IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 October 2023, accepted 5 November 2023, date of publication 13 November 2023,
date of current version 17 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3332512

RESEARCH ARTICLE

A Comparison of Re-Sampling Techniques for Detection of Multi-Step Attacks on Deep Learning Models

MUHAMMAD HASSAN JAMAL[1], NAILA NAZ[1], MUAZZAM A. KHAN KHATTAK[1,2], (Senior Member, IEEE), FAISAL SAEED[3], SAAD NASSER ALTAMIMI[4], AND SULTAN NOMAN QASEM[4], (Senior Member, IEEE)

[1]Department of Computer Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan
[2]ICESCO Chair Big Data Analytics and Edge Computing, Quaid-i-Azam University, Islamabad 45320, Pakistan
[3]DAAI Research Group, College of Computing and Digital Technology, Birmingham City University, B4 7XG Birmingham, U.K.
[4]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia

Corresponding author: Muhammad Hassan Jamal (mhassan@cs.qau.edu.pk)

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-RG23052).

ABSTRACT The increasing dependence on data analytics and artificial intelligence (AI) methodologies across various domains has prompted the emergence of apprehensions over data security and integrity. There exists a consensus among scholars and experts that the identification and mitigation of Multi-step attacks pose significant challenges due to the intricate nature of the diverse approaches utilized. This study aims to address the issue of imbalanced datasets within the domain of Multi-step attack detection. To achieve this objective, the research explores three distinct re-sampling strategies, namely over-sampling, under-sampling, and hybrid re-sampling techniques. The study offers a comprehensive assessment of several re-sampling techniques utilized in the detection of Multi-step attacks on deep learning (DL) models. The efficacy of the solution is evaluated using a Multi-step cyber attack dataset that emulates attacks across six attack classes. Furthermore, the performance of several re-sampling approaches with numerous traditional machine learning (ML) and deep learning (DL) models are compared, based on performance metrics such as accuracy, precision, recall, F-1 score, and G-mean. In contrast to preliminary studies, the research focuses on Multi-step attack detection. The results indicate that the combination of Convolutional Neural Networks (CNN) with Deep Belief Networks (DBN), Long Short-Term Memory (LSTM), and Recurrent Neural Networks (RNN) provides optimal results as compared to standalone ML/DL models. Moreover, the results also depict that SMOTEENN, a hybrid re-sampling technique, demonstrates superior effectiveness in enhancing detection performance across various models and evaluation metrics. The findings indicate the significance of appropriate re-sampling techniques to improve the efficacy of Multi-step attack detection on DL models.

INDEX TERMS Deep learning (DL), machine learning (ML), multi-step attacks, synthetic minority over-sampling technique (SMOTE), borderline SMOTE, SMOTEENN, SMOTETomek.

I. INTRODUCTION

The current practice of examining datasets to extract significant results is motivated by the ubiquitous presence of data, which is fundamentally transforming numerous organizations [1]. The remarkable expansion can be attributed to the significant rise in online interactions across various platforms and devices, alongside the continuous proliferation of digital technologies and the Internet of Things (IoT) ecosystem [2]. The integration of artificial intelligence (AI) techniques into the analytical process is a pivotal element contributing to the transformative impact of data analytics.

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

Since 2000, there has been a notable increase in the resources allocated toward the research and advancement of AI which will have long-term effects on the data analytics industry. Organizations need powerful AI tools and frameworks to efficiently manage large data sets and carry out complex ML operations to perform extensive data processes [3]. ML and DL are two subfields of AI that help computers analyze large amounts of data to find patterns and insights in order to make future predictions [4]. Efficient algorithms made possible by ongoing AI development have allowed organizations to extract inaccessible information data to achieve useful insights. The importance of data analytics has increased the need to take precautions to safeguard sensitive information. With so much data being collected and analyzed, there is a real risk of cyber attacks and data leaks [5]. As the domain of data analytics continues to advance, it becomes imperative to tackle security issues to uphold the reliability of insights derived from data. Implementing privacy safeguards for data, using strong encryption techniques, enforcing tight access rules, and applying efficient detection procedures are all necessary to protect sensitive data from unauthorized access or attacks [6].

Attacks encompass any illegitimate attempt aimed at compromising the confidentiality, integrity, or availability of a network's system, hardware, software, or data. These attacks are initiated by malignant individuals or entities with the intention of stealing information, sabotaging systems, inducing service disruptions, or inflicting other forms of harm. The attacks can be varied, either they are single-step or multi-step attacks [7]. A single-step attack involves a straightforward and focused attempt to exploit a weakness in a system or network, in which the attacker typically employs a single method to breach the target's security defenses. For example, Password Guessing, SQL Injection, Clickjacking, etc. are types of single-step attacks [8]. The term ''Multi-step attack'' refers to a type of cyber-attack that involves multi-steps or stages to achieve the attacker's goal. Instead of relying on a single vulnerability or method, such attackers employ a combination of techniques, or strategies to achieve their objective. These attacks are typically more sophisticated and difficult to detect or mitigate than single attacks. Examples of Multi-step attacks include HTTP DDoS, ICMP Flood, Web Crawler, etc [9]. In this study, different re-sampling techniques have been compared on a Multi-step cyber attack dataset using different evaluation measures to check the effectiveness of several traditional ML algorithms.

The applications are used for technological purposes in different areas such as intrusion detection, attack diagnosis, fraud detection, etc., and their data is organized for initial research. If the datasets are imbalanced then they must be manipulated before the modeling procedure [10]. For imbalanced datasets, the focus is on minority classes because the majority classes are not of interest, which leads to an under-fit or an over-fit model. In this scenario, in order to overcome this issue a re-sampling technique would have to be implemented before modeling the data. Re-sampling uses the sample data in an effective method as accuracy can be increased and the ambiguity of a population variable can be assessed [11].

Re-sampling methods are classified as * Over-sampling. * Under-sampling. * Hybrid re-sampling.

The predominant form of re-sampling is over-sampling, which augments the original data set with supplementary instances from the minority class. There are several types of over-sampling methods employed in data analysis, including the Synthetic Minority Over-Sampling Technique (SMOTE) and the Borderline SMOTE, among other techniques [12]. The SMOTE algorithm involves the random selection of a minority class instance, followed by the identification of its K-nearest neighbors. From these neighbors, one is chosen, and a new synthetic instance belonging to the minority class is generated and appended to the training set. In contrast, Borderline SMOTE is characterized by its association with the decision boundary between the minority and majority classes. Synthetic data is exclusively created along this boundary to reduce the occurrence of misclassifications [13]. Under-sampling is a method that reduces the proportion of the majority class until it is comparable to the minority class proportion [14]. Over-sampling methods are often preferred to under-sampling, as removing instances of the majority class may lead to missing out on necessary insights and repeating the data which is the reason for overfitting [15]. The primary method of under-sampling is employed to randomly choose and eliminate the instances of the majority class, therefore generating the training set. In the Edited Nearest Neighbour (ENN), instances that result in misclassification of the majority class are removed. In Tomek, the cross-pairs, i.e., the pairs of different binary classes with the least distance are selected because these instances define the class boundary; this link between the cross-pair is called Tomek-link [16]. Under-sampling techniques used extensively in the existing literature include Edited Nearest Neighbor (ENN), Neighborhood Cleaning Rule, Tomek, Instance Hardness Threshold (IHT), One side selection, etc. Hybrid re-sampling is a method in which over-sampling and under-sampling methods are combined to overcome the drawbacks of the individual models [17]. In this technique, an over-sampling method is applied to data initially, followed by the under-sampling method. SMOTETomek, and SMOTEENN, are two examples of hybrid re-sampling. The process of hybrid re-sampling methods comprises two steps. In the first step, over-sampling methods are applied, and under-sampling is applied in the second step [18]. The initial stage of the SMOTEENN technique closely resembles that of the SMOTETomek approach, wherein SMOTE is employed to generate synthetic instances for the minority class. However, the subsequent step in the SMOTETomek method diverges, but it utilizes ENN to eliminate observations from both the minority and majority classes that result in misclassifications [19]. We utilized four balancing techniques in this research: Synthetic Minority Over-sampling Technique (SMOTE), Borderline SMOTE,

