Imbalance Datasets in Malware Detection: A Review of Current Solutions and Future Directions

Hussain Almajed, Abdulrahman Alsaqer, Mounir Frikha
Department of Computer Networks and Communications, College of Computer Sciences and Information Technology
King Faisal University, Al-Ahsa, 31982, Saudi Arabia

***Abstract*—Imbalanced datasets are a significant challenge in the field of malware detection. The uneven distribution of malware and benign samples is a challenge for modern machine learning based detection systems, as it creates biased models and poor detection rates for malicious software. This paper provides a systematic review of existing approaches for dealing with imbalanced datasets in malware detection such as data-level, algorithm-level, and ensemble methods. We explore different techniques including Synthetic Minority Oversampling Technique, deep learning techniques including CNN and LSTM hybrids, Genetic Programming for feature selection, and Federated Learning. Furthermore, we assesses the strengths, weakness, and areas of application of each approach. Computational complexity, scalability, and the practical applicability of these techniques remains as challenges. Finally, the paper summarizes promising directions for future research like lightweight models and advanced sampling strategies to further improve the robustness and practicality of malware detection systems in dynamic environments.**

***Keywords*—Malware detection; machine learning; imbalance datasets; oversampling; SMOTE**

I. Introduction

Cybersecurity is a critical area in today's world and malware detection is a critical area of cybersecurity, because malicious software is proliferating at a rapid rate, and it is getting more sophisticated [1], [2]. Moreover, Malware detection solutions are essential given the urgent need to solve the issue. However, detecting malware more effectively has become increasingly difficult because of the complexity of modern malware and the volume of data being generated. In response to this challenge, Machine Learning (ML) techniques have risen in prominence by learning malware patterns and determining their difference from benign software [3]. But the problem of imbalanced datasets is a major obstacle in developing effective malware detection systems. This comes from having a dataset used to train ML models that contain a much lower number of malware samples than data associated with benign samples, leading to biased models that cannot adequately detect malicious activity.

In this paper, we present techniques for addressing imbalanced datasets in malware detection and evaluate their effectiveness through a systematic review.

II. Background

A. Imbalanced Datasets in Malware Detection

Malware detection datasets are imbalanced when the malware samples (minority class) have a very small distribution compared to benign samples (majority class) [4], [3], [5]. As a result, model predictions become skewed towards the majority class and ignore important minority samples, which are most often the focus in cybersecurity. Fig. 1 demonstrates the imbalanced data distribution.

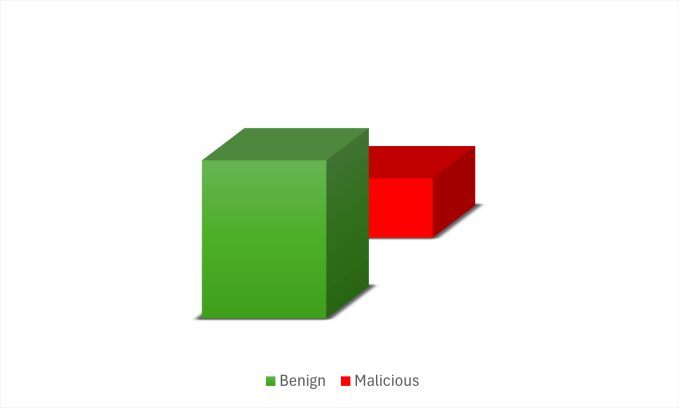

Fig. 1. Visual representation of an imbalanced malware dataset.

B. Balanced Datasets in Malware Detection

Malware detection datasets that are balanced are those which have approximately the same number of samples in the majority class (non-malicious data) and minority class (malicious data) [4], [2]. This balance prevents classifiers from biasing towards any one class, and results in more accurate detection of both non-malicous and malicous data. Fig. 2 demonstrates the balanced data distribution where both data are equal in the count.

C. Challenges of Imbalanced Datasets

Imbalanced datasets raise the following challenges:

- Biased Prediction: Datasets with imbalanced classes, therefore, often lead to classifiers that are skewed towards the majority class, and would often then perform poorly on the minority class [4], [2].
- Poor Generalization: Insufficient training examples lead to failure of the classifiers to generalize well on minority class predictions [3].
- Metric Misleading: As high accuracy can be obtained by ignoring the minority class, standard accuracy measures become unreliable [2], [6].
- Class Overlapping: Classes of imbalanced datasets might overlap, and there will be no clear boundaries

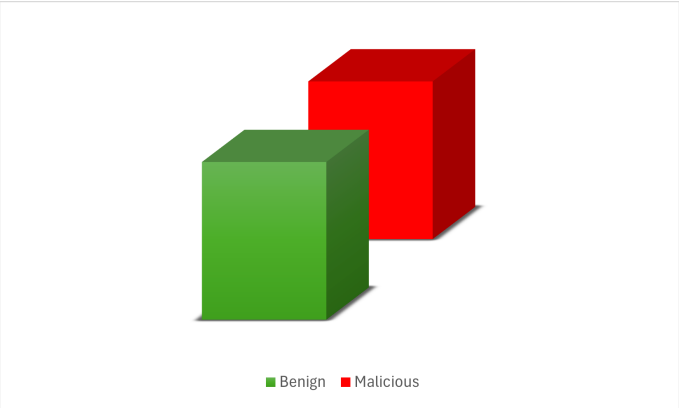

Fig. 2. Visual representation of balanced malware datasets.

TABLE I. OVERVIEW OF DATA-LEVEL METHODS

Method	Definition	Advantages	Limitations
Over-Sampling [4]	It add synthetic examples to the minority class to balance the dataset.	Balances class distribution without losing existing data.	Risk of overfitting and computational overhead in managing large synthetic datasets [6].
Under-Sampling [6]	Reduces the majority class samples to balance the dataset by either randomly removing examples or applying heuristic methods.	Simplifies the dataset, and encourages the model to focus equally on both classes.	Loss of potentially valuable data from the majority class.

Fig. 3. Illustration of OverSampling and UnderSampling methods for handling imbalance datasets.

TABLE II. SUMMARY OF ALGORITHM-LEVEL METHODS

Method	Definition	Advantages	Limitations
Cost-Sensitive Learning [7]	Assigns higher misclassification costs to minority classes.	Improves focus on minority samples, and aligns learning with real-world impact.	Requires precise cost estimation; may still bias towards majority class.
Thresholding [7], [2]	To balance the class distribution, the decision threshold is adjusted.	Simple Implementation, No data loss.	heavily depends on the choice of the optimal threshold value and not be effective for all types.
One class classification [7], [2]	It learns from one class (typically the minority class) and seeks to identify instances that belong to this class, rejecting all others.	Useful for high-dimensional datasets and more robust to noisy data.	More complex to implement and not generalize well to new, unseen data.

separating classes, which complicates distinguishing between majority and minority samples [6], [2].

- Overfitting and Underfitting: On the other hand, oversampling the minority class results to overfit while undersampling the majority class results to underfit [6].

D. Approaches to Address Imbalanced Datasets

In this section, different techniques are introduced to address the imbalanced dataset problem in malware detection. We broadly categorize these approaches into data level methods, algorithm level methods, and ensemble methods that solve the imbalance problem from different angles.

1) Data-Level Methods: Data level approaches try to balance the class distribution by changing the data before applying any ML algorithm [6]. Table I shows the data-level method.

The Fig. 3 shows the illustration of over-sampling and under-sampling.

2) Algorithm-Level Methods: Algorithm level methods adapt existing learning algorithms to make them more sensitive to imbalanced data [7]. Unlike these methods, they do not change the dataset but rather change the training process. Common algorithm-level methods show in Table II.

3) Ensemble Methods: It is a combination of multiple classification techniques from the above mentioned categories and can be seen as a wrapper of other methods such as nsembling which is widely used as a classification technique [7].The method consists of pretraining and fine tuning on the original imbalanced dataset. Also, it combines predictions from multiple models to increase robustness and decrease bias [3], [4]. Bagging and Boosting are techniques. where it Combines the strengths of individual classifiers for better overall performance and reduces the impact of minority class under representation by focusing on difficult to classify samples [6], [4].

