Journal of
Sensor and Actuator Networks

MDPI

Article

Advanced Diagnosis of Cardiac and Respiratory Diseases from Chest X-Ray Imagery Using Deep Learning Ensembles

Hemal Nakrani, Essa Q. Shahra *, Shadi Basurra, Rasheed Mohammad, Edlira Vakaj and Waheb A. Jabbar *

Faculty of Computing, Engineering and Built Environment, Birmingham City University, Birmingham B4 7RQ, UK; hemal.nakrani@mail.bcu.ac.uk (H.N.); shadi.basurra@bcu.ac.uk (S.B.); rasheed.mohammad@bcu.ac.uk (R.M.); edlira.vakaj@bcu.ac.uk (E.V.)
* Correspondence: essa.shahra@bcu.ac.uk (E.Q.S.); waheb.abdullah@bcu.ac.uk (W.A.J.)

Abstract: Chest X-ray interpretation is essential for diagnosing cardiac and respiratory diseases. This study introduces a deep learning ensemble approach that integrates Convolutional Neural Networks (CNNs), including ResNet-152, VGG19, EfficientNet, and a Vision Transformer (ViT), to enhance diagnostic accuracy. Using the NIH Chest X-ray dataset, the methodology involved comprehensive preprocessing, data augmentation, and model optimization techniques to address challenges such as label imbalance and feature variability. Among the individual models, VGG19 exhibited strong performance with a Hamming Loss of 0.1335 and high accuracy in detecting Edema, while ViT excelled in classifying certain conditions like Hernia. Despite the strengths of individual models, the ensemble meta-model achieved the best overall performance, with a Hamming Loss of 0.1408 and consistently higher ROC-AUC values across multiple diseases, demonstrating its superior capability to handle complex classification tasks. This robust ensemble learning framework underscores its potential for reliable and precise disease detection, offering significant improvements over traditional methods. The findings highlight the value of integrating diverse model architectures to address the complexities of multi-label chest X-ray classification, providing a pathway for more accurate, scalable, and accessible diagnostic tools in clinical practice.

Keywords: X-ray; VGG19; ViT; ResNet152; ensemble learning; deep learning

Academic Editor: Lei Shu

Received: 11 February 2025
Revised: 14 April 2025
Accepted: 16 April 2025
Published: 18 April 2025

Citation: Nakrani, H.; Shahra, E.Q.; Basurra, S.; Mohammad, R.; Vakaj, E.; Jabbar, W.A. Advanced Diagnosis of Cardiac and Respiratory Diseases from Chest X-Ray Imagery Using Deep Learning Ensembles. *J. Sens. Actuator Netw.* **2025**, *14*, 44. https://doi.org/10.3390/jsan14020044

1. Introduction

Chest X-rays serve as a cornerstone of medical imaging, playing a pivotal role in the diagnosis and management of a wide array of thoracic conditions, including respiratory diseases, cardiac anomalies, and even malignancies. Their utility has become particularly pronounced in recent years, especially during the COVID-19 pandemic, which underscored the urgent need for rapid and accurate detection of lung-related diseases. Chest X-rays remain one of the most accessible and cost-effective diagnostic tools globally, providing critical insights into the structural and pathological states of the thorax. However, traditional manual interpretation of these images is not without limitations. Factors such as the increasing workload of radiologists, diagnostic variability across practitioners, and the inherent subjectivity of manual assessments present significant challenges. Additionally, the growing demand for rapid and precise diagnoses further exacerbates these limitations, particularly in resource-constrained healthcare systems [1,2].

The growing complexities in chest X-ray interpretation demand innovative solutions that enhance efficiency, accuracy, and scalability. Advances in Artificial Intelligence (AI),

particularly deep learning models such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have demonstrated transformative potential in medical imaging. These models excel at detecting intricate patterns that may be challenging for the human eye, leading to more precise and consistent diagnoses. Moreover, AI-driven systems can process vast amounts of imaging data rapidly, making them well-suited for high-demand clinical environments where timely decision-making is critical [3,4].

A significant strength of AI models lies in their ability to adapt and improve over time. As medical imaging databases expand, these systems refine their predictive accuracy and generalizability, enhancing diagnostic performance. This adaptability is particularly valuable in resource-limited settings, such as rural hospitals and clinics, where access to expert radiologists is limited. By providing standardized, data-driven insights, AI-powered tools help mitigate diagnostic variability and reduce reliance on scarce medical expertise, ensuring more consistent and equitable healthcare delivery [5,6]. Despite these advancements, implementing AI in medical diagnostics is not without its challenges. Issues such as dataset imbalance, model interpretability, and computational complexity remain significant barriers to widespread adoption. The integration of CNNs and ViTs into a unified framework offers a promising avenue for addressing these challenges. CNNs excel in hierarchical feature extraction, while ViTs, with their self-attention mechanisms, are adept at capturing long-range dependencies within images. Combining these strengths in a hybrid model has the potential to create a system that is both robust and precise, capable of addressing a wide spectrum of diagnostic requirements [7,8].

In this paper, we propose a novel hybrid deep learning framework that integrates advanced CNN architectures (ResNet-152, VGG19, EfficientNet) and ViTs to address the critical challenges of chest X-ray interpretation. By leveraging the complementary strengths of these models, the proposed ensemble meta-model enhances classification precision and robustness, even in the face of dataset imbalances and co-occurring disease labels. The framework is rigorously evaluated using the NIH Chest X-ray dataset, one of the largest and most comprehensive public repositories in this domain. Key contributions of this work include:

- Novel Ensemble Architecture: We propose a unique ensemble framework that integrates Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), effectively leveraging their complementary strengths to improve multi-label classification accuracy. This hybrid design offers a new paradigm for combining spatial and contextual representations in medical imaging.
- Comprehensive and Scalable Methodology: Our pipeline includes advanced preprocessing, targeted data augmentation, and iterative model refinement, specifically optimized for large-scale clinical datasets. This ensures that the proposed approach is not only accurate but also robust and generalizable.
- Empirical Evaluation and Model Insights: The study provides an in-depth comparative analysis of individual models and the ensemble configuration, highlighting their respective strengths and guiding the ensemble's design. The ensemble consistently achieves superior metrics, including lower Hamming Loss and higher ROC-AUC, validating the efficacy of the integration strategy.
- Clinical and Societal Impact: With its high accuracy and reliability, the proposed system demonstrates strong potential for deployment in real-world clinical environments, particularly in resource-constrained settings. This contributes to ongoing efforts to improve diagnostic equity and accessibility through AI-powered solutions.

This research not only advances the state-of-the-art in AI-driven medical imaging but also lays the groundwork for the integration of such technologies into clinical practice. By addressing the limitations of existing approaches and providing a scalable, efficient,

and accurate diagnostic solution, this study contributes to the broader goal of enhancing healthcare delivery through technological innovation [9].

The structure of this paper is as follows: Section 2 presents the systematic literature review, Section 3 details the proposed methodology, Section 4 discusses the results, and Section 5 concludes the study while outlining potential future work.