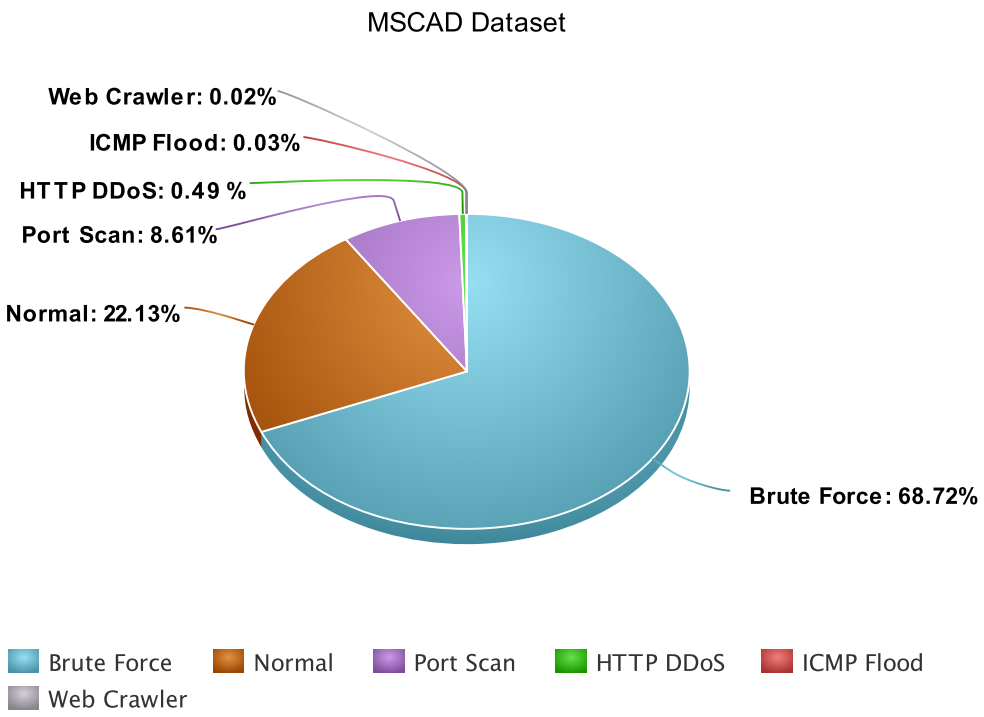

FIGURE 1. The data properties.

SMOTENN, and SMOTETomek in ML to guarantee that our dataset was representative of the total population.

The primary goal of ML is to enable computers to derive semantic understanding from data through the identification of patterns and relationships. This is achieved by constructing models based on these patterns, which are subsequently utilized to draw inferences or make predictions [20]. ML encompasses three distinct forms of learning, namely supervised learning, unsupervised learning, and reinforcement learning. Algorithms and models are developed that allow machines to learn from provided training data and improve their performance over time without being explicitly programmed [21]. Researchers have implemented different traditional models to observe the data and used multiple re-sampling techniques to enhance the model's performance [22]. This work focuses on the significance of the re-sampling technique and evaluates the influence of this technique on the traditional model's performance. We have used a Multi-step cyber attack dataset, which contains six classes: Brute Force, Normal Class, Port Scan, HTTP DDoS, ICMP Flood, and Web Crawler [23]. Several traditional ML models have been employed to measure the effectiveness of these models against the above-mentioned attacks.

The main contributions of this research study are:

- The performance of different machine learning and deep learning models is evaluated for Multi-step cyber attacks without applying any resampling techniques.
- The multiple resampling techniques (SMOTE, BorderlineSMOTE, SMOTEENN, and SmoteTomek) with more than 10 different ML/DL models were evaluated as well.
- The best-performing models are identified based on accuracy, precision, recall, F-1, and G-mean scores.
- The study emphasizes the detection of Multi-step attacks, which has not been the primary focus in previous research.
- Convolutional neural networks generally perform better by combining them with other different ML/DL models for Multi-step attacks.

The research work is organized into five sections: the first section is the introduction, and the second section is the literature review, which highlights the existing work conducted in the specified area. The third section discusses the methodology selected for this research. The results are discussed in the fourth section, and lastly, the research work is summarized in the conclusion section.

II. LITERATURE REVIEW

AI has become an indispensable tool for conducting various analyses, including descriptive, predictive, and prescriptive analytics, offering businesses deeper insights and more informed decision-making capabilities [24]. Moreover, the significance of real-time data analysis has grown substantially. The availability of advanced data streaming and processing technologies empowers organizations to analyze data as it is generated, enabling them to respond swiftly to changing circumstances and make well-informed decisions

in real-time [25]. For that purpose, we have explored the existing literature which has been summarized in Table 1.

In [23], the author examines the necessity of a reliable and current dataset for effectively training Intrusion Detection Systems (IDS) to identify and respond to cyberattacks. The researchers present a unique Multi-step cyber-attack dataset (MSCAD) comprising two distinct assault scenarios characterized by the involvement of numerous phases. The MSCAD was utilized to train the Intrusion Detection System (IDS) and evaluate its effectiveness based on G-mean and Area Under Curve (AUC). The author proceeds to discuss other essential open-source and publicly available datasets that are based on specific criteria. The authors place great emphasis on the relevance of employing the latest machine learning techniques in the development of detection models, as well as the value of selecting pertinent features for the identification of Multi-step attacks. The importance of employing appropriate criteria for assessing the efficacy of the detection model on an imbalanced dataset is also emphasized.

The author [26] presents a novel unbalanced learning approach called AWGSENN, which aims to address the issue of imbalanced data. In the AWGSENN algorithm, the starting weights are determined based on two factors: the count of majority class neighbors and the distance between the minority instance and its neighboring instances. Subsequently, a probability density function is formulated, which exhibits a shape resembling that of the Gaussian distribution, to generate new examples in a nonlinear manner. The data is processed using the edited closest neighbor criterion to eliminate instances that exhibit overlapping and noisy occurrences. The proposed methodology is evaluated using a total of 37 datasets obtained from the KEEL data collection. The performance of the suggested strategy is compared against five different over-sampling strategies and two hybrid techniques. The empirical investigation findings provide evidence that their method exhibits superior performance compared to the current state-of-the-art, as indicated by considerable improvements in both the G-mean and the Area Under the Curve metrics. The results obtained from the Wilcoxon signed-rank test provide evidence that their method exhibits superior performance compared to competing re-sampling strategies. The proposed system is employed for the purpose of detecting Android malware, and the findings demonstrate the potential efficacy of this approach.

In the discourse surrounding COVID-19 and its variants, [27] highlights that individuals with specific preexisting medical disorders, such as thyroid illness, Hepatitis C virus (HCV) infection, breast tissue disease, chronic dermatitis, and other serious ailments, may face an elevated susceptibility to catching the virus. The prompt and precise diagnosis of these illnesses is of utmost importance. There is an urgent need for the implementation of preventive testing measures in populous nations such as India. An imbalance in classification arises when classifiers accurately classify instances belonging to the majority class but inaccurately classify instances from the minority class. The act of mislabelling in the context of human life is deemed unacceptable. A variety of data balancing strategies were employed across many ML algorithms to mitigate the issue of misclassification and enhance accuracy in datasets of this nature. The findings are promising, indicating a significant enhancement in accuracy. By obtaining precise diagnoses, the proposed study can assist patients in preventing the acquisition of potentially fatal diseases and infections.

