based on the other. When visual features are used as queries and textual features as keys and values, each image region effectively ”asks” which semantic concepts from the text are most relevant to it, helping visual representations become more semantically informed. Likewise, when textual features are used as queries and visual features as keys and values, each word in the clinical report ”looks for” the visual regions that support its words generation, ensuring that the text is grounded in visual evidence. This bidirectional attention promotes richer and complementary feature alignment and improves the accuracy and interpretability of the generated clinical reports.

Then the fused multimodal representation is processed by a Long Short-Term Memory (LSTM) decoder, which generates a sequence of hidden states. These are mapped to the vocabulary space via two fully connected layers, with dropout applied to mitigate overfitting. The overall output consists of the generated report text along with attention weight matrices, which can be used for interpretability or further analysis. More precisely, the transformed embeddings are passed as input to the decoder RNN layer (e.g., LSTM network) to generate a sequence of words that forms the report. At each time step t, the LSTM processes the input x_t (transformed embedding), along with the hidden state h_{t-1} and memory (cell state) m_{t-1} from the previous time step, and produces a new hidden state h_t as follows:

$$h_t = \text{LSTM}\left(x_t, h_{t-1}, m_{t-1}\right), \tag{4}$$

where h_t is the hidden state at time step t, x_t is the input at time step t, and h_{t-1}, m_{t-1} represent the hidden state and memory from the previous time step [20]. The LSTM generates the report sequentially, using updated attention weights at each time step.

After generating the report, the predicted sequence of words is compared to the ground truth using a loss function (e.g., Cross Entropy Loss), where it is computed as follows:

$$\text{Loss} = -\sum_{t=1}^{T}\sum_{c=1}^{C} y_{t,c} \cdot \log(p_{t,c}), \tag{5}$$

where $y_{t,c}$ represents the ground truth probability distribution (or one-hot encoding) for the target word or token at time step t, and $p_{t,c}$ is the predicted probability for that word or token at time step t. The loss provides feedback for the attention mechanism, as the model learns to focus on more relevant parts of the data during each step of training.

The gradients of the loss with respect to the model parameters (including the attention weights and LSTM weights) are calculated and used to update the model parameters via gradient descent. The update rule for the attention weights can be defined as follows:

$$W \leftarrow W - \eta \cdot \frac{\partial \text{Loss}}{\partial W}, \tag{6}$$

where η is the learning rate, and $\frac{\partial \text{Loss}}{\partial W}$ is the gradient of the loss with respect to the attention weights. This iterative process ensures that attention weights are progressively refined at each step of training, improving the model's ability to focus on the most relevant features. Additionally, the `clip_grad_norm_` function is applied to prevent exploding gradients, ensuring stable learning during the update process.

3.1. Hyperparameter settings

In this study, we used a Genetic Algorithm (GA) to fine-tune hyperparameters and enhance model performance. Genetic algorithms utilize principles from natural selection and evolution, where a set of candidate solutions progresses through generations to increase their fitness. This evolutionary approach is particularly effective in navigating

complex hyperparameter spaces where traditional optimization methods may struggle. Here, GA was used to modify essential hyperparameters including learning rate, batch size, dropout rate, and number of layers

4. Results

4.1. Dataset

The data we used in this work was supplied by the Indiana University Hospital Network and is divided into two parts [21]:

1. **X-rays**: Chest X-ray images with two types of view, including frontal and lateral views. The dataset includes a total of 7,472 x-ray images. These images are representative of chest radiographs from different individuals.
2. **Reports**: Different reports, in XML, for patients, where each XML document represents one patient and includes clinical information associated with that patient. The XML tag <parent Images id=”image-id”> is used for the specific images that relate to the report. The id attribute within it contains the name of the image, which aligns with the PNG images. A report or one XML document may pertain to a series of images.

The XML files consist of four key features:

- **Comparison**: This section provides information about the methods for serial follow-up.
- **Indication**: Put here patient-relevant clinical information, typically when the radiologist is writing the report.
- **Findings**: This attribute represents information related to the x-ray image, whether there are abnormal findings in specific regions like the right or left lung.
- **Impression**: This can be inferred from the indicators and findings; this represents the results of the medical report, whether the lungs are clear or if there are pulmonary conditions.

4.1.1. Text Preprocessing

Feature extraction from the XML files and preparation for downstream analysis had been done for AIM-X as follows: see Table 1. Data points without associated images or impressions were excluded to ensure integrity and reliability in the dataset. After that, the textual data went through several strict steps of preprocessing in order to get rid of the extraneous elements that obscure clarity, consistency, and suitability for subsequent analysis:

- Extra spaces were eliminated.
- Words like “XX” and “XXX” were removed.
- Stopwords were removed, and numbers were converted into words (e.g., 8 to “eight,” & to “and”).
- All special characters, parentheses, and brackets were removed.
- All text was converted to lowercase, and contractions like “isn’t” and “doesn’t” were expanded.
- Words like “dt,” “rt,” were also removed.

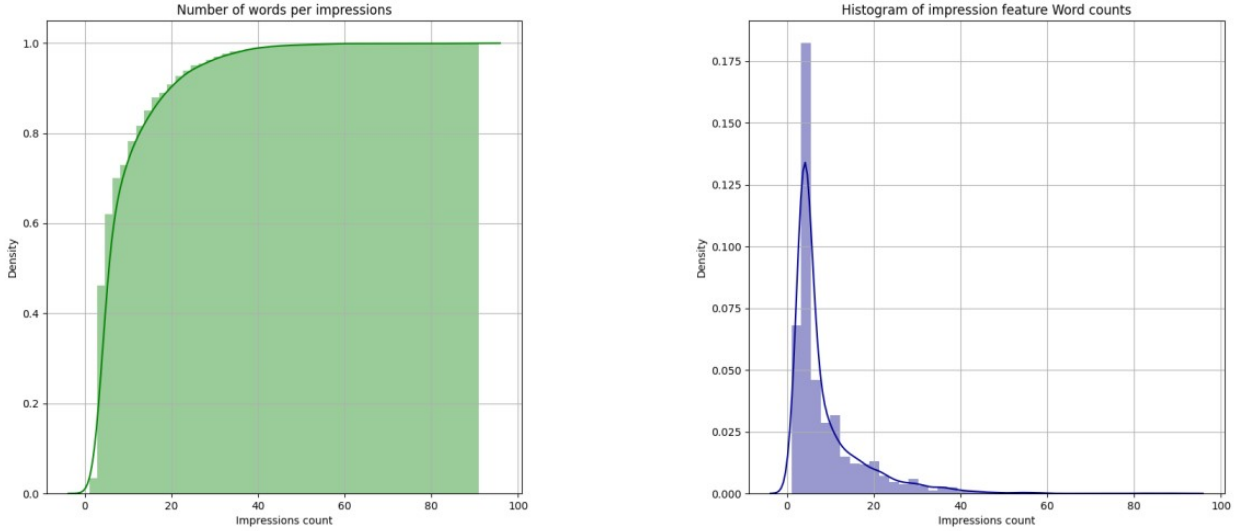
The impression feature is one of the key features of the dataset as it summarizes, in important ways, the data that are often employed to give diagnostic insight to medical data. The CDF plot (Figure 2a) shows how the cumulative probability of the values of Impression change across the dataset, thereby providing insight into the general distribution and concentration of the values. (Figure 2b) provides further detail about the distribution of the Impression values by emphasizing the density and frequency of such values across intervals. This in-depth analysis will enable insight into how often certain Impression values occur and how they cluster in the whole dataset, an integral component in the conveyance of consistency and variation within medical impressions [22].

4.1.2. Image Preprocessing

X-ray images were resized to a standard shape (224x224 pixels) and pixel values were also normalized between 0 and 1 to facilitate faster convergence during training. The approaches taken to augment the data were done so in

Table 1: This is an example of the final dataset that we used for splitting and incorporating into the framework.

Images	Text
CXR1_1_IM-0001-3001.jpg; CXR1_1_IM-0001-4001.jpg	Normal chest
CXR1000_IM-0003-1001.jpg; CXR1000_IM-0003-3001.jpg; CXR1000_IM-0003-2001.jpg	One increased opacity right upper lobe associated atelectasis may represent focal consolidation mass lesion atelectasis recommend chest CT evaluation. Two opacity overlying left fifth rib may represent focal airspace disease
CXR1008_IM-0009-2001.jpg; CXR1008_IM-0009-4004.jpg	No acute cardiopulmonary abnormality
CXR1015_IM-0001-2001.jpg; CXR1015_IM-0001-1001.jpg; CXR1015_IM-0013-1001.jpg; CXR1015_IM-0013-2001.jpg	Bibasilar opacities right greater than left, features suggest combination consolidation and atelectasis



(a) (b)

Fig. 2: The Impression feature's distribution is visualized through (a) the Cumulative Distribution Function (CDF) and (b) the Probability Density Function (PDF). The CDF plot shows the cumulative probability of the Impression feature values, while the PDF illustrates the density and frequency of these values across the dataset. Together, these plots provide a comprehensive overview of the distribution characteristics of the Impression feature.

order to enhance the generalizability of the model by increasing diversity in the dataset, to encourage it to learn robust invariant features, rather than to memorise the training dataset.

4.2. Performance Evaluation

We adopted a few measures to evaluate the model's performance, based on Bilingual Evaluation Understudy Score (BLUE), and Recall-Oriented Understudy for Gisting Evaluation (ROUGE). They were utilised to evaluate the generated report/text against the ground truth reports. In this study, we concentrated on the "impression" section of the medical reports, as it represents the most critical component for conveying diagnostic insights. Consequently, all evaluation measures were applied exclusively to the "impression" portion of each data point to ensure a focused and

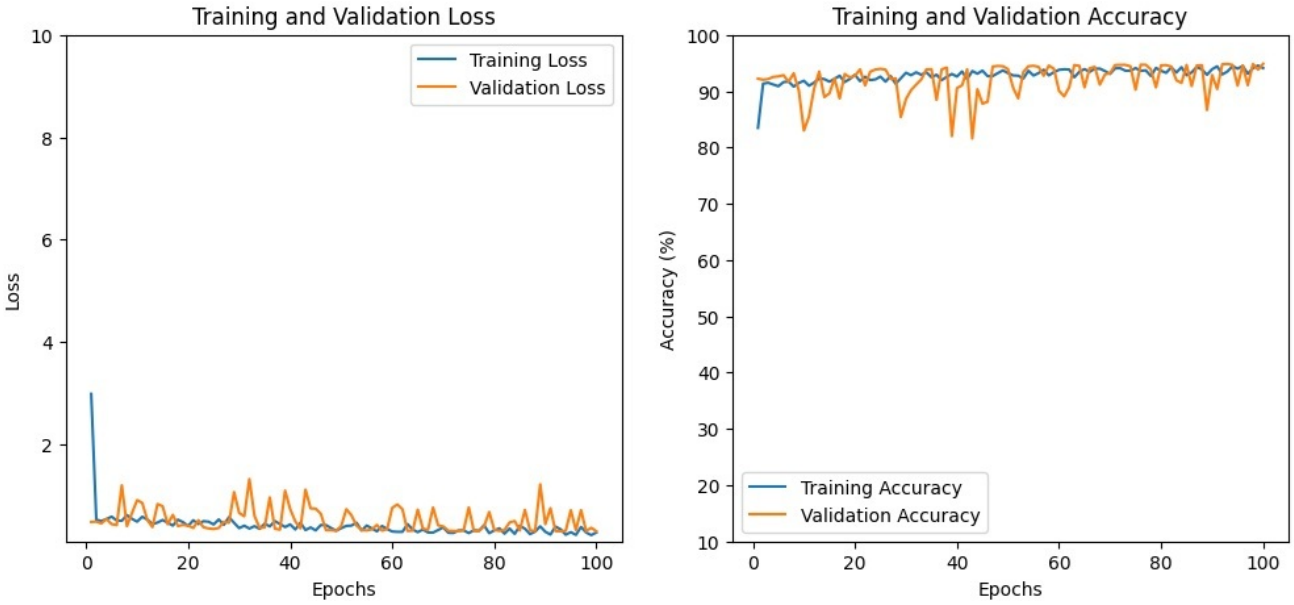


Fig. 3: **Training and Validation Curves of the AIM-X Model**. The plot demonstrates the model's high performance, with validation accuracy converging between 92-94.5%, indicating robust generalization. Simultaneously, the loss function (e.g., CrossEntropyLoss) steadily decreases, confirming effective optimization and minimized prediction errors. These results highlight the model's stability and suitability.

precise assessment of the model's performance in generating contextually relevant and clinically accurate outputs. AIM-X model (Fig 3) demonstrates the high performance of our medical report generation, achieving high accuracy (e.g., 94.5%) on validation data while maintaining low loss values, indicating robust learning and generalization. The training curves show stable convergence, with no signs of overfitting, suggesting effective regularization.

5. Comparison with related methods

We perform a comparison with several models that were previously proposed for the generation of medical reports and the captioning of images. These include the state-of-the-art methodologies including a benchmark method for image captioning, More precisely, we compared our model with attention-based mechanisms such as attend and tell (AdaTT) [20]. TieNet [23] was also used in the comparison. Finally, a reinforcement learning-based model, called the CCR model, was included, which is optimized for accuracy and inherently boosts precision [24]. All these models were applied to the dataset using the same set of evaluation metrics for consistency. The resulting scores for each model across the respective metrics are reported in Table 2, providing a comprehensive assessment of their performance.

Moreover, as an ablation study, we showed the performance of our baseline models, following a conventional design:

- The first baseline model (v1) uses the image modality's query to interact with the text modality.
- The second baseline model (v2) employs the text modality's query to interact with the visual modality.

We included these baseline models in the comparison to mainly demonstrate the robustness of our dual cross-attention mechanism. As shown in the evaluation results Table 2, these comparisons highlight the superior performance of AIM-X model over related and baselines models.

AIM-X outperforms related models in chest X-ray report generation by effectively integrating multi-modal attention mechanisms that combine image features with language model embeddings. Unlike TieNet, which produces more generic reports; Show and Tell, which generates descriptive but non-clinical outputs; Visual Sentinel, which employs adaptive yet non-medical attention; and CX-ReportGen, which relies on template-driven reports, AIM-X dynamically assesses the relevance of both visual and textual information. This enables it to produce more contextualized, patient-

Table 2: Performance of various models in terms of NLG metrics. BL-n represents the BLEU score using up to 4-grams, while RG-L stands for ROUGE-L. AIM-X shows the improvement across all NLG metrics relative to the Baseline versions (1,2).

Model	**NLG Metrics**				
	BL_1	BL_2	BL_3	BL_4	RG_L
AdaTT [20]	0.220	0.127	0.089	0.069	0.308
TieNet [23]	0.330	0.194	0.124	0.081	0.311
CCR [24]	0.162	0.084	0.055	0.036	0.244
Baseline V1	0.083	0.035	0.026	0.023	0.082
Baseline V2	0.207	0.105	0.069	0.058	0.199
AIM-X	**0.727**	**0.6595**	**0.601**	**0.469**	**0.727**

specific, and clinically meaningful reports. The model's multi-scale attention refinement further helps capture critical findings, resulting in a more accurate and flexible reporting framework.

6. Conclusions

This paper proposes a novel architecture, AIM-X, designed for generating diagnostic reports from multimodal medical data. AIM-X focuses on the integration of X-ray images with their corresponding textual reports, leveraging an enhanced attention mechanism as its core innovation. This mechanism enables the model to selectively prioritize critical features in both the visual and textual data, ensuring that the generated diagnostic reports are both contextually relevant and clinically accurate. By effectively aligning visual and textual data, the attention mechanism enhances the coherence and relevance of AIM-X's outputs.

AIM-X framework uses a CNN-based architecture for the extraction of robust image features, through DenseNet, and a Transformer-based model for text processing. The integration of these features, via a dual-cross attention mechanisms, enables AIM-X to achieve highly accurate and rich-in-context reports. This fusion not only improves diagnostic accuracy but also enhances the interpretability of AI-driven decisions, making AIM-X a valuable tool for medical diagnostics.

In a nutshell, the attention mechanism is the central contribution of AIM-X, which enables the effective fusion of multimodal data to improve both the accuracy and interpretability of medical report generation. Future work will focus on refining this attention mechanism, extending its applicability to other medical imaging modalities, and incorporating feedback from clinical professionals to further enhance its real-world impact.

References

[1] Xu, X., Li, J., Zhu, Z., Zhao, L., Wang, H., Song, C., Chen, Y., Zhao, Q., Yang, J. & Pei, Y. A Comprehensive Review on Synergy of Multi-Modal Data and AI Technologies in Medical Diagnosis. *Bioengineering*. **11**, 219 (2024)

[2] Lipkova, J., Chen, R., Chen, B., Lu, M., Barbieri, M., Shao, D., Vaidya, A., Chen, C., Zhuang, L., Williamson, D. & Others Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*. **40**, 1095-1110 (2022)

[3] Li, C., Yang, Y., Liang, H. & Wu, B. Transfer learning for establishment of recognition of COVID-19 on CT imaging using small-sized training datasets. *Knowledge-Based Systems*. **218** pp. 106849 (2021)

[4] Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Densely connected convolutional networks. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 4700-4708 (2017)

[5] Ashraf, R., Habib, M., Akram, M., Latif, M., Malik, M., Awais, M., Dar, S., Mahmood, T., Yasir, M. & Abbas, Z. Deep Convolution Neural Network for Big Data Medical Image Classification. *IEEE Access*. **8** pp. 105659-105670 (2020)

[6] Yang, X., Ye, Q., Cai, G., Wang, Y. & Cai, G. PD-ResNet for classification of Parkinson's disease from gait. *IEEE Journal Of Translational Engineering In Health And Medicine*. **10** pp. 1-11 (2022)

[7] Yenduri, G., Ramalingam, M., Selvi, G., Supriya, Y., Srivastava, G., Maddikunta, P., Raj, G., Jhaveri, R., Prabadevi, B., Wang, W. & Others Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*. (2024)

[8] Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *Npj Digital Medicine*. **4**, 86 (2021,5,20).