2. Literature Review

The prominence of Convolutional Neural Networks (CNNs) in medical image analysis is well-established [4]. CNNs excel in extracting intricate features from images, making them ideal for tasks like identifying pathologies in chest X-rays [2]. The integration of transfer learning has further refined the application of CNNs. By adapting models like ResNet, VGGNet, and EfficientNet, which were pre-trained on vast, diverse datasets, researchers have achieved remarkable accuracy in disease detection [7]. These models, pre-trained on large-scale datasets, bring a level of sophistication and a depth of learning that is otherwise challenging to achieve in medical-specific datasets [5]. The study by [3] addresses the challenge of accurately diagnosing multiple thoracic diseases from chest X-rays, which are characterized by varied and overlapping visual features. The objective is to evaluate and compare the performance of different deep learning architectures—CNN, ResNet, and Vision Transformers (ViT)—in multi-label classification tasks. The study fine-tunes two variants of Vision Transformers, pre-trained on ImageNet, alongside CNN and ResNet models, using the NIH Chest X-ray dataset. The results indicate that the pre-trained ViT models outperform both CNN and ResNet architectures, achieving higher accuracy in diagnosing multiple diseases simultaneously. This highlights the potential of ViTs, particularly when pre-trained on large, diverse datasets, to significantly enhance diagnostic accuracy in medical imaging [3]. The integration of transfer learning into medical image fusion has yielded significant advancements, as demonstrated in work by [10]. The authors in [10] utilize a modified VGG19 model (TL_VGG19) to enhance multi-modal medical image fusion by preserving fine-grained details and ensuring brightness and contrast consistency using the Equilibrium Optimization Algorithm (EOA). This approach improves image quality for clinical diagnostics but has limitations, including potential performance issues on diverse datasets and high computational overhead. Despite these challenges, TL_VGG19 shows promise in improving image fusion for medical applications. Additionally, research in lung disease classification from X-ray images has been prolific. Studies utilizing VGGNet, for example, have shown exceptional accuracy in classifying COVID-19, a critical advancement during the pandemic [8]. The Xception model, recognized for its depth and efficiency, has been particularly effective when combined with image enhancement techniques such as contrast limited adaptive histogram equalization (CLAHE) [6]. This approach not only improves image quality but also augments the model's capacity to discern subtle features indicative of various lung conditions [11].

The emergence of Visual Transformers (ViTs) represents a paradigm shift in medical image classification [12]. Utilizing self-attention mechanisms, ViTs offer a novel approach to processing images, different from the traditional convolutional methods [13]. They have shown impressive results, often on par with or surpassing CNNs, particularly in handling larger datasets and capturing global image features. Studies have highlighted ViTs' adaptability and scalability, making them a promising technology for future exploration [14–16]. The study by [17] addresses the complex challenge of classifying multiple diseases from chest X-rays, which follow a long-tailed distribution. The CXR-LT challenge encourages research on classifying 26 clinical findings from a dataset of over 350,000 chest X-rays, addressing challenges in disease prevalence imbalance. It provides a benchmark dataset and highlights top-performing solutions that tackle label imbalance and co-occurrence.

The study demonstrates improved classification performance and explores the potential of vision-language models for few- and zero-shot learning in medical image classification [17]. The study by [18] introduces HydraViT, a hybrid model combining a CNN with a transformer-based context encoder and a multi-branch output module to improve multi-label chest X-ray classification. By leveraging transformers for long-range dependencies and specialized branches for label co-occurrence, HydraViT outperforms existing methods, increasing classification accuracy by 1.2–1.4%. It enhances sensitivity to both individual diseases and their relationships, demonstrating its effectiveness in complex medical image analysis. The authors in [19] tackles multi-disease classification in chest X-rays, addressing the challenge of long-tailed disease distribution. They propose an optimized ensemble framework using class-specific residual attention (CSRA) and head-tail separation techniques to manage imbalance. Extensive experimentation identified the best model components, leading to significant performance improvements. The framework ranked among the top in the CXR-LT competition, highlighting the need for tailored approaches in imbalanced medical imaging. Another study by [20] introduces CheXFusion, a transformer-based fusion module designed to improve multi-label chest X-ray classification by integrating multi-view image features using self-attention and cross-attention mechanisms. It tackles challenges like long-tailed disease distribution and label co-occurrence while incorporating data balancing and self-training techniques. CheXFusion achieved state-of-the-art performance (mAP 0.372) on the MIMIC-CXR test set, winning first place in the ICCV CVAMD 2023 CXR-LT Shared Task, highlighting the importance of multi-view integration in medical imaging.

The Research Gap: Looking ahead, based on the papers reviewed in Table 1, the field is poised for further groundbreaking advancements. The exploration of hybrid models that blend the strengths of CNNs, ViTs, and other architectures offers a path towards more robust and accurate diagnostic tools. The potential of AI in medical imaging extends beyond disease classification, with applications in predictive analytics, treatment planning, and personalized medicine becoming increasingly feasible.

Table 1. Summary of research papers used in the literature review.

Ref.	Method Used	Dataset Used	Advantages	Challenges
[3]	Comparative Study of CNN, ResNet, and ViTs	NIH Chest X-ray dataset	ViTs achieve highest accuracy; pre-training on large datasets enhances performance	High computational resources; requires extensive pre-training.
[10]	A novel image fusion method using TL_VGG19 network combined with Equilibrium Optimization Algorithm (EOA)	C1: MRI-PET (269 pairs), C2: MRI-SPECT (357 pairs), C3: MRI, CT, PET, SPECT (1424 images)	Enhanced image synthesis quality with improved brightness, contrast, and sharpness; demonstrated highest performance in six evaluation metrics including QA, QAB/F, and QC. Preserved intricate details from input images.	High computational cost due to complex model training; potential limitations in generalizing to datasets not included in the transfer learning.

Table 1. *Cont.*

Ref.	Method Used	Dataset Used	Advantages	Challenges
[17]	Multi-Label Classification with Long-Tailed Distribution	CXR-LT dataset (350,000 chest X-rays)	Addresses label imbalance and co-occurrence; utilizes vision-language models for future tasks	Handling rare diseases; managing large-scale datasets
[18]	Hybrid CNN-Transformer with Multi-Branch Output	ChestX-ray14 dataset (112,120 images)	Captures long-range dependencies; handles label co-occurrence; improves classification accuracy by 1.2–1.4%	High computational complexity; managing adaptive weights for co-occurrence relationships
[19]	Optimized Ensemble Framework with CSRA	MIMIC-CXR-LT dataset	Improves classification performance; handles long-tailed distribution effectively	Complexity of ensemble methods; managing computational resources
[20]	Transformer-based Fusion with Self-Attention and Cross-Attention	MIMIC-CXR dataset	Enhances multi-view classification; achieves state-of-the-art performance; handles class imbalance	Managing computational complexity; optimizing data balancing
[13]	Vision Transformer (ViT) with self-attention mechanism	COVID-19 Dataset and COVID-19 Radiography Dataset	Outperforms CNN-based models; achieves 97% accuracy and 94% F1-score on Radiography dataset; effective in capturing global context	Limited generalizability to other chest conditions; requires fine-tuning of ViT architecture and intensive preprocessing

3. Proposed Methodology

Building upon the insights from our systematic literature review, we propose a comprehensive approach to address the intricate challenges in multi-label image classification of chest X-rays as shown in Figure 1. This methodology aims not only to enhance diagnostic precision but also to adeptly handle the complexity of concurrent conditions often present in these images.

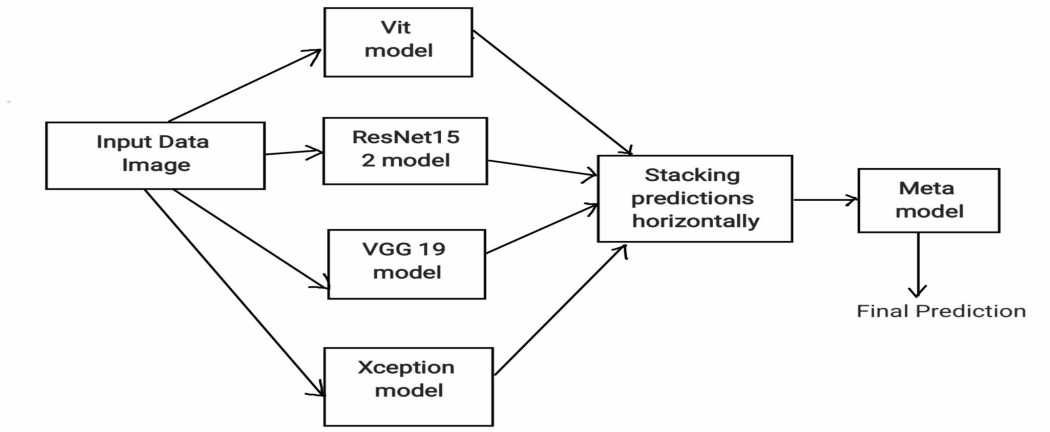

Figure 1. Proposed methodology approach.