The author has introduced a novel approach in [28] to improve the Synthetic Minority Over-sampling Technique (SMOTE) by integrating it with the Kalman filter. It is frequently observed that there exists a significant disparity in the number of samples among different groups. Therefore, the accuracy of the predictor is negatively impacted. The Synthetic Minority Over-sampling Technique (SMOTE) is widely recognized as one of the prominent algorithms utilized to address imbalanced datasets by generating synthetic data. Nevertheless, the classifier's subpar performance cannot simply be attributed to data imbalance. Multiple research papers have demonstrated the significance of using noisy samples in the misclassification of data sets. Handling large data sets can be computationally demanding. Hence, it is imperative to optimize the efficiency of data management. The proposed technique, known as Kalman-SMOTE (KSMOTE), aims to diminish the dataset's magnitude by eliminating noisy samples from the resultant dataset following the application of SMOTE. The dataset consists of both the original data and artificially created samples. The accuracy of the model has been confirmed by multiple datasets. In comparison to contemporary methodologies, their model demonstrates superior performance in experimental evaluations.

The study conducted by [29] addresses the issue of learning from imbalanced data, a significant concern in various domains. This imbalance can impede the efficacy of conventional learning approaches that assume a uniform data distribution. A significant disparity in the size of two groups gives rise to a phenomenon sometimes referred to as a ''class imbalance problem.'' The utilization of imbalanced data in ML would result in biased conclusions and inaccurate predictions due to the significant gap between the majority and minority classes. Numerous algorithms have been developed in response to the increasing popularity of this topic of study. Nevertheless, the algorithm's objective assessment is limited. The objective of this study is to conduct a comparative analysis of five prominent data sampling strategies, namely SMOTE, ADASYN, Borderline SMOTE, SMOTETomek, and RUSBoost, with respect to their effectiveness in addressing class imbalance problems. The efficacy of each technique is thoroughly examined based on evaluation indicators, and a comparative analysis is conducted.

The authors in [30] obtained a dataset on Portuguese bank marketing from the UCI ML repository and conducted an analysis to evaluate several re-sampling techniques on this imbalanced dataset. The prevailing trend in the corporate

TABLE 1. Summary of literature review.

References	Year	Dataset Name	Methodology	Evaluation Measures	Attack Type	K-Fold	Resampling Technique
[23]	2022	MSCAD	DT, RF	G-mean, AUC	Multi	Yes	Yes
[26]	2019	KEEL Repository	AWGSEEN	G-mean, AUC	Single	No	Yes
[27]	2022	Medical Datasets	LDA, DT, SVM NB, KNN, ANN	Accuracy, Precision, Recall, F-1 score	Single	No	Yes
[28]	2022	UCI	DT, RF	Accuracy, Precision, Recall, F-1 score	Single	No	Yes
[29]	2013	UCI	SVM	Accuracy, Precision, Recall, F-1, G-mean	Single	No	Yes
[30]	2022	Portuguese Bank	LR	Accuracy, Precision, Recall, F-1, AUC	Single	No	Yes
[31]	2021	PLCO NLST	LR, RF, SVM	AUC	Single	Yes	Yes
[32]	2023	Credit card fraud	KNN, LR, LDA, NB, CART	Accuracy, Precision, Recall	Single	No	Yes
[33]	2019	SDP Research	DT, LR, NB, RF, SVM	Recall	Single	Yes	Yes
[34]	2023	MSCAD	KNN	Accuracy, Precision, Recall, F-1 score	Multi	No	No
[35]	2023	MSCAD	CNN-DBN	Accuracy, Precision, Recall, F-1 score	Multi	No	Yes
This Study	2023	MSCAD	DT, RF, NB DNN, LSTM, GRU, CNN, RNN, CNN-LSTM, CNN-RNN, CNN-DBN	Accuracy, Precision, Recall, F-1, G-mean	Multi	Yes	Yes

sphere is utilizing ML techniques to extract valuable insights from readily available data, aiming to provide effective solutions for organizations. To effectively implement ML in several domains, including banking, medical diagnosis, and fraud detection, it is imperative to address the issue of data imbalance. One often utilized approach to address this issue is the utilization of re-sampling techniques. Scientists from many global institutions have collaborated to devise a collection of distinct re-sampling methodologies. The study examines the impact of various re-sampling techniques on the efficacy of a conventional ML model, namely logistic regression. The objective of the study is to forecast the subscription status of a customer's term deposit by binary categorization. Various metrics, such as the Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC), Accuracy, Precision, Recall, and F-1 score, are employed to evaluate the performance of the model. The empirical findings indicate that the K-Means SMOTE model had the highest performance with an accuracy rate of 94%, followed by the SMOTEENN model with an accuracy rate of 92%. Logistic regression was used to apply these models to the Portuguese banking dataset.

The problem of data skew in the training of ML algorithms is investigated by [31]. They targeted the two most well-known, and commonly considered to be representative, lung cancer datasets; PLCO and NLST. Both data sets have an imbalance ratio of 24.7, where the majority class is over-represented in comparison to the minority class. The goal is to make use of these data sets to make future lung cancer cases more predictable. Results from 23 class imbalance strategies (such as re-sampling and hybrid systems) are compared to those from three widely used classifiers (logistic regression, random forest, and LinearSVC) to determine which strategies are more effective in resolving class imbalance. The re-sampling procedure has the capability to utilize a total of eleven distinct under-sampling techniques, including RUS, etc. Additionally, it incorporates seven various over-sampling approaches, such as SMOTE, etc. Furthermore, it incorporates two integrated sampling methods, namely SMOTEENN and SMOTETomek. Hybrid systems encompass various methodologies, among which Balanced Bagging is included. The findings demonstrate that the utilization of class imbalance learning techniques enhances the model's efficacy in classification tasks.

Under-sampling is associated with a higher standard deviation (SD) compared to other forms of imbalanced sampling, whereas over-sampling is linked to a lower standard deviation. Over-sampling has been found to enhance the model's Area Under the Curve (AUC) and is considered a dependable approach in data analysis. The random forest algorithm implemented in ROS demonstrates superior predicting capabilities when applied to diverse datasets related to lung cancer.

To efficiently address the issue of fraudulent card usage, [32] propose a methodology for its mitigation. The incidents of credit card scams have had a significant impact on the established economic structure of the market, leading to a disruption in its functioning. Moreover, it has resulted in a loss of trust among several key stakeholders, including financial institutions and consumers. Card fraud results in the annual loss of billions of dollars. In recent times, the distribution of information pertaining to fraudulent card transactions has been imbalanced due to a significant surge in the volume of valid transactions. When there is a significant disparity in the number of samples between two classes, the presence of imbalanced data arises, which requires resolution prior to tackling the issue of fraud. Hence, the task of categorizing fraud has grown increasingly challenging due to the potential biases toward the dominant group. The primary objective of this study is a two-way approach: firstly, to examine the problem of biased data by employing hybrid re-sampling and over-sampling preprocessing techniques; and secondly, to mitigate fraudulent activities. To assess the efficacy of the proposed framework relative to established algorithms such as KNN, LR, LDA, NB, and CART, various metrics pertaining to accuracy, precision, and recall are employed. The empirical evidence demonstrates that the model aims to detect instances of fraudulent transactions even in scenarios when the dataset exhibits a significant imbalance. Furthermore, the predictive capacity of the system in accurately determining desired classes has significantly improved, now achieving an accuracy rate of 99.9%.