E. Motivation

This systematic literature review is motivated by the necessity of improving malware detection capabilities in the presence of:

- The Growing Threat of Malware: Malware attacks have been increasing in frequency and sophistication, making risks to individuals and organizations. As stated by the report of AV-atlas, where over three millions new malware were found in the first two weeks of November 2024 [8]
- Importance of Effective Malware Detection: Undetected malware can lead to the loss of sensitive information, financial implications, operational disruption.
- Challenges with Imbalanced Datasets: Non-malicious samples outnumber malware samples, leading to model bias and high false negatives.
- Need to Address Data Imbalance: To enhance security, improve malware detection accuracy and strengthens overall cybersecurity defenses.

F. Problem Statement

Imbalanced datasets in malware detection pose a big problem for ML models, which leads to biased detection systems that fail to well detect malware [6], [5]. Failure in the identification of the minority class leads to models that perform poorly when it comes to classifying benign against malicious samples, this being due to the current imbalance between benign and malicious samples. This work is a systematic literature review to investigate and assess existing solutions to solve this problem, and to gain insights to develop better methods to deal with imbalanced datasets in malware detection.

G. Scope

The scope of this SLR is to review the literature on imbalance in datasets for malware detection. It includes data level, algorithm level and ensemble methods used to handle the imbalanced datasets. The scope is to evaluate these methods, to identify the challenges and limitations of applying them, and to suggest potential directions for future research. In addition, the review will point out how different solutions have been used in the case of malware detection and their pros and cons.

H. Objective

The objectives of this research are as follows:

- Conduct a Comprehensive Literature Review: In order to systematically review the existing literature regarding how to handle imbalanced datasets in malware detection.
- Investigate Current Solutions: In order to identify and evaluate different approaches used to tackle imbalanced datasets.
- Assess Effectiveness: Focusing on metrics such as accuracy and F1-score, these approaches will further be assessed for their effectiveness in improving malware detection.
- Identify Challenges and Gaps: The challenges, limitations, and gaps of existing methods dealing with imbalanced datasets in malware detection will be identified.
- Suggest Future Directions: propose several directions that could become future research paths in regard to imbalanced datasets in malware detection.

By addressing these objectives, this review aims to offer a clear understanding of the current landscape of imbalanced dataset in malware detection.

III. Research Methodology

We follow a systematic approach to review the existing literature on imbalance datasets in malware detection, following the Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA) guidelines. It includes defining the research questions, selecting databases, developing search strings, establishing of inclusion exclusion criteria, and applying a quality assessment framework. The methodology is organized as follows:

A. Data Sources and Search Strategy

To ensure comprehensive coverage of relevant studies, the search was conducted across the following academic databases:

- IEEE Xplore
- MDPI
- SpringerLink
- ScienceDirect

The keywords used for the selection based on the related research objectives:

("Imbalanced Datasets") AND ("Malware Detection")

Only peer-reviewed journal articles and conference papers published between 2020 and 2024 were considered to capture recent developments.

B. Inclusion and Exclusion Criteria

To filter search results for relevant studies, we established the following inclusion and exclusion criteria:

1) Inclusion Criteria:

- Studies that focus on imbalanced datasets and malware detection
- Peer-reviewed journal articles, conference papers.
- Studies that provide empirical results or evaluations using datasets relevant to imbalanced datasets.
- Publications written in English.
- Propose novel methods or provide empirical evaluations.

2) Exclusion Criteria:

- Studies not related to imbalanced datasets and malware detection.
- Publications that only provide theoretical models without empirical validation.
- Non-peer-reviewed sources such as theses, white papers, and editorials.

C. Study Selection Process

The study selection process adhered to the PRISMA framework, proceeding in three stages:

- Initial Screening: All retrieved articles were screened by titles and abstracts to exclude irrelevant studies and choose those meeting the inclusion criteria for full-text review.
- Full-Text Review: Full texts of selected articles were reviewed to determine their relevance and quality. Excluded articles that did not provide detailed information on balancing techniques, datasets, or empirical evaluations.
- Data Extraction and Coding: A standardized form was used to extract data from the final set of articles, including balancing techniques, datasets, and

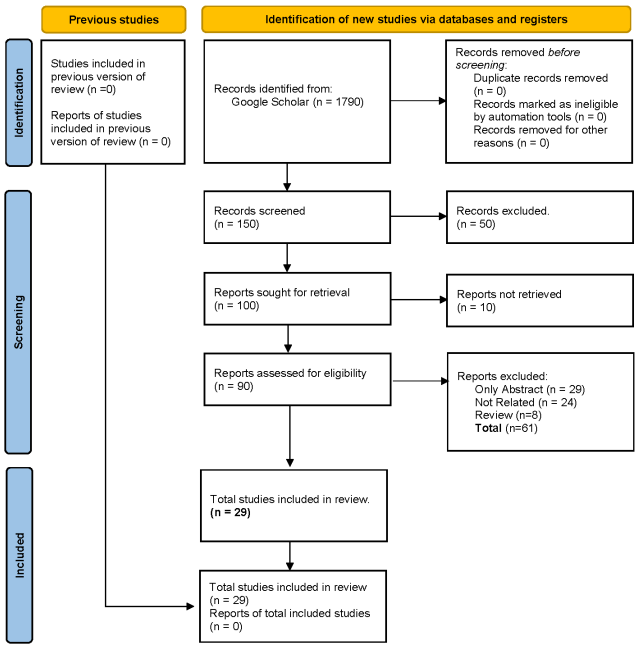

Fig. 4. PRISMA flow diagram summarizing the study selection process.

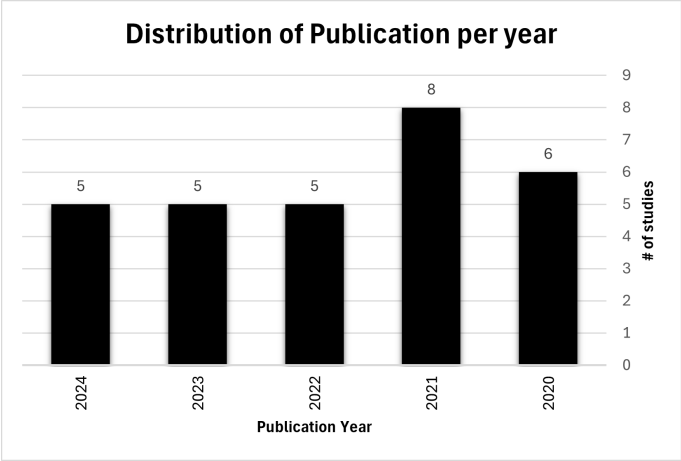

Fig. 5. Distribution of publications included in the review based in year.

evaluation metrics, as well as identify challenges and contributions.

Fig. 4 shows the PRISMA flow diagram.

Fig. 5 shows the distribution of the number of papers selected for this SLR per year.

IV. Literature Review

The problem of imbalanced datasets in malware detection in Android devices is addressed by Dehkordy and Rasoolzadegan [9]. They obtained a dataset from Drebin and AMD datasets containing 9,223 applications, and was heavily preprocessed to reduce the features from 1,262 to 78 for faster learning. The authors used SMOTE (Synthetic Minority Oversampling Technique), undersampling, and a hybrid approach to solve the imbalanced issue. To improve the accuracy of detection they employed dataset preprocessing, ranking of features and using multiple classifiers like K-nearest neighbors (KNN), Support Vector Machines (SVM) and Iterative Dichotomiser 3 (ID3). The best results were obtained by a combination of KNN with SMOTE, with an accuracy of 98.69%. However, the study limited to false positive rates of 2.09% to 4.77% and an approach that is only applicable to a limited number of malware families, which limits the model's generalizability. Guan et al. [10] propose n Class Imbalance Learning (CIL) approach to address the class imbalance problem for Android malware detection. It applies the K-means clustering-based under-sampling, which retains the representative majority samples, and then the SMOTE algorithm to generate the synthetic minority samples. A Random Forest (RF) classifier is then trained on the combined dataset. The dataset used for evaluation consists of 10,182 malware samples from VirusShare and 127 benign samples, with a class imbalance ratio of 1:80. They showed that the CIL method outperforms other traditional methods such as SMOTE and random under-sampling. In general, CIL shows good generalizability to other imbalanced datasets, and it is a promising solution to the class imbalance problem in malware detection.