3.1. Dataset Overview

The NIH Chest X-ray dataset a pivotal resource in our research, is a comprehensive public collection available on Kaggle. This dataset contains images classified into 15 distinct diagnostic categories, as shown in Figure 2, covering a broad range of thoracic pathologies, which play a critical role in training and validating the deep learning models used in our study. Our work specifically utilizes the NIH ChestX-ray14 dataset, an extensive and publicly available collection of frontal-view chest radiographs curated by the National Institutes of Health Clinical Center. This dataset comprises 112,120 chest X-ray images from 30,805 unique patients, each annotated with 14 different thoracic disease labels as well as a "No Finding" class. The images are stored in PNG format with a spatial resolution of 1024 × 1024 pixels, providing ample detail for the extraction of clinical features by deep learning models.

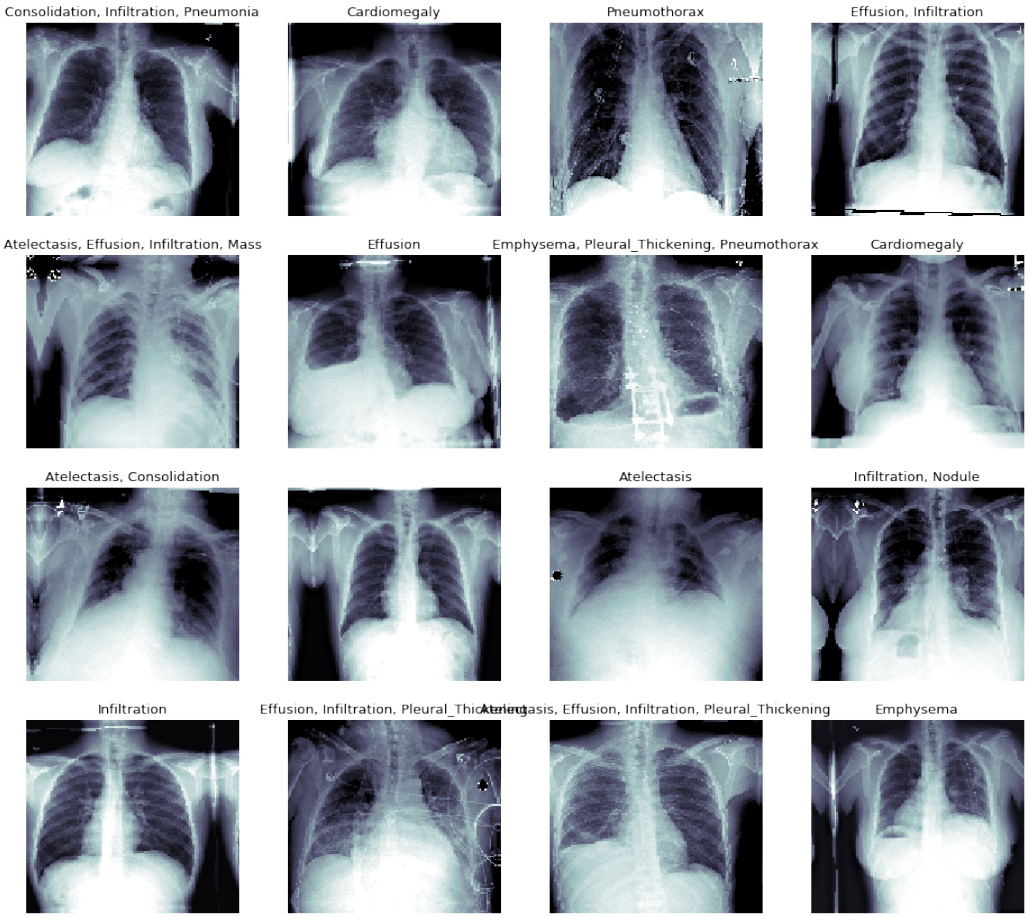

Figure 2. Dataset labels.

The disease labels within the dataset were generated through Natural Language Processing (NLP) techniques which automatically extracted information from radiology reports, achieving an accuracy of over 90%. This high level of accuracy makes the dataset particularly well-suited for weakly supervised learning tasks. The diseases included cover a wide array of thoracic conditions, such as Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia. For the purposes of our study, we performed a binary classification task, distinguishing between images that show signs of disease and those labeled as "No Finding". To achieve this, we filtered the dataset into two classes: the

Positive Class–images annotated with one or more disease labels–and the Negative Class–images labeled as "No Finding" without any pathological annotations.

To address class imbalance, we carefully constructed a balanced dataset by randomly sampling an equal number of images from both classes. Each image was resized to 224 × 224 pixels to meet the input size requirements for the deep learning architectures utilized in our models as shown in Figure 3. In addition, the dataset includes a separate file containing bounding box annotations for approximately 1000 images. However, these annotations were not employed in this study, as our focus was solely on the classification task rather than localization.

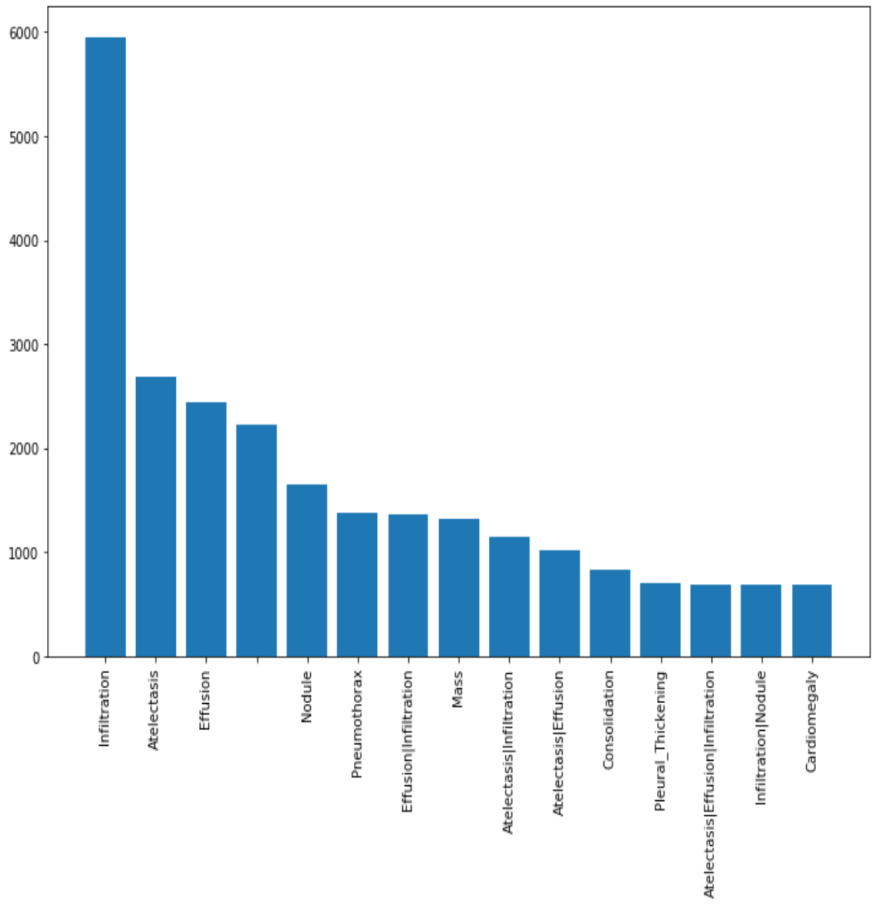

Figure 3. Balance datasets.