According to [33], the Software Defect Prediction (SDP) models encounter challenges in accurately identifying defective instances, mostly attributed to the presence of significantly skewed data. In recent times, several solutions have been proposed to address the issue of class imbalance. Among these strategies, the over-sampling strategy has gained significant recognition. This approach ensures a consistent proportion of defective to non-detective examples by intentionally generating new instances with defects. Nevertheless, these methodologies would generate artificial samples that exhibit a limited range of variation and a significant amount of irrelevant data. Therefore, the authors proposed the implementation of a Cluster-based Over-sampling with noise filtering (KMFOS) approach as a recommended solution for addressing the problem of class imbalance in SDP. KMFOS generates new instances with faults by employing an interpolation technique that combines existing instances from two distinct clusters. Subsequently, the newly generated instances with defects have an unequal distribution throughout the dataset containing defects. The noise occurrences are mitigated by implementing the Closest List Noise Identification (CLNI) technique, which expands the cluster-based over-sampling approach. A comprehensive assessment was performed on a total of 24 projects to compare the effectiveness of KMFOS with various over-sampling techniques, namely SMOTE, Borderline SMOTE, ADASYN, random over-sampling (ROS), K-means SMOTE, SMOTE C IPF, SMOTE C ENN, and SMOTE C Tomek Links. In addition, they assess the performance of KMFOS in comparison to other state-of-the-art methods for addressing class imbalance, including the Balanced Bagging classifier, RUSboost classifier, Instance Hardness Threshold, and cost-sensitive strategies. Based on the conducted studies, it has been determined that the KMFOS achieves superior performance in terms of Recall compared to both conventional over-sampling strategies and other class-imbalance approaches. Hence, the utilization of the KMFOS technique proves to be a proficient approach to generating comprehensive data for SDP, thereby enhancing the effectiveness of forecasting models.

In [35], the author addresses the utilization of machine learning techniques for the purpose of identifying and detecting intrusions within Industrial Control Systems (ICS). The authors elucidate the process of cleaning, normalizing, and balancing data as a crucial step in the pre-processing phase for machine learning. They presented a hybrid methodology for detecting assaults and evaluated its effectiveness in comparison to established deep-learning techniques. The researchers reached the conclusion that the efficacy of the security mechanism can be greatly enhanced through the integration of several techniques in deep learning and machine learning.

The author in [34] effectively illustrates the necessity of employing advanced protocols in order to identify and mitigate the growing prevalence of intricate and frequent breaches. The authors present a detection methodology that relies on a K-nearest neighbors (KNN) classifier and ensemble approaches as a means of safeguarding intelligent environments against intrusions. The authors considered the importance of comprehending the vulnerabilities present inside a network and understanding the underlying motivations that drive attackers. The many forms of cyberattacks, such as malware, zero-day exploits, and DoS attacks, are also emphasized. The proposed method exhibits a high level of accuracy in identifying instances of cyber-attacks due to its utilization of a dataset that is deemed reliable and credible. This article explains the significance of enhancing the detection system within enterprises as a means to mitigate the risk of cyber-attacks.

III. PROPOSED METHODOLOGY

This research consists of three sections, namely, Data, ML/DL Models, and Comparative Analysis. The process that we followed is illustrated in Figure 1. Jupyter Notebook and

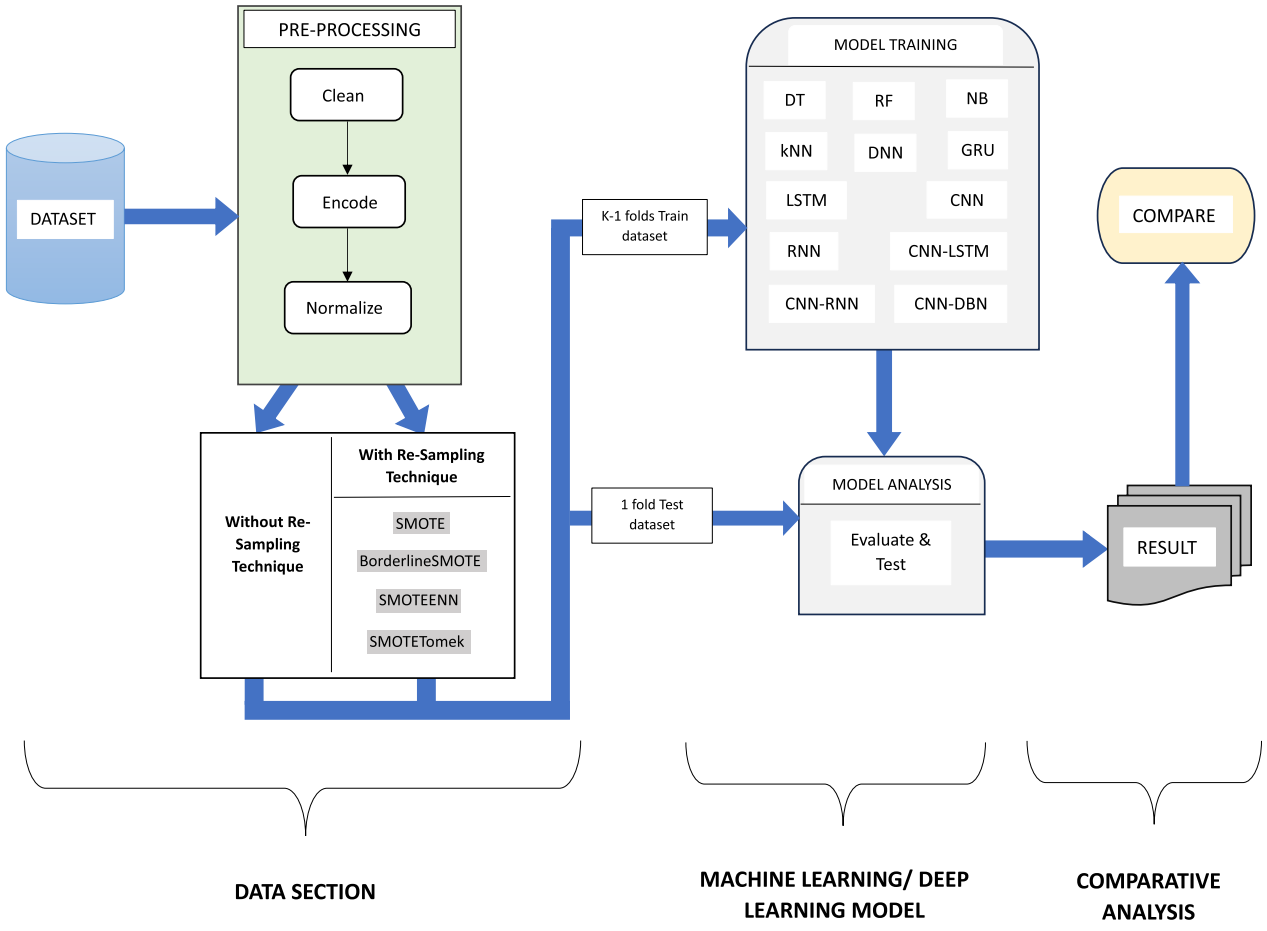

FIGURE 2. The workflow of study.

Python 3.10 were used to perform the experiments on the dataset using different traditional models.

A. DATA SECTION

The first section of this research is the data section, which also has three parts, namely the dataset, which is the collection of data, pre-processing, and re-sampling of the data, as mentioned below.

We have selected a Multi-step cyber attack dataset (MSCAD) for Multi-step attack detection. The dataset was collected through a testbed, which consisted of an attack network and a victim network. The MSCAD dataset is special because it includes instances of Multi-step cyber attacks, which are complex attacks that require several stages or processes to accomplish their goals. The Multi-step cyber attack dataset (MSCAD) was introduced due to the shortcomings of other datasets, which are unable to effectively deal with Multi-step attacks. MSCAD is homogeneous and contains the following types of attacks;

- The goal of a Brute Force attack is to gain unauthorized access to a system by trying every possible password or encryption key unless an appropriate one can be obtained. Such attacks may result in severe data violations, illegitimate system access, and even damage an individual's wealth or reputation.
- An HTTP DDoS (Hypertext Transfer Protocol-Distributed denial of service attack involves making an overwhelming amount of HTTP requests to a website or online application. This may cause temporary website slowdowns or pauses, but the long-term effects are usually insignificant.
- An ICMP (Internet Control Message Protocol) flood attack involves crashing the network with a high number of ICMP packets and is one type of DDoS attack that targets a network. This attack might interrupt network communication, compromising normal network functions.
- Port scanning is usually the first stage of an intensive attack. A port scan attack is an attacker's reconnaissance approach for finding open ports on an intended system. Port scans may find possible shortcomings that can be leveraged in succeeding attacks.
- A web crawler, also referred to as a web scraper, is an application that automatically examines websites,

gathers data, and follows links in order to organize and gather information. Unauthorized data scraping by malevolent web crawlers may breach website terms of service and cause issues relating to privacy.