[9] Tunstall, L., Von Werra, L. & Wolf, T. Natural language processing with transformers. (" O'Reilly Media, Inc.",2022)
[10] Serrano, S. & Smith, N. Is attention interpretable?. *ArXiv Preprint ArXiv:1906.03731*. (2019)
[11] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. & Summers, R. Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *IEEE CVPR*. **7** pp. 46 (2017)
[12] Tutek, M. & Šnajder, J. Toward practical usage of the attention mechanism as a tool for interpretability. *IEEE Access*. **10** pp. 47011-47030 (2022)
[13] Li, M., Jiang, Y., Zhang, Y. & Zhu, H. Medical image analysis using deep learning algorithms. *Front. Public Health*. **11** pp. 1273253 (2023,11)
[14] Wiegreffe, S. & Pinter, Y. Attention is not not explanation. *ArXiv Preprint ArXiv:1908.04626*. (2019)
[15] Nakaura, T., Yoshida, N., Kobayashi, N., Shiraishi, K., Nagayama, Y., Uetani, H., Kidoh, M., Hokamura, M., Funama, Y. & Hirai, T. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. *Japanese Journal Of Radiology*. **42**, 190-200 (2024)
[16] Li, Q. Harnessing the power of pre-trained vision-language models for efficient medical report generation. *Proceedings Of The 32nd ACM International Conference On Information And Knowledge Management*. pp. 1308-1317 (2023)
[17] Jing, B., Xie, P. & Xing, E. On the automatic generation of medical imaging reports. *ArXiv Preprint ArXiv:1711.08195*. (2017)
[18] Cao, Y., Cui, L., Zhang, L., Yu, F., Li, Z. & Xu, Y. MMTN: multi-modal memory transformer network for image-report consistent medical report generation. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **37**, 277-285 (2023)
[19] Raminedi, S., Shridevi, S. & Won, D. Multi-modal transformer architecture for medical image analysis and automated report generation. *Scientific Reports*. **14**, 19281 (2024)
[20] Lu, J., Xiong, C., Parikh, D. & Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 375-383 (2017)
[21] Indiana University Hospital Network,*Chest X-Rays with Radiology Reports* [Dataset],Version 1.0, National Library of Medicine Open-i,https://openi.nlm.nih.gov/faq.php,CC BY-NC-ND 4.0.
[22] Tuerlinckx, F. The efficient computation of the cumulative distribution and probability density functions in the diffusion model. *Behavior Research Methods, Instruments, & Computers*. **36**, 702-716 (2004)467d
[23] Wang, X., Peng, Y., Lu, L., Lu, Z. & Summers, R. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 9049-9058 (2018)
[24] Liu, G., Hsu, T., McDermott, M., Boag, W., Weng, W., Szolovits, P. & Ghassemi, M. Clinically accurate chest x-ray report generation. *Machine Learning For Healthcare Conference*. pp. 249-269 (2019)

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 June 2024, accepted 2 July 2024, date of publication 8 July 2024, date of current version 16 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3424526

RESEARCH ARTICLE

GraphX-Net: A Graph Neural Network-Based Shapley Values for Predicting Breast Cancer Occurrence

ABDULLAH BASAAD[1], SHADI BASURRA[1], EDLIRA VAKAJ[1], MOHAMMED ALESKANDARANY[2], AND MOHAMMED M. ABDELSAMEA[3]
[1]School of Computing and Digital Technology, Birmingham City University, B4 7XG Birmingham, U.K.
[2]School of Human Sciences, University of Derby, DE22 3AW Derby, U.K.
[3]Department of Computer Science, University of Exeter, EX4 4QF Exeter, U.K.

Corresponding author: Mohammed M. Abdelsamea (m.abdelsamea@exeter.ac.uk)

This work was supported by Antser Ltd., U.K.

ABSTRACT Breast cancer is a major health problem worldwide, and an accurate prediction of its recurrence is crucial to early detection of recurrence and personalized treatment. In recent years, various AI techniques have been applied to predict cancer recurrence with increasingly high accuracy. Graph Neural Networks (GNNs) have emerged as powerful tools for analyzing structured data, including knowledge networks. In this study, we explore the application of GNN-based node classification to predict the occurrence of relapse status in breast cancer patients. We propose a noval model, namely GraphX-Net, Shapley Value-based Graph Neural Network. GraphX-Net leverages a graph representation by analyzing public breast cancer gene expression data, where nodes represent patients and edges capture relationships between them based on various factors such as tumor cellularity, histological subtype, hormone therapy, patient vital status, primary tumor laterality, type of breast surgery, and other clinicopathological parameters. Our approach integrates diverse data sources and utilizes graph convolutional layers to generate informative node embeddings. The model uniquely calculates node feature contributions using Shapley values, sets nodes' thresholds, and considers the total node neighboring effects. These innovations enable GraphX-Net to achieve state-of-the-art performance in predicting breast cancer recurrence, highlighting its potential as a powerful tool for improving patient outcomes.

INDEX TERMS Graph neural networks, XAI, node classification, breast cancer, occurrence prediction.

I. INTRODUCTION

The use of node classification graphs to predict the occurrence of relapse status in breast cancer has emerged as a promising avenue in the field of cancer research. Relapse or recurrence of breast cancer after treatment continues to be a major concern for both patients and healthcare providers. It has long been reported that morbidity and mortality observed in patients with breast cancer primarily result from disease progression through local-regional or systemic dissemination, as well as the occurrence of recurrences post-treatment, or a combination of these factors [1]. Upon development of metastatic deposits outside the primary location, the prognosis of patients is dramatically worsened, and cure would be unlikely [2]. The ability to accurately predict the probability of relapse can greatly help in treatment planning, monitoring, and ultimately improving patient outcomes [3], [4]. Researchers have achieved notable advancements in the development of predictive models of relapse status in breast cancer through the effective use of node classification graphs [5]. One particularly interesting research study was explored the predictive potential of K-Banhatti and Zagreb type degree-based topological indices in quantitative structure–property relationship (QSPR) analysis for medications used to treat type-I and type-II diabetes. These indices were computed for 14 anti-diabetes drug

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

molecules using edge and vertex partitioning techniques. Using these topological indices, the researchers developed QSPR regression models to predict the physicochemical properties of the drugs under study [6]. Another interesting study discussed the use of topological indices (TIs) to predict the physical and biological properties of drugs used to treat Alzheimer's disease. Degree-based TIs were generated using edge partitioning for drugs such as Tacrine, Donepezil, and Rivastigmine. A QSPR model was developed using linear regression to predict characteristics such as boiling point, flash point, molar volume, molecular weight, complexity, and polarizability. The findings suggest that TI can be valuable tools in drug discovery and design for the treatment of Alzheimer's disease [7].

This study focused on a node classification graph representing the interconnectivity between nodes, where each node corresponds to an individual patient. This representation encapsulates the intricate relationships and interdependencies among various factors that contribute to relapse, encompassing clinical variables, histopathological features, and treatment regimens. Through the utilisation of machine learning algorithms on these graphs, researchers can discern patterns and features that possess predictive value for relapse occurrence. A notable advantage of employing node classification graphs lies in their capacity to accommodate the heterogeneous nature of breast cancer data [8]. Breast cancer is a multifaceted ailment influenced by diverse variables that interact with one another [9]. By representing these variables as features of the nodes and their relationships as edges, node classification graphs offer a comprehensive framework for modelling and analysing these intricate connections.

In recent years, several studies have been dedicated to improving the precision of predicting relapse recurrence in breast cancer through the use of various techniques. One such approach involves the integration of Multi-Omics Data, which encompasses the incorporation of multiple types of omics data, including genomics, transcriptomics, proteomics, and epigenomics [10]. By amalgamating information from various molecular levels, researchers endeavour to identify biomarkers and molecular signatures that are associated with relapse. This comprehensive integration facilitates the development of models that predict the occurrence of relapses. Another avenue for providing valuable information on tumour characteristics and improving relapse prediction is the incorporation of imaging and radiationomics. Using medical imaging techniques such as mammography, magnetic resonance imaging, or PET-CT scans, researchers can gain pertinent information on tumour properties [11]. Radiomic models, which involve the extraction of a multitude of features from the imaging data, enable the capture of subtle patterns and textures that are indicative of relapse. These extracted features are subsequently employed to train machine learning models, thereby enhancing the accuracy of relapse prediction. Furthermore, the inclusion of clinical and pathological information [12] and the analysis of longitudinal data are instrumental in elucidating the dynamic changes associated with relapse [13]. These additional factors contribute valuable information to advance our overall understanding of relapse recurrence in breast cancer.

This study aims to tackle the challenges associated with understanding the relational nature of patient data and determining the contribution of inputs within a neural network framework. To address these challenges, we propose a novel approach based on Graph Neural Networks (GNNs), we called GraphX-Net, that harnesses Shapley (SHapley Additive exPlanations) values to evaluate the significance of features and generate an explainable graph representation. This is achieved by employing Explainable Artificial Intelligence (XAI) techniques to enhance transparency and interpretability in predicting relapse. Specifically, we used the XGBClassifier algorithm, known for its interpretability, to train a machine learning model. To gain insight at both local and global levels, we incorporate Shapley values as a means of quantifying the contribution of each feature to relapse prediction for individual patients, as well as the entire dataset. To classify unlabelled nodes and establish edges between nodes, we adopt a weighted approach that considers the contributions of features to measure probabilities. Furthermore, we initialised two graph convolutional layers as message passing layers to aggregate, transform, and update node representations, incorporating the information learned from the underlying graph structure.

Here are the main contributions of GraphX-Net:

- The GraphX-Net utilizes GNNs, which are well-suited for capturing complex relationships and dependencies within graph-structured data. In the context of breast cancer, representing patients and their relevant features as nodes in a graph, with the edges representing the relationships between patients based on similarity.
- GraphX-Net can effectively capture local and global information about patients and their features, offering more accurate predictions compared to traditional methods.
- GraphX-Net incorporates Shapley values to provide the interpretability and explainability of the predictions. Using Shapley values, it can identify the importance of different features in the prediction of relapse, allowing clinicians and researchers to understand the underlying factors driving the predictions.
- GraphX-Net provides a powerful tool for clinicians and researchers in the field of breast cancer by combining GNNs with Shapley values. It offers accurate predictions of relapse risk while also providing interpretable insights into the underlying factors driving those predictions.

II. RELATED WORK

Breast cancer constitutes a multifaceted disease influenced by a number of factors influencing its course and likelihood of recurrence. Traditional machine learning methodologies frequently disregard the inherent interdependencies and

associations between individual patients, thus restricting their predictive efficacy [14]. On the contrary, graph neural networks (GNN) present a compelling avenue to harness the intricate network of relationships between patients and their associated attributes, thus improving the precision of the prediction of breast cancer relapse [15]. In this sectionb13, we explore the application of GNNs in predicting breast cancer relapse, highlighting their potential to improve prognostic accuracy and help in personalized treatment strategies.

Recent studies have explored deep learning's ability to automatically learn intricate data representations, proving its effectiveness in extracting essential features for classification tasks. For example, proposed six network models for abnormal breast detection in mammograms, with the aim of improving previously developed AI methods. Among these models, a Net-5 model achieved the best results and outperformed 15 methods, including biogeography-based optimization (BBO), wavelet energy entropy (WEE), cross-validation (CV), k-nearest neighbor (kNN) algorithm and others [16]. The combination of BDR-CNN and GCN in the Net-5 model, which extracted image level and relation awareness features, respectively, contributed to its superior performance [17]. An automated diagnosis process consisting of three steps was developed to classify cancerous patches within entire slides to predict slide-level cancer metastases. A deep Neural Network (DNN) with a DBSCAN clustering algorithm was used in the second stage to reveal morphological features of cancerous regions [18].

The graph convolutional neural network (GCNN) was used to showcase 20 gene signatures to predict the likelihood of relapse in patients with breast cancer (BRCA). The prognostic and diagnostic capabilities of these genes were rigorously tested against other established algorithms and biomarkers, establishing the superior performance of the GCNN genes [19].

Some machine learning (ML) algorithms were constructed and assessed, including Random Forest, Boosting, and Stacking. These models were built to compare their performance with the Graph Neural Network model. Prior to evaluating the models, the data was split and SMOTE (Synthetic Minority Oversampling Technique) was utilized to address the class imbalance. Additionally, a feature selection technique, such as using a forest-based method [20], was applied to identify important features. These steps were carried out before proceeding with the evaluation of each model to assess its performance.

Random Forest is a popular ensemble learning method that is used for both classification and regression tasks in machine learning. It is a combination of multiple decision trees, where each tree is trained on a random subset of training data and features. The final prediction is determined by aggregating the predictions of all individual trees [21]. Boosting is an ensemble learning technique in machine learning where multiple weak learners (often decision trees) are combined to create a strong learner. Unlike random forests, which train multiple trees independently, boosting trains weak learners sequentially, with each subsequent learner focusing on the mistakes made by the previous ones. The key idea behind boosting is to iteratively improve the overall model by giving more weight to the misclassified or difficult examples [22]. Stacking, also known as stacked generalization, is a machine learning ensemble technique that combines multiple models to improve predictive performance. It involves training multiple base models on a dataset and then training a metamodel that takes the predictions of the base models as input to make the final prediction [23]

Having discussed related work, here we introduce a novel Graph Neural Network model, called GraphX-Net, that leverages Shapley values to construct an efficient neural graph tailored to the data. This neural graph is designed to fit entirely in memory and represents the graph data using graph information components, including node features and edges. In the context of relapse breast cancer data, where edge weights are unavailable, we utilise these values effectively to connect patient nodes and create distinct clusters based on their Shapley value contributions. Given the subtle differences in contributions, we introduce two additional parameters, the node threshold and node connections, to manage the relationships among nodes. The integration of Graph Neural Networks and Shapley values enhances the accuracy and interpretability of relapse prediction for breast cancer patients, addressing the critical need for accurate and explainable models in the field of breast cancer prognosis. GraphX-Net not only achieves high prediction accuracy but also offers valuable insights into the specific features and interactions within the patient's treatment graph that contribute to the relapse risk. This unique contribution enables clinicians and researchers to understand the underlying mechanisms that drive relapse in breast cancer and to make informed decisions regarding treatment and follow-up strategies. Our extensive experiments and comparative analysis underscore the substantial progress achieved by the GraphX-Net model integrated with Shapley values in the realm of breast cancer prognosis. These findings pave the way for better patient care and personalized medicine.

III. GRAPHX-NET

The GraphX-Net model incorporates a fusion of machine learning algorithms and deep learning algorithms. We first transformed it into a graph representation, where each patient is represented as a node within the graph, while the edges symbolise the interconnections between patients. The determination of node edges is based on the assessment of input contributions. By analysing the positive or negative contributions of each input, the corresponding node labels can be identified.

The utilisation of Feature Shapley Values Fig. 1 enables quantification of each feature's contribution to the predictions made by GraphX-Net model. This computation involves iteratively training the model, while systematically excluding

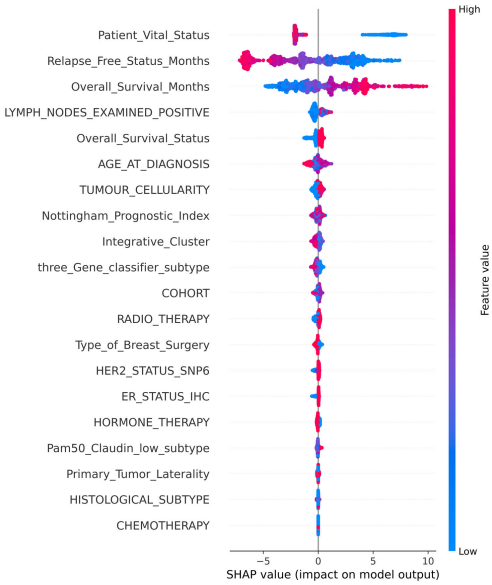


FIGURE 1. The important features according to their Shapley values contributions that play a critical role in understanding and interpreting the impact of features on a machine learning model's output.

distinct features during each iteration. Through a comparison of prediction differences between the complete model and models with excluded features, Shapley values were attributed to individual features. These values accurately depict the incremental importance of each feature in the prediction process.

In SHAP (SHapley Additive exPlanations) summary plots, colors convey important information about the value of the features and their impact on the model's predictions. Specifically, the colors in a SHAP summary plot are used to indicate the magnitude and direction of the feature values. Here's a detailed explanation of what the colors represent [24]:

- **Feature Values**:
  - **Red/Hot Colors**: Indicate high values of the feature.
  - **Blue/Cool Colors**: Indicate low values of the feature.
- **Impact on Model Prediction**:
  - The position of the points along the x-axis represents the SHAP value. Positive SHAP values (to the right) increase the predicted outcome, while negative SHAP values (to the left) decrease it.
- **Combination of Color and Position**:
  - By combining color and position, we can infer how different values of a feature affect the prediction.

The SHAP summary plot plays a crucial role in discerning the most significant features and their respective range of effects throughout the dataset. By examining the Feature Shapley Values within the context of the GraphX-Net model, valuable insights and interpretations could be obtained. It aids in the identification of influential features and facilitates an

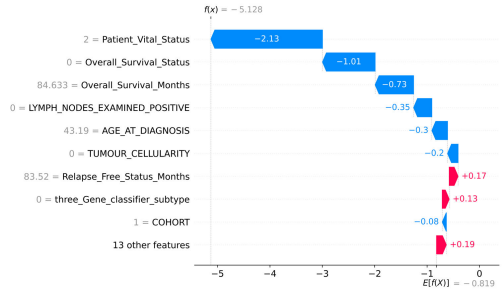


FIGURE 2. Waterfall SHAP values illustrate how SHAP values impact individual predictions in GraphX-Net model.

understanding of how these features influence the predictions generated by the GNN. Through the visualisation and analysis of the Feature Shapley Values Fig. 2, a deeper comprehension of the underlying relationships and dependencies within the graph can be attained. Furthermore, this process helped identify potential biases or limitations present within the model [24].

- f(x): is the prediction after considering all features.
- E[f(x)]: is the mean prediction.
- The blue bar shows how much a particular feature decreases the value of the prediction.
- The red bar shows how much a particular feature increases the value of the prediction.

While GNN architecture diagram in itself does not directly depict the Feature Shapley Values, incorporating these values into the analysis and interpretation of the GraphX-Net model greatly enhanced comprehension. It provides valuable insights into the relative importance of different features within the prediction process.

1) GRAPH CONSTRUCTION

The beginning of the graph construction process involved the application of a graph construction algorithm or technique, specially designed to establish edges between nodes, following predefined criteria. This key task is of paramount importance, as it forms the basis for the creation of pertinent connections between patients within the network. In this endeavor, we utilized the collective contributions of all input attributes encapsulated within each node to discern the intricate relationships that govern the adjacency's between the focal node and its counterparts in the graph.

$$f(N_c) = \sum_{i=0}^{n} (l_1 sv, l_2 sv, l_3 sv, \ldots.., l_n sv) \tag{1}$$

This equation calculates the Shapley value scores for specific features associated with each node i from 0 to n. Here, $l_i sv$ represents the Shapley value for feature i, and the sum aggregates these values to quantify the contribution of all relevant features across the nodes.

Through the computation of contributions for all inputs within each observation, accounting for both positive and negative contributions, we can determine the probabilities

assigned to each node's classification as either 0 (indicating "No") or 1 (indicating "Yes"). Moreover, by considering the average contribution of adjacent nodes, we can assess their respective adjacencies.