Overall, the NIH ChestX-ray14 dataset serves as a rich resource for thoracic disease detection, providing a diverse range of clinical images and disease annotations that are crucial for developing robust and accurate deep learning models.

3.2. Approach and Model Selection

Our approach involves integrating advanced neural network architectures, each tailored for the classification of chest X-ray images:

Transfer Learning: A cornerstone of our methodology, transfer learning enables the adaptation of pre-trained models such as ResNet, VGGNet, and EfficientNet to our specific task, thereby enhancing efficiency and accuracy.

ViT Model: Employing the Vision Transformer (ViT) for its self-attention mechanism offers a novel perspective in image processing, particularly effective in capturing global features crucial for medical diagnostics [21].

ResNet152 Model: Chosen for its deep architecture and residual connections, ResNet152 effectively overcomes the vanishing gradient problem and is adept at learning diverse features from medical images [22].

VGG19 Model: VGG19, known for its uniform architecture and small convolutional filters, excels at capturing detailed image features, making it a suitable candidate for identifying complex patterns in chest X-rays [23].

Xception Model: The Xception model's use of depthwise separable convolutions provides an efficient yet adaptable approach to feature extraction, enhancing the model's performance in medical image classification [24].

3.3. Model Training Overview

To effectively classify thoracic pathologies from chest X-ray images, four state-of-the-art deep learning architectures were fine-tuned using transfer learning: VGG19, ResNet152, Xception, and the Vision Transformer (ViT). In each case, the original classification head was removed, and a custom classifier was appended to enable multi-label classification over 14 disease classes. All models shared a consistent training strategy, which is summarized in Table 2.

Table 2. Common preprocessing and training setup.

Parameter	Value
Input Shape	Images resized from 256 × 256 × 3 to 224 × 224 × 3
Normalization	Pixel values scaled to the range [0, 1] using Rescaling (1./255)
CLAHE	Contrast limited adaptive histogram equalization
Data Augmentation	Horizontal flipping, Rotation, Shearing, Zooming
Loss Function	Binary Cross-Entropy (suitable for multi-label classification)
Optimizer	Adam optimizer with default learning rate
Metrics	Binary Accuracy and Mean Absolute Error (MAE)
Training Duration	Maximum of 10 epochs
Batch Size	32

3.3.1. VGG19 Model

The model uses VGG19 as a base, pre-trained on ImageNet with the top layer removed (include_top=False). The layers of the base model are frozen to retain learned features and avoid overfitting. A custom head is added with: Flatten layer to reshape the output, Dense(512, ReLU) for non-linearity, and Dense(14, Sigmoid) for multi-label classification.

After training for 4 epochs, the model achieved a validation accuracy of 88.5%. This approach leverages transfer learning, using pre-trained features from VGG19 and adapting them for the specific task with a custom head.

3.3.2. ResNet152 Model

The model uses ResNet152 pre-trained on ImageNet with the top layer removed and the base model's layers frozen. A custom head with Flatten, Dense(512, ReLU), and Dense(14, Sigmoid) is added for multi-label classification. EarlyStopping is employed to prevent overfitting, and the model achieves a validation accuracy of 88.4% after 4 epochs. This approach leverages transfer learning with ResNet152 for feature extraction and a custom head for the specific task.

3.3.3. Xception Model

The model uses Xception as the base, pre−trained on ImageNet, with the top layer removed (include_top=False). The layers of the base model are frozen to retain the learned features from ImageNet. A custom head consisting of Flatten, Dense(512, ReLU), and Dense(14, Sigmoid) is added for multi-label classification. EarlyStopping is used to halt training if there is no improvement in validation accuracy. The model achieved 88% validation accuracy by epoch 7, demonstrating effective transfer learning with Xception.

3.3.4. Vision Transformer (ViT) Model

The model uses the pre-trained TFViTModel with frozen layers and input preprocessed to match the ViT format. A custom head with MaxPooling1D, Flatten, Dense(512, ReLU), and Dense(14, Sigmoid) was added for multi-label classification. EarlyStopping was employed to prevent overfitting. The model achieved a validation accuracy slightly below 88.5%.

This unified training procedure enabled a fair comparison among the models while ensuring reproducibility and consistency across all experiments.

3.4. Horizontal Stacking and Hybrid Architecture

Our approach employs a sophisticated ensemble technique known as horizontal stacking [25], applied after the individual training of models like ViT, ResNet152, VGG19, and Xception as shown in Figure 4. This technique involves aligning the prediction outputs of each model side-by-side to form a comprehensive feature set. Specifically, the predictions from each model are concatenated horizontally, resulting in an expanded feature vector. This vector effectively captures diverse perspectives and learned patterns from each individual model, leading to a richer representation of the data [26].

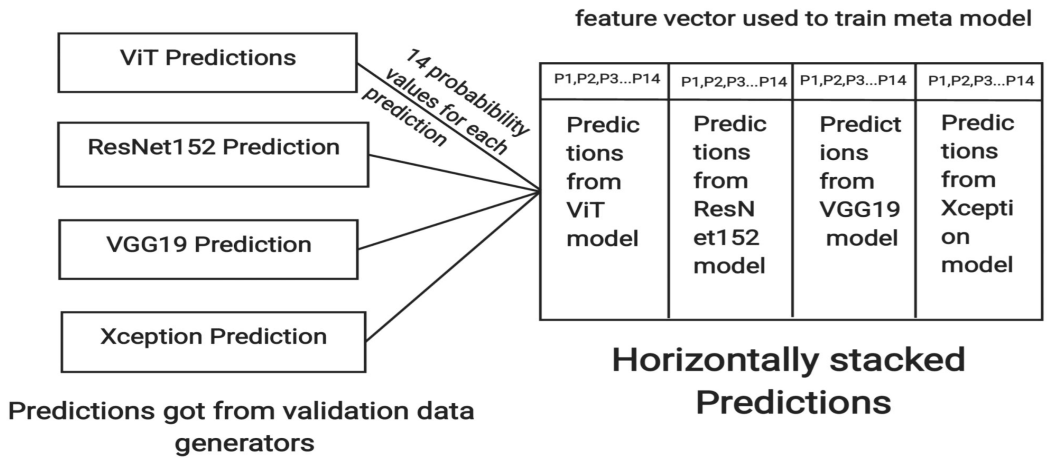

Figure 4. Horizontal stacking.

Horizontal stacking plays a critical role in our hybrid architecture. It enables the synthesis of various learned features and insights into a unified representation. This representation is instrumental in capturing the complexities and subtleties inherent in chest X-ray images. By combining the strengths of each model, this hybrid architecture aims to enhance diagnostic accuracy, particularly in identifying multiple co-existing conditions in X-rays.

3.5. Meta-Model Ensembling

Upon creating the horizontally stacked feature vectors, we introduce a meta-model to learn from this aggregated data. The meta-model serves as the final decision-maker, trained

on the stacked predictions from all the primary models [27]. This approach allows the meta-model to benefit from the distinct learning trajectories and specialized insights of each base model [28]. In practice, for any given chest X-ray image, the image is first evaluated by each of the primary models (ViT, ResNet152, VGG19, and Xception). The predictions from these models are then horizontally stacked, forming a feature vector that encapsulates a broad spectrum of analytical perspectives. The meta-model is trained on these vectors, effectively learning to weigh and integrate the varying predictions. This ensembling with the meta-model represents a significant step towards realizing a robust, reliable system for chest X-ray classification. It not only improves the accuracy of predictions by leveraging the strengths of various architectures but also provides a level of redundancy and validation through its multi-model approach. The meta-model's ability to synthesize insights enhances the overall system's reliability, making it a powerful tool in medical diagnostics. Through these advanced techniques of horizontal stacking and meta-model ensembling, our proposed methodology sets the stage for significant advancements in medical image analysis. It underscores our commitment to leveraging cutting-edge AI techniques to enhance healthcare outcomes and support medical professionals in their diagnostic endeavors.