The dataset consists of 128,799 instances and 67 features. The total number of instances in each class label is shown in Figure 1. The figure depicts that the dataset contains 22.13% Normal, 68.71% Brute Force, 8.61% Port Scan, 0.49% HTTP DDoS, 0.03% ICMP Flood, and 0.02% Web Crawling instances [35]. The testbed for this dataset was built on the basis of two scenarios. Port scanning was initially performed to identify open ports. In the first instance, the web crawler was launched alongside a brute force attack against any host on the victim network. As a result, three hosts in the victim network were impacted. The Radware tool was utilized in the second case to perform an HTTP DDoS attack and an ICMP Flood. In this case, three different hosts were compromised. It should be emphasized that during both cases, ordinary traffic was initiated within the victim network [23].

After selecting the dataset, it was then pre-processed as it could potentially contain some redundant or empty values. For that purpose, the dataset was cleaned initially, which focuses on removing any complex or incomplete data. There are numerous approaches available for removing redundant data. After cleaning, all the labeled data is presented in the form of a string, which should be converted into an integer so that it can be used for further experiments. To achieve that goal, we used an encoding mechanism that eventually converts those strings into the form of integers so that they can be used to model the data. This encoding mechanism enumerates the categorical features using the label encoding method. The data is then normalized to predict the results by correlating the features. This technique is used for a distinct range of values, such as when one attribute has a value between 0 and 1, and the other attribute contains a value ranging from 10 to 1000. This dataset also has a diverse range of values, and this problem is handled by executing normalization. Once the pre-processing is completed, the dataset is then used to train the model. For that purpose, four different types of re-sampling techniques consisting of over-sampling, under-sampling, and hybrid re-sampling were used on multiple different traditional ML approaches which are explained in Section I: Introduction.

B. MACHINE LEARNING/ DEEP LEARNING MODELS

ML is a specialized domain of AI that is primarily concerned with the development of algorithms and models capable of enabling machines to discern patterns within datasets and derive meaningful insights from them. As these models are exposed to novel data, they acquire knowledge from prior instances and progressively enhance their efficacy [36]. DL is a specific sub-field of ML that focuses on developing models capable of learning hierarchical representations of data. These models are particularly effective in dealing with complex and large-scale datasets, where traditional ML algorithms may struggle to extract meaningful patterns [37]. This learning approach consists of two steps. The process of model evolution focuses on the validation and training of the data, whereas the model analysis involves the evaluation and testing of the dataset. We used multiple re-sampling techniques to model the data using several traditional models for this experiment. First, we used 80% of the data from the dataset for training, while 20% of the data was considered for testing the model, which hadn't provided a satisfactory performance. To address this issue, we preferred K-fold cross-validation. The models were compared using metrics such as Accuracy, Precision, Recall, F-1, and G-mean scores. Finally, the data was stored for subsequent analysis. The workflow used in this research study is illustrated in Figure 2.

The experiments are performed on every above-mentioned model for each re-sampling technique to consider the models for comparison. Furthermore, the model is evaluated and the potential bias in the data is mitigated through the implementation of a 5 K-fold cross-validation technique [38]. For DL models, we tested them via Adam optimizer, and the training process was done for five epochs. These models had 13276 instances which were tested with a batch size of 32 and verbose of 1 and the loss was calculated via sparse categorical entropy. As, in these models, we worked on multi-class classification, in this context, the Softmax activation function is used for the output layer.

The K-nearest Neighbors (KNN) technique is a valuable tool in machine learning for classification and regression applications [39]. Its simplicity of implementation and high level of accuracy make it particularly advantageous in these contexts [40]. The K-nearest neighbors (KNN) mathematical model encompasses three main steps: selecting an appropriate distance metric, identifying the K-nearest neighbors, and generating predictions based on these neighbors which are discussed in equation 1.

$$\hat{y} = argmax_c \sum_{i=1}^{k} \delta(y_i, c) \tag{1}$$

- $\hat{y}$ is the predicted class label.
- k is the number of nearest neighbors.
- c indicates the class labels.
- y_i is the class label of the i-th nearest neighbor.
- $\delta(y_i, c)$ is the Kronecker delta function, which is 1 if $y_i = c$ (i.e. if the i-th nearest neighbor relates to class c), or else 0.

A Deep Belief Network (DBN) is a deep learning architecture consisting of multiple layers of Restricted Boltzmann Machines (RBMs). Restricted Boltzmann Machines (RBMs) utilize stochastic generative neural networks as a means of unsupervised learning and feature extraction [41], [42]. It is imperative to provide clear definitions for Restricted Boltzmann Machines (RBMs) and the associated training algorithms in order to formulate the mathematical representation of a Deep Belief Network (DBN), which is

depicted in equation 2 below.

$$P(v, h^1, h^2, \ldots, h^n) = \frac{1}{Z} \exp(-E(v, h^1) - E(h^1, h^2) - \ldots - E(h^{N-1}, h^N)) \quad (2)$$

- $P(v, h^1, h^2, \ldots, h^N)$ is the joint probability distribution of visible and hidden layers.
- Z is the partition function assuring probabilities sum to be equal to 1.
- E signifies the energy function linked to every pair of layers.
- v illustrates the visible layer and h^i embodies the i-th hidden layer in the DBN.

The Convolutional Neural Network (CNN) architecture is predominantly utilized in the fields of identification and computer vision [43]. The mathematical framework of Convolutional Neural Networks (CNNs) encompasses many hierarchical levels and computational procedures which are represented in equation 3.

$$Output = Activation(\sum\nolimits_{i=1}^{N}(W_i * Input_i) + b) \quad (3)$$

- *Output* represents the features produced by the layer.
- *Activation* is the activation function employed element-wise.
- N is the number of filters (kernels) in the layer.
- W_i depicts the i-th filter's weight.
- $Input_i$ is the i-th input feature.
- b is the bias term.
- $*$ represents the convolution operation between the filter and input feature.

The utilization of prior temporal information in its hidden state enables a Recurrent Neural Network (RNN) to effectively process sequential input. In order to formally express a recurrent neural network (RNN) [44] in mathematical terms, it is necessary to establish equations that delineate the process of updating the hidden state at each time step, taking into account both the input and the prior hidden state [45]. The RNN is represented in equation 4 as.

$$h_t = tanh(W_{hx}x_t + W_{hh}h_{t-1} + b_h) \quad (4)$$

- h_t is the hidden state at time step t.
- x_t is the input at time step t.
- W_{hx} and W_{hh} are weight matrices.
- b_h is the bias vector.
- *tanh* is the hyperbolic tangent activation function.

C. COMPARATIVE ANALYSIS

This section comprises all the results of every re-sampling method for evaluation metrics. The traditional models were used on the Multi-step attacks dataset to compare the re-sampling techniques. The G-mean score has been illustrated via graphical representation to compare the results of every efficient performing model for each re-sampling technique and the results of all the metrics are presented in the tables.

IV. RESULTS & DISCUSSION

The dataset used in this research contains 128799 total records and each record had 67 attributes. The data was pre-processed, validated, and trained during the model development. The experimental setup is explained, and the comparison of the models has been analyzed with the help of the results produced. The results of the state-of-the-art models are compared using re-sampling techniques, and the overall outcomes are summarized. The results are depicted in the tables below and the performance of the models has been evaluated and compared using performance metrics namely Accuracy, Precision, Recall, F-1, and G-mean score. In this experiment, Windows 11 is used with an I5-11th Gen Laptop which contains a 2.4 GHz processor and 8GB of RAM.

We completed the experiments based on four approaches. In the first experiment, we compared the model's results without using any re-sampling technique. The second experiment was performed on other models using four strategies: SMOTE, BorderlineSMOTE, SMOTEENN, and SMOTETomek. We compared the results of [23], [34], and [35], and some other models on which we have performed the experiments to compare their performance. The models were then evaluated and the results were compared by using the Multi-step cyber attack dataset.