Imbalanced datasets in malware detection for edge computing in Android based Internet of Things (IoT) environments is addressed by Khoda et al. [11]. The authors propose two methods a dynamic class weighting technique and modified Fuzzy-SMOTE for synthetic oversampling. The first approach generates valid synthetic malware samples preserving the malicious functionality, the second approach dynamically adjusts class weights during training to improve malware detection. The evaluation show over 9% improvement in F1 score over traditional imbalanced learning techniques. 50,000 Android applications and 500 malware samples in the dataset. The fuzzy approach is limited by the requirement of careful tuning, while the dynamic class weighting method is less sensitive to such parameters.

The challenge of detecting Android ransomware in an imbalanced dataset is addressed by ALMOMANI et al. [12]. A hybrid evolutionary approach using Binary Particle Swarm Optimization (BPSO) and SVM is employed for feature selection and classification to improve classification performance by effective optimization. The SMOTE was used to balance the dataset. Sensitivity, specificity and g-mean were used to evaluate the model, scoring 96.4%, 98.7% and 97.5%, respectively. The dataset has 10,153 Android applications out of 500 ransomwares. However, the dataset is small making it difficult to generalize, especially for new ransomware variants.

Hemalatha et al. [13] suggest a DenseNet model based on Deep Learning (DL), with a class balanced categorical cross entropy loss to overcome class imbalances. Malware binaries are transformed into grayscale images and malware detection is framed as a multi-class image classification problem. The experiments were performed on Malimg, BIG 2015, and MaleVis datasets with high accuracies of 98.23%, 98.46%, and 98.21%, respectively; and 89.48% on the unseen Malicia dataset. However, the model lacks in accuracy on unseen data, and struggles with novel malware (zero day attacks). Future work could involve improving generalization to deal with zero day attacks.

Goyal and Kumar [14] discuss malware detection and the effect of data imbalance. To balance the dataset, the researchers use random under-sampling to reduce 42,797 malware samples to 1,079 benign samples. They compared different ML classifiers (KNN, Decision Tree, RF). RF achieved the best accuracy of 98.94% on the imbalanced dataset, and 90.38% on the balanced dataset. They show the impact of data imbalance on model accuracy, and that more reliable results can be obtained from balanced datasets. The study concludes that balanced datasets are necessary to reduce bias and increase reliability, and future research could include further investigation of more sophisticated balancing techniques to improve the applicability of the model to real world scenarios.

Salas and Geus [15] addresses the challenge of class imbalance. The authors propose the MobileNet Fine-Tuning (MobileNet FT) model, a fine tuned version of MobileNet that utilizes bicubic interpolation and class weight estimation techniques. Experimental results show that the proposed model reaches accuracy rates for different datasets, such as Microsoft Big 2015 (98.71%), Malimg (99.08%), MaleVis (96.04%), and a new Fusion dataset (98.04%). These results show that the model is robust to a range of malware families. The approach is also shown to have limitations, such as a degradation in performance as the number of malware families increases and problems with unseen malware. This motivates further investigation into more adaptive models to improve scalability and robustness to new threats.

Almaleh et al. [16] suggest to improve the detection of malware in Windows using a hybrid method. They uses logistic regression with Recurrent Neural Network (RNN) to detect malware from Application programming interfaces (API) call sequences. The study presents a solution to the problem of imbalanced datasets through the use of an undersampling technique that creates a balanced dataset of 2,158 samples of malicious and non-malicious samples. They initialize the RNN weights using logistic regression for improved model accuracy. For the balanced dataset, the model reached an accuracy of 83%, and for the imbalanced, an accuracy of 98%. Limitations include a relatively small trained dataset after balancing due to its restriction on generalizability. Future work could build on the model for other operating systems and overcome these limitations so that the model is more applicable and robust in different scenarios.

Yu Ding et al. [17] proposed self-attention based approach, considering malware ASM files as text sequences to distinguish the malware families. The imbalance dataset technique used to represent ASM files as integer vectors and use a self attention neural network to improve minority class recognition. The sequence classification accuracy is improved by capturing internal dependencies within sequences using this approach. The model is evaluated using the Microsoft Malware Classification Challenge dataset, and shows a robustness to different datasets with 98.48% accuracy and 89.66% F1 score for Simda class. However, small sample recognition problems are not completely solved, and the interpretability of the model is restricted. The future work could include improving early detection of new malware and make neural networks easier to interpret for practical application in cybersecurity.

Moti et al. [18] handles the problem of imbalanced dataset for malware detection. The synthetic samples for minority classes are generated using a hybrid model composed of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks with Sequence Generative Adversarial Networks(SeqGAN), so that the dataset is balanced. The classification accuracy is improved to 98.99% using this approach. They evaluate on a Microsoft dataset from a Kaggle competition that contains nine malware families. While the high accuracy, the model depends only on opcode sequences without preprocessing, which can restrict its feature diversity. Moreover, the training overhead of SeqGAN is not high and the model still needs to be adapted to different datasets and zero day threats.

The problem of Android malware detection is addressed by Almomani et al. [19]. They present a vision based DL model that converts Android Application Package (APK) bytecodes to visual images and uses CNNs for classification. They evaluate the model on an imbalanced dataset (14,733 malware and 2,486 benign samples), without using data augmentation. The main contribution is the development of 16 fine tuned CNN algorithms that are efficiently able to classify malware, which demonstrates 99.40% accuracy on balanced datasets and 98.05% on imbalanced datasets. The study highlighted reduced computational cost due to no longer requiring the manual feature extraction. The limitations include dependence on pre trained CNN weights and uncertain adaptability to new malware types or other datasets.

In the problem of detecting macro malware in Microsoft document files, Mimura [20] tackles the problem of highly imbalanced datasets. They propose a method to combine Doc2Vec and Latent Semantic Indexing (LSI) with four classifiers (SVM, RF, Multilayer Perceptron (MLP), CNN) to increase the accuracy of malware detection. The highest F-measure of 0.99 indicated a high accuracy. The dataset consisted of more than 30,000 samples from VirusTotal and Stack Overflow with an imbalanced distribution favoring benign samples. Limitations include possible lack of generalizability from dataset composition, and future work is to collect more data for robust evaluation. The results of the study demonstrate that LSI is robust to class imbalance and promising results in practical malware detection applications.

Nikale and Purohit [21] addresses the issue of class imbalance in the dataset. The authors used dynamic features such as system calls and binder calls to classify Android APKs into five families: Ransomware, smsware, adware, scareware and benign. The dataset consists of 525 APK samples from different sources, including Contagio Mobile and Google Play Store. The research introduced SMOTE, Adaptive Synthetic sampling (ADASYN), and balanced cost. They tested various classifiers, and the highest accuracy of 91% was obtained with Extreme Gradient Boosting (XGBoost) combined with SMOTE. However, the study is limited by small dataset, hence, restricted in generalizing its findings. Furthermore, constraint is noted on the focus on a fixed set of dynamic features without investigating more specific behavioral characteristics.

SAWADOGO et al. [22] evaluate the impact of data imbalance on eleven ML algorithms, the authors use CICMalDroid 2020, a malware dataset. For comparison, they created two subsets: one imbalanced and the other balanced. They claim that traditional evaluation metrics (Accuracy, Precision and Recall) are inappropriate for imbalanced datasets, while

Balanced Accuracy and Geometric Mean are more appropriate. The results show that algorithms such as AdaBoost and SVM perform very poorly on imbalanced data, whereas Extra Trees and RF are less sensitive. Therefore the authors suggest to use balanced evaluation metrics to better represent the model performance on imbalanced dataset. A study limitation is that a single dataset was used. Additionally, future work should study more sophisticated learning techniques on more diverse datasets to improve the robustness of Android malware detection models.

In this paper, Haluška et al. [23] compare 16 data preprocessing methods for imbalanced classification problems, with a focus on cybersecurity. The authors extensively used six cybersecurity datasets and 17 other public imbalanced datasets from different domains as benchmarks. Overall, the performance of oversampling methods is better than that of undersampling methods, and the standard SMOTE algorithm gives a substantial performance boost. Experiment results indicate that SMOTE and its variants, e.g. generalization of SMOTE and SVM SMOTE, work well in various datasets and metrics, e.g. PR AUC, ROC AUC, and P-ROC AUC. This study shows that to improve predictive performance on multiple tasks in cybersecurity and other domains, it is essential to choose appropriate preprocessing methods and carefully consider method choice and hyperparameter tuning.