4. Results and Discussion

4.1. Evaluation Metrics

This study addresses a *multi-label classification* problem, where each chest X-ray image may simultaneously exhibit multiple thoracic pathologies. Accordingly, evaluation metrics were carefully selected to align with the structure of the problem, ensuring that performance insights are both meaningful and reliable.

4.1.1. Metrics Used During Model Training

To monitor training convergence and stability, we employed the following metrics:

- Binary Accuracy: Measures correctness of predictions on a per-label basis. This is suitable for multi-label classification, where each class is treated independently.
- Mean Absolute Error (MAE): Calculates the average absolute difference between predicted probabilities and actual labels, reflecting the model's confidence calibration.

These metrics were used solely for training supervision and were tracked using the `EarlyStopping` and `ModelCheckpoint` callbacks during model fitting.

4.1.2. Metrics Used for Model Comparison

To compare the performance of the trained models on the validation dataset, we used metrics more aligned with multi-label evaluation:

- Hamming Loss: Measures the fraction of incorrectly predicted labels (false positives and false negatives) over the total number of labels. It penalizes partial misclassifications and is more robust than standard accuracy in multi-label settings [29,30].
- ROC-AUC Curve: Evaluates the discriminative ability of each model across all classes using the Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC). As a threshold-independent metric, it offers insights into classification performance regardless of the decision threshold [31].

Table 3 shows the comparison of these metrcis: Hamming Loss and AUC across the models.

Table 3. Comparison of Hamming Loss and AUC values across models.

Disease/Model	ViT	ResNet	VGG19	Xception	Meta Model
Atelectasis	0.53	0.56	0.56	0.49	0.62
Cardiomegaly	0.57	0.51	0.52	0.49	0.61
Consolidation	0.59	0.59	0.58	0.51	0.63
Edema	0.66	0.68	0.68	0.54	0.74
Effusion	0.56	0.54	0.55	0.48	0.64
Emphysema	0.47	0.48	0.49	0.40	0.61
Fibrosis	0.41	0.36	0.37	0.52	0.68
Hernia	0.42	0.43	0.41	0.50	0.65
Infiltration	0.52	0.52	0.53	0.51	0.55
Mass	0.48	0.46	0.46	0.51	0.55
Nodule	0.44	0.44	0.43	0.57	0.61
Pleural_Thickening	0.48	0.43	0.44	0.52	0.63
Pneumonia	0.54	0.53	0.51	0.49	0.53
Pneumothorax	0.49	0.46	0.46	0.47	0.63
Hamming Loss	0.24258	0.18126	0.13355	0.16085	0.140803

4.1.3. Rationale for Excluding Traditional Metrics

Although *Precision*, *Recall*, and *F1-score* are standard metrics for binary and multi-class classification, they were deliberately excluded in this work for the following reasons:

- These metrics require *micro*, *macro*, or *weighted* averaging to adapt them for multi-label settings, which can obscure class-wise performance and inflate aggregated results in the presence of class imbalance.
- In datasets like chest X-rays, where labels are non-exclusive and often sparse, aggregate metrics can be misleading or overly optimistic/pessimistic depending on label prevalence.
- Interpreting these metrics becomes non-trivial, particularly when multiple pathologies co-occur or when certain classes are under-represented.

To maintain clarity, reproducibility, and clinical relevance, we focused on metrics that are better suited to multi-label learning tasks and are commonly adopted in the medical imaging literature.

4.2. Model Performance

Trained ViT Model: The Vision Transformer model showed a Hamming Loss of 0.24258, indicating around 24.25% incorrect label predictions. Its ROC-AUC curve varied significantly across different conditions, suggesting room for optimization (see Figure 5).

Trained ResNet152 Model: The ResNet152 model exhibited a lower Hamming Loss of 0.18126 compared to the ViT model, indicating improved accuracy. However, the ROC-AUC values revealed inconsistencies across various conditions as shown in Figure 6.

Trained VGG19 Model: VGG19 outperformed the previous models with a Hamming Loss of 0.13355, suggesting better alignment with the true labels. The ROC-AUC values pointed to its strong discriminative power, especially for conditions like Edema as shown in Figure 7.

Trained Xception Model: Xception's performance was moderate with a Hamming Loss of 0.16085. Its ROC-AUC values across diseases suggest it performs near the level of random guessing for most conditions, indicating a need for further model refinement as shown in Figure 8.

Trained Meta Model: The Meta model demonstrated the best performance among all models, with the lowest Hamming Loss (0.140803) and higher ROC-AUC values, showcasing its superior discriminatory capabilities as shown in Figure 9.

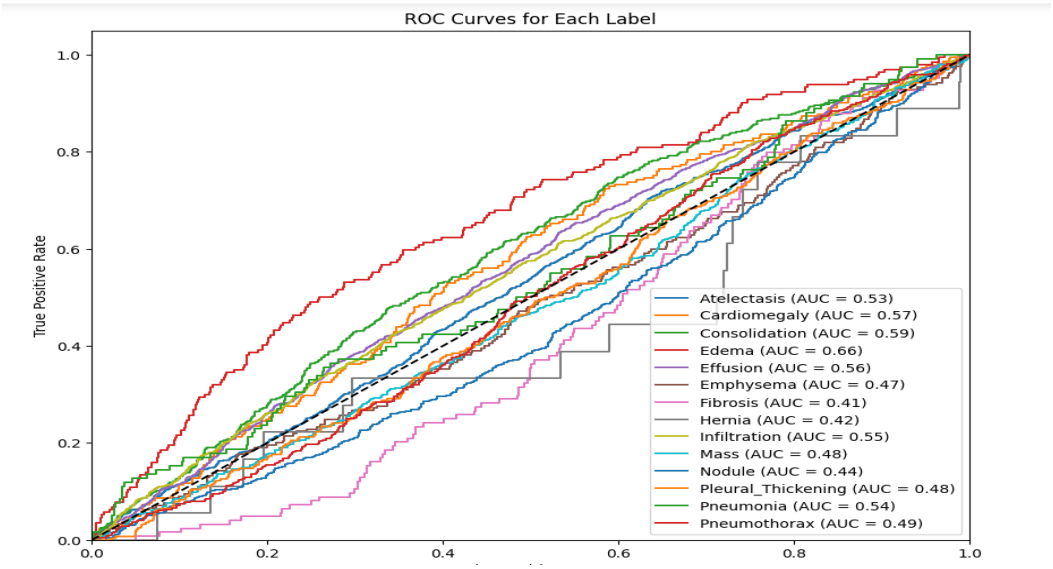

Figure 5. Model ROC curve.