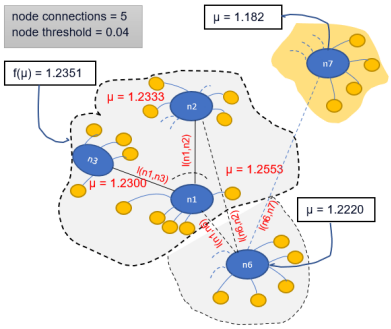


FIGURE 3. This diagram depicts the process of connecting a node by considering the information from its neighboring nodes. The figure highlights that the information of clusters is influenced by both the node threshold value and the requirement that the number of nodes in each cluster not exceed the node connections parameter.

By performing calculations on the contributions of all inputs for each observation, encompassing both positive and negative contributions, we derived the probabilities associated with the classification of each node as 0 (representing "No") or 1 (representing "Yes"). Additionally, we evaluated the adjacencies of neighboring nodes based on their respective contributions. For example, it suggests that the node should be connected to a group where the average contribution is either less than or equal to a predefined threshold. This threshold serves as an integer parameter that is incorporated into the grouping of nodes. To determine the connectivity of a new node to a group, the difference between the average contributions of the group and the contribution of the new node was assessed against the threshold. If this difference was less than or equal to the threshold, and the total number of nodes in the group remains within the permitted group size, the new node should be connected to the group as depicted in Fig. 3. The figure uses different colors to indicate clusters of nodes. Nodes within the gray area all belong to a single cluster, where the mean contributions meet the node threshold value. In contrast, the node in the yellow area, although connected to the gray cluster, belongs to a different cluster. This is because its mean contribution does not align with the overall mean of the gray cluster.

2) FEATURE EXTRACTION

Here we used XGBoost method as the basis for calculating the relative contribution of each feature to the model. This is based on the gain metric to evaluate the importance of a feature by considering its contribution between individual trees within the model [25]. A higher gain value, relative to other features, signifies greater significance in generating predictions. Consequently, the gain metric served as a pivotal attribute for interpreting the relative importance of each feature. By scrutinizing the gain values, we can effectively pinpoint the most influential features within the model as shown in Table 2.

Algorithm 1 Calculate the Total Contributions of Each Node

Input: lst_shap_values ▷ List of all Shapley values
Input: df ▷ Data frame
Output: df ▷ Data frame updated

```
for each i, row ∈ df do
    pos_cont ⇐ 0
    neg_cont ⇐ 0
    for each c ∈ df.cols do
        f_cont ⇐ lst_shap_values[c.values[i]
        if f_cont > 0 then
            pos_cont ⇐ f_cont + pos_cont
        end if
        if f_cont < 0 then
            neg_cont ⇐ f_cont + neg_cont
        end if
    end for
    df.at[i, 'pos_cont'] ⇐ pos_cont
    df.at[i, 'neg_cont'] ⇐ neg_cont
end for
return df
```

Algorithm 2 Create Graph's Adjacency Matrix

Input: df ▷ Data frame
Output: adjacency Matrix

```
for each row in df do
    neg_cont ⇐ row.neg_cont
    pos_cont ⇐ row.pos_cont
    classi ⇐ row.relapse_free_status
    patient_id ⇐ row.patient_id
    node_ad ⇐ get_node_adjacents(neg_cont,
        pos_cont, classi, patient_id)
    adjacency matrix ⇐ node_ad
end for
return adjacency Matrix
```

Patient attributes and clinical data are of significant importance in the prediction of relapse in breast cancer cases. Various features, including patient demographics, medical history, genetic information, and treatment records, are carefully extracted and associated with each patient node. These features were obtained using the gain metric, which allows for the identification of relevant attributes to predict relapse. These features are incorporated as node features in the graph representation. This integration enabled a thorough analysis of the breast cancer relapse graph by leveraging the insights gained from the XGBoost model. The presence of these features as attributes associated with each node offers valuable information for various graph-based tasks, such as node classification, anomaly detection, and graph clustering.

By incorporating the insights obtained from the XGBoost model within the graph context, a deeper understanding of

Algorithm 3 Get Node Adjacents

Input: node_connections ▷ Cluster nodes = 5
Input: node_threshold ▷ Node's threshold parameter = 0.04
Input: neg_cont ▷ Node negative contributions
Input: pos_cont ▷ Node positive contributions
Input: classi ▷ Node classification
Output: edge_indexes

for each row **in** df **do**
  if (classi = 0 **and** row.relapse_free_status = 0) **then**
    Add node, negative contributions into an array
  end if
  if (classi = 1 **and** row.relapse_free_status = 1) **then**
    Add node, positive contributions into an array
  end if
end for
if (classi == 0) **then**
  Sorting negative contributions
  for each i, v **in** neg.dic **do**
    if (neg_cont < $vandi \neq$ patient_id) **then**
      if (adjacents $\geq$ node_connections) **or** (count > 1 **and** (v - (sum/count)) > node_threshold) **then**
        Exit Loop
      end if
    end if
    adjacentsArray $\Leftarrow$ node
    count $\Leftarrow$ count + 1
    sum $\Leftarrow$ sum + v
  end for
end if
if (classi == 1) **then**
  Sorting positive contributions
  for each i, v **in** pos.dic **do**
    if (pos_cont < $vandi \neq$ patient_id) **then**
      if (adjacents $\geq$ node_connections) **or** (count > 1 **and** (v - (sum/count)) > node_threshold) **then**
        Exit Loop
      end if
    end if
    adjacentsArray $\Leftarrow$ node
    count $\Leftarrow$ count + 1
    sum $\Leftarrow$ sum + v
  end for
end if
return adjacentsArray

the underlying patterns and relationships within the breast cancer relapse graph can be achieved. Each node within the graph corresponds to a patient, with the colour assigned to each node indicating the occurrence of relapse for that particular patient. It is worth mentioning that due to the size of the data set, only a sample of patients is displayed in this visualisation.

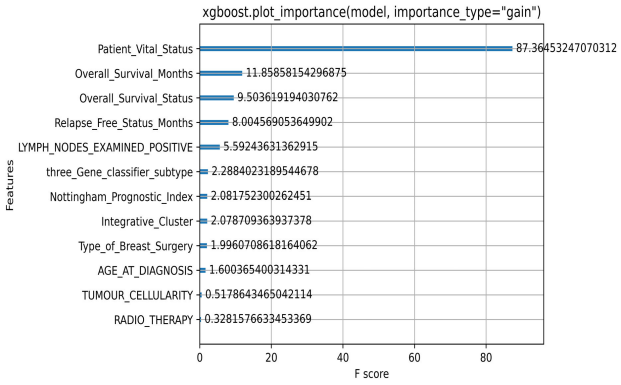


FIGURE 4. Visualizing Feature Importance with XGBoost Gain.

3) GNN MODEL ARCHITECTURE

The selection of an appropriate architecture for the neural network graph model is of paramount importance to achieve accurate predictions for breast cancer relapse. Models such as Graph Convolutional Networks and other GNN variants are specifically designed to leverage the inherent connectivity patterns within the graph data. These models excel at learning meaningful representations by both the features of individual nodes and the information from their neighbouring nodes. Through iterative information aggregation, these models can effectively capture local and global dependencies, enabling robust predictions for breast cancer relapse.

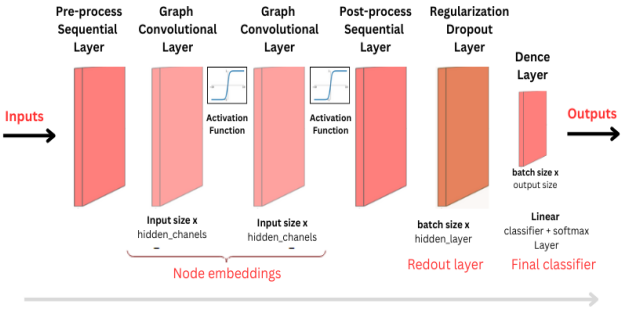


FIGURE 5. GraphX-Net model architecture. The model applies preprocessing using feed- forward network to the node features to generate initial node representations. Also applies two graph convolutional layers, with skip connections, to the node representation to produce node embeddings. And finally feed the node embeddings in a Softmax layer to predict the node class. Each graph convolutional layer added captures information from a further level of neighbours.

The graph data is represented by the graph_info tuple, which consists of the following three elements:

- node features: This is a NumPy array [num nodes, num features] that includes the node features. In this dataset, the nodes are the patient IDs, and the node features are the features that we have extracted in the feature extraction section.
- edges: This is [num edges, num edges] NumPy array representing a sparse adjacency matrix of the links

between the nodes. In this example, the links are the citations between the papers.
- edge weights (Optional): No weight relationships in the graph between patients with relapsed breast cancer.

4) IMPLEMENTATION

The model architecture has been implemented as in Fig. 5. The graph convolutional layers perform the following steps:

- **preparation**: the input node representations were processed using a feed-forward Network to generate a message. To simplify the processing, a linear transformation was applied to the node representations. It includes three layers: BatchNormalization, Dropout, and Dense with the activation function Gelu.
  1) BatchNormalization: it normalizes the input data along the dimension of the features. It helps stabilize and accelerate training by reducing the internal covariate shift. The input here should be a tensor type which is a multi-dimensional array of numbers. The formula for batch normalization can be expressed as follows:

$$y = \frac{x - \text{mean}(x)}{\sqrt{\text{var}(x) + \epsilon}} \times \gamma + \beta \tag{2}$$

  As proved by [26]. Where:
  - x is the input tensor.
  - mean(x) and var(x) are the mean and variance of x in the batch. The variance indicates how far x is from the mean.
  - γ and β are learnable scaling and shifting parameters, respectively.
  - ϵ is a small constant to avoid division by zero.
  2) Dropout: Dropout is a regularization technique that randomly sets a fraction of inputs to zero during training. This helps prevent overfitting and enhances generalization. The dropout formula is simple, where ρ is the dropout rate:

$$\text{Dropout}(x) = x \times \text{mask} \tag{3}$$

$$\text{mask} \sim \text{Bernoulli}(1 - p)$$

  This is approved by [27]. In this equation:
  - x represents the input tensor.
  - mask is a random binary mask with the same shape as x that is drawn from a Bernoulli distribution with a probability of $1-p$, where p is the dropout rate.
  3) Dense with Gelu Activation: The Dense layer performs a linear transformation on the input data and applies the Gelu activation function. Gelu is a variant of the ReLU activation and is defined as follows:

$$\text{Gelu}(x) = 0.5x\left(1+\tanh\left(\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right)\right) \tag{4}$$

  As proved by [28]

- **Aggregate**: The messages of the neighbours of each node are aggregated concerning the edge weights using a permutation invariant pooling operation, such as sum, mean, and max, to prepare a single aggregated message for each node. In the message-passing framework of GNNs, node features are update d by aggregating messages from neighboring nodes. Using permutation invariant pooling operations ensures that the aggregated message is the same regardless of the order in which neighbors are considered.
  - Sum Pooling: $(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i$
  - Mean Pooling: $(x_1, x_2, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Max Pooling:$(x_1, x_2, .., x_n) = \max_i x_i$

$$\text{Aggregate(messages)} = \sum_{i \in \text{neighbors}} \text{messages}_i \tag{5}$$

  This equation represents the aggregation process, where messages are a set of messages prepared for a given node, and the sum iterates over the neighboring nodes of the current node to aggregate their messages.
- **Update**: The node representations and aggregated messages of both shapes [num nodes, representation dim] are combined and processed to produce the new state of the node representations (node embeddings). The node representations and aggregated messages are concatenated to create a sequence and then processed by Concatenating the node representations and aggregated messages.

$$\text{Update(node embedding, aggregated messages)} = \text{Concatenate(node embedding, aggregated messages)} \tag{6}$$

  This equation represents the update process, where the node embedding is updated by adding the aggregated messages from neighbouring nodes and passing the result through the sigmoid activation function.

$$\text{Reset gate: } z_t = \sigma(W_z \cdot x_t + U_z \cdot h_{h-1} + b_z)$$
$$\text{Update gate: } r_t = \sigma(W_r \cdot x_t + U_r \cdot h_{h-1} + b_r)$$
$$\text{Candidate hidden state: } \overline{h}_t = \tanh(W_h \cdot x_t + U_h \cdot (r_t \odot h_{t-1}) + b_h)$$
$$\text{Hidden state: } h_t = (1 - z_t) \odot (h_{t-1} + z_t \odot \overline{h}_t)$$

  As proved by [29]. In these equations:
  - x_t represents the input at time step t.

- h_{t-1} represents the hidden state (output) from the previous time step $t-1$
- z_t is the reset gate, controlling which parts of the previous state should be ignored.
- r_t is the update gate, controlling how much of the previous state to keep.
- $\bar{h}_t$ is the candidate hidden state that combines the input and the reset-gated previous state.
- h_t is the updated hidden state at time step t.
- W_z, W_r,W_h are learnable weight matrices.

GRU (Gated Recurrent Unit) has been used in the GraphX-Net model. Combining different layers, including Graph Convolutional Layers, Pooling Layers, and Recurrent Layers such as GRU. We can typically feed the node features and graph structure (e.g., adjacency matrix or graph Laplacian) into the GraphX-Net model. The GRU layer can be utilised to capture temporal dependencies among neighbouring nodes and update node embeddings iteratively during the message-passing process [25].

$$h_n^{(k+1)} = \text{UPDATE}_u^{(k)} \left(h_u^{(k)}, \text{AGGREGATE}_u^{(k)} \left(\{h_v^{(k)}, \forall v \in N(u)\}\right)\right) \quad (7)$$

As proved by [30]. In this equation:

- $h_n^{(k+1)}$ represents the node embedding of node n at the $(k+1)-th$ layer.
- $h_u^{(k)}$ represents the node embedding of node u at the $k-th$ layer.
- $h_v^{(k)}$ represents the embedding of the node of the neighboring node v of the node u in the $k-th$ layer.
- $\text{UPDATE}_u^{(k)}$ is the update function that takes the current node embedding $h_u^{(k)}$ and the aggregated messages from neighboring nodes as inputs and updates the node embedding of node u to $h_v^{(k)}$ for the $(k+1)-th$ layer.
- $\text{AGGREGATE}_u^{(k)}$ is the aggregation function that takes the node embeddings of neighboring nodes $h_v^{(k)}, \forall v \in N(u)$ and combines them to obtain aggregated messages for node u in the $k-th$ layer.

$$h_u^{(k+1)} = \text{UPDATE}_u^{(k)}(h_u^{(k)}, m_{N(u)}^{(k)}) \quad (8)$$

As proved by [30]. In this equation:

- $h_u^{(k)}$ represents the node embedding of node u at the $k-th$ layer.
- $m_{N(u)}^{(k)}$represents the aggregated messages from the neighboring nodes $N(u)$ of node u at the $k-th$ layer.
- $\text{UPDATE}_u^{(k)}$ is the update function that takes the current node embedding ($h_u^{(k)}$ and the aggregated messages $m_{N(u)}^{(k)}$ as inputs and updates the node embedding to $h_u^{(k+1)}$ for the $(k+1)-th$ layer.

In the context of message-passing iterations in a Graph Neural Network, we utilise the terms UPDATE and AGGREGATE to refer to arbitrary differentiable functions, typically implemented as neural networks. These functions are responsible for updating the node embeddings and aggregating information from the graph neighbourhood. The aggregated messages from the neighbouring nodes $N(u)$ of node u at the k-th layer are denoted as $m_{N(u)}^{(k)}$.

In the message-passing process of a Graph Neural Network, the update function UPDATE combines the message $m_{N(u)}^{(k)}$ with the previous embedding $h_u^{(k-1)}$ of node u to generate the updated embedding $h_u^{(k)}$. At the initial iteration ($k=0$), the embeddings are set to the input features for all nodes, i.e., $h_u^{(0)} = x_u, \forall u \in V$, where x_u represents the input features of node u. After running K iterations of the GNN message passing, we can use the output of the final layer to define the embeddings for each node, i.e.,

$$z_u = h_u^{(k)}, \forall u \in V \quad (9)$$

This is proved by [31]. In this equation:

- z_u represents the embedding vector for node u.
- $h_u^{(k)}$ represents the node embedding of node u at the $k-th$ layer.
- $\forall u \in V$ means the equation applies to all nodes u in the set of nodes V.

The superscripts, such as k, were used to differentiate the embeddings and functions at different iterations of the message-passing process. This enabled us to track and distinguish evolving embeddings and functions throughout the message-passing iterations in the GNN models [32].

The main steps of our GraphX-Net model can be described as follows:

- Apply reprocessing using feedforward to the node features to generate initial node representations.
- Apply one or more graph convolutional layers, with skip connections, to the node representation to produce node embeddings.
- Apply post-processing using feed-forward network to the node embeddings to generate the final node embeddings.
- Feed the node embeddings in a Softmax layer to predict the node class.

With each additional graph convolutional layer, the model was able to capture information from an extended range of neighbouring nodes.

- Each graph convolutional layer in the model captured and integrated information from neighbouring nodes.
- The layer applies a convolutional operation on the node representations, taking into account the connections and relationships defined by the graph structure.
- The convolutional operation involved aggregating and combining information from neighbouring nodes to update the representations of each node (Eq. 9).

IV. PATIENTS' COHORT AND EXPERIMENTAL SETUP

This study utilised a large cohort of invasive breast cancer with long-term clinical follow-up and complete

clinicopathological data. This dataset consists of1980 breast cancer cases within the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort. The METABRIC study protocol, detailing the molecular profiling methodology in a cohort of 1980 breast cancer samples is described by [33]. Patient demographics are summarised in Appendix VI. Data was downloaded from the cBioportal data repository and it include of clinical records, patient demographics, tumor characteristics, treatment details, and follow-up data sourced from trusted repositories [34].

A. DATA PRE-PROCESSING

Here the original data have undergone multiple preprocessing steps to ensure suitability for analysis. Initially, the data were presented in a tabular format, requiring various data-wrangling operations. One of the challenges encountered was dealing with missing values, which could not be uniformly imputed across the dataset. Consequently, these missing values were excluded from the analysis to ensure the compatibility of the data. In addition, certain parameters were categorized to enhance their utility in subsequent analyses. Inconsistencies were reviewed and rectified to maintain data integrity.

Following data cleaning, we further performed pre-processing tasks were performed, including feature selection and normalisation. Feature selection helped identify the most informative variables for relapse prediction, reducing dimensionality and improving model performance. Normalization ensured that features were on a common scale, preventing bias due to varying magnitudes. These cleaning and pre-processing steps formed a crucial foundation for our subsequent analysis and modelling efforts, ensuring reliable and meaningful results.