Alzammam et al. [24] provide a comparative view of different approaches to the problem of imbalanced multi-class classification in malware detection using CNN. The study focuses on evaluating the effectiveness of different methods, such as cost-sensitive learning, oversampling, and cross-validation, to mitigate the imbalance issue in three publicly available malware datasets. For instance, the study shows that oversampling outperforms other methods in boosting the accuracy and F1-score of the CNN model on every dataset, while the proposed model achieves substantial improvements in accuracy and F1-score when oversampling, with accuracy reaching as high as 99.94% for the Malimg dataset. Finally, this research highlights the importance of dataset characteristics when selecting a method to correct for imbalance, as well as other data factors (including noise and overlapping) and the complexity of applying pre-trained models to malware classification.

Phung and Mimura [25] suggest a way to detect malicious JavaScript by using ML, but specifically dealing with the class imbalance problem. Once the balance between the benign and malicious datasets is adjusted through an oversampling technique, the authors use them to train a classifier for prediction. Experimental results indicate that the proposed method can effectively detect new malicious JavaScript with higher accuracy and efficiency (0.72 recall with Doc2Vec). With the same training and test time per sample, this outperforms the baseline method by 210% in terms of recall score on the dataset with over 30,000 samples: 21,745 benign samples from popular websites and 214 malicious samples from PhishTank, along with an additional 8000 malicious samples from GitHub. The research limits itself to various resampling techniques without exploring or comparing them in a more comprehensive way.

In mobile malware detection, Khoda et al. [26] propose a novel way to handle the problem of imbalanced datasets via synthetic oversampling. This method proposes the addition of features to existing malware samples to generate synthetic malware samples that are valid and retain the malicious functionality. They test the approach using a Deep Neural Network (DNN) on the Drebin Android malware dataset. Results indicate that the proposed method achieves higher precision, recall and F1 score than the oversampling and undersampling techniques in general, and especially at lower imbalance ratios. The performance of the proposed model is much more accurate than previous methods, achieving an F1 score of 94.2% at 10% imbalance ratio and accuracy of 98.8%. The dataset used is the Drebin dataset which has 50,000 apps and 500 malicious apps as the minority class.

Reshi and Singh suggest a new method to handle the imbalance issue in malware datasets through the use of Variational Autoencoder (VAE) [27]. The proposed solution uses VAEs to extract and compress features from the given data and, thus, the model is able to learn features that are resistant to noise and distinguish between real malware and other types of noise. In addition, VAEs improve the data augmentation technique by generating synthetic malware samples from the learned latent space to help overcome the imbalanced problem. This approach expands the training set and, therefore, improves the model's ability to generalize and increase the detection rate for new or less frequent variants of malware. The research contributes by enhancing the VAEs by combining them with CNNs for malware detection. The proposed model gives an accuracy of 98% on the Malimg dataset, which is better than the baseline model. The dataset used in this work is the Malimg dataset, which has a highly unbalanced class distribution. The main drawback of the proposed work is that the integrated VAE-CNN model is relatively complicated and may need appropriate resource allocation and hyperparameter optimization.

Faridun and Im propose a novel malware detection approach using the TabNetClassifier, which is a DL architecture designed explicitly for tabular data analysis [28]. In this research, they enhance malware detection by utilizing the TabNetClassifier in conjunction with the SMOTE to address class imbalance in datasets. Initially, the dataset of 138,047 Portable Executable (PE) header samples is trained using the TabNetClassifier, which is imbalanced with 41,323 benign and 96,724 malware samples. SMOTE is applied to balance the dataset and improves model performance significantly. The main contribution of this research is to show the success of combining TabNetClassifier with SMOTE in improving malware detection accuracy and sensitivity. After applying SMOTE, the model achieves an accuracy of 99.10%, precision of 99.03%, and recall of 99.19%.

Li et al. [29] present a novel method for malicious family classification based on multimodal fusion and weight self-learning. The method deals with the problem of imbalanced datasets and concept drift in malware family classification. This approach integrates multiple features of byte, format, statistic, and semantic types to improve the robustness of the classification model. Experimental results show high efficiency and small resource overhead in classifying highly imbalanced malware family datasets while delivering very good classification performance. The dataset used consists of some types of malware, namely ransomware, Trojans, viruses, and malicious

mining programs. However, the research is limited by the reliance on static analysis and may not find the dynamic behaviors of malware.

In order to improve ransomware detection and classification, Onwuegbuche et al. [30] propose a three-stage feature selection method. This method applies chi-square (CHI2), Duplicated Features (DUF), and Constant Features (COF) filter feature selection techniques to reduce the dimensionality of the dataset, taking into account the different importance of different feature groups. Further, the study addresses the class imbalance problem by employing the SMOTE and cost-sensitive ML methods. The performance of this method is evaluated on the Elderan ransomware dataset and several ML models (XGBoost, Logistic Regression, RF, Decision Trees, and SVM). The results demonstrate that the proposed feature selection method leads to a 10% average improvement in binary classification and 21.79% in multi-class classification over previous studies. Among binary classifiers, XGBoost with cost-sensitive learning and SMOTE is the best with 98.78% balanced accuracy, while the best multi-class classifier is the RF model with cost-sensitive learning achieving 61.94% balanced accuracy.

Andelic et al. [31] deal with the problem of malware detection in imbalanced datasets. The authors suggest combining a Genetic Programming Symbolic Classifier (GPSC) with dataset oversampling techniques to increase the detection accuracy. They apply the GPSC algorithm to an open dataset containing hybrid features consisting of binary hexadecimal and Dynamic Link Library (DLL) calls of Windows executables. The dataset is initially imbalanced, containing 301 malicious and 72 non-malicious samples. In order to address this imbalance, the authors use oversampling techniques such as ADASYN, BorderlineSMOTE, KMeansSMOTE, SMOTE, and SVMSMOTE, and they train the GPSC with Five-Fold Coss-Validation(5FCV) and Random Hyperparameter Value Search (RHVS) method to select the best combination of hyperparameters. The classification accuracy of the proposed method is 0.9962. GPSC is used to generate Symbolic Expressions (SEs) that can be easily applied to and implemented into malware detection models, overcoming the limitations of traditional ML models, which are hard to interpret and transform into mathematical equations.

According to Çayır et al. [32], a new ensemble model called the Random CapsNet Forest (RCNF) is proposed to tackle the imbalance in malware type classification. The authors use the Capsule Network (CapsNet) architecture to preserve spatial information without the use of pooling layers and incorporate the bootstrap aggregating (bagging) technique to form an ensemble model. The idea is to reduce the variance of CapsNet models and improve the robustness of classification by using this approach, which is tested on two highly imbalanced datasets, Malimg and BIG2015, where the RCNF model is also shown to outperform other competitors with fewer trainable parameters. It achieves an F-Score of 98.20% for the BIG2015 dataset and 96.61% for the Malimg dataset. Advantages noted regarding the simplicity of the architecture and the ability to train from scratch without the need for transfer learning.

LIN et al. [33] present a ML framework based on a VAE and a MLP that helps overcome the problem of imbalanced datasets in intrusion detection systems (IDS). An efficient range-based sequential search algorithm is included in the framework to determine the optimal sequence length for data segmentation from multiple sources, including network packets and system logs. Experimental results on HDFS dataset demonstrate that the proposed method achieves an F1 score of around 97% and recall rate of 98%, better than other solutions. Imbalanced datasets are treated using the proposed approach, which increases the F1-score by up to 35% and the recall rate by 27%. In addition, the work also points out the necessity of the appropriate data segmenting and the possibility of the proposed model detecting the new attack variants. The dataset used is the HDFS dataset, a public system log dataset.

In the context of Federation Learning (FL), ransomware detection, and attribution, Vehabovic et al. [34] address an essential problem of dataset imbalance. The authors suggest a modification of the FL scheme where the weighted cross entropy loss function is used to combat bias in datasets distributed across various clients. This approach is particularly applicable since ransomware data distribution and quantity can differ significantly also across different locations and companies. The performance of the proposed FL scheme is evaluated using an up-to-date repository of Windows-based ransomware families and benign applications. The results indicate that the weighted cross entropy loss function approach can mitigate the effect of dataset imbalance, especially in the case of binary ransomware detection with an average accuracy of 94.67%, but the study also points out the difficulties of multi-class attribution with imbalanced datasets, which results in more decline in performance compared to the balanced datasets.