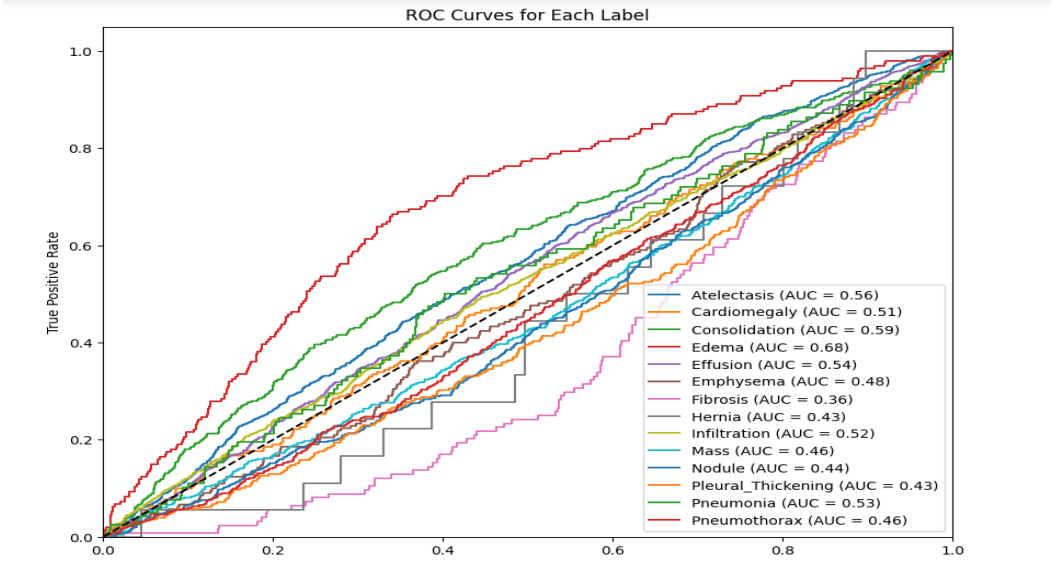

Figure 6. ResNet model ROC curve.

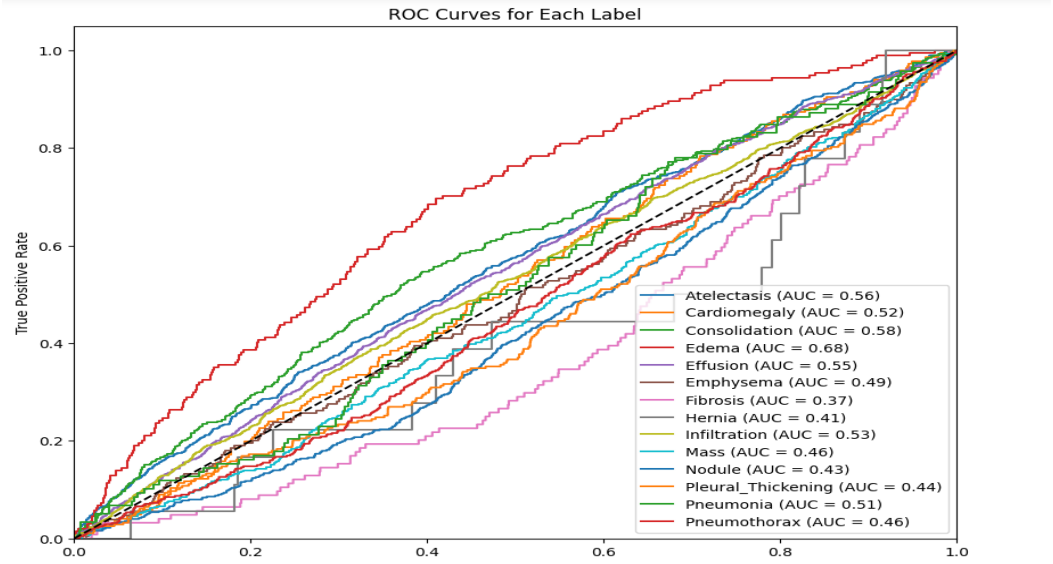

Figure 7. Vgg model ROC curve.

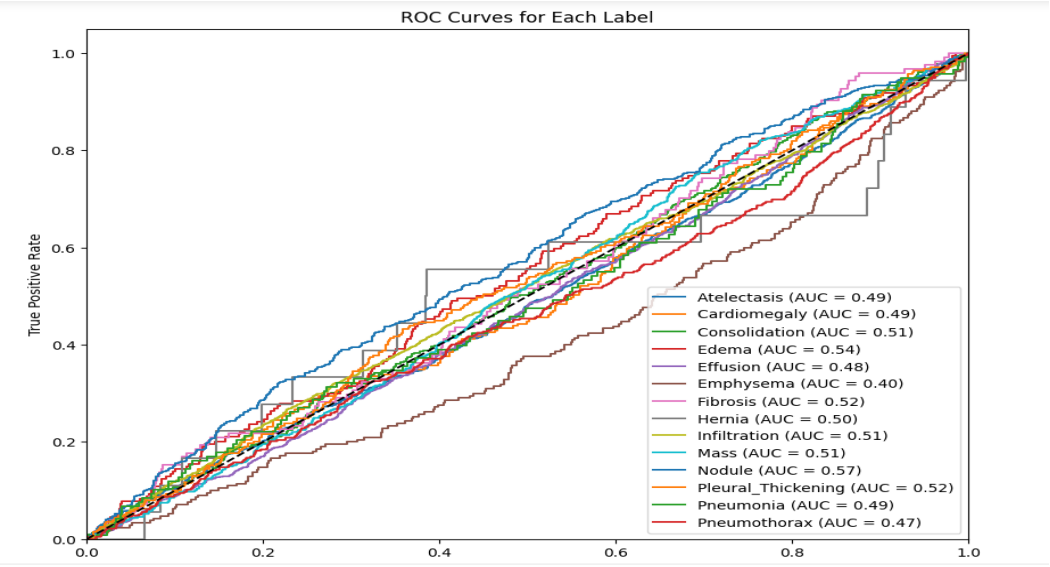

Figure 8. Xception model ROC curve.

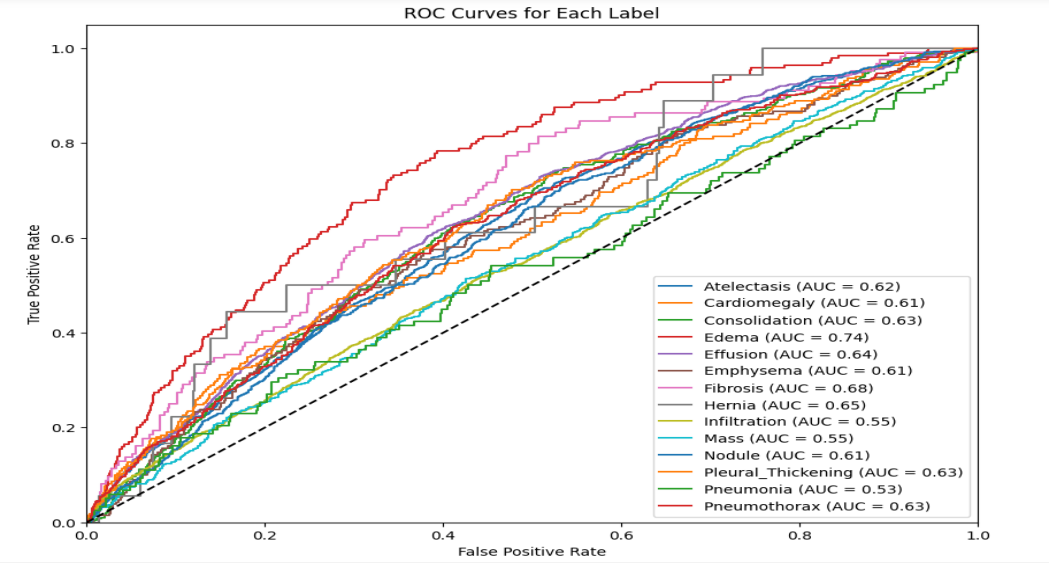

Figure 9. Meta model ROC curve.

4.3. Comparative Analysis and Implications

Our analysis revealed that while individual models like ViT, ResNet152, VGG19, and Xception have their strengths, the Meta model emerged as the most reliable, especially in discriminating complex conditions in chest X-rays. This suggests that an ensemble approach, which leverages the strengths of individual models, could be more effective in clinical diagnostic settings. The findings highlight the need for ongoing enhancements and the exploration of new models to further improve diagnostic accuracy.