TABLE 1. Experiment findings on feature selection and other metrics.

Illustration	METABRIC cross-sectional data
Size of the training set	935
Size of the test set	584
Total class count	2
Total number of features in the dataset	24
Number of trainings data features after correlation	10
Number of features in the optimal set	9
Learning rate	1.2×10^{-2}
Hidden units	[32,32]
Dropout rate	0.5
Epochs	300
Batch size	128

The processing of the METABRIC dataset includes crucial steps to facilitate effective analysis and modelling. Initially, the selection of the target variable, especially the 'Relapse free status' plays a pivotal role. Feature selection techniques are then applied to identify the most influential attributes affecting this clinical outcome. To mitigate class imbalance,

TABLE 2. The most significant 9 input features in the Relapse breast cancer dataset.

#	Feature
1.	Patient's Vital Status
2.	AGE_AT_DIAGNOSIS
3.	Relapse Free Status (Months)
4.	Overall Survival (Months)
5.	LYMPH_NODES_EXAMINED_POSITIVE
6.	Overall Survival Status
7.	TUMOUR CELLULARITY
8.	Nottingham Prognostic Index
9.	Integrative Cluster

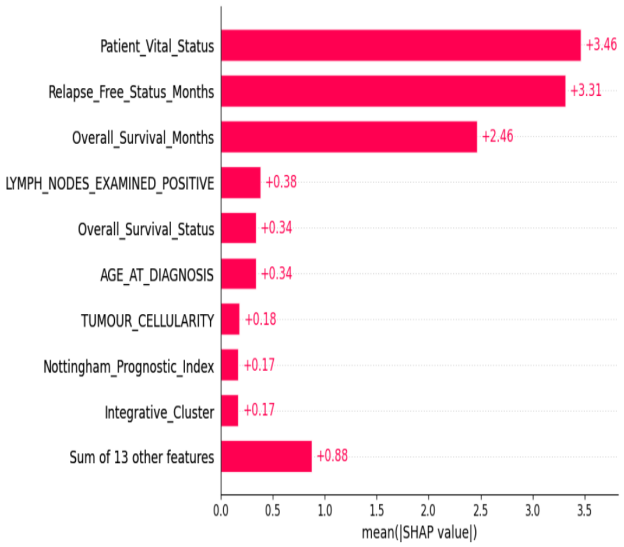


FIGURE 6. An illustrative SHAP bar plot is employed to evaluate global feature importance, offering a comprehensive assessment of the influence of each feature on the model's predictions. The plot showcases how each feature contributes to the model's output, enabling a deeper understanding of their impact and aiding in model interpretation, thereby enhancing the transparency and reliability of the predictive framework.

the Synthetic Minority Oversampling Technique (SMOTE) was employed, creating synthetic samples for the minority class. The ExtractTreeClassifier class can extract valuable features from the dataset, providing insights beyond raw attributes. Following feature selection and extraction, the dataset was partitioned into training and test sets. Subsequently, the importance of the features was determined using methods such as the XGBoost classifier and the Shapley values, offering a clear understanding of the impact of each feature. Classical machine learning models, including Boosting, Random forest, and Stacking algorithms, were constructed and assessed for predictive accuracy. To explore feature interactions in depth, a graph-based approach was adopted, where features are represented as nodes and their relationships as edges. Graph neural networks were used for node classification in this feature graph, providing a holistic perspective of feature interactions and their influence on the target variable. This comprehensive approach improves our understanding of breast cancer prognosis and treatment decisions.

In summary, METABRIC dataset processing involves meticulous target selection, feature engineering, SMOTE for

class balance, advanced feature analysis with XGBoost and Shapley values, classical ML modelling, and GNN-based feature interaction exploration. These steps collectively empowered robust predictions and a deeper understanding of breast cancer patient outcomes.

B. TRAINING SETTINGS

For node classification, the GraphX-Net model was trained using GNN. Specifically, we focus on the architecture of the model, which includes pre-process and post-process layers to enhance the model's performance Fig. 5.

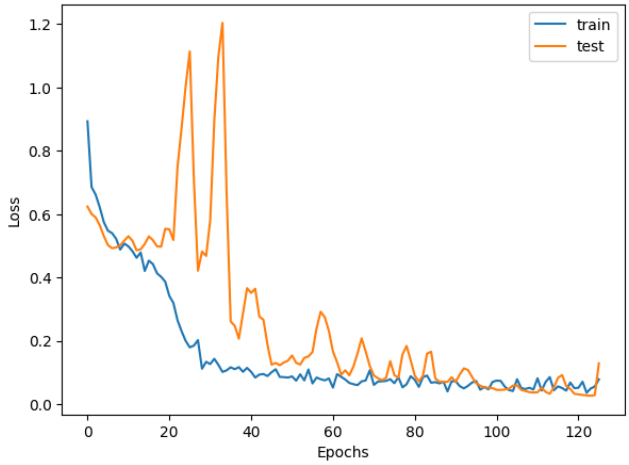


FIGURE 7. Analysing the Reduction in Loss Across Epochs to Optimise the Performance of GraphX-Net in Breast Cancer Relapse Prediction.

C. GRAPHX-NET TRAINING

1) Pre-process Layer: to prepare the input data and set the stage for effective learning, which was constructed using three distinct layers: BatchNormalisation, Dropout, and Dense with Gelu activation.
   a) BatchNormalisation: to normalise the node features during training, thereby reducing the internal covariate shift. This layer helps stabilise the learning process and accelerates convergence, ultimately improving the model's performance.
   b) Dropout: The Dropout layer introduces regularisation by randomly dropping out a fraction of the nodes during each training epoch. This is to prevent overfitting and to ensure that the model learns robust representations that generalise well to unseen data.
   c) Dense with Gelu activation: Apply linear transformation followed by a Gelu activation function to introduce nonlinearity. Gelu is a Gaussian Error Linear Unit activation that is known to perform well in deep learning models and has a smooth gradient, allowing for more stable training [35].
2) GraphConvLayers Module:
   The backbone of our node classification model lies in the two graph convolutional layers. Graph convolutional layers have been designed specifically to handle graph-structured data and can capture local and global dependencies among nodes. It aggregates information from neighbouring nodes to update the node embeddings. It leverages the graph's adjacency matrix to weigh the importance of each neighboring node, capturing the graph's structural dependencies. Each layer refines the node representations, allowing the model to gradually understand the underlying graph topology.
3) Post-process Layer:
   The postprocess layer was the final part of our GNN-based node classification model, which is responsible for refining the learned representations and generating the final node predictions. Similar to the pre-processing layer, the post-process layer consisted of three layers: BatchNormalisation, Dropout, and Dense with Gelu activation. The model outputs a probability distribution for each node, indicating the likelihood of belonging to each class.

Training a GNN model for node classification using GraphConvLayer module involved a thoughtful design of

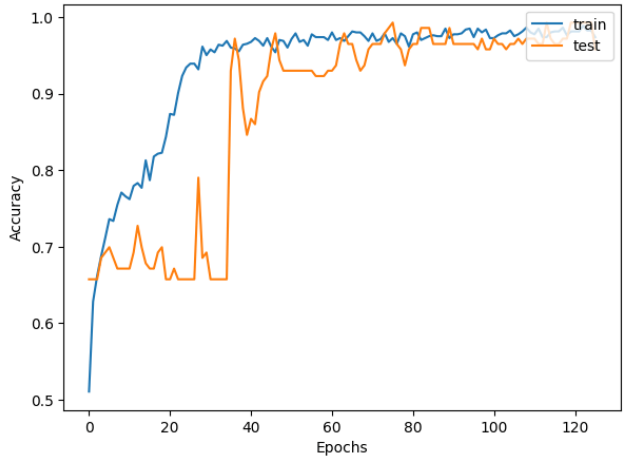


FIGURE 8. Improving Accuracy Across Training Epochs: Optimization of GraphX-Net in Breast Cancer Relapse Prediction.

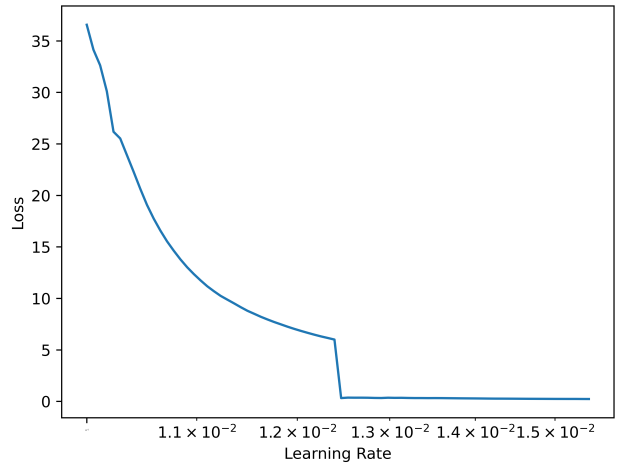


FIGURE 9. Exploring the Impact of Learning Rate on Model Training: An In-Depth Analysis of Learning Rate vs. Loss.

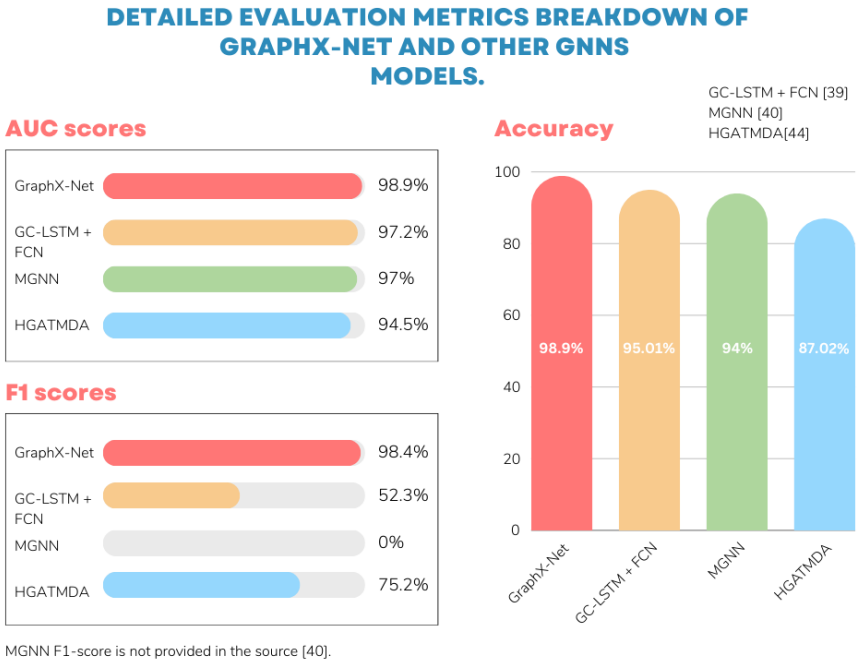


FIGURE 10. Detailed evaluation metrics breakdown of GraphX-Net and other GNNs models.

the model architecture. The incorporation of pre-process and post-process layers, each consisting of BatchNormalisation, Dropout, and Dense with Gelu activation, enhanced the model's ability to capture meaningful features and structural dependencies present in the graph. With this powerful approach, we can achieve accurate node classification results in various applications, enabling us to harness the potential of GNNs in solving real-world problems effectively.

D. PERFORMANCE EVALUATION

To evaluate the GraphX-Net model, its performance was thoroughly assessed to assess its ability to generalize effectively to unseen data. Several evaluation techniques are commonly used for this purpose. The key techniques used in the evaluation of GraphX-Net included:

- **Train-Validation-Test Split**: Data were split into trains 0.07 and 0.15 for validation and test data, respectively. The GraphX-Net model showed a good performance with a test accuracy of 98.90
- **Learning Rate Adjustment:** To detect the best learning rate value, Lambda Callback package has been used to calculate the best LR value. With callback actions, we recorded all loss values.
  By analysing the graph, we can identify the optimal learning rate, which is typically the learning rate where the loss decreases the fastest before diverging or oscillating. Here in our case, the graph shows that the best LR value is between 1.2 x 10^{-2} and 1.3 x 10^{-2}
- **Cross-Validation**: Performed k-fold cross-validation to obtain more robust performance estimates.
- **Task-Specific Metrics**: As shown in Table 1.

TABLE 3. Classical machine learning model evaluation metrics. K-fold cross-validation across all models = 3.

Models	Metrics			
	AUC score	F1 score	Balanced accuracy	Cross-validation
RF	76.15	76.47	75.96	83.30
Boosting	75.82	75.88	75.68	82.24
Stacking	75.72	75.20	75.68	80.74

E. MULTIPART FIGURES

Figures compiled of more than one sub-figure presented side-by-side, or stacked. If a multipart figure is made up of multiple figure types (one part is lineart, and another is grayscale or color), the figure should meet the stricter guidelines.

Due to the extensive process involved in implementing both Graph Neural Network (GNN) and classical machine learning (ML) approaches, as well as the subsequent evaluation of these models, it is crucial to present the results in a clear and organized manner. The complexity of these methodologies requires a thorough examination and comparison to ensure a complete understanding of their performance. To facilitate this, the results have been systematically compiled and are illustrated in the Table 4. This table provides a detailed comparison of various aspects, allowing for an easy and effective assessment of the strengths and limitations of each approach.

Classical machine learning techniques often fail to provide reliable results for complex problems such as breast cancer recurrence prediction. A study on the WPBC dataset, consisting of 198 patients (151 nonrecurring and 47 recurring cases), reported accuracy rates of 78.5%, 73.8%, and 67.2% for Support Vector Machine (SVM), Random Forest, and Decision Tree, respectively [44]. Similarly, our study

TABLE 4. Comparison between GraphX-Net Model and Classical Machine Learning Approaches.

Aspect	GraphX-Net Model	Classical Machine Learning Approaches
Dataset	Same dataset	Relapse breast cancer dataset
Ease of Implementation	Moderate	High
Interpretability	Moderate	High
Performance with Graph-Structured Data	Superior	Limited
Accuracy and Performance Metrics	Displayed in Fig. 10	Displayed in Table 3
Hierarchical Understanding	Leverages hierarchical nature of GNNs for deeper understanding of structural patterns	Lacks ability to capture hierarchical nature of data
Capability to Capture Complex Interactions	High	Limited
Predictive Capabilities	Higher, with powerful predictive capabilities	Lower compared to GraphX-Net
Scalability	Scales better for large and complex datasets	Scales well for small to moderately large datasets
Flexibility in Feature Engineering	Automatically captures relevant features through graph structure	Requires extensive manual feature engineering
Computational Complexity	Higher due to the complexity of graph computations	Generally lower, dependent on the simplicity of the algorithm
Training Time	Longer, due to the complex nature of graph computations	Typically shorter, depending on the algorithm
Handling Missing Data	More robust to missing data due to the interconnected nature of graphs	Requires imputation or preprocessing
Robustness to Noisy Data	More robust due to the ability to capture and utilize relationships within the graph	Varies with the algorithm, often sensitive to noise
Examples of Algorithms Used	Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), GraphSAGE	Decision Trees, Random Forests, Support Vector Machines (SVM), Logistic Regression
Insight into Data Trends	Provides valuable insights by uncovering complex patterns and relationships within the data	Limited to predefined features and relationships
Use Cases Beyond Current Study	Increasingly adopted in areas requiring analysis of relational data, such as social networks, biology, and chemistry	Widely used in various fields such as finance, healthcare, and marketing

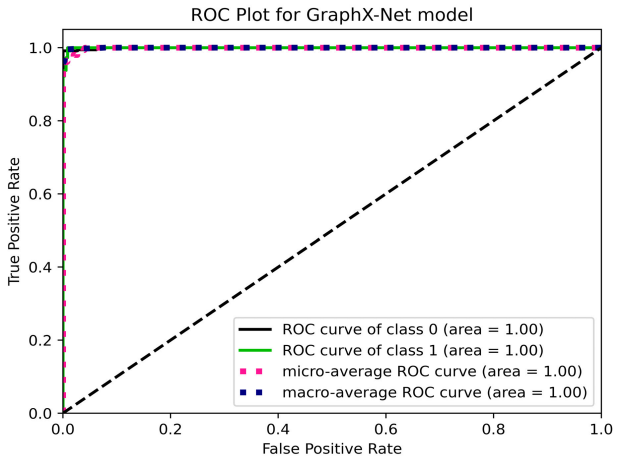


FIGURE 11. ROC Analysis: GraphX-Net Performance in Discriminating Positive and Negative Cases for the model. The classifier's curve goes straight up to (0, 1) and then straight to the right to (1, 1) which is the perfect classifier output.

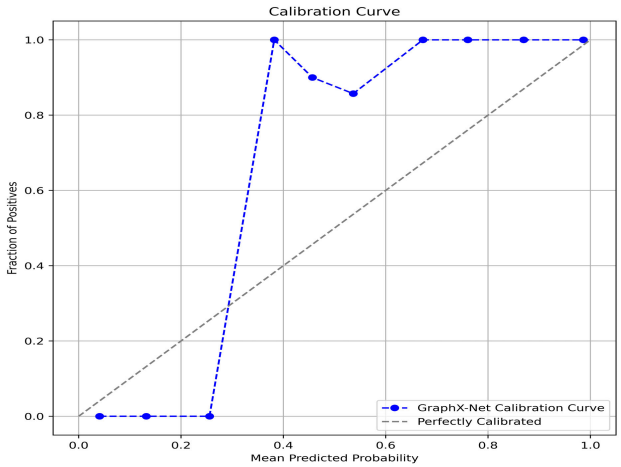


FIGURE 12. GraphX-Net Model Calibration Analysis: Predicted Probabilities vs. Actual Outcomes. The predicted probabilities are divided into several bins or intervals. For each bin, the actual observed event rate is calculated. This is the fraction of true positive cases within that bin.

as shown in Table 3 used Random Forest, Boosting, and Stacking yielded accuracies of 75.96%, 75.68%, and 75.68%. In contrast, our GraphX-Net model as in Fig.10, which utilizes Graph Neural Networks (GNNs), achieved a significantly higher accuracy of 98.9%, showcasing its superior ability to predict breast cancer recurrence accurately

by effectively extracting crucial information from medical data.