In the first experiment, we did not use any re-sampling strategy and the results are displayed in Table 2. The results show that the accuracy of KNN and CNN-DBN of [35] is marginally better than other models. The precision score for KNN, CNN-RNN, and CNN-DBN of [35], the recall score for KNN, CNN-RNN, and CNN-DBN of [35], the F-1 score for KNN, CNN, CNN-RNN and CNN-DBN of [35], and the G-mean score for KNN and CNN-DBN outperforms other models respectively. For SMOTE, we compared all the results and we have found that the hybrid CNN-DBN model of [35] and the KNN model performed well when compared to other models in Table 3. If we look at DL models, the CNN-DBN outperformed other DL models with an accuracy of 0.976. CNN, LSTM, CNN-LSTM, and CNN-RNN also showed reasonable results, but their performance is less accurate when they are compared to the CNN-DBN model. For other traditional ML models, KNN provides the best accuracy of 0.992, precision of 0.993, recall of 0.992, F-1 score of 0.991, and G-mean score of 0.991.

TABLE 2. Comparison results without re-sampling.

References	Model	Accuracy	Precision	Recall	F-1	G-Mean
[34]		0.825	0.942	0.911	0.926	x
[23]	DT	x	x	x	x	0.710
	RF	x	x	x	x	0.650
[35]	NB	0.787	0.462	0.627	0.313	0.451
	DNN	0.687	0.114	0.166	0.135	x
	LSTM	0.990	0.708	0.668	0.682	x
	GRU	0.842	0.404	0.331	0.344	x
	CNN	0.993	0.813	0.770	0.788	x
	RNN	0.990	0.777	0.753	0.763	x
	CNN+DBN	0.996	0.820	0.779	0.801	0.813
Results for this Study	KNN	0.997	0.927	0.823	0.850	0.751
	CNN+LSTM	0.994	0.812	0.759	0.781	0.792
	CNN+RNN	0.995	0.817	0.772	0.791	0.804

TABLE 3. Comparison results for SMOTE oversampling.

References	Model	Accuracy	Precision	Recall	F-1	G-Mean
[34]		0.825	0.942	0.911	0.926	x
[23]	DT	x	x	x	x	0.780
	RF	x	x	x	x	0.790
[35]	NB	0.641	0.644	0.641	0.573	0.399
	DNN	0.166	0.027	0.166	0.047	x
	LSTM	0.970	0.972	0.971	0.971	0.971
	GRU	0.865	0.886	0.865	0.862	0.851
	CNN	0.967	0.968	0.967	0.967	0.966
	RNN	0.954	0.956	0.954	0.955	0.954
	CNN+DBN	0.976	0.976	0.975	0.975	0.975
Results for this Study	KNN	0.992	0.993	0.992	0.992	0.991
	CNN+LSTM	0.975	0.976	0.975	0.975	0.974
	CNN+RNN	0.970	0.971	0.970	0.970	0.970

TABLE 4. Comparison results for BorderlineSMOTE oversampling.

References	Model	Accuracy	Precision	Recall	F-1	G-Mean
[34]		0.825	0.942	0.911	0.926	x
[23]	DT	x	x	x	x	0.780
	RF	x	x	x	x	0.780
Results for this Study	NB	0.525	0.626	0.526	0.483	0.404
	DNN	0.167	0.028	0.166	0.048	x
	LSTM	0.866	0.878	0.866	0.867	0.862
	GRU	0.665	0.718	0.665	0.660	0.636
	CNN	0.844	0.864	0.844	0.845	0.838
	RNN	0.842	0.853	0.843	0.843	0.837
	CNN+DBN	0.873	0.882	0.874	0.875	0.870
	KNN	0.973	0.974	0.973	0.973	0.973
	CNN+LSTM	0.871	0.878	0.871	0.872	0.867
	CNN+RNN	0.858	0.864	0.858	0.858	0.853

The results for BoderlineSMOTE explain that the CNN-DBN of [35] and KNN produced massive results as displayed in Table 4. For DL, the CNN-DBN model has the best accuracy, precision, recall, F-1, and G-mean scores of 0.873, 0.882, 0.874, 0.875, and 0.870 respectively when compared with other DL models and when we compare the results of KNN with [23] and [34] and other traditional ML algorithms, there is a huge difference as KNN has better evaluation results.

For the performance of SMOTEENN, the results of KNN are at their peak with an accuracy of 0.999. But when we compare other results the CNN-DBN and CNN-LSTM models have better scores. If we compare CNN-DBN and CNN-LSTM, there is a marginal difference in results as CNN-DBN produced accuracy precision, recall, F-1, and G-mean scores of 0.986, 0.985, 0.986, 0.986, and 0.985 respectively and CNN-LSTM provided scores of 0.985, 0.985, 0.984, 0.984, and 0.984 respectively. Hence this shows that the CNN-DBN still outperforms all the other models and CNN-LSTM due to its longer evaluation procedure as depicted in Table 5.

We have compared the results of SMOTETomek in Table 6. The DL models and other traditional ML model scores are listed in this table. The KNN usually produced effective results when we compared the overall results. But if we ignore KNN, there are some interesting results presented in the table 6. The CNN-DBN of [35] produced accuracy and G-mean scores of 0.980 and 0.978. When the results are compared to CNN-LSTM and CNN-RNN's Accuracy and G-mean, the results are 0.978, 0.977, and 0.974, 0.975 respectively. This explains that these models can provide us with comparable outcomes when applied to the Multi-step attack dataset.

TABLE 5. Comparison results for SMOTEENN hybrid re-sampling.

References	Model	Accuracy	Precision	Recall	F-1	G-Mean
[34]		0.825	0.942	0.911	0.926	x
[23]	DT	x	x	x	x	0.790
	RF	x	x	x	x	0.790
Results for this Study	NB	0.644	0.641	0.646	0.578	0.363
	DNN	0.169	0.027	0.167	0.049	x
	LSTM	0.977	0.978	0.976	0.977	0.976
	GRU	0.927	0.929	0.928	0.927	0.925
	CNN	0.978	0.979	0.978	0.979	0.978
	RNN	0.919	0.929	0.918	0.920	0.914
	CNN+DBN	0.986	0.985	0.986	0.986	0.985
	KNN	0.999	0.999	0.999	0.999	0.999
	CNN+LSTM	0.985	0.985	0.984	0.984	0.984
	CNN+RNN	0.980	0.980	0.979	0.979	0.979

TABLE 6. Comparison results for SMOTETomek hybrid re-sampling.

References	Model	Accuracy	Precision	Recall	F-1	G-Mean
[34]		0.825	0.942	0.911	0.926	x
[23]	DT	x	x	x	x	0.830
	RF	x	x	x	x	0.820
[35]	NB	0.643	0.648	0.643	0.574	0.401
	DNN	0.167	0.278	0.166	0.047	x
	LSTM	0.968	0.970	0.969	0.968	0.973
	GRU	0.914	0.917	0.914	0.913	0.910
	CNN	0.971	0.972	0.972	0.971	0.971
	RNN	0.953	0.955	0.954	0.954	0.953
	CNN+DBN	0.980	0.979	0.979	0.979	0.978
Results for this Study	KNN	0.995	0.994	0.994	0.994	0.994
	CNN+LSTM	0.978	0.979	0.978	0.978	0.977
	CNN+RNN	0.974	0.975	0.974	0.974	0.975

The empirical results explain that the SMOTEENN re-sampling technique is the best re-sampling technique for all the algorithms while using the Multi-step cyber attack dataset. It is evident from the results that the second-best technique is both SMOTE and SMOTETomek for this study as some of the model's results are better in SMOTE and vice versa. These re-sampling techniques results are listed in the tables below.