4.4. Error Analysis and Insights

To better understand the limitations of the proposed ensemble model and its individual components, an error analysis was conducted on the misclassified cases. This section provides an in-depth look at the patterns in errors, potential biases in the model predictions, and limitations inherent in the dataset or methodology.

- Patterns in Misclassification:
  - The Vision Transformer (ViT) model, despite its strong performance in detecting Edema, struggled significantly with conditions such as Hernia and Fibrosis [8].

These cases often presented subtle features that might not be captured well by global attention mechanisms.
  - ResNet-152 and VGG19 showed improved consistency across multiple conditions but were prone to errors in diseases with overlapping visual features, such as Infiltration and Mass [4].
- Dataset Biases:
  - An imbalance in the NIH Chest X-ray dataset was observed, with certain conditions like Pneumonia and Effusion being overrepresented compared to rarer diseases such as Hernia or Fibrosis [7].
- Model Limitations:
  - The Xception model, while balanced in performance, lacked the specificity required to differentiate closely related thoracic anomalies, leading to near-random guessing for some rare disease labels [6].
- Potential Bias Sources:
  - Class Imbalance: The long-tailed distribution in the dataset caused a disproportionate focus on majority classes during training [7].
  - Co-Occurrence of Labels: Certain diseases co-occurred frequently (e.g., Infiltration and Pneumonia), which led the models to overgeneralize, reducing their ability to handle cases where only one condition was present [8].
  - Image Quality Variability: Variations in image resolution and noise across the dataset further impacted the model's ability to generalize [5].

4.5. Comparison with Existing Literature

To evaluate the effectiveness of the proposed models, particularly the meta-ensemble architecture, a comprehensive comparison was conducted with existing works focused on multi-label classification of chest X-ray images.

Earlier research, such as CheXNet, employed DenseNet-121 on the ChestX-ray14 dataset, achieving an AUC of approximately 0.76 [32]. Similarly, Wang et al. [33] reported benchmark performances using ResNet50, with AUCs ranging from 0.70 to 0.75 across various thoracic diseases. Bharati et al. [34] utilized hybrid CNN architectures and achieved comparable accuracy, while Kim et al. [35] further demonstrated the effectiveness of multi-class CNN classifiers in similar tasks.

More recently, transformer-based models have gained significant attention. Jiang et al. introduced MXT, a pyramid ViT-based model, which achieved per-label AUCs between 0.62 and 0.75 [16], while Okolo et al. developed IEViT, an enhanced ViT model, to improve attention and performance in multi-label classification tasks [15]. Additionally, Uparkar et al. [14] showed that ViTs can outperform traditional CNNs, such as ResNet and VGG19.

In comparison, our proposed Meta model, which combines ViT, ResNet152, VGG19, and Xception through a stacking ensemble, achieved the following results:

- Average AUC: 0.635 across 14 thoracic diseases;
- Best per-class AUC: 0.74 for Edema;
- Lowest Hamming Loss: 0.1408.

While some prior methods report slightly higher AUCs for specific pathologies, our model demonstrates consistent and balanced performance across all 14 labels. The reduction in Hamming Loss further indicates improved generalization and lower misclassification across labels.

Recent literature highlights the advantages of ensemble learning, particularly stacking, in enhancing robustness and predictive accuracy [19,26,28,36–38]. Our meta-model

aligns with these findings, effectively integrating multiple base models to outperform individual networks.

5. Conclusions and Future Work

5.1. Conclusions

This study comprehensively evaluated deep learning architectures, including the Vision Transformer (ViT), ResNet152, VGG19, Xception, and an ensemble Meta model, for multi-label chest X-ray image classification. The evaluation metrics, Hamming Loss and ROC-AUC, provided insights into the models' diagnostic capabilities. Among the individual models, VGG19 achieved the lowest Hamming Loss (0.1335) and excelled in detecting conditions like Edema, while ViT demonstrated strengths in tasks such as Hernia classification but had a higher overall Hamming Loss of 0.2426. Xception exhibited balanced but unremarkable performance across conditions, with a Hamming Loss of 0.1609. The ensemble Meta model outperformed all individual architectures, achieving the lowest Hamming Loss of 0.1408 and the highest ROC-AUC values across multiple diseases, underscoring its superior reliability and diagnostic precision. These results emphasize the ensemble Meta model's potential for clinical applications, where accurate and robust disease detection is crucial. This research highlights the importance of combining diverse model architectures to address the complexities of multi-label chest X-ray classification and sets a foundation for further advancements in developing AI-powered diagnostic tools. Key observations from this study include:

- The ViT model demonstrated moderate effectiveness, with notable performance in detecting conditions like Edema, but struggled with others like Hernia.
- ResNet152 and VGG19 showed similar levels of accuracy in their predictions, with VGG19 slightly outperforming in terms of lower Hamming Loss.
- Xception exhibited a balanced performance across various conditions but did not show exceptionally high discriminative ability for any specific disease.
- The Meta model emerged as the most effective, achieving the lowest Hamming Loss and consistently higher ROC-AUC values across different diseases, indicating its superior discriminatory capabilities.
- These findings highlight the Meta model's potential for clinical application, where accuracy and reliability are crucial.

5.2. Future Work

While some architectures may not have fully demonstrated their potential in this study, the extensive analysis provides a solid foundation for future research. Understanding each model's unique strengths and limitations is crucial in guiding subsequent model selection and refinement. Future research should continue to explore ensemble models and emerging architectures, focusing on enhancing diagnostic accuracy and efficiency in interpreting chest X-rays. This could significantly elevate patient care standards, reducing the diagnostic burden on healthcare professionals. Continued advancements in AI and machine learning, combined with a deeper understanding of medical imaging, are poised to transform the landscape of medical diagnostics. The pursuit of more accurate, efficient, and accessible diagnostic tools remains a promising and vital avenue in healthcare technology [39].

Author Contributions: H.N., E.Q.S. and S.B. conceived of the presented idea. H.N., E.Q.S., S.B., R.M., E.V. and W.A.J. developed the theory and performed the computation. H.N. planned and carried out the simulations. E.Q.S., S.B., R.M., E.V. and W.A.J. verified the analytical method. H.N. wrote the draft of the manuscript with input from all authors. E.Q.S., S.B., R.M., E.V. and W.A.J. revised and