As a form of unsupervised learning, Shi et al. [35] explore using One-Class Classification (OCC) to detect malware in the Internet of Things (IoT) domain. To combat dimensionality and information loss, the authors suggest that categorical features should be changed into numerical formats by using the Term Frequency-Inverse Document Frequency (TF-IDF) method. they compare the performance of OCC models, such as Isolation Forest and deep autoencoder, trained on benign NetFlow samples alone. It is shown that these models achieve 100% recall with precision rates greater than 80% and 90% on a number of test datasets, highlighting the adaptability of unsupervised learning to time-evolving malware threats in IoT, and making an important contribution to the study of malware detection in IoT, particularly when labeled malicious data are scarce. TF-IDF is used for feature transformation, and the comparison of various OCC algorithms leads to valuable improvements in the IoT security framework.

Using ML techniques, in particular, the use of genetic programming for feature selection, Al-Harahsheh et al. [36] propose how to enhance malware detection. The researchers built a malware detection model in which the features are selected by a genetic programming algorithm, and then a set of parallel classifiers is used to enhance detection accuracy at a lower cost. The proposed model employs five feature selection methods: Filter-based, wrapper-based, Chi-Square, Genetic Programming Mean (GPM), and Genetic Programming Mean Plus (GPMP). Experimental results demonstrate that the GPMP method (which uses fewer features than the Filter-based method) results in better accuracy and F1-score values of 0.881066 and 0.867546, respectively. The research

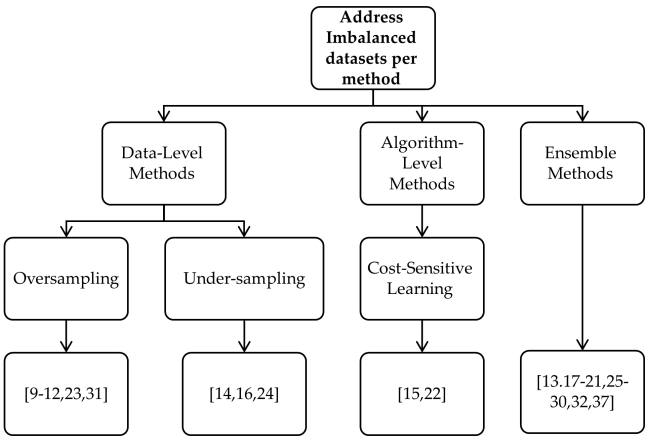

Fig. 6. Taxonomy of the literature review per method used.

indicates that genetic programming is able to select features to improve the performance of malware detection effectively. A number of classifiers, such as RF, Random Tree, and SVM, were used to compare the performance of the proposed feature selection methods.

Al-Khshali et al. [37] propose a new technique employing subspace learning-based OCC methods to detect malware. In this work, they address the issues of class imbalance and the curse of dimensionality in the application of traditional ML algorithms for malware detection. In order to overcome these problems, the researchers introduce a pipeline that uses subspace learning techniques such as Subspace Support Vector Data Description (SSVDD) and Graph Embedded Subspace Support Vector Data Description (GESSVDD). The proposed framework solves multiple problems at once, including class imbalance and the curse of dimensionality. The results show promising performance, with a True Positive Rate (TPR) of 100% for subspace-learning-based OCC. The datasets used in this study include (Benign and Malicious PE Files, ClaMP, and Malware Analysis Datasets by Oliveira) which are diverse but representative of a wide variety of malware types.

Table III shows a summary of the literature review conducted previously.

Table IV shows the limitations and contributions of studies conducted.

A. Taxonomy of the Research

Fig. 6 shows the taxonomy clearly categorizes the different techniques utilized by the studies to mitigate the imbalances, with the data level set of methods focusing on manipulating the dataset distribution, the algorithm level which trains towards the adaption of the learning algorithms and the ensemble where a combination of a variety of methods is used to get better results.

V. Discussion of the Literature Review

The literature review reveals that many ML approaches applied to address imbalanced datasets in malware detection. Many studies indicate handling class imbalance is important to improve detection performance.

1) Oversampling Techniques:

- In several studies SMOTE was commonly employed to increase the minority class representation, leading to improvements in detection rates for families of malware like adware, ransomware or smsware [9] [21] [30]. In the case of imbalanced learning, SMOTE was found to be effective at improving metrics such as F-measure and MCC [9]. Complex oversampling methods gave incremental improvements, indicating the need to balance computational cost with performance gains [23].
- However, oversampling methods like the conventional SMOTE have some limitations that need to be overcome or minimized through the use of more advanced forms like the Modified Fuzzy-SMOTE whose oversampling strategies produces synthetic data that is more reflective of real data distribution . However, these techniques are still inefficient with large, high-dimensional data sets when they are applied.

2) Feature Selection Techniques:

- The improvement in model efficiency and accuracy was greatly aided by feature selection techniques. The genetic programming based feature selection methods, including GPMP, demonstrated that selecting fewer but more relevant features can reduce computational complexity and improve classifier performance [36]. Multi stage feature selection was used in other studies to select features such as API calls, registry operations, and directory logs, which improves model interpretability and classification performance [30]. Moreover, swarm intelligence based optimization, and in particular BPSO was also efficient to choose the best features in order to achieve a large performance gain when dealing with highly imbalanced data for Android malware detection [12].
- The problem of selecting features often demands an expert's input in the process for feature selection. Perhaps, even more, automated approaches, such as feature selection by applying AI methods, could be more beneficial for this step.

3) ML Approaches:

- RF and SVM are used frequently as they are robust, and can handle non linear relationships in data. For instance, Guan et al.[10] achieved significant accuracy improvements by combining RF with SMOTE, especially in datasets with a high imbalance ratio (1:80). In another study, in a custom malware detection dataset, the RF model showed robustness with an accuracy of 98.94% [14]. However, As dataset size increases, RF achieves high accuracy on imbalanced datasets, but it's scalability becomes an issue. Moreover, RF fails to capture complexities of feature interactions unless it is heavily tuned.
- SMOTE was used in combination with SVM and showed 97.83% accuracy on Android malware datasets, as reported by Dehkordy and Rasoolzadegan [9]. However, Sawadogo et al. [22] point out that