V. CONCLUSION

This article highlights the significant importance of mitigating imbalanced datasets within the domain of DL models for Multi-step cyber attack detection. The usefulness of three re-sampling strategies, namely over-sampling, under-sampling, and hybrid re-sampling, in addressing the issues associated with imbalanced data distributions has been examined through a comparative study. The analysis has yielded valuable insights and the findings from the conducted experiments on the Multi-step cyber attack dataset indicate that the hybrid approach, SMOTEENN, shows superior performance compared to both over-sampling and under-sampling techniques. This is evidenced by the considerable enhancement in detection accuracy observed across a range of ML and DL models. The study emphasizes the necessity of developing resilient and flexible defense systems to counter complex cyber attacks, as conventional methods may prove inadequate. Considering the dynamic nature of cyber attacks, it is imperative to employ novel methodologies that can effectively address the challenges associated with imbalanced datasets and provide a robust defense against Multi-step attacks to safeguard the integrity and security of data-driven systems. This research also enhances the comprehension of the influence of re-sampling strategies on the performance

of DL models. Consequently, it paves the way for future progress in the realm of cyber security and facilitates the development of AI-based defense systems that are more robust and reliable.

CONFLICTS OF INTEREST

There are no potential conflicts of interest.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) for funding this work through the Research Group grant no. IMSIU-RG23052.

REFERENCES

[1] W. Liang, K.-C. Li, J. Long, X. Kui, and A. Y. Zomaya, "An industrial network intrusion detection algorithm based on multifeature data clustering optimization model," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 2063–2071, Mar. 2020.

[2] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of big data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263–286, Jan. 2017.

[3] Y. K. Dwivedi et al., "Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *Int. J. Inf. Manage.*, vol. 57, Apr. 2021, Art. no. 101994.

[4] A. Salih, S. T. Zeebaree, S. Ameen, A. Alkhyyat, and H. M. Shukur, "A survey on the role of artificial intelligence, machine learning and deep learning for cybersecurity attack detection," in *Proc. 7th Int. Eng. Conf. Res. Innov. Amid Global Pandemic (IEC)*, Feb. 2021, pp. 61–66.

[5] I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to explainable artificial intelligence in Industry 4.0: A survey on what, how, and where," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5031–5042, Aug. 2022.

[6] H. Omotunde and M. Ahmed, "A comprehensive review of security measures in database systems: Assessing authentication, access control, and beyond," *Mesopotamian J. Cyber Secur.*, vol. 2023, pp. 115–133, Aug. 2023.

[7] T. Alyas, K. Alissa, M. Alqahtani, T. Faiz, S. A. Alsaif, N. Tabassum, and H. H. Naqvi, "Multi-cloud integration security framework using honeypots," *Mobile Inf. Syst.*, vol. 2022, Aug. 2022, Art. no. 2600712.

[8] H. Okhravi, W. W. Streilein, and K. S. Bauer, "Moving target techniques: Leveraging uncertainty for cyber defense," *Lincoln Lab. J.*, vol. 22, no. 1, pp. 100–109, 2016.

[9] Ö. Sen, C. Eze, A. Ulbig, and A. Monti, "On holistic multi-step cyberattack detection via a graph-based correlation approach," in *Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids (SmartGridComm)*, Oct. 2022, pp. 380–386.

[10] G. Mohindru, K. Mondal, and H. Banka, "Different hybrid machine intelligence techniques for handling IoT-based imbalanced data," *CAAI Trans. Intell. Technol.*, vol. 6, no. 4, pp. 405–416, Dec. 2021.

[11] G. Yue, P. Wei, Y. Liu, Y. Luo, J. Du, and T. Wang, "Automated endoscopic image classification via deep neural network with class imbalance loss," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.

[12] S. Piri, D. Delen, and T. Liu, "A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets," *Decis. Support Syst.*, vol. 106, pp. 15–29, Feb. 2018.

[13] M. Almseidin and M. Alkasassbeh, "An accurate detection approach for IoT botnet attacks using interpolation reasoning method," *Information*, vol. 13, no. 6, p. 300, Jun. 2022.

[14] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *Proc. 11th Int. Conf. Inf. Commun. Syst. (ICICS)*, Apr. 2020, pp. 243–248.

[15] M. S. E. Shahabadi, H. Tabrizchi, M. K. Rafsanjani, B. B. Gupta, and F. Palmieri, "A combination of clustering-based under-sampling with ensemble methods for solving imbalanced class problem in intelligent systems," *Technol. Forecasting Social Change*, vol. 169, Aug. 2021, Art. no. 120796.

[16] A. Guzmán-Ponce, R. M. Valdovinos, J. S. Sánchez, and J. R. Marcial-Romero, "A new under-sampling method to face class overlap and imbalance," *Appl. Sci.*, vol. 10, no. 15, p. 5164, Jul. 2020.

[17] C. Lin, C.-F. Tsai, and W.-C. Lin, "Towards hybrid over- and under-sampling combination methods for class imbalanced datasets: An experimental study," *Artif. Intell. Rev.*, vol. 56, no. 2, pp. 845–863, Feb. 2023.

[18] S. Ahmed, A. Mahbub, F. Rayhan, R. Jani, S. Shatabda, and D. M. Farid, "Hybrid methods for class imbalance learning employing bagging with sampling techniques," in *Proc. 2nd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solution (CSITSS)*, Dec. 2017, pp. 1–5.

[19] U. R. Salunkhe and S. N. Mali, "A hybrid approach for class imbalance problem in customer churn prediction: A novel extension to under-sampling," *Int. J. Intell. Syst. Appl.*, vol. 10, no. 5, pp. 71–81, May 2018.

[20] E. Viegas, A. O. Santin, and V. Abreu Jr., "Machine learning intrusion detection in big data era: A multi-objective approach for longer model lifespans," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 1, pp. 366–376, Jan. 2021.

[21] J. Alzubi, A. Nayyar, and A. Kumar, "Machine learning from theory to algorithms: An overview," *J. Phys., Conf. Ser.*, vol. 1142, Nov. 2018, Art. no. 012012.

[22] P. Li, M. Abdel-Aty, and J. Yuan, "Real-time crash risk prediction on arterials based on LSTM-CNN," *Accident Anal. Prevention*, vol. 135, Feb. 2020, Art. no. 105371.

[23] M. Almseidin, J. Al-Sawwa, and M. Alkasassbeh, "Generating a benchmark cyber multi-step attacks dataset for intrusion detection," *J. Intell. Fuzzy Syst.*, vol. 43, no. 3, pp. 3679–3694, Jul. 2022.

[24] I. H. Sarker, "Data science and analytics: An overview from data-driven smart computing, decision-making and applications perspective," *Social Netw. Comput. Sci.*, vol. 2, no. 5, p. 377, Sep. 2021.

[25] E. D. Zamani, C. Smyth, S. Gupta, and D. Dennehy, "Artificial intelligence and big data analytics for supply chain resilience: A systematic literature review," *Ann. Oper. Res.*, vol. 327, no. 2, pp. 605–632, Aug. 2023.

[26] Y. Pang, L. Peng, Z. Chen, B. Yang, and H. Zhang, "Imbalanced learning based on adaptive weighting and Gaussian function synthesizing with an application on Android malware detection," *Inf. Sci.*, vol. 484, pp. 95–112, May 2019.

[27] V. Kumar, G. S. Lalotra, and R. K. Kumar, "Improving performance of classifiers for diagnosis of critical diseases to prevent COVID risk," *Comput. Electr. Eng.*, vol. 102, Sep. 2022, Art. no. 108236.

[28] G. S. Thejas, Y. Hariprasad, S. S. Iyengar, N. R. Sunitha, P. Badrinath, and S. Chennupati, "An extension of synthetic minority oversampling technique based on Kalman filter for imbalanced datasets," *Mach. Learn. Appl.*, vol. 8, Jun. 2022, Art. no. 100267.

[29] G. Goel, L. Maguire, Y. Li, and S. McLoone, "Evaluation of sampling methods for learning from imbalanced data," in *Proc. 9th Int. Conf. Intell. Comput. Theories (ICIC)*, Nanning, China. Berlin, Germany: Springer, 2013, pp. 392–401.