V. DISCUSSION AND CONCLUSION

Herein, we introduced a novel Graph Neural Network model, called GraphX-Net, that leverages Shapley values to construct an efficient neural graph tailored to the data. This neural graph is designed to fit entirely in memory and represents the graph data using graph information components, including node features and edges. In the context of relapse breast cancer data, where edge weights are unavailable, we utilise these values effectively to connect patient nodes and create distinct clusters based on their Shapley value contributions. Given the subtle differences in contributions, we introduce two additional parameters, node threshold and node connections, to manage the relationships among nodes. The integration of Graph Neural Networks and Shapley values enhances the accuracy and interpretability of relapse prediction for breast cancer patients, addressing the critical need for accurate and explainable models in the field of breast cancer prognosis. GraphX- Net not only achieves high prediction accuracy but also offers valuable insights into the specific features and interactions within the patient's treatment graph that contribute to the relapse risk. This unique contribution enables clinicians and researchers to understand the underlying mechanisms that drive relapse in breast cancer and make informed decisions regarding treatment and follow-up strategies. Through rigorous experimentation and comparison with existing approaches, our results demonstrate the significant advancements of GraphX-Net model with Shapley values brings to the field of breast cancer prognosis, opening up new avenues for improved patient care and personalised medicine.

Our study introduces the GraphX-Net model, a novel approach that uses graph neural networks (GNNs) to predict breast cancer relapse. Demonstrating its effectiveness in utilizing graph structures for classification, the GraphX-Net model offers valuable insights into graph connectivity, node clustering, and layout visualization, facilitating the conversion of traditional tabular datasets into graphs. By incorporating Shapley values, a deeper understanding of feature importance and underlying factors influencing predictions could be gained, revealing biological mechanisms and potential risk factors associated with breast cancer relapse. The interpretability of the model proves vital for clinical decision-making and personalized treatment strategies. Its accurate relapse predictions highlight the potential for clinical translation, enabling real-world healthcare applications and risk assessment for improved patient outcomes. Overall, our research advances precision medicine, fostering tailored and targeted breast cancer therapies worldwide.

The proposed GraphX-Net model has significant practical implications in the field of computational biology and precision medicine. From a computational perspective, the utilization of Graph Neural Networks (GNNs) allows for the effective handling of structured data, capturing intricate dependencies that traditional models might overlook. This advanced capability ensures more accurate predictions, which are crucial for early intervention and personalized treatment strategies. By achieving state-of-the-art performance metrics, GraphX-Net demonstrates its potential to be a valuable tool in clinical decision-making processes, ultimately contributing to improved patient outcomes. The findings of this study not only advance the field of bioinformatics but also pave the way for future research to leverage GNNs for various predictive modeling tasks in healthcare and beyond.

One promising future research direction involves the integration of longitudinal patient data to enhance the predictive power of the model. By incorporating temporal information, such as changes in gene expression or treatment responses over time, the model could capture dynamic patterns and provide even more accurate predictions of breast cancer recurrence. Additionally, exploring the application of GraphX-Net to other types of cancer or diseases with similar multi-factorial etiologies could further validate its utility and extend its benefits to a broader range of medical conditions.

TABLE 5. Clinicopathological characteristics in the METABRIC cohort.

Variables	N (%)
Age at diagnosis[Median (range)]	61.8 (21.93-96.29)
Tumour size [Median (range)]	23 (1, 182)
NPI [Median (95% CI)]	4.46 (4.41-4.51)
Survival [Median (Months, 95% CI)]	149 (141-159)
Axillary lymph nodes status	
0	1035
1	337
2	171
3	114
>3	314
Axillary lymph nodes status	
Positive	1497
Negative	438
Null	42
PAM50 subtype	
Basal	330
HER2	238
Luminal A	715
Luminal B	489
Normal-like	199
Not classified	6
Adjuvant systemic therapy (AT)	
No AT	305
Hormone therapy (HT)	1216
Chemotherapy	416
Hormone + chemotherapy	192

VI. CODE AVAILABILITY

The code is available in GraphX-Net's Github repository at "https://github.com/abdullahbasaad/GraphX-Net.git" The repository contains Python scripts to generate and evaluate explanations using performance metrics and also visualize explanationa.

APPENDIX

SUPPLEMENTARY TABLE S1

See Table 5.

REFERENCES

[1] S. A. Rabbani and A. P. Mazar, "Evaluating distant metastases in breast cancer: From biology to outcomes," *Cancer Metastasis Rev.*, vol. 26, nos. 3–4, pp. 663–674, Dec. 2007.

[2] A. Nicolini, R. Giardino, A. Carpi, P. Ferrari, L. Anselmi, S. Colosimo, M. Conte, M. Fini, G. Giavaresi, P. Berti, and P. Miccoli, "Metastatic breast cancer: An updating," *Biomed. Pharmacother.*, vol. 60, pp. 548–556, Nov. 2006.

[3] A. I. Riggio, K. E. Varley, and A. L. Welm, "The lingering mysteries of metastatic recurrence in breast cancer," *Brit. J. Cancer*, vol. 124, no. 1, pp. 13–26, Jan. 2021.

[4] L. Hutchinson and V. T. DeVita, "Focus issue on biomarkers," *Nature Rev. Clin. Oncol.*, vol. 7, no. 6, p. 295, Jun. 2010.

[5] L. Chen, J. Hodskins, S. Chokshi, J. Croley, M. Stevens, G. Pasley, K. Huller, J. Reynolds, H. Weiss, and S. Massarweh, "P5-13–24: A predictive model of early systemic disease relapse after standard adjuvant therapy for breast cancer," *Cancer Res.*, vol. 71, no. 24, pp. P5–13, Dec. 2011.

[6] A. Ullah, S. Jabeen, S. Zaman, A. Hamraz, and S. Meherban, "Predictive potential of K-Banhatti and zagreb type molecular descriptors in structure–property relationship analysis of some novel drug molecules," *J. Chin. Chem. Soc.*, vol. 71, no. 3, pp. 250–276, Mar. 2024.

[7] W. Ahmed, K. Ali, S. Zaman, and A. Raza, "Molecular insights into anti-Alzheimer's drugs through predictive modeling using linear regression and QSPR analysis," *Modern Phys. Lett. B*, vol. 38, no. 27, Sep. 2024, Art. no. 2450260.

[8] W.-M. Song, T. Huang, S. Yoo, E. Lee, Y. Zhao, L. Wang, Z. Tu, X. Dai, H. Irie, J. Zhu, and B. Zhang, "Abstract 363: Planar filtered gene regulatory networks in breast cancer," *Cancer Res.*, vol. 74, no. 19, p. 363, Oct. 2014.

[9] U. Testa, G. Castelli, and E. Pelosi, "Breast cancer: A molecularly heterogenous disease needing subtype-specific treatments," *Med. Sci.*, vol. 8, no. 1, p. 18, Mar. 2020.

[10] L. Yuan, L. Guo, Y. Zhang, K. Han, A. Nandi, and B. Honig, "De-Huang integration of multi-omics data for gene regulatory network inference and application to breast cancer," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, pp. 782–791, 2019.

[11] W. Luo, Q. Huang, X. Huang, H. Hu, F. Zeng, and W. Wang, "Predicting breast cancer in breast imaging reporting and data system (BI-RADS) ultrasound category 4 or 5 lesions: A nomogram combining radiomics and BI-RADS," *Sci. Rep.*, vol. 9, 2019, Art. no. 11921.

[12] C. Acharya, D. Hsu, C. Anders, A. Anguiano, K. Salter, K. Walters, R. Redman, S. Tuchman, C. Moylan, S. Mukherjee, W. Barry, H. Dressman, G. Ginsburg, K. Marcom, K. Garman, G. Lyman, J. Nevins, and A. Potti, "Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer," *J. Amer. Med. Assoc.*, vol. 299, no. 13, pp. 1574–1587, 2008.

[13] Y. Zhu, E. Tzoras, A. Matikas, J. Bergh, A. Valachis, I. Zerdes, and T. Foukakis, "Expression patterns and prognostic implications of tumor-infiltrating lymphocytes dynamics in early breast cancer patients receiving neoadjuvant therapy: A systematic review and meta-analysis," *Frontiers Oncol.*, vol. 12, 2022, Art. no. 999843.

[14] A. Lafourcade, M. His, L. Baglietto, M.-C. Boutron-Ruault, L. Dossus, and V. Rondeau, "Factors associated with breast cancer recurrences or mortality and dynamic prediction of death using history of cancer recurrences: The French E3N cohort," *BMC Cancer*, vol. 18, no. 1, p. 171, Dec. 2018.

[15] A. Bode and Z. Dong, "Precision oncology-the future of personalized cancer medicine?" *NPJ Precis. Oncol.*, vol. 1, no. 2, 2017.

[16] C. Zuo, J. Qian, S. Feng, W. Yin, Y. Li, P. Fan, J. Han, K. Qian, and Q. Chen, "Deep learning in optical metrology: A review," *Light, Sci. Appl.*, vol. 11, no. 1, p. 39, Feb. 2022.

[17] Y.-D. Zhang, S. C. Satapathy, D. S. Guttery, J. M. Górriz, and S.-H. Wang, "Improved breast cancer classification through combining graph convolutional network and convolutional neural network," *Inf. Process. Manage.*, vol. 58, no. 2, Mar. 2021, Art. no. 102439.

[18] A. Vulli, P. N. Srinivasu, M. S. K. Sashank, J. Shafi, J. Choi, and M. F. Ijaz, "Fine-tuned DenseNet-169 for breast cancer metastasis prediction using FastAI and 1-cycle policy," *Sensors*, vol. 22, no. 8, p. 2988, Apr. 2022.

[19] A. Jha, G. Verma, Y. Khan, Q. Mehmood, D. Rebholz-Schuhmann, and R. Sahay, "Deep convolution neural network model to predict relapse in breast cancer," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 351–358.

[20] D. Ollech and K. Webel, "A random forest-based approach to identifying the most informative seasonality tests," Deutsche Bundesbank Discuss. Paper, 2020.

[21] S. Rigatti, "Random forest," *J. Insurance Med.*, vol. 47, no. 1, pp. 31–39, 2017.

[22] D.-C. Feng, Z.-T. Liu, X.-D. Wang, Y. Chen, J.-Q. Chang, D.-F. Wei, and Z.-M. Jiang, "Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach," *Construct. Building Mater.*, vol. 230, Jan. 2020, Art. no. 117000.

[23] K. Ting and I. Witten, "Issues in stacked generalization," *J. Artif. Intell. Res.*, vol. 10, pp. 271–289, 1999.

[24] S. Lundberg. (2023). *Others SHAP (SHapley Additive exPlanations) Documentation*. Accessed: May 27, 2024. [Online]. Available: https://shap.readthedocs.io/en/latest/index.html

[25] S. Ben Jabeur, N. Stef, and P. Carmona, "Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering," *Comput. Econ.*, vol. 61, no. 2, pp. 715–741, Feb. 2023.

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 448–456. [Online]. Available: https://proceedings.mlr.press/v37/ioffe15.html

[27] S. Ioffe and Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 448–456. [Online]. Available: https://proceedings.mlr.press/v37/ioffe15.html

[28] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[29] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 92–102, Apr. 2018.

[30] F. Scarselli, M. Gori, A. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[31] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, p. 11. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf

[32] L. Wu, P. Cui, J. Pei, and L. Zhao, *Graph Neural Networks: Foundations, Frontiers, and Applications*, 1st ed., Singapore: Springer, Jan. 2023.

[33] C. Curtis et al., "The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups," *Nature*, vol. 486, pp. 346–352, Jun. 2012.

[34] J. Gao, B. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz, "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Sci. Signaling*, vol. 6, p. pl1, 2013.

[35] M. Lee, "Gelu activation function in deep learning: A comprehensive mathematical analysis and performance," 2023, *arXiv:2305.12073*.

[36] M. Wang, D. Zheng, Z. Ye, Q. Gan, M. Li, X. Song, J. Zhou, C. Ma, L. Yu, Y. Gai, T. Xiao, T. He, G. Karypis, J. Li, and Z. Zhang, "Deep graph library: A graph-centric, highly-performant package for graph neural networks," 2019, *arXiv:1909.01315*.

[37] K. Akensert. (2021). *Graph Attention Network (GAT) for Node Classification*. [Online]. Available: https://keras.io/examples/graph/gat_node_classification/

[38] A. Presekal, A. Ştefanov, V. Rajkumar, and P. Palensky, "Attack graph model for cyber-physical power systems using hybrid deep learning," *IEEE Trans. Smart Grid*, vol. 14, pp. 4007–4020, 2023.

diagnostics

MDPI

Article

A BERT-GNN Approach for Metastatic Breast Cancer Prediction Using Histopathology Reports

Abdullah Basaad [1], Shadi Basurra [1], Edlira Vakaj [1], Ahmed Karam Eldaly [2] and Mohammed M. Abdelsamea [2,*]

[1] School of Computing and Digital Technology, Birmingham City University, Birmingham B4 7XG, UK; abdullah.basaad@mail.bcu.ac.uk (A.B.); shadi.basurra@bcu.ac.uk (S.B.); edlira.vakaj@bcu.ac.uk (E.V.)
[2] Department of Computer Science, University of Exeter, North Park Road, Exeter EX4 4QF, UK; a.karam@ucl.ac.uk
* Correspondence: m.abdelsamea@exeter.ac.uk

Abstract: Metastatic breast cancer (MBC) continues to be a leading cause of cancer-related deaths among women. This work introduces an innovative non-invasive breast cancer classification model designed to improve the identification of cancer metastases. While this study marks the initial exploration into predicting MBC, additional investigations are essential to validate the occurrence of MBC. Our approach combines the strengths of large language models (LLMs), specifically the bidirectional encoder representations from transformers (BERT) model, with the powerful capabilities of graph neural networks (GNNs) to predict MBC patients based on their histopathology reports. This paper introduces a BERT-GNN approach for metastatic breast cancer prediction (BG-MBC) that integrates graph information derived from the BERT model. In this model, nodes are constructed from patient medical records, while BERT embeddings are employed to vectorise representations of the words in histopathology reports, thereby capturing semantic information crucial for classification by employing three distinct approaches (namely univariate selection, extra trees classifier for feature importance, and Shapley values to identify the features that have the most significant impact). Identifying the most crucial 30 features out of 676 generated as embeddings during model training, our model further enhances its predictive capabilities. The BG-MBC model achieves outstanding accuracy, with a detection rate of 0.98 and an area under curve (AUC) of 0.98, in identifying MBC patients. This remarkable performance is credited to the model's utilisation of attention scores generated by the LLM from histopathology reports, effectively capturing pertinent features for classification.

Keywords: LLM; GNN; XAI; node classification; MBC; extra trees classifier; univariate selection; random forest classifier; BERT

Citation: Basaad, A.; Basurra, S.; Vakaj, E.; Eldaly, A.K.; Abdelsamea, M.M. A BERT-GNN Approach for Metastatic Breast Cancer Prediction Using Histopathology Reports. *Diagnostics* **2024**, *14*, 1365. https://doi.org/10.3390/diagnostics14131365

Academic Editor: Dechang Chen

Received: 16 May 2024
Revised: 21 June 2024
Accepted: 25 June 2024
Published: 27 June 2024

1. Introduction

Metastatic breast cancer (MBC) represents a crucial challenge in the realm of oncology and is the main cause of death for patients with breast cancer. It needs a significant contribution to the burden of cancer-related mortality among women worldwide [1]. Despite advances in detection and treatment modalities, the metastatic spread of breast cancer remains a critical clinical challenge, necessitating innovative approaches to improve early identification and intervention [2]. Metastasis, the process by which cancer cells disseminate from the primary tumour site to distant organs or tissues, represents a crucial stage in cancer progression. Statistics underscore the gravity of metastatic breast cancer, with approximately 30% of early-stage breast cancer cases eventually progressing to metastatic disease [3]. Moreover, MBC accounts for the majority of breast cancer-related deaths, highlighting the urgent need for improved diagnostic and therapeutic strategies.

Cancer is a multifaceted disease that presents significant challenges in terms of diagnosis and management. Proper diagnosis is essential for devising effective treatment strategies and improving patient outcomes. Historically, histopathological examination has played a

central role in cancer diagnosis, providing valuable information on cellular and tissue-level changes associated with malignancies [4,5]. Despite the onset of molecular and genomic techniques, histopathology remains indispensable in cancer diagnosis and management, providing complementary information that contributes to a comprehensive understanding of the disease [6]. Integrating histopathological data with molecular profiles and imaging modalities enables a holistic approach to cancer prognosis, facilitating personalised treatment strategies tailored to individual patient characteristics [7]. Histopathology involves the microscopic examination of tissue specimens to assess their morphological characteristics, including cell morphology, tissue architecture, and the presence of pathological abnormalities such as tumour formation [8]. This analysis enables pathologists to identify the type of cancer, assess its aggressiveness or grade, determine its stage, and provide crucial prognostic information for treatment planning [9]. Over the years, histopathological techniques have evolved significantly, with advancements in staining methods, imaging technologies, and molecular diagnostics enhancing the accuracy and precision of cancer diagnosis [10].

Histopathological reports serve as indispensable tools in the prognosis and management of cancer, including metastatic breast cancer. These reports offer detailed insights into the histological characteristics of tumour specimens, providing crucial information regarding tumour type, grade, hormone receptor status, and the presence of metastasis. The comprehensive analysis of histopathological findings plays a pivotal role in guiding treatment decisions and prognostic assessments for patients with metastatic breast cancer [11].

Leveraging the power of large language models (LLMs), such as bidirectional encoder representations from transformers (BERT), presents a novel avenue for extracting meaningful insights from histopathology reports. Using sophisticated natural language processing techniques, LLMs can describe complex textual data, extracting relevant features and contextual information crucial to cancer prognoses [12]. Integrating LLMs into the prognostic pipeline holds promise for enhancing the accuracy and efficiency of metastatic breast cancer detection.

This work combines LLMs with graph neural networks (GNNs) and offers a robust framework for cancer prediction tasks, using embeddings that capture the semantic meaning of words and phrases in histopathological reports. This means that words with similar meanings or contexts are likely to have similar embeddings, even if they appear in different parts of the report, and GNNs can uncover hidden patterns and dependencies crucial for accurate disease prediction [12]. The application of GNNs in conjunction with LLMs represents a complementary approach to metastatic breast cancer detection, harnessing textual and structural information for improved prognostic accuracy [13]. Although this proposed system offers an initial step towards improving metastatic breast cancer prediction, further validation and refinement are imperative to confirm its efficacy and reliability. Subsequent investigations, including clinical trials and real-world validation studies, will be essential to assess the system's performance in diverse patient populations and clinical settings, ultimately paving the way for its integration into routine clinical practice. Thus, the main contributions of this work can be summarised as shown in Figure 1.

1. Integrating advanced natural language processing (NLP) techniques with graph neural networks (GNNs) and proposing a novel approach for comprehensive analysis of histopathology reports.
2. Using attention scores from LLMs to construct interpretable and context-aware graph representations.
3. Incorporating feature extraction methods to identify critical variables for predicting breast cancer metastases, which enhances the model's predictive accuracy and interoperability.
4. Providing clinicians with a robust approach that can provide valuable insights into the underlying data, thus facilitating more informed decision-making and improving patient outcomes.