TABLE III. SUMMARY OF LITERATURE REVIEW

Author	Year	Best balancing techniques	Scope	Dataset	Metrics Result
Dehkordy and Rasoolzadegan [9]	2021	SMOTE	Android malware detection on imbalanced datasets.	Drebin Dataset and AMD Dataset	KNN, SVM, ID3 Accuracy: 98.69%, 97.83%, 97.59%
Guan et al. [10]	2021	SMOTE	Android malware detection on imbalanced datasets.	VirusShare (10,182 malware, 127 benign apps)	RF, KNN, NB, SVM
Khoda et al. [11]	2021	Modified Fuzzy-SMOTE	Malware detection in edge computing.	Drebin Dataset, AndroZoo and Google Play Store	DNN (F1 Score 99%)
Almomani et al. [12]	2021	SMOTE	Android Ransomware Detection in Imbalanced Data.	Custom Dataset	SMOTE-tBPSO-SVM (Specificity 98.7%)
Hemalatha et al. [13]	2021	Reweighted class-balanced loss function	Malware detection on imbalanced datasets.	Malimg	DenseNet-based (Accuracy 98.46%)
Goyal and Kumar [14]	2020	Random Under-Sampling	Malware detection using ML classifiers.	Custom dataset (42,797 malware - 1,079 benign)	RF (Accuracy 98.94%)
Salas and Geus [15]	2024	Bicubic interpolation, Class weight estimation, ReduceLROnPlateau	Malware classification using DL.	Malimg	MobileNet FT (Accuracy 99.08%)
Almaleh et al. [16]	2023	Undersampling	Malware detection in Windows.	Custom dataset (42,797 malware - 1,079 benign)	LR & RNN (Accuracy 98%)
Yu Ding et al. [17]	2020	Novel classification approach	Malware classification.	BIG 2015	Self-attention Neural Network (Accuracy 98.48%)
Zahra Moti et al. [18]	2020	SeqGAN	Malware detection.	Microsoft dataset	CNN-LSTM (Accuracy 98.99%)
Almomani et al. [19]	2022	-	Android malware detection on imbalanced datasets.	Leopard Android dataset (14,733 malware - 2,486 benign)	Xception CNN (Accuracy 99.40% balanced - 98.05% imbalanced)
Mimura [20]	2020	Word frequency-based feature selection	Malware detection in Microsoft document files.	VirusTotal and Stack Overflow	SVM (F-measure 99%)
Nikale and Purohit [21]	2023	SMOTE, ADASYN and Balanced Cost	Android malware detection on imbalanced datasets.	Contagio, Koodous and APKPure	XGBoost (Accuracy 91%)
Sawadogo et al. [22]	2022	-	Android malware detection on imbalanced. datasets	CICMalDroid	RF
Halu ˇ ska et al. [23]	2022	SMOTE	Imbalanced classification in Cybersecurity and other domains.	23 datasets (6 cybersecurity datasets and 17 public imbalanced datasets)	PR AUC, ROC AUC, and P-ROC AUC are 6.283, 6.174, and 4.087, respectively.
Alzammam et al. [24]	2020	Oversampling	Imbalanced multi-class malware classification.	Malimg, Microsoft, and VirusTotal	Accuracy:99.94%, 98.31%, 96.06% respectively.
Phung and Mimura [25]	2021	Oversampling + ML	Static analysis for detecting malicious JavaScript.	Imbalanced dataset over 30,000 samples (PhishTank + Github).	Recall score of 72% with Doc2Vec model, outperforming baseline method by 210%.
Khoda et al. [26]	2020	Oversampling + DNN	Mobile malware detection with imbalanced data.	Drebin Android malware dataset (50,000 apps, 500 malicious)	F1-score of 94.2% and accuracy of 98.8% at 10% imbalance ratio.
Reshi and Singh [27]	2024	VAEs for generating synthetic samples	Enhance Malware detection in imbalanced datasets using DL.	Malimg dataset	Accuracy 98%
Faridun and Im [28]	2024	SMOTE + TabNetClassifier	Malware detection using tabular data analysis and addressing class imbalance issues using DL.	138,047 PE header samples (41,323 benign and 96,724 malware samples)	Accuracy: 99.10%, F1-Score: 99.11% after applying SMOTE.
Li et al. [29]	2022	Multimodal fusion and Weighted Soft Voting	Malware family classification in Intelligent Transportation Systems.	Microsoft BIG-15	Accuracy: 99.2%, Macro-F1 score: 98.1%.
Onwuegbuche et al. [30]	2023	(SMOTE, Cost-sensitive learning) + ML	Ransomware detection and classification using ML models.	Elderan ransomware dataset	98.78% (binary - XGBoost with cost-sensitive learning and SMOTE), 61.94% (multi-class - RF model using cost-sensitive learning).
Andelic et al. [31]	2023	ADASYN	Improving malware detection in imbalanced datasets using GPSC and oversampling.	uci malware detection (301 malicious and 72 non-malicious).	Accuracy: 99.62%.
Çayır et al. [32]	2021	Ensemble learning with bagging RCNF.	Image-based malware family classification.	Malimg and BIG2015	F-Score: 96.61%, 98.20% respectively.
LIN et al. [33]	2022	DL(VAE) + ML(MLP)	Intrusion Detection in Heterogeneous Networks.	HDFS logs	F1 score: 97%, Recall rate: 98%
Vehabovic et al. [34]	2023	Federated learning (FL) with Weighted cross-entropy loss function.	Ransomware detection and attribution using FL.	9 ransomware families (140 malicious samples each) and 2,000 benign Windows applications.	Binary detection: 94.67%, Multi-class attribution: 84.15%.
Shi et al. [35]	2024	OCC (ML:Isolation Forest, DL:Deep Autoencoder)	IoT Malware Detection.	IoT-23 dataset	F1-score: (Isolation Forest: 88%), (Deep Autoencoder: 95%).
Al-Harahsheh et al. [36]	2021	SMOTE over-sampling technique+GPMP feature selection + ML	Malware detection using ML classifiers with genetic programming for feature selection.	Ten different datasets	GPMP method with RF achieves an accuracy of 97.95% and an F1-score of 96.35%.
Al-Khshali et al. [37]	2024	Subspace Learning-Based OCC (SSVDD and GESSVDD), uses ML	Malware Detection.	3 datasets(Benign & Malicious PE Files, ClaMP, Malware Analysis Datasets: PE Section Headers by Oliveira)	100% TPR for subspace-learning-based OCC.

SVM's performance really drops when faced with larger and diverse datasets because SVM is not scalable due to its reliance on kernel based methods, and is sensitive to hyperparameter settings on real world imbalanced datasets.

4) DL Approaches:

- Studies explored DL approaches. Researchers investigated hybrid models like CNN + LSTM with Generative Adversarial Networks (GAN) to generate new synthetic samples to balance datasets in order to detect minority classes [18]. Colored and grayscale image representations of Android malware were used for detecting Android malware with vision based CNN, like Xception which reduces the number of manual feature extraction phases and increases scalability [19]. Fine tuning MobileNet through transfer learning techniques emphasized that CNNs are effective in mitigating overfitting and enhancing generalization particularly in resource constrained environments [15].
- However, these models are computationally intensive and, therefore, unsuitable for real-time processing or on devices with low computational capabilities. Essentially, if these models were simplified or pruned, then they would most likely be more useful in terms of time and versatility.

A. One-Class Classification Models

- Innovative one class classification models proved to be effective in cases of lack of labeled data, especially in IoT environments. Isolation Forest and deep autoencoders showed high adaptability to new malware threats with 100% recall by using TF-IDF and n-grams

TABLE IV. LIMITATIONS AND CONTRIBUTIONS OF LITERATURE REVIEW

Ref	Contribution	Limitation
[9]	Improved malware detection through dataset preprocessing, feature ranking, and the use of multiple classifiers.	Limited malware family coverage and a notable false positive rate.
[10]	A hybrid Class-Imbalance Learning (CIL) method using clustering-based under-sampling combined with SMOTE.	Requires careful tuning of clustering parameters and may not generalize well to newer malware types.
[11]	Modified Fuzzy-SMOTE to handle data imbalance in malware detection.	Requires careful tuning of parameters.
[12]	Combining BPSO and SVM integrated with SMOTE, to effectively detect Android ransomware.	The dataset's small size and imbalance limit the model's generalizability.
[13]	DenseNet-based model for malware detection which visualizes malware binaries as grayscale images.	Struggles with detecting zero-day malware and has reduced accuracy for unseen malware classes.
[14]	Compares the performance of ML classifiers on balanced versus imbalanced datasets.	Potential bias which reduces the generalizability of the findings to real-world scenarios.
[15]	Combining multiple datasets into a new Fusion dataset for enhanced diversity.	Struggles with generalizing to unseen malware types.
[16]	A hybrid malware detection model combining logistic regression for weight initialization with RNN to improve detection capabilities of API call sequences.	Small balanced dataset reduces the model's ability to generalize effectively to diverse, large-scale scenarios.
[17]	Effectively addresses the issue of imbalanced datasets by treating malware ASM files as text sequences.	Struggles with the recognition of small-sample malware families and lacks interpretability for practical cybersecurity use.
[18]	CNN-LSTM hybrid model combined with SeqGAN to address class imbalance in malware detection.	High computational overhead due to SeqGAN training and reliance solely on opcode sequences which limits feature diversity.
[19]	Vision-based DL model utilizing 16 fine-tuned CNNs to detect Android malware efficiently without manual feature extraction.	Relies on pre-trained CNNs which limits adaptability to new malware types.
[20]	Detecting macro malware using Doc2vec and LSI combined with ML classifiers to address imbalanced dataset.	Dataset may not fully represent real-world conditions.
[21]	Proposed a familial classification model for Android malware using dynamic features while addressing dataset imbalance issues.	Small dataset and focused only on basic dynamic features, limiting broader applicability.
[22]	Investigates the impact of imbalanced datasets on the performance of various ML models for Android malware detection.	Uses a single dataset and evaluates a limited number of traditional ML algorithms.
[23]	Comprehensive benchmark of 16 data preprocessing methods for imbalanced classification	Slowness of Python-based implementations and need to subsample and perform feature selection on larger datasets.
[24]	Comparative analysis of techniques to address imbalanced datasets.	Complexity in using pre-trained models, and the need to consider dataset characteristics and other data factors.
[25]	Proposed an oversampling-based algorithm that improves recall score.	Does not explore other resampling techniques or compare them comprehensively.
[26]	Proposed a technique for generating synthetic malware samples that preserve malicious functionality.	Limited to Android malware and may not be directly applicable to other types of malware.
[27]	Proposed a novel approach combining VAEs with CNNs to address data imbalance	Complexity of the integrated VAE-CNN model, requiring careful resource management and hyperparameter tuning.
[28]	Combining TabNetClassifier with SMOTE to enhance malware detection accuracy.	Does not explore applications on other types of malware datasets
[29]	Proposed a novel approach combining multimodal fusion with weight self-learning + XGBoost to improve classification accuracy and mitigate concept drift.	Relies on static analysis, may not capture dynamic behaviors of malware.
[30]	Proposed a three-stage feature selection method and addressed class imbalance using SMOTE and cost-sensitive learning.	Limited to older ransomware families, dataset size, and specific balancing techniques.
[31]	Application of GPSC with oversampling techniques to achieve high classification accuracy and generate interpretable symbolic expressions.	Small dataset size and potential for oversampling techniques to introduce noise or overfitting.
[32]	First application of CapsNet in malware type classification, and ensemble model of CapsNet for imbalanced datasets.	Number of estimators limited to 10 RCNF due to increasing trainable parameters and significant training time.
[33]	ML framework that combines a VAE and a MLP to address imbalanced datasets and detect attack variants.	Potential for overfitting due to model complexity and need for further evaluation on other datasets.
[34]	Modified FL scheme to mitigate dataset imbalance.	Performance decline in multi-class attribution with imbalanced datasets.
[35]	Demonstrated the effectiveness of OCC in IoT malware detection using unlabeled benign data. Introduced TF-IDF for feature transformation.	Reliance on a specific dataset and potential need for further validation across more diverse IoT environments.
[36]	Proposed GPMP that uses genetic programming to select relevant features, leading to improved detection accuracy and reduced computational cost.	Lack of detailed analysis of the computational cost of the proposed method compared to others.
[37]	Adapting subspace learning techniques to OCC for malware detection.	Need for further exploration of subspace learning techniques in cybersecurity.