edited the manuscript. E.Q.S. and S.B. supervised the project. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Geroski, T.; Filipović, N. Artificial Intelligence Empowering Medical Image Processing. In *In Silico Clinical Trials for Cardiovascular Disease: A Finite Element and Machine Learning Approach*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 179–208.
2. Huang, M.L.; Liao, Y.C. A lightweight CNN-based network on COVID-19 detection using X-ray and CT images. *Comput. Biol. Med.* **2022**, *146*, 105604. [CrossRef] [PubMed]
3. Jain, A.; Bhardwaj, A.; Murali, K.; Surani, I. A Comparative Study of CNN, ResNet, and Vision Transformers for Multi-Classification of Chest Diseases. *arXiv* **2024**, arXiv:2406.00237.
4. Liu, C.; Cao, Y.; Alcantara, M.; Liu, B.; Brunette, M.; Peinado, J.; Curioso, W. TX-CNN: Detecting tuberculosis in chest X-ray images using convolutional neural network. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 2314–2318. [CrossRef]
5. Rahman, T.; Chowdhury, M.E.H.; Khandakar, A.; Islam, K.R.; Islam, K.F.; Mahbub, Z.B.; Kadir, M.A.; Kashem, S. Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray. *Appl. Sci.* **2020**, *10*, 3233. [CrossRef]
6. Patil, P.; Patil, H. X-ray Imagining Based Pneumonia Classification using Deep Learning and Adaptive Clip Limit based CLAHE Algorithm. In Proceedings of the 2020 IEEE 4th Conference on Information & Communication Technology (CICT), Chennai, India, 3–5 December 2020; pp. 1–4. [CrossRef]
7. Basu, S.; Mitra, S.; Saha, N. Deep Learning for Screening COVID-19 using Chest X-Ray Images. In Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, ACT, Australia, 1–4 December 2020; pp. 2521–2527.
8. Alsaati, M.A.Y. Diagnosis of COVID-19 in X-ray Images using Deep Neural Networks. *Int. Res. J. Multidiscip. Technov.* **2024**, *6*, 232–244. [CrossRef]
9. Madani, A.; Moradi, M.; Karargyris, A.; Syeda-Mahmood, T. Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 1038–1042. [CrossRef]
10. Do, O.C.; Luong, C.M.; Dinh, P.H.; Tran, G.S. An efficient approach to medical image fusion based on optimization and transfer learning with VGG19. *Biomed. Signal Process. Control.* **2024**, *87*, 105370. [CrossRef]
11. Yimer, F.; Tessema, A.; Simegn, G. Multiple Lung Diseases Classification from Chest X- Ray Images using Deep Learning approach. *Int. J. Adv. Trends Comput. Sci. Eng.* **2021**, *10*, 1–7.
12. Onalaja, J.; Shahra, E.Q.; Basurra, S.; Jabbar, W.A. Image Classifier for an Online Footwear Marketplace to Distinguish between Counterfeit and Real Sneakers for Resale. *Sensors* **2024**, *24*, 3030. [CrossRef]
13. Saeed, M.; Ullah, M.; Khan, S.D.; Cheikh, F.A.; Sajjad, M. Vit based covid-19 detection and classification from cxr images. *Electron. Imaging* **2023**, *35*, VDA-407. [CrossRef]
14. Uparkar, O.; Bharti, J.; Pateriya, R.K.; Gupta, R.K.; Sharma, A. Vision Transformer Outperforms Deep Convolutional Neural Network-based Model in Classifying X-ray Images. *Procedia Comput. Sci.* **2023**, *218*, 2338–2349. [CrossRef]
15. Okolo, G.I. IEViT: An enhanced vision transformer architecture for chest X-ray image classification. *Comput. Methods Programs Biomed.* **2022**, *226*, 107141. [CrossRef]
16. Jiang, X.; Zhu, Y.; Cai, G.; Zheng, B.; Yang, D. MXT: A New Variant of Pyramid Vision Transformer for Multi-label Chest X-ray Image Classification. *Cogn. Comput.* **2022**, *14*, 1362–1377. [CrossRef]
17. Holste, G.; Zhou, Y.; Wang, S.; Jaiswal, A.; Lin, M.; Zhuge, S.; Yang, Y.; Kim, D.; Nguyen-Mau, T.H.; Tran, M.T.; et al. Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge. *Med. Image Anal.* **2024**, *97*, 103224. [CrossRef] [PubMed]
18. Öztürk, Ş.; Turalı, M.Y.; Çukur, T. Hydravit: Adaptive multi-branch transformer for multi-label disease classification from chest X-ray images. *arXiv* **2023**, arXiv:2310.06143. [CrossRef]
19. Jeong, J.; Jeoun, B.; Park, Y.; Han, B. An Optimized Ensemble Framework for Multi-Label Classification on Long-Tailed Chest X-ray Data. In Proceedings of the ICCV Workshop on Computer Vision for Automated Medical Diagnosis (CVAMD), Paris, France, 2 October 2023.

20. Kim, D. CheXFusion: Effective Fusion of Multi-View Features using Transformers for Long-Tailed Chest X-Ray Classification. *arXiv* **2023**, arXiv:2308.03968.
21. Krishnan, K.S.; Krishnan, K.S. Vision transformer based COVID-19 detection using chest X-rays. In Proceedings of the 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), Solan, India, 7–9 October 2021; pp. 644–648.
22. Showkat, S.; Qureshi, S. Efficacy of Transfer Learning-based ResNet models in Chest X-ray image classification for detecting COVID-19 Pneumonia. *Chemom. Intell. Lab. Syst.* **2022**, *224*, 104534. [CrossRef]
23. Ikechukwu, A.V.; Murali, S.; Deepu, R.; Shivamurthy, R. ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images. *Glob. Transitions Proc.* **2021**, *2*, 375–381. [CrossRef]
24. Gupta, A.; Anjum; Gupta, S.; Katarya, R. InstaCovNet-19: A deep learning classification model for the detection of COVID-19 patients using Chest X-ray. *Appl. Soft Comput.* **2021**, *99*, 106859. [CrossRef]
25. Xie, J.; Xu, B.; Chuang, Z. Horizontal and vertical ensemble with deep representation for classification. *arXiv* **2013**, arXiv:1306.2759.
26. Mallick, J.; Talukdar, S.; Ahmed, M. Combining high resolution input and stacking ensemble machine learning algorithms for developing robust groundwater potentiality models in Bisha watershed, Saudi Arabia. *Appl. Water Sci.* **2022**, *12*, 77. [CrossRef]
27. Acar, E.; Rais-Rohani, M. Ensemble of metamodels with optimized weight factors. *Struct. Multidiscip. Optim.* **2009**, *37*, 279–294. [CrossRef]
28. Lu, M.; Hou, Q.; Qin, S.; Zhou, L.; Hua, D.; Wang, X.; Cheng, L. A Stacking Ensemble Model of Various Machine Learning Models for Daily Runoff Forecasting. *Water* **2023**, *15*, 1265. [CrossRef]
29. Stemerman, R.; Arguello, J.; Brice, J.; Krishnamurthy, A.; Houston, M.; Kitzmiller, R.R. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. *JAMIA Open* **2021**, *4*, ooaa069. [CrossRef]
30. Tsoumakas, G.; Katakis, I.; Vlahavas, I. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 667–685.
31. Grandini, M.; Bagli, E.; Visani, G. Metrics for multi-class classification: An overview. *arXiv* **2020**, arXiv:2008.05756.
32. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv* **2017**, arXiv:1711.05225.
33. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3462–3471. [CrossRef]
34. Bharati, S.; Podder, P.; Mondal, M.R. Hybrid deep learning for detecting lung diseases from X-ray images. *Informatics Med. Unlocked* **2020**, *20*, 100391. [CrossRef]
35. Kim, S.; Rim, B.; Choi, S.; Lee, A.; Min, S.; Hong, M. Deep learning in multi-class lung diseases' classification on chest X-ray images. *Diagnostics* **2022**, *12*, 915. [CrossRef] [PubMed]
36. Ashraf, S.N.; Mamun, M.A.; Abdullah, H.M.; Alam, M.G.R. SynthEnsemble: A Fusion of CNN, Vision Transformer, and Hybrid Models for Multi-Label Chest X-Ray Classification. In Proceedings of the 2023 26th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 13–15 December 2023. [CrossRef]
37. Marikkar, U.; Atito, S.; Awais, M.; Mahdi, A. LT-ViT: A Vision Transformer for Multi-Label Chest X-ray Classification. *arXiv* **2023**, arXiv:2311.07263.
38. Sajed, S.; Sanati, A.; Garcia, J.E.; Rostami, H.; Keshavarz, A.; Teixeira, A. The effectiveness of deep learning vs. traditional methods for lung disease diagnosis using chest X-ray images: A systematic review. *Appl. Soft Comput.* **2023**, *147*, 110817. [CrossRef]
39. Ravi, V.; Narasimhan, H.; Pham, T.D. A cost-sensitive deep learning-based meta-classifier for pediatric pneumonia classification using chest X-rays. *Expert Syst.* **2022**, *39*, e12966. [CrossRef]