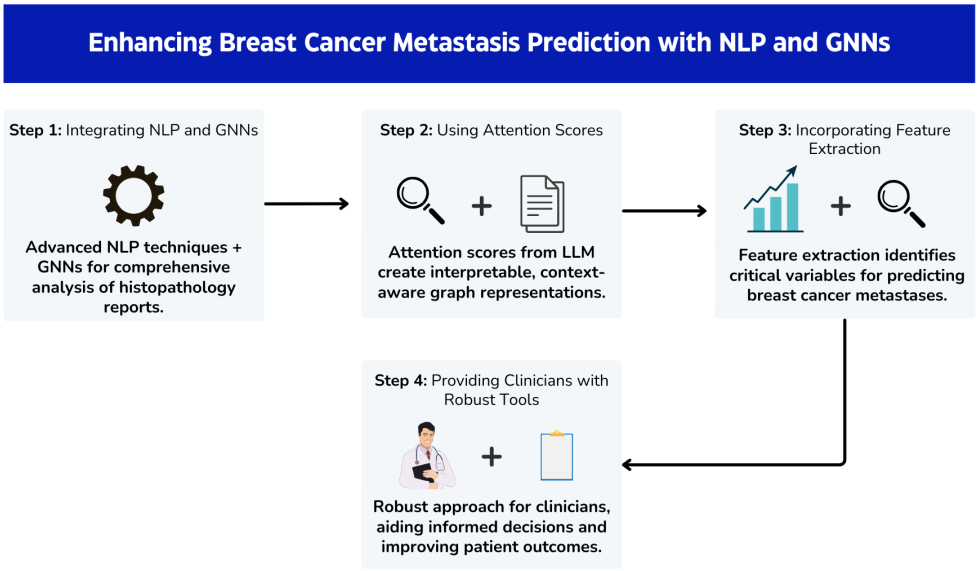


Figure 1. Enhancing breast cancer metastasis prediction with NLP and GNNs.

The remainder of this article is organised as follows. Section 2 provides an overview of relevant studies on metastatic diseases. Section 3 details the proposed methodology of this study, including the utilisation of bidirectional encoder representations from transformers (BERT) and their application in analysing and gaining insights from histopathology reports. This section also discusses the use of attention mechanisms to extract information related to specific entities within the dataset. Additionally, it covers the refinement of attention to facilitate graph construction and feature extraction processes. Section 3.7 presents the architecture of the BG-MBC model, including the neural network structures and various layers employed in the graph neural network. This section also discusses the model training process and its capability for interpretable predictions. Section 3.8 is dedicated to the integration of the BG-MBC model with a large language model (LLM) approach. Following that, Sections 3.9–3.11 discuss graph building, baseline models, and the architecture of the BG-MBC model, respectively. The results section, Section 4, includes data pre-processing, comparisons with existing approaches, and model evaluation. Finally, Sections 5 and 6 present the discussion and conclusions, respectively.

2. Related Work

In recent years, the application of machine learning (ML) models has gained considerable traction in the field of oncology, particularly in the context of predicting metastases and improving patient outcomes. Several studies have investigated the use of ML techniques to tackle the challenges associated with metastatic disease. One notable study by Esteva et al. [14] employed deep learning algorithms to analyse histopathology images and predict the presence of metastatic breast cancer in lymph nodes with high accuracy. The study demonstrated the potential of convolutional neural networks (CNNs) in detecting metastatic disease from digitised tissue slides, offering a promising approach for automated histopathological analysis. Liu et al. [15] proposed an automated framework using a convolutional neural network (CNN) to detect and localise tumours in gigapixel microscopy images, achieving state-of-the-art results on the Camelyon16 dataset. The method detects 92.4% of tumours at eight false positives per image, outperforming the previous best (82.7%) and human pathologists (73.2%). They achieved over 97% AUC scores and identified mislabelling in the dataset, potentially reducing false negative rates significantly. In addition to histopathological analyses, researchers have explored the utility of clinical data and molecular markers to predict metastatic outcomes. For example, Li et al. [16] developed a predictive model for distant metastasis in breast cancer patients using a combination

of clinical features and gene expression profiles. The study achieved accurate prognostic predictions by integrating machine learning algorithms with gene expression data, allowing the early identification of patients at high risk of metastasis. Another study [17] evaluates various predictive models for determining the time to tumour recurrence in breast cancer patients, achieving an accuracy of up to one year. Analysing 198 patients, it was found that 40% were predicted to experience tumour recurrence within the first year of diagnosis. The study employed several classification models, including spectral clustering, DBSCAN, and k-means, alongside prediction models such as support vector machines (SVMs), decision trees, and random forests. The results highlighted the efficacy of these machine learning techniques, with SVMs achieving the highest accuracy at 78.7% in predicting the time to tumour recurrence or patient recovery. In [18], the authors proposed a multimodal deep learning method that integrates whole-slide H&E images (WSIs) and clinical data to assess relapse and metastasis risk in HER2-positive breast cancer patients. The images, resized to 512 × 512 pixels, were analysed using a deep convolutional neural network (CNN) to extract features, which were combined with clinical data. The model achieved an AUC of 0.76 in two-fold cross-validation and 0.72 on independent testing with TCGA data, demonstrating its potential for prognosis prediction despite demographic and experimental differences.

Previous studies have shown significant limitations and deficiencies, highlighting the need for a more robust approach. For instance, one study analysed 198 breast cancer patients [17] and used various machine learning techniques, such as spectral clustering, DBSCAN, k-means, SVMs, decision trees, and random forests, achieving a maximum accuracy of 78.7% with SVMs for predicting tumour recurrence within one year. However, this study was limited by a small sample size and moderate prediction accuracy, indicating the necessity for more reliable and scalable methods.

Another study used a multimodal deep learning approach [18] that integrated whole-slide H&E images (WSIs) and clinical data to predict relapse and metastasis in patients with HER2-positive breast cancer. Although it achieved promising AUC scores of 0.76 in cross-validation and 0.72 on independent TCGA data, there were still challenges in its generalisability across different patient demographics and experimental setups. These studies did not fully exploit the potential of integrating advanced language models and graph neural networks to enhance predictive power and interoperability.

Here, our proposed BG-MBC model addresses these limitations by combining large language models (LLMs), specifically the BERT model, with graph neural networks (GNNs) to analyse histopathology reports for metastatic breast cancer (MBC) prediction. By leveraging BERT embeddings to capture semantic information and constructing nodes from patient medical records, our model integrates graph information derived from BERT. It employs univariate selection, extra trees classifier, and Shapley values to identify the most crucial features of 676 generated embeddings, enhancing predictive accuracy.

3. Materials and Methods

Our proposed approach leverages bidirectional encoder representations from transformers (BERT) to extract attentions and embeddings from histopathology reports, as shown in Figure 2. The attentions are utilised to define graph edges, capturing the relationships between different entities acted by medical numbers. Meanwhile, the embeddings extracted from the reports serve as node features within the graph. Combining these attentions and embeddings, a graph was constructed to represent the histopathology data. Subsequently, this graph representation was passed into the BERT graph metastatic breast cancer (BG-MBC) model. This model is designed to handle graph-structured data and is capable of effectively incorporating the complex relationships captured by the graph edges and node features. By utilising the attentions and embeddings output by BERT as inputs to the BG-MBC model, the aim is to enhance the performance of breast cancer metastasis prediction, leveraging the rich contextual information provided by the histopathology reports.

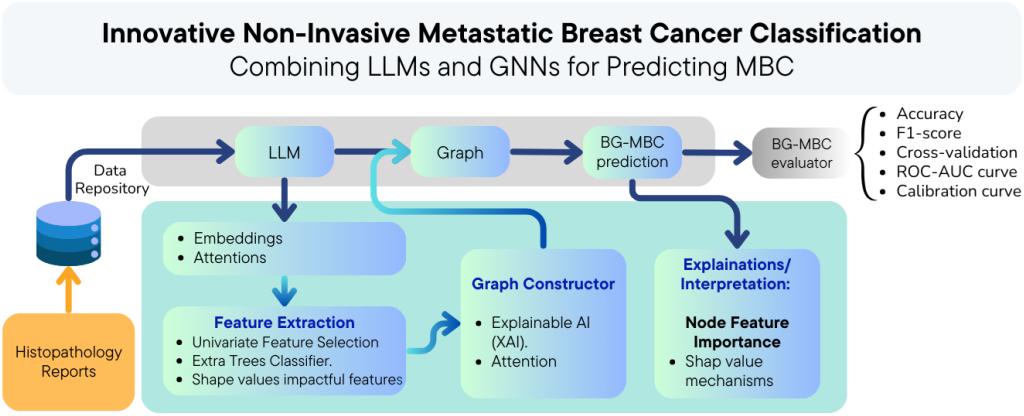


Figure 2. The BG-MBC model employs natural language processing (NLP) to extract crucial features from histopathology reports. Attention mechanisms construct a graph from these embeddings, which is then processed by a graph neural network (GNN). Shapley values are used to interpret and explain its predictions.

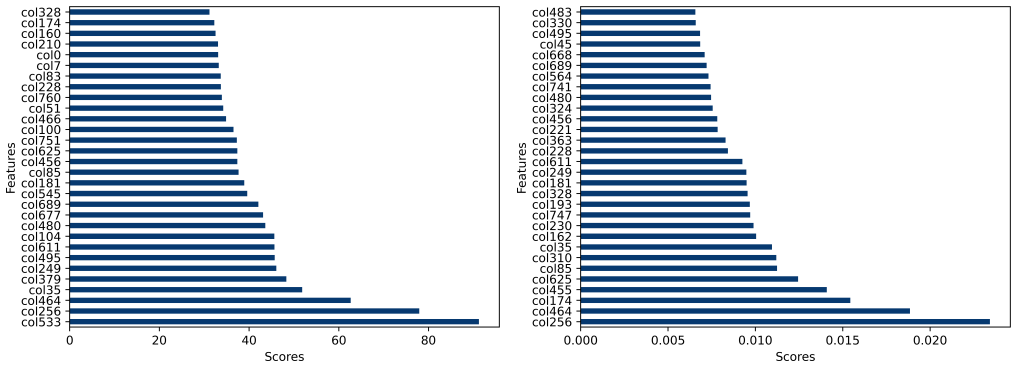
3.1. Dataset

Data were acquired from BIACH and RI in a semi-structured Excel file comprising two main columns: patient identifiers (medical record number—MR No) and histopathology reports (Hist_report). The dataset consists of 25,652 entries in .csv format (raw text). The Hist_report column includes a variety of pathological observations, such as clinical details, specimen descriptions, microscopic observations, impressions, and gross findings [19]. The histopathology diagnoses pertain to metastasis, biopsy, and lymph node reports. This paper focuses on detecting metastasis in patients among the different types of diagnoses.

3.2. Data Pre-Processing

Various data repositories have been used, containing medical records, histopathology reports, and corresponding diagnostic results. These datasets were merged into a single file to facilitate streamlined analysis. Due to the unstructured nature of the original datasets, which span multiple CSV files, it became imperative to consolidate the data into a single file for streamlined analysis. Leveraging Google Collab during the pre-processing stage, the data are formatted into a tabular structure, facilitating the merging process to consolidate all necessary information into one comprehensive dataset. Following this, extensive cleaning procedures were executed, which involved feature selection and normalisation. Feature selection aids in identifying the most informative variables for metastasis prediction, thereby reducing dimensionality and enhancing model performance. On the other hand, normalisation is crucial in ensuring that features are uniformly scaled, mitigating bias issues from differing magnitudes across variables.

The feature selection process employed three distinct algorithms: univariate selection, extra trees classifier, and SHAP values. Univariate feature selection is a technique widely used in machine learning and data analysis, identifying the most relevant feature indices from a dataset based on their individual statistical properties, as illustrated in Figure 3a. Additionally, the extra trees classifier is utilised to extract important features from the dataset, providing insights into the relative importance of each feature and their impact on model predictions. Figure 3b displays the feature importance scores obtained from the extra trees classifier, highlighting each feature's contribution to predicting the target variable.



(**a**) (**b**)

Figure 3. (**a**) Results of univariate feature selection and (**b**) feature importance analysis using extra trees classifier.

Furthermore, Shapley additive explanations (SHAP) values are leveraged to emphasise the feature selection process. SHAP values offer a comprehensive understanding of feature importance in machine learning models, providing insights into the contribution of each feature to individual predictions. Unlike traditional methods, SHAP values offer a nuanced perspective on feature importance, enabling practitioners to identify influential features and their impact on specific predictions. This level of granularity enhances interpretability and trust in model predictions, ultimately improving decision-making processes [20,21]. Figure 4 visualises the influence of each feature on predictions, playing a crucial role in interpreting the results.

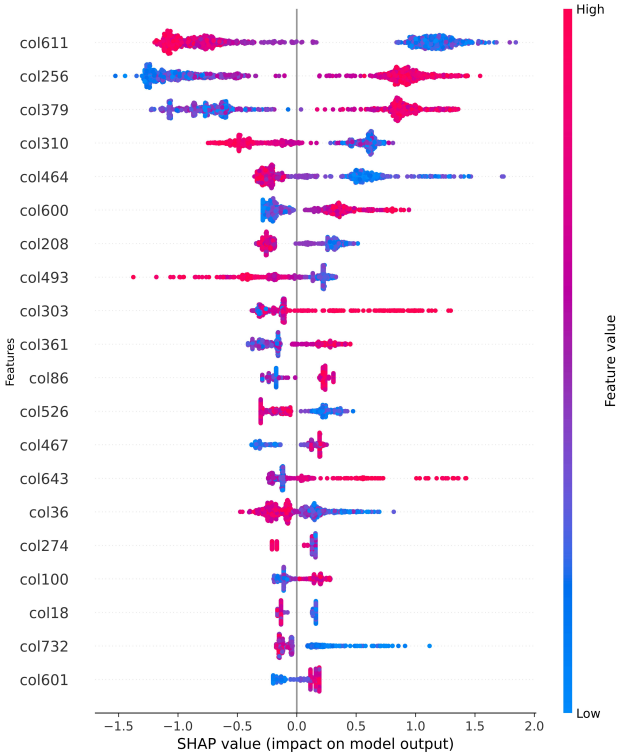


Figure 4. Shapley additive explanations (SHAP): the SHAP values most impactful on features of the model.

By utilising these approaches, the selection of the most relevant features is ensured. Out of the 767 columns generated as text embeddings, the first 30 most-relevant columns were extracted for further processing, as shown in Table 1.

Table 1. Experiment findings on feature selection.

Illustration	MBC Cross-Sectional Data
Size of the training set	813
Size of the test set	187
Total class count	2
Total number of features in the dataset	767
Number of training data features after correlation	30
Number of nodes	1000
Number of node features in the optimal set	30
Number of edges	4984

3.3. Bidirectional Encoder Representations from Transformers

BERT is a transformer-based model developed by Google AI that is designed to generate deep bidirectional representations of input text. The model consists of an encoder architecture that processes input sequences in both forward and backward directions, allowing it to capture contextual information effectively [12]. BERT produces two primary outputs: embeddings and attention scores. In embeddings, BERT generates contextual word embeddings for each token in the input sequence. These embeddings capture the semantic meaning of each token in the context of the entire input sequence. In addition, BERT also produces attention scores that indicate the importance or relevance of each token to every other token in the input sequence. These attention scores are represented as a matrix A, where A_{ij} denotes the attention weight assigned to token i when processing token j.

3.4. Tokenisation and Embedding Extraction for Histopathology Reports

To analyse histopathology reports using BERT, the following steps are typically followed. Histopathology reports are first tokenised into individual words or subword units using BERT's tokeniser. This process involves breaking down the text into tokens and mapping them to corresponding indices in BERT's vocabulary. After tokenisation, the tokenised sequences are fed into the BERT model, which generates embeddings for each token in the input sequence, capturing the contextual information of the entire report. The embeddings from the last hidden state layer of BERT are often used for downstream tasks as they contain rich contextual information. Additionally, attention scores can be extracted to understand the relationships between tokens in the report. Once embeddings are obtained for all tokens in the report, the mean embeddings are calculated for all tokens. Mathematically, this can be represented as

$$\text{Mean Embedding} = \frac{1}{N}\sum_{i=1}^{N} E_i, \tag{1}$$

where N is the total number of tokens in the report and E_i represents the embedding for token i [12,22].

3.5. Building Attentions per Node in the Dataset

Considering the varying lengths of histopathology reports, padding and truncation parameters are applied to represent them as fixed-size feature vectors. Padding ensures that shorter sequences match the length of the longest sequence by adding special tokens, while truncation cuts longer sequences to a predefined length. This standardises the length of all sequences [23]. Each report is processed to extract 767 features, utilising the embeddings from the last hidden state layer of BERT, which encapsulate rich contextual

information [12,22]. These embeddings for each token are concatenated or aggregated to form a feature vector of length 767. Mathematically, the feature vector F_i for token i can be represented as

$$E_i = [E_{i1}, E_{i2}, \ldots, E_{iN}], \tag{2}$$

where E_{ij} represents the (j-th) dimension of the embedding for token i, and N is the total number of dimensions in the embedding space (e.g., 768 for BERT-base).

3.6. Refining Attention Mechanisms

BERT typically utilises multiple attention heads to capture different aspects of contextual relationships between tokens. The attention scores from each attention head can be aggregated by taking the mean across all heads to obtain a single attention score for each token. This process is repeated for each token in the report, resulting in a matrix of attention scores with dimensions (number of tokens) $\times$ 767. Mathematically, the attention score A_{ij} for token i and dimension j can be calculated as

$$A_{ij} = \frac{1}{H} \sum_{h=1}^{H} A_{ij}^{(h)}, \tag{3}$$

where H is the total number of attention heads, and $A_{ij}^{(h)}$ is the attention score for token i and the dimension j from attention head h.

3.6.1. Graph Construction

The initial phase of graph construction entails calculating the edges connecting nodes based on predefined criteria. This foundational step holds utmost significance, laying the concept for establishing meaningful relationships among medical records within the network by using the combined attributes and their attentions (the attention mechanism in the BERT output refers to the ability of the model to selectively focus on relevant parts of the input text when making predictions. It assigns different weights to each word or token in the input sequence, allowing the model to prioritise important information and ignore irrelevant parts. This attention mechanism helps BERT achieve better understanding and performance in natural language processing tasks) [24] contained within each node to explore the complexity of the connections that define the adjacencies between nodes in the graph. By computing attention across all inputs within each observation using Algorithm 1, the ascertained probabilities are assigned to each node's classification as either 0 (no metastasis) or 1 (metastasis patient).