for feature transformation [35]. Furthermore, subspace learning based OCC models, such as Graph Embedded Subspace SVDD (GESSVDD), demonstrated excellent scalability and the capability of preventing the curse of dimensionality with a True Positive Rate of 100% [37].

- It is possible that using both of these methods will improve the performance of detection systems in identifying malware while at the same time protecting the privacy of users at the same time. However, the consistency across FL different systems remains a problem.

B. Federated Learning and Capsule Networks

- FL as a promising approach to ransomware detection, with privacy maintained through distributed training without distribution of raw data. Weighted cross entropy loss improved detection performance of minority classes across different client nodes, making FL a compelling solution for real world scenarios where data centralization is infeasible [34]. In addition, CapsNet with their capability of spatial feature preserving were proven effective in imbalanced malware type classification with high F-scores using fewer trainable parameters than conventional CNNs [32].

C. Variational Autoencoders and Symbolic Classifiers

- VAE used to address data imbalance problem by generating synthetic samples to balance malware and benign instances, which helped greatly improve CNN model generalization and the accuracy was improved from 90% to 98% [27]. GPSCs showed its inter-

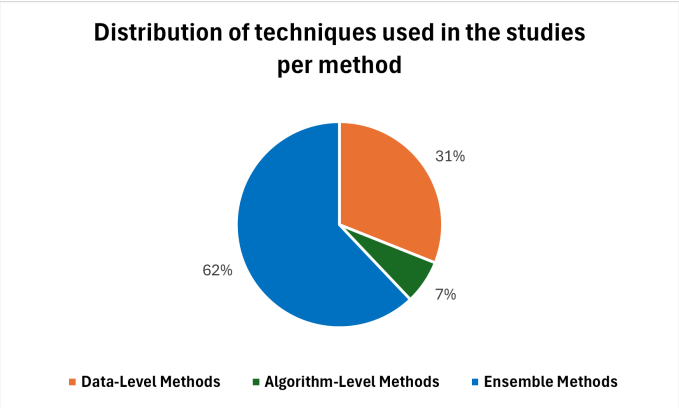

Fig. 7. Distribution of techniques used across studies.

pretability, using symbolic expressions along with oversampling techniques such as SMOTE to adequately address imbalance while still maintaining high precision and recall [31].

D. Datasets

- Malimg, BIG 2015, and Drebin are examples of the datasets widely applied in malware detection research. Although these datasets have been used to assess models, the lack of variability enhances the datasets, and the imbalance ratios are not real-world. Moreover, they were chosen because they were prevalent in the reviewed studies and relevant to malware detection research. These are benchmarks in the field, often cited due to their diversity in malware types and their real world class imbalance. For example, the Drebin dataset with more than 50,000 Android applications is one of the most used datasets to validate imbalanced learning techniques [26]. The Malimg dataset also covers a wide variety of malware families, and so is suited to the evaluation of DL and ensemble methods [13], [27].

The literature showed a wide range of techniques to address the problem of imbalanced datasets on malware detection. SMOTE, feature selection, DL, GANs, one class classification, and FL each had its own benefits, including improving detection rates and recall, maintaining privacy, and interoperability. Moreover, the literature that the ensemble method is the most used by the studies compared to other methods, as shown in Fig. 7. Hence, the importance of these advances to improving the robustness and efficiency of malware detection systems considering continuous evolution and diversification of threats.

VI. Challenges and Open Directions

Although imbalanced datasets for malware detection have been addressed, there are still open issues.

- Oversampling and Computational Challenges: Many oversampling techniques such as SMOTE and its variants, are effective, but they can also cause computational overhead and overfitting, particularly for complex synthetic data generation [23] [26]. A crucial need still remains to achieve the balance between computational efficiency and performance improvements, especially when working with resource constrained environments or large datasets.
- Feature Selection Complexity: Feature selection is another challenge. However, techniques such as Genetic Programming based feature selection and multi stage feature prioritization have been shown to be successful, but can significantly complicate the training process and require a great deal of fine tuning [30] [36]. However, the challenge to integrate such methods into real world scenarios where computational resources may be limited still remains. More efficient and automated feature selection approaches are also needed, that can lessen reliance on domain specific knowledge without compromising accuracy.
- DL and Hybrid Model Limitations: Despite their potential, DL models typically require largescale computations and are easily overfit over unbalanced data without a necessary regularization. While CNN-LSTM combined with GANs and CapsNet are effective, they bring along additional layers of complexity that hinder their practical deployment [18] [32]. Also, FL provides privacy preserving capabilities but comes at the cost of synchronization issues and model consistency across distributed nodes [34].
- Dataset: Future work should therefore aim at developing larger datasets that are more general and include samples of rare types of malware as well as more realistic conditions. Shared databases could be federations hence creating federated datasets that would otherwise share data securely.

Further research should be conducted to develop lightweight and computationally efficient models that can be used in real time malware detection environment. Transfer learning and FL are promising, but more work is needed to make them work effectively with imbalances without a huge computational overhead. Addressing these challenges will be key to making malware detection systems robust and practical for dynamic and diverse threat landscapes.

VII. Conclusion

This paper provides a comprehensive review of the problems that exist in imbalanced datasets in malware detection, presenting an investigation of existing solutions including data level, algorithm level, and ensemble methods. Key approaches to overcome data imbalance issues in malware detection are identified including SMOTE, DL hybrids like CNN and LSTM, and advanced strategies like FL. The review shows that the methods increase the malware detection rate significantly. However, issues remain like computational overhead, overfitting, and limited generalizability of these models to unseen malware types. Specifically, Modified Fuzzy-SMOTE and FL deal with some of these challenges by generating more realistic synthetic data and preserving privacy in distributed training environments.

This paper identifies a taxonomy that classifies the varied methodologies used to deal with class imbalance in malware

detection. The results show that ensemble methods are the most effective across various scenarios, especially for improving detection accuracy and robustness. The study also provides tradeoffs between computational efficiency and model performance, and provides a guide for future developments. Proposed future research may focus on creating advanced techniques and frameworks that address the difficulties of imbalanced datasets in detecting malware. The combination of Modified Fuzzy-SMOTE with feature selection methods generate realistic synthetic samples for machine learning algorithms while improving the robustness of RF and SVM classifiers. Moreover, optimizing hybrid models like CNN-LSTM becomes viable for real-time malware detection through optimization processes that include parameter sharing and model pruning mechanisms. Also, real-time data distribution adaptation in dynamic cost-sensitive learning algorithms leads to enhanced performance across malware families. Finally, the inclusion of real-world data variations with diverse samples within expanded datasets helps models achieve better generalization capabilities and maintain robustness within dynamic operational settings.