[30] B. P. Ashwini, R. M. Savithramma, and R. Sumathi, "A comparative study of re-sampling techniques on machine learning model performance," *Grenze Int. J. Eng. Technol.*, vol. 8, no. 1, pp. 227–234, 2022.

[31] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021.

[32] A. A. El-Naby, E. E.-D. Hemdan, and A. El-Sayed, "An efficient fraud detection framework with credit card imbalanced data in financial services," *Multimedia Tools Appl.*, vol. 82, no. 3, pp. 4139–4160, Jan. 2023.

[33] L. Gong, S. Jiang, and L. Jiang, "Tackling class imbalance problem in software defect prediction through cluster-based over-sampling with filtering," *IEEE Access*, vol. 7, pp. 145725–145737, 2019.

[34] K. M. A. Alheeti, A. Alzahrani, O. H. Jasim, D. Al-Dosary, H. M. Ahmed, and M. S. Al-Ani, "Intelligent detection system for multi-step cyber-attack based on machine learning," in *Proc. 15th Int. Conf. Develop. eSystems Eng. (DeSE)*, Jan. 2023, pp. 510–514.

[35] M. H. Jamal, M. A. Khan, S. Ullah, M. S. Alshehri, S. Almakdi, U. Rashid, A. Alazeb, and J. Ahmad, "Multi-step attack detection in industrial networks using a hybrid deep learning architecture," *Math. Biosci. Eng.*, vol. 20, no. 8, pp. 13824–13848, 2023.

[36] M. Soori, B. Arezoo, and R. Dastres, "Artificial intelligence, machine learning and deep learning in advanced robotics: A review," *Cognit. Robot.*, vol. 3, pp. 54–70, Apr. 2023.

[37] Z. Tian, J. Li, L. Liu, H. Wu, X. Hu, M. Xie, Y. Zhu, X. Chen, and W. Ou-Yang, "Machine learning-assisted self-powered intelligent sensing systems based on triboelectricity," *Nano Energy*, vol. 113, Aug. 2023, Art. no. 108559.
[38] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, "Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]," *IEEE Comput. Intell. Mag.*, vol. 13, no. 4, pp. 59–76, Nov. 2018.
[39] A. Rashid, M. J. Siddique, and S. M. Ahmed, "Machine and deep learning based comparative analysis using hybrid approaches for intrusion detection system," in *Proc. 3rd Int. Conf. Advancements Comput. Sci. (ICACS)*, Feb. 2020, pp. 1–9.
[40] I. Triguero, D. García-Gil, J. Maillo, J. Luengo, S. García, and F. Herrera, "Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data," *WIREs Data Mining Knowl. Discovery*, vol. 9, no. 2, p. e1289, Mar. 2019.
[41] N. Ahmadi, M. Nilashi, B. Minaei-Bidgoli, M. Farooque, S. Samad, N. O. Aljehane, W. A. Zogaan, and H. Ahmadi, "Eye state identification utilizing EEG signals: A combined method using self-organizing map and deep belief network," *Sci. Program.*, vol. 2022, Feb. 2022, Art. no. 4439189.
[42] H. Zhang, Y. Li, Z. Lv, A. K. Sangaiah, and T. Huang, "A real-time and ubiquitous network attack detection based on deep belief network and support vector machine," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 3, pp. 790–799, May 2020.
[43] S. Sharma and K. Guleria, "Deep learning models for image classification: Comparison and applications," in *Proc. 2nd Int. Conf. Advance Comput. Innov. Technol. Eng. (ICACITE)*, Apr. 2022, pp. 1733–1738.
[44] Y. Fu, F. Lou, F. Meng, Z. Tian, H. Zhang, and F. Jiang, "An intelligent network attack detection method based on RNN," in *Proc. IEEE 3rd Int. Conf. Data Sci. Cyberspace (DSC)*, Jun. 2018, pp. 483–489.
[45] S. Capobianco, L. M. Millefiori, N. Forti, P. Braca, and P. Willett, "Deep learning methods for vessel trajectory prediction based on recurrent neural networks," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 6, pp. 4329–4346, Dec. 2021.

MUHAMMAD HASSAN JAMAL received the bachelor's and master's degrees from Quaid-i-Azam University, Islamabad, Pakistan, in 2017 and 2023, respectively. His research interests include machine learning, deep learning, intrusion detection, data science, and artificial intelligence.

NAILA NAZ received the M.Sc. and M.S. (IST) degrees from Quaid-i-Azam University, Islamabad, Pakistan, in 2019 and 2022, respectively. Her research interests include intrusion detection, dataset analysis, machine learning/deep learning, and information retrieval.

MUAZZAM A. KHAN KHATTAK (Senior Member, IEEE) received the Ph.D. degree in computer sciences as a sandwich program from International Islamic University, Islamabad, Pakistan, and the University of Missouri, Kansas City, MO, USA, in 2011, and the Ph.D. degree from UMKC, in 2016. He joined the School of Electrical Engineering and Computer Science (SEECS), NUST University, Islamabad, as an Assistant Professor, in 2013, and was later promoted to the Associate Dean and a Tenured Associate Professor, in 2017. He was with the School of Computer Science, University of Ulm, Germany, and the School of Computer and Electrical Engineering, University of Missouri, Columbia, MO, as a Research Fellow. He was with the Networking and Multimedia Laboratory, UMKC, as a Research Fellow. He is currently a Tenured Associate Professor with the Department of Computer Science, Quaid-i-Azam University, Islamabad. His research interests include wireless network sensors, body area networks, image processing, image compression, image encryption, and data network security.

FAISAL SAEED received the B.Sc. degree in computers (information technology) from Cairo University, Egypt, and the M.Sc. degree in information technology management and the Ph.D. degree in computer science from Universiti Teknologi Malaysia (UTM), Malaysia, in 2010 and 2013, respectively. He is currently a Senior Lecturer with the Computing and Data Science Department, School of Computing and Digital Technology, Birmingham City University (BCU), U.K. He is leading the Smart Health Laboratory, Data Analytics and AI Research Group, BCU. Previously he was an Assistant/Associate Professor with Taibah University, Saudi Arabia, from 2017 to 2021, and a Senior Lecturer with the Department of Information Systems, Faculty of Computing, UTM, from 2014 to 2017. He managed several projects funded by the Minister of Higher Education in Malaysia and Saudi Arabia. Recently he has been leading the Knowledge Transfer Partnership (KTP) fund sponsored by Innovate U.K. He has published several papers in indexed journals and international conferences. His research interests include data mining, artificial intelligence, machine learning, information retrieval, and health informatics. He served as the general chair for international conferences and the guest editor for indexed journals.

SAAD NASSER ALTAMIMI received the Ph.D. degree in cybersecurity from the University of Glasgow, U.K. He is currently an Assistant Professor with the Department of Information Systems, Imam Mohammad Ibn Saud Islamic University, Saudi Arabia. He brings profound expertise to the field of information security. His research interests include end-user behavioral security, InfoSec awareness and compliance, privacy, and the IoT security. His contributions play a pivotal role in enhancing digital safeguarding.

SULTAN NOMAN QASEM (Senior Member, IEEE) received the B.Sc. degree in computer science from Mustansiriyah University, Iraq, in 2002, and the M.Sc. and Ph.D. degrees in computer science from Universiti Teknologi Malaysia (UTM), Malaysia, in 2008 and 2011, respectively. He was a Senior Lecturer with the Faculty of Computing, UTM, from May 2011 to October 2012. He was an Assistant Professor with Imam Mohammad Ibn Saud Islamic University, Saudi Arabia, from 2012 to 2020, where he is currently an Associate Professor with the Department of Computer Science, College of Computer and Information Sciences. He has published several in peer-reviewed reputed journals, book chapters, and conference proceedings. He has directed many funded research projects. His research interests include applied artificial intelligence, machine learning, data science, multi-objective evolutionary algorithms, metaheuristic optimization algorithms, and health informatics. He has served as the program committee member for various international conferences and a reviewer for various international journals.