Algorithm 1 Calculate the output for report

Input: text ▷ List of all Histopathology reports
Input: df ▷ Data frame
Output: attentions, embeddings

```
for each text ∈ df['Report'] do
    add text tokinisation to the output
    add text attention to attentions
    add text embedding to embeddings
end for
return attentions, embeddings
```

By computing attention scores for all features in each node using Algorithm 1, the probabilities associated have been obtained by classifying each node as 0 (no metastasis) or 1 (metastasis). Additionally, it examines the connections between neighbouring nodes based on their respective attention scores using Algorithm 2. For instance, it indicates that a node should be linked to a group where the average attention score is either less than or equal to a predefined threshold. The threshold is fine-tuned experimentally by

systematically varying its value and evaluating the model's performance across different thresholds. This involves selecting a range of threshold values and then training and testing the model with each value to observe how it affects the model's performance metrics, such as accuracy, precision, recall, or F1 score. Finally, by determining all of these criteria, the adjacency metric building comes into place in Algorithm 3.

Algorithm 2 Calculate the attentions per node

Input: attentions ▷ attentions
Output: attention scores per node

for each **attention matrix** ∈ **attentions do**
    Stacked all the layers per node
    Find the average attention per node
end for
return attention scores per node

Algorithm 3 Get node adjacents

Input: df ▷ Data frame
Input: attention ▷ Node attention
Input: cls ▷ Node class
Input: node threshold = 200
Input: node connections = 5
Output: adjacentsArray

for each **row** ∈ **df do**
    if cls = 0 **and** result = 0 **then**
        add attention to the zeroes dictionary
    end if
    if cls = 1 **and** result = 1 **then**
        add attention to the ones dictionary
    end if
end for
if cls = 0 **then**
    Sorting zeroes dictionary
    dic ⟸ zeroes dictionary
else
    Sorting ones dictionary
    dic ⟸ ones dictionary
end if
for each i, val in dic **do**
    if (attention - val) $\leq$ node threshold **then**
        if adjacents $\geq$ node connections **then**
            Exit
        end if
        add the node to the adjacency matrix
        increment node counter
        adjust the total mean of nodes
    end if
end for
return adjacentsArray

Based on these evaluations, the threshold value that yields the best overall performance or achieves the desired balance between different evaluation metrics is chosen as the optimal threshold. This threshold serves to establish connections between entities based on their similarities, facilitating the grouping of entities according to their attention levels. This threshold, which is an integer parameter, is integrated into the node grouping process. To determine whether a new node should connect to a group, a comparison should be

made to find out the difference between the group's average attention scores and the new node's attention score against the threshold. When the disparity is within or equal to the set threshold and the cumulative count of nodes within the group does not exceed the permitted group size, the new node is linked to the existing group, as illustrated in Figure 5.

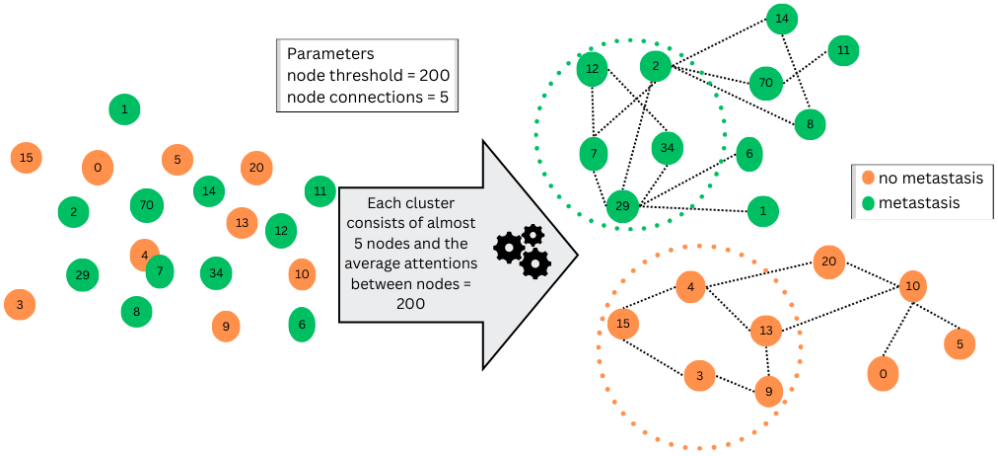


Figure 5. This diagram depicts connecting a node by considering the information from its neighbouring nodes. The figure highlights that the cluster information is influenced by both the node threshold value and the requirement that the number of nodes in each cluster not exceed the node connections parameter.

3.6.2. Feature Extraction

The BERT model has revolutionised the analysis of textual data, including histopathology reports in the domain of medical diagnosis. The process involves tokenising the raw text of histopathology reports, converting it into a format that can be understood by machine learning models. BERT, being a state-of-the-art language model, is particularly effective in capturing the contextual information and semantic meaning of the text [12].

After tokenising the histopathology reports with BERT, the model produces embeddings for each token, encapsulating their semantic context. These embeddings serve as input features for subsequent tasks, such as classification. In the context of histopathology reports, these embeddings convert the text data into a high-dimensional space, capturing the intricate relationships and patterns. These embeddings are particularly beneficial for predicting BCM. However, given the substantial number of embeddings generated (in this instance, 767 columns), spotting the most informative features to guarantee precise prediction is essential. This is where feature selection techniques come into play. Three feature selection algorithms are employed, namely univariate selection, extra trees classifier, and the Shapely values, which are the most impactful features on the model, with many other methods that could be used like recursive feature elimination [25]. These algorithms analyse the embeddings generated by BERT and select the most relevant features that contribute significantly to the predictive performance of the model. The aim is to reduce the dimensionality of the feature space while retaining the most discriminate information. The selected features are then used to build the BG-MBC model. In this model, each patient number acts as an entity in the graph and is considered a node in the graph, and the selected features serve as node attributes. The attentions obtained from BERT are utilised to build the adjacency matrix, capturing the relationships between different medical records based on their semantic similarities and contextual dependencies, as explained in the algorithms mentioned above. This approach effectively leverages the power of BERT embeddings and attention mechanisms to extract meaningful information from histopathology reports and build a predictive model for cancer metastasis. By incorporating GNNs, the model can capture complex relationships and dependencies present in the data, leading to improved predictive performance [12].

3.7. BG-MBC Model

The BG-MBC model architecture is a type of neural network that is specifically designed to operate on graph-structured data. It leverages GNN techniques to process and analyse the relationships and dependencies present in the data. The BG-MBC model architecture is summarised in Figure 6.

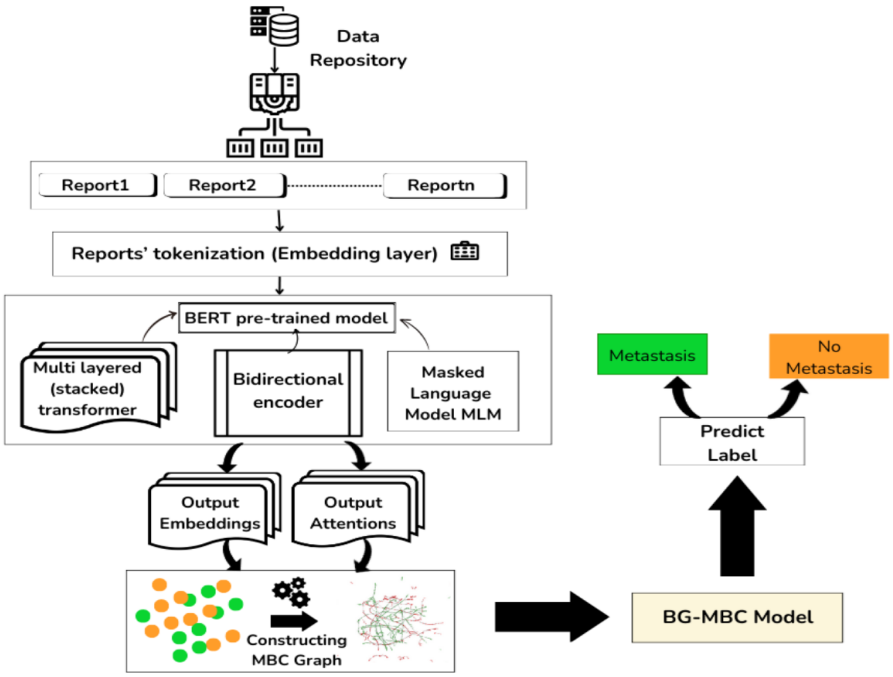


Figure 6. The diagram illustrating the process of utilising a BERT model to generate embeddings and attentions from histopathology reports. The embeddings and attentions are then used to construct a graph representation of the data. This graph is then passed to a BG-MBC model, which integrates GNN techniques for BCM prediction based on the structured data derived from the BERT model outputs.

3.7.1. Input Layer

The input to the BG-MBC model consists of a graph representation of the data, where each node in the graph corresponds to a data point (e.g., a medical record in the context of metastasis breast cancer diagnosis). The features associated with each node include information derived from embeddings and attentions obtained from pre-processing steps, such as those generated by a BERT model from histopathology reports.

3.7.2. Graph Convolutional Layers

The core component of the BG-MBC model is the graph convolutional layers. These layers apply convolutional operations directly to the graph structure, allowing the model to aggregate information from neighbouring nodes. This enables the model to capture the local structure and dependencies within the graph, essential for tasks such as node classification or regression. The node classification layer of the BG-MBC model is crucial in the context of breast cancer prognosis. At the output of the model, this layer is responsible for assigning a label to each node in the graph, which suggests a positive or negative diagnosis of breast cancer metastasis. Typically, the node classification layer employs softmax activation to output probability distributions over the possible classes. During this stage, the model utilises two convolutional layers to aggregate information from node neighbours and nodes based on various options. In the aggregation phase, the model offers three options for updating node state. One option involves computing the sum of elements from a tensor along segments specified by a segment indices tensor, resulting in an aggregated message of type "sum". Alternatively, the model can calculate the mean or

maximum of elements for aggregation. The aggregated message is then used to update node information.

3.7.3. Output Layer

The output layer serves as the concluding segment of our GNN-based node classification model, refining the learned representations and producing the ultimate node predictions. It consists of three different layers: batch normalisation, dropout, and dense with GELU activation. The output layer enhances the model's predictive capabilities.

3.7.4. Training Procedure

The BG-MBC model is trained using a supervised learning approach, where the model learns to predict the correct labels for the nodes in the graph based on the input features and the graph structure. The training process involves optimising a loss function, such as sparse categorical cross-entropy loss, using Adam optimisation.

3.7.5. Interpretability

One of the key advantages of the BG-MBC model architecture is its interpretability. By leveraging Shapley values, BG-MBC enhances the transparency and explainability of its predictions, allowing the model to identify the importance of various features in predicting metastasis. This insight enables clinicians and researchers to understand the key factors influencing the predictions, thus providing a deeper understanding of the underlying drivers of metastatic breast cancer. Furthermore, analysing the learned parameters and the information propagation through the graph reveals the features and relationships that contribute most significantly to the model's predictions. This level of interpretability is crucial for metastasis prediction as it ensures clinical acceptance and trust by clarifying the reasoning behind the model's decisions.

3.8. BG-MBC and LLM (BERT) Models Integration

Here, the focus is on the integration of the large language model (LLM) BERT with BG-MBC to develop a comprehensive framework for analysing histopathology reports and predicting metastatic breast cancer, as described in Figure 6.

The last hidden layer in neural network models, especially in architectures like BERT, holds significant importance in various natural language processing tasks. In the context of analysing histopathology reports for predicting metastatic breast cancer, understanding the role of the last hidden layer becomes crucial for extracting meaningful insights [12,26]. In particular, the last hidden layer captures rich semantic information about the input text. It encodes contextualised representations of words and phrases, essential for understanding medical reports' nuances and intricacies. This semantic representation forms the foundation for extracting relevant features and identifying patterns associated with metastatic breast cancer [27].

On the other hand, by analysing the activations and embeddings in the last hidden layer, it would be possible to extract important features that contribute to the classification of histopathology reports [28]. These features may encompass specific keywords, phrases, or contextual cues indicative of cancer metastasis. Extracting meaningful features from this layer enables the model to focus on the most relevant aspects of the input data [29]. The extraction of relevant features can be represented as

$$F = \text{ExtractFeatures}(E), \tag{4}$$

where E denotes the embeddings generated by the BERT model for a histopathology report.

Moreover, BERT and similar models incorporate attention mechanisms, allowing them to dynamically weigh the importance of different words and context tokens in the input sequence. The attention scores generated in the last hidden layer highlight the salient aspects of the input text, guiding the model's decision-making process. Understanding these attention patterns helps in identifying the key components of histopathology reports

that contribute to the prediction of metastatic breast cancer [22]. The calculation of mean attention scores across layer heads is given by

$$\bar{A} = \frac{1}{H} \sum_{i=1}^{H} A_i, \tag{5}$$

where A represents the attention scores obtained from the BERT model, and H is the number of layer heads. On the other hand, in order to calculate the attention means per node that is considered, the below calculation can be applied:

$$\text{Attention}_{\text{node}} = \frac{1}{H} \sum_{i=1}^{N} \text{Attention}_{\text{layer}_i}, \tag{6}$$

where $\text{Attention}_{\text{node}}$ represents the attention mean score for a particular node, H is the number of layers, and $\text{Attention}_{\text{layer}_i}$ denotes the attention score for the i-th layer.

Finally, in many cases, fine-tuning or transfer learning techniques are applied to pretrained models like BERT. The last hidden layer serves as a crucial component during these processes as it contains the most refined representations learned from extensive pretraining on large texts of histopathology reports. Fine-tuning allows the model to adapt its parameters to the specific task of predicting cancer metastasis, leveraging the knowledge encoded in the last hidden layer [30,31].

We analyse the histopathology reports using the BERT model to generate embeddings, from which relevant features are extracted to provide insights into the contents of the reports. This involves identifying key variables or dimensions within the embeddings that are indicative of metastatic breast cancer. Additionally, the extraction of important attention involves focusing on the last hidden state layer of the BERT model to extract important attention scores from each report. For each embedding variable, calculating the mean across all layer heads yields a comprehensive measure of attention for each report. This process enables us to quantify the relevance of different parts of the histopathology reports.

3.9. Graph Building

We use the embeddings and attention scores to construct a unidirectional and homogeneous graph. Each medical record serves as a node in the graph, with edges representing relationships based on attention scores between reports. This graph encapsulates the structural and semantic information present in the data. The construction of the graph involves defining nodes and edges based on attention scores.

$$G = (V, E), \tag{7}$$

where V represents the set of nodes (medical records) and E represents the set of edges (connections based on attention scores).

3.10. Baseline Model

The baseline model comprises two feedforward networks implemented using Keras deep learning framework. Each network consists of a Keras sequential model composed of three stacked layers: batch normalisation, dropout, and dense layers with GELU activation function. These network layers are interconnected via skip connection layers. The baseline model takes input data, which are fed into the first feedforward layer, processed through the Keras sequential model, and then passed to the second feedforward network. The two networks are linked together by skip connection layers. This process is iterated four times, resulting in a Keras model with output logits.

3.11. BG-MBC Model Architecture

The output graph is passed into the BG-MBC model, which is a GNN architecture designed for node classification tasks. In Figure 7, the BG-MBC model begins with a

pre-processing layer (input layer), which includes operations such as batch normalisation, dropout, and dense layers. These operations help standardise and prepare the graph data for further processing. Following the pre-processing layer, the BG-MBC model incorporates two graph convolutional layers. These layers aggregate and update node messages by considering information from neighbouring nodes and learning representations that capture the graph's structure and semantics. Finally, the output from the graph convolutional layers passes through a post-processing layer (output layer), which includes batch normalisation, dropout, and dense layers. These operations refine the node representations and prepare them for the final node classification task.

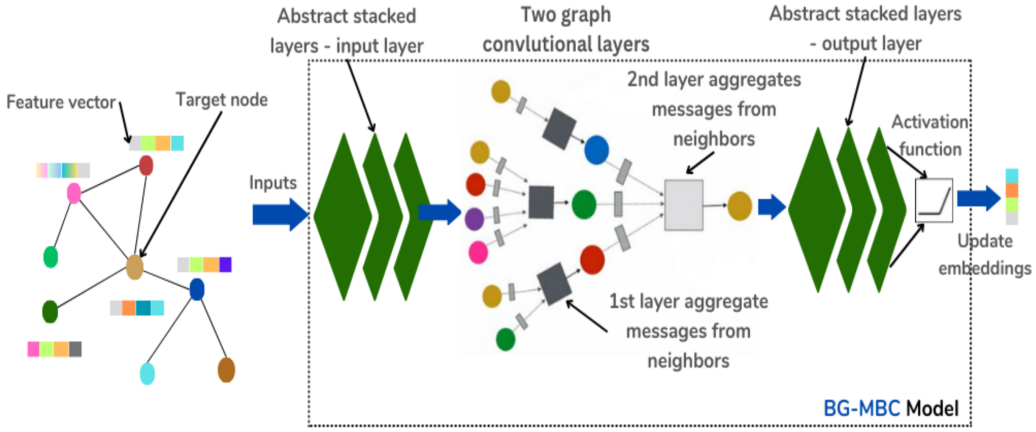


Figure 7. BG-MBC model architecture. The model consists of three main layers: the first and last layers are feedforward network layers, each containing three stacked layers. In between these feedforward layers, there are two graph convolutional layers responsible for generating and aggregating messages, which are then used to update the target node embeddings.

$$Y = \text{BG-MBC}(X). \tag{8}$$

where X represents the input data (graph) and Y represents the output (node classifications).

The BG-MBC model consists of pre-processing, graph convolutional, and post-processing. The operation of the graph convolutional layers involves aggregating and updating node messages as follows:

$$H_{new}^{(l)} = \text{GCNLayer}(H^{(l)}), \tag{9}$$

where $H^{(l)}$ represents the node representations at layer l and $H_{new}^{(l)}$ represents the updated node representations. The post-processing layer applies additional transformations to the node representations before the final classification, as follows:

$$H_{final} = \text{PostprocessingLayer}(H_{new}^{(L)}), \tag{10}$$

where H_{final} represents the final node representations.

By integrating LLM BERT embeddings and attention scores with BG-MBC, a robust framework was created for analysing histopathology reports and predicting metastatic breast cancer. This approach leverages the strengths of both techniques and provides a holistic understanding of the underlying data, leading to improved diagnostic accuracy and patient outcomes.

4. Experimental Results

The experimental results showcase the effectiveness of the approach applied in predicting metastasis in breast cancer patients. This section encapsulates the data sources utilised and the diverse processes and tools employed to prepare the data for subsequent analysis.

4.1. Comparison with Existing Approaches

This work compares the proposed approach with three existing methods in the literature, in addition to the baseline model discussed in the previous section. The first method, which was proposed by Ting et al. [32], focused on convolutional neural network (CNN) enhancement for breast cancer classification. They utilised CNN architecture for classification tasks but did not explicitly incorporate graph-based representations or attention mechanisms. The second method, proposed by Thwin et al. [33], involved an attention-based ensemble network for breast cancer classification. They employed ensemble techniques along with attention mechanisms to improve classification performance. The third method, proposed by Mullooly et al. [34], applied CNNs to analyse breast biopsies and correlate tissue characteristics with mammographic breast density. While attention mechanisms were utilised in the second method, our method specifically focuses on leveraging BERT embeddings and attentions to construct a graph representation of histopathology reports. This approach enables a deeper understanding of the relationships between entities within the data.

4.2. Performance Evaluation

The dataset is partitioned into training (80%), and testing (20%). The Lambda Callback package is utilised to determine the optimal learning rate value. By monitoring loss values during training with various learning rates, the best learning rate value is identified in Table 2. The BG-MBC model parameters were determined through a systematic approach involving empirical testing, cross-validation, and best practices from the existing literature. Each parameter was selected to optimise the model's learning efficiency while minimising overfitting and ensuring robust generalisation to new data.

Table 2. Values of BG-MBC model hyperparameters.

Model Hyperparameters	Value
hidden units	[34, 34]
learning rate	0.01
dropout rate	0.2
num epochs	300
batch size	128
k-fold	3

To determine the optimal learning rate, the Lambda Callback package was utilised to record loss values during training. By analysing the resulting graph, the optimal learning rate was identified as the rate at which the loss decreased most rapidly before diverging or oscillating, which was between 1.2×10^{-2} and 1.3×10^{-2}.

For the dropout rate, the literature commonly suggests starting with values between 0.2 and 0.5. Empirical testing within this range showed that a dropout rate of 0.2 provided the best balance between model complexity and performance.

Early stopping was employed to prevent overfitting by halting the training process when the model's performance on a validation set ceased to improve. This ensures the model does not over-train on the training data and maintains its ability to generalise to new data.

Additionally, the model included two unique parameters: node threshold and node connections, which manage the relationships among nodes. These parameters were fine-tuned through trial and error to achieve optimal performance.

Finally, BG-MBC performance is assessed by using a set of evaluation criteria, including receiver operating characteristic (ROC) graphs between true positive rate and false positive rate, area under ROC curve (AUC), F1 scores, accuracy, and the cross-validation metric.

Overall, this comprehensive approach to parameter selection, including the use of cross-validation and empirical testing, ensured that the BG-MBC model achieved high accuracy and generalisability in predicting metastatic breast cancer.

4.3. Results

Table 3 shows the performance of the proposed approach and existing methods for metastatic breast cancer prediction using histopathology reports. It could be observed that the proposed approach, BG-MBC, achieves the best performance compared to existing methods. In particular, the BG-MBC achieves an accuracy score of 0.98. The reason why it outperformed existing methods is that it incorporates advanced graph-based techniques, such as GNNs and large language models, in constructing the graph. Moreover, it effectively captures the relationships and dependencies within the data. The model's superior performance suggests that it can better model complex patterns and make more accurate predictions compared to existing methods. On the other hand, the baseline model achieves an accuracy score of 0.94. Although still achieving decent performance, it is short compared to BG-MBC, indicating that there is room for improvement in terms of capturing the underlying data structure and relationships. CNNI_BCC [32] achieves an accuracy score of 0.94, which is slightly lower than BG-MBC but still competitive. On the other hand, the model of Thwin et al. [33] achieves an accuracy score of 0.90, indicating that it may not capture the underlying patterns as effectively as BG-MBC or CNNI_BCC. Finally, the deep convolutional neural network model [34] achieves an accuracy score of 0.93, which is slightly lower than BG-MBC but still relatively good.

Table 3. Detailed evaluation metrics breakdown of BG-MBC and existing methods. Best results are highlighted in bold, and second-best results are underlined. All scores given as '-' are not provided in their sources.

Models	AUC Score	F1 Score	Balanced Accuracy	Cross-Validation
BG-MBC	**0.98**	**0.98**	**0.98**	**0.99**
Baseline model	0.90	0.90	<u>0.94</u>	<u>0.92</u>
CNNI_BCC [32]	0.90	-	0.90	-
DL-CNN [33]	0.91	-	0.90	0.89
DCNN [34]	<u>0.96</u>	-	0.93	-

Figure 8 shows the BG-MBC calibration curve. The calibration curve is a valuable tool used in machine learning to assess the reliability of the predicted probabilities of a model. It compares the predicted probabilities of a binary classifier with the actual observed probabilities. Calibration curves offer information on the consistency and accuracy of model predictions, helping to identify potential biases or overconfidence. They are particularly useful in applications where accurate probability estimates are essential, such as medical diagnoses or risk assessments [35]. As a result, Figure 8 suggests that the model's probability estimates are trustworthy and can be used confidently for the metastatic breast cancer task and to make decisions based on probabilities. The alignment indicates that the model's predicted probabilities are reliable and well-calibrated. When using these probabilities to make decisions or set thresholds, it is expected that the model's predictions are accurate and representative of the true underlying probabilities.

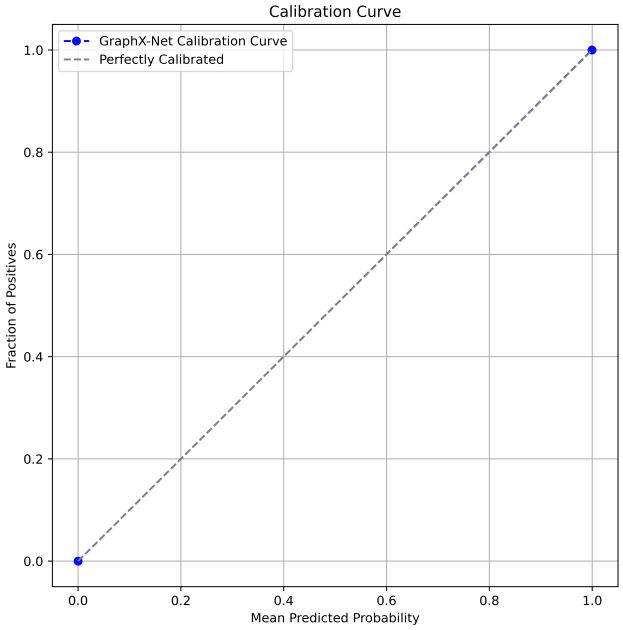


Figure 8. BG-MBC calibration curve demonstrating the alignment with the 45-degree diagonal line, indicating perfect calibration between predicted probabilities and actual outcomes.

5. Discussion

The combination of BG-MBC and large language models (LLMs) like BERT offers a promising approach for gaining insights from histopathology reports in the context of metastatic breast cancer diagnoses. By leveraging the strengths of both techniques, this integration enhances our understanding of the underlying data and improves predictive accuracy. The BERT model is capable of capturing the semantic meaning of each token in the context of the entire input sequence (histopathology reports) by utilising the attention scores obtained from the BERT model and extracting valuable information about the relationships between different tokens and phrases within the reports. The attention scores from the BERT model serve as a foundation for constructing the graph representation of the data. By considering the attention weights between tokens, it would be easy to identify the important connections and dependencies within the text. This graph representation allows us to capture the complex relationships between different features and their impact on the diagnosis of metastatic breast cancer. Incorporating node threshold attributes and connections in building the graph determines the significance of each node, while the connections between nodes reflect the relationships and dependencies between reports. By setting appropriate thresholds and establishing connections based on attention scores, the graph accurately represents the underlying structure of the data. The embeddings generated by the BERT model provide a rich representation of the semantic meaning of each token in the histopathology reports by using these embeddings as node features in the graph and incorporating the contextual information captured by the LLM into the GNN model. This enhances the model's ability to understand the relationships between different features and make accurate predictions. As a limitation of this work, histopathology reports offer a static view of tissue characteristics, potentially missing dynamic changes crucial for accurate diagnosis. Moreover, interpretation can be subjective, leading to inconsistencies among pathologists and potential misclassification. Consequently, integrating histopathology data with other clinical information could address some of these limitations, enhancing the accuracy and reliability of the classifier.

6. Conclusions

Using histopathology reports to build a machine learning classifier for metastatic breast cancer (MBC) diagnosis offers several potential benefits. First, these reports provide

detailed and standardised information about tissue samples, making them valuable for diagnosis. Second, leveraging histopathology data with advanced machine learning algorithms has the potential to achieve high accuracy in prediction tasks. In this paper, we introduce a hybrid mode using graph neural network (GNN) and bidirectional encoder representations from transformers (BERT) for MBC prediction, which we called BG-MBC, to analyse histopathology reports. By leveraging the complementary strengths of both techniques, this approach can improve our understanding of the disease and contribute to more effective diagnostic and treatment strategies in the medical field. In this study, we introduce a novel hybrid model, BG-MBC, which integrates graph neural networks (GNNs) and bidirectional encoder representations from transformers (BERT) to analyse histopathology reports for MBC prediction. The BG-MBC model leverages the unique strengths of both GNN and BERT to improve our understanding of metastatic breast cancer and foster more effective diagnostic and treatment strategies.

The following is a summary of the novelty and practical impact of the work:

1. Combining GNN and BERT: The integration of GNN and BERT is innovative in the context of medical data analysis. BERT excels at understanding the semantic relationships within text data, while GNN is adept at capturing the structural relationships in graph-structured data. By combining these two approaches, BG-MBC can utilise both the linguistic context from the histopathology reports and the intricate connections between different medical features.
2. Enhanced prediction accuracy: By leveraging BERT's ability to capture the semantic meaning of the text and GNN's capacity to model relationships and dependencies between data points, BG-MBC provides a more holistic analysis. This dual approach leads to higher prediction accuracy, making it a powerful tool for MBC diagnosis.
3. Comprehensive data representation: The model uses BERT to generate embeddings that capture the semantic richness of the text in the reports. These embeddings are then used as node features in a graph constructed based on the attention scores, reflecting important connections within the text. This comprehensive representation enhances the model's understanding of the complex interplay between different medical features.
4. Improved understanding of disease: By accurately modelling the relationships and dependencies within the data, BG-MBC can provide deeper insights into the underlying patterns and factors contributing to metastatic breast cancer. This understanding can guide clinicians in making more informed decisions regarding diagnosis and treatment.

In summary, the BG-MBC model represents a significant advancement in the field of medical diagnostics by combining the complementary strengths of GNN and BERT. This innovative approach not only enhances predictive accuracy but also provides valuable insights into the disease, ultimately contributing to more effective diagnostic and treatment strategies. Future research can further explore the potential of this approach by investigating different LLM architectures, refining the graph construction process, and evaluating the model's performance on larger and more diverse datasets.

Author Contributions: Conceptualisation, A.B., S.B., E.V., A.K.E. and M.M.A. ; methodology, A.B., S.B., E.V., A.K.E. and M.M.A.; software, A.B.; validation, A.B., S.B., E.V. and M.M.A.; formal analysis, A.B.; investigation, M.M.A.; resources, A.B., S.B., E.V. and M.M.A.; data curation, A.B.; writing—original draft preparation, A.B., A.K.E. and M.M.A.; writing—review and editing, A.B., S.B., E.V., A.K.E. and M.M.A.; visualisation, A.B.; supervision, S.B., E.V. and M.M.A.; project administration, M.M.A.; funding acquisition, S.B., E.V. and M.M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Antser Ltd., UK.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data used in the research can be found in the reference [19].

Conflicts of Interest: The authors declare that this study received funding from Antser Ltd., UK. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large language model
GNN	Graph neural network
XAI	Explainable artificial intelligence
MBC	Metastatic breast cancer
BERT	Bidirectional encoder representations from transformers

References

1. Redig, A.; McAllister, S. Breast cancer as a systemic disease: A view of metastasis. *J. Intern. Med.* **2013**, *274*, 113–126. [CrossRef] [PubMed]
2. Bertucci, F.; Ng, C.; Patsouris, A.; Droin, N.; Piscuoglio, S.; Carbuccia, N.; Soria, J.; Dien, A.; Adnani, Y.; Kamal, M.; et al. Genomic characterization of metastatic breast cancers. *Nature* **2019**, *569*, 560–564. [CrossRef] [PubMed]
3. Siegel, R.; Miller, K.; Fuchs, H.; Jemal, A. Cancer statistics, 2022. *CA Cancer J. Clin.* **2022**, *72*, 7–33. [CrossRef] [PubMed]
4. Hoda, S.; Cheng, E. *Robbins Basic Pathology*; Oxford University Press: Cary, NC, USA, 2017.
5. Aziz, N. Cancer survivorship research: State of knowledge, challenges and opportunities. *Acta Oncol.* **2007**, *46*, 417–432. [CrossRef] [PubMed]
6. Gurcan, M.; Boucheron, L.; Can, A.; Madabhushi, A.; Rajpoot, N.; Yener, B. Histopathological image analysis: A review. *IEEE Rev. Biomed. Eng.* **2009**, *2*, 147–171. [CrossRef] [PubMed]
7. Nowak-Sliwinska, P.; Scapozza, L.; Altaba, A. Drug repurposing in oncology: Compounds, pathways, phenotypes and computational approaches for colorectal cancer. *Biochim. Biophys. Acta-(BBA)-Rev. Cancer*. **2019**, *1871*, 434–454. [CrossRef] [PubMed]
8. Rosai, J. *Rosai and Ackerman's Surgical Pathology E-Book*; Elsevier Health Sciences: Gainesville, FL, USA, 2011.
9. Elmore, J.; Longton, G.; Pepe, M.; Carney, P.; Nelson, H.; Allison, K.; Geller, B.; Onega, T.; Tosteson, A.; Mercan, E.; et al. A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis. *J. Pathol. Inform.* **2017**, *8*, 12. [CrossRef] [PubMed]
10. Colling, R.; Pitman, H.; Oien, K.; Rajpoot, N.; Macklin, P.; CM-Path AI in Histopathology Working Group; Bachtiar, V.; Booth, R.; Bryant, A.; Bull, J.; et al. Artificial intelligence in digital pathology: A roadmap to routine use in clinical practice. *J. Pathol.* **2019**, *249*, 143–150. [CrossRef] [PubMed]
11. Hicks, D.; Tubbs, R. Assessment of the HER2 status in breast cancer by fluorescence in situ hybridization: A technical review with interpretive guidelines. *Hum. Pathol.* **2005**, *36*, 250–261. [CrossRef] [PubMed]
12. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional Transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
13. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 4–24. [CrossRef] [PubMed]
14. Esteva, A.; Kuprel, B.; Novoa, R.; Ko, J.; Swetter, S.; Blau, H.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
15. Liu, Y.; Gadepalli, K.; Norouzi, M.; Dahl, G.; Kohlberger, T.; Boyko, A.; Venugopalan, S.; Timofeev, A.; Nelson, P.; Corrado, G.; et al. Detecting cancer metastases on gigapixel pathology images. *arXiv* **2017**, arXiv:1703.02442.
16. Hamamoto, R.; Takasawa, K.; Machino, H.; Kobayashi, K.; Takahashi, S.; Bolatkan, A.; Shinkai, N.; Sakai, A.; Aoyama, R.; Yamada, M.; et al. Application of non-negative matrix factorization in oncology: One approach for establishing precision medicine. *Briefings Bioinform.* **2022**, *23*, bbac246. [CrossRef] [PubMed]
17. Gupta, S. Prediction time of breast cancer tumor recurrence using Machine Learning. *Cancer Treat. Res. Commun.* **2022**, *32*, 100602. [CrossRef] [PubMed]
18. Yang, J.; Ju, J.; Guo, L.; Ji, B.; Shi, S.; Yang, Z.; Gao, S.; Yuan, X.; Tian, G.; Liang, Y.; et al. Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 333–342. [CrossRef] [PubMed]
19. Botlagunta, M.; Botlagunta, M.; Myneni, M.; Lakshmi, D.; Nayyar, A.; Gullapalli, J.; Shah, M. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. *Sci. Rep.* **2023**, *13*, 485. [CrossRef] [PubMed]
20. Munn, M.; Pitman, D. *Explainable AI for Practitioners*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2022.
21. Lundberg, S.; Lee, S. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4768–4777.
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [CrossRef]

23. Tao, M.; Feng, Y.; Zhao, D. A frustratingly easy improvement for position embeddings via random padding. In Proceedings of the CCF International Conference On Natural Language Processing And Chinese Computing, Foshan, China, 12–15 October 2023; pp. 301–313.
24. Clark, K.; Khandelwal, U.; Levy, O.; Manning, C. What does BERT look at? An analysis of BERT's attention. *arXiv* **2019**, arXiv:1906.04341.
25. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
26. Tenney, I.; Das, D.; Pavlick, E. BERT rediscovers the classical NLP pipeline. *arXiv* **2019**, arXiv:1905.05950.
27. Lu, Y.; Li, Z.; He, D.; Sun, Z.; Dong, B.; Qin, T.; Wang, L.; Liu, T. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv* **2019**, arXiv:1906.02762.
28. Puccetti, G.; Miaschi, A.; Dell'Orletta, F. How Do BERT Embeddings Organize Linguistic Knowledge? In Proceedings of the Deep Learning Inside out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Online, 10 June 2021; pp. 48–57.
29. Tang, R.; Lu, Y.; Liu, L.; Mou, L.; Vechtomova, O.; Lin, J. Distilling task-specific knowledge from BERT into simple neural networks. *arXiv* **2019**, arXiv:1903.12136.
30. Howard, J.; Ruder, S. Universal language model fine-tuning for text classification. *arXiv* **2018**, arXiv:1801.06146.
31. Clark, K.; Luong, M.; Le, Q.; Manning, C. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv* **2020**, arXiv:2003.10555.
32. Ting, F.; Tan, Y.; Sim, K. Convolutional neural network improvement for breast cancer classification. *Expert Syst. Appl.* **2019**, *120*, 103–115. [CrossRef]
33. Thwin, S.; Malebary, S.; Abulfaraj, A.; Park, H. Attention-Based Ensemble Network for Effective Breast Cancer Classification over Benchmarks. *Technologies* **2024**, *12*, 16. [CrossRef]
34. Mullooly, M.; Ehteshami Bejnordi, B.; Pfeiffer, R.; Fan, S.; Palakal, M.; Hada, M.; Vacek, P.; Weaver, D.; Shepherd, J.; Fan, B.; et al. Application of convolutional neural networks to breast biopsies to delineate tissue correlates of mammographic breast density. *NPJ Breast Cancer* **2019**, *5*, 43. [CrossRef] [PubMed]
35. Hastie, T.; Tibshirani, R.; Friedman, J.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009.