FUNDING

This work was funded by King Faisal University, Saudi Arabia. [Project No. GRANT KFU250100].

ACKNOWLEDGMENT

This work was supported through the Annual Funding track by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Project No. GRANT KFU250100].

CONFLICTS OF INTEREST

All authors declare no conflict of interest.

REFERENCES

[1] A. Alharbi, A. H. Seh, W. Alosaimi, H. Alyami, A. Agrawal, R. Kumar, and R. A. Khan, "Analyzing the impact of cyber security related attributes for intrusion detection systems," *Sustainability*, vol. 13, no. 22, p. 12337, 2021.

[2] H. Ali, M. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq, "Imbalance class problems in data mining: A review," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 3, pp. 1560–1571, 2019.

[3] S. Rane, S. Yadav, Y. Hambir, A. Gupta, and E. Kapoor, "Ai-powered malware detection: Leveraging machine learning for enhanced cybersecurity," *Nanotechnology Perceptions*, pp. 1331–1347, 2024.

[4] K. M. Hasib, M. S. Iqbal, F. M. Shah, J. A. Mahmud, M. H. Popel, M. I. H. Showrov, S. Ahmed, and O. Rahman, "A survey of methods for managing the classification and solution of data imbalance problem," *arXiv preprint arXiv:2012.11870*, 2020.

[5] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of classification methods on unbalanced data sets," *Ieee Access*, vol. 9, pp. 64 606–64 628, 2021.

[6] M. Saini and S. Susan, "Tackling class imbalance in computer vision: a contemporary review," *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 1279–1335, 2023.

[7] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of imbalanced data: review of methods and applications," in *IOP conference series: materials science and engineering*, vol. 1099, no. 1. IOP Publishing, 2021, p. 012077.

[8] AV-ATLAS, *AV-ATLAS Malware Analysis Portal*, https://portal.av-atlas.org/malware Accessed: 2024-11-20. [Online]. Available: https://portal.av-atlas.org/malware

[9] D. T. Dehkordy and A. Rasoolzadegan, "A new machine learning-based method for android malware detection on imbalanced dataset," *Multimedia Tools and Applications*, Apr. 2021. [Online]. Available: http://dx.doi.org/10.1007/s11042-021-10647-z

[10] J. Guan, X. Jiang, and B. Mao, "A method for class-imbalance learning in android malware detection," *Electronics*, vol. 10, no. 24, p. 3124, Dec. 2021. [Online]. Available: http://dx.doi.org/10.3390/electronics10243124

[11] M. E. Khoda, J. Kamruzzaman, I. Gondal, T. Imam, and A. Rahman, "Malware detection in edge devices with fuzzy oversampling and dynamic class weighting," *Applied Soft Computing*, vol. 112, p. 107783, Nov. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.asoc.2021.107783

[12] I. Almomani, R. Qaddoura, M. Habib, S. Alsoghyer, A. A. Khayer, I. Aljarah, and H. Faris, "Android ransomware detection based on a hybrid evolutionary approach in the context of highly imbalanced data," *IEEE Access*, vol. 9, p. 57674–57691, 2021. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2021.3071450

[13] J. Hemalatha, S. Roseline, S. Geetha, S. Kadry, and R. Damaševičius, "An efficient densenet-based deep learning model for malware detection," *Entropy*, vol. 23, no. 3, p. 344, 2021. [Online]. Available: http://dx.doi.org/10.3390/e23030344

[14] M. Goyal and R. Kumar, "Machine learning for malware detection on balanced and imbalanced datasets," in *2020 International Conference on Decision Aid Sciences and Application (DASA)*. IEEE, 2020, p. 867–871. [Online]. Available: http://dx.doi.org/10.1109/DASA51403.2020.9317206

[15] M. P. Salas and P. L. De Geus, "Deep learning applied to imbalanced malware datasets classification," *Journal of Internet Services and Applications*, vol. 15, no. 1, p. 342–359, Sep. 2024. [Online]. Available: http://dx.doi.org/10.5753/jisa.2024.3907

[16] A. Almaleh, R. Almushabb, and R. Ogran, "Malware api calls detection using hybrid logistic regression and rnn model," *Applied Sciences*, vol. 13, no. 9, p. 5439, Apr. 2023. [Online]. Available: http://dx.doi.org/10.3390/app13095439

[17] Y. Ding, S. Wang, J. Xing, X. Zhang, Z. Qi, G. Fu, Q. Qiang, H. Sun, and J. Zhang, "Malware classification on imbalanced data through self-attention," in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, Dec. 2020, p. 154–161. [Online]. Available: http://dx.doi.org/10.1109/TrustCom50675.2020.00033

[18] Z. Moti, S. Hashemi, and A. N. Jahromi, "A deep learning-based malware hunting technique to handle imbalanced data," in *2020 17th International ISC Conference on Information Security and Cryptology (ISCISC)*. IEEE, Sep. 2020, p. 48–53. [Online]. Available: http://dx.doi.org/10.1109/ISCISC51277.2020.9261913

[19] I. Almomani, A. Alkhayer, and W. El-Shafai, "An automated vision-based deep learning model for efficient detection of android malware attacks," *IEEE Access*, vol. 10, p. 2700–2720, 2022. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2022.3140341

[20] M. Mimura, "An improved method of detecting macro malware on an imbalanced dataset," *IEEE Access*, vol. 8, p. 204709–204717, 2020. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2020.3037330

[21] S. P. Swapna Augustine Nikale, "Android malware detection and familial classification using dynamic features for imbalanced dataset," *European Chemical Bulletin*, vol. 12, no. 7, pp. 1508–1518, 2023.

[22] Z. Sawadogo, G. Mendy, J. M. Dembele, and S. Ouya, "Android malware detection: Investigating the impact of imbalanced data-sets on the performance of machine learning models," in *2022 24th International Conference on Advanced Communication Technology (ICACT)*. IEEE, Feb. 2022, p. 435–441. [Online]. Available: http://dx.doi.org/10.23919/ICACT53585.2022.9728833

[23] R. Haluška, J. Brabec, and T. Komárek, "Benchmark of data preprocessing methods for imbalanced classification," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 2970–2979.

[24] A. Alzammam, H. Binsalleeh, B. AsSadhan, K. G. Kyriakopoulos, and S. Lambotharan, "Comparative analysis on imbalanced multi-class classification for malware samples using cnn," in *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)*. IEEE, 2020, pp. 1–6.

[25] N. M. Phung and M. Mimura, "Detection of malicious javascript on an imbalanced dataset," *Internet of Things*, vol. 13, p. 100357, 2021.

[26] M. E. Khoda, J. Kamruzzaman, I. Gondal, T. Imam, and A. Rahman, "Mobile malware detection with imbalanced data using a novel synthetic oversampling strategy and deep learning," in *2020 16th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 2020, pp. 1–6.

[27] H. H. Reshi and K. Singh, "Enhancing malware detection using deep learning approach," in *2024 International Conference on Automation and Computation (AUTOCOM)*. IEEE, 2024, pp. 497–501.

[28] R. Faridun and E. G. Im, "Enhancing malware detection with tabnet-classifier: A smote-based approach," in *Proceedings of the Korea Information Processing Society Conference*. Korea Information Processing Society, 2024, pp. 294–297.

[29] S. Li, Y. Li, X. Wu, S. Al Otaibi, and Z. Tian, "Imbalanced malware family classification using multimodal fusion and weight self-learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 7642–7652, 2022.

[30] F. C. Onwuegbuche, A. D. Jurcut, and L. Pasquale, "Enhancing ransomware classification with multi-stage feature selection and data imbalance correction," in *International Symposium on Cyber Security, Cryptology, and Machine Learning*. Springer, 2023, pp. 285–295.

[31] N. Andelic, S. Baressi Segota, and Z. Car, "Improvement of malicious software detection accuracy through genetic programming symbolic classifier with application of dataset oversampling techniques," *Computers*, vol. 12, no. 12, p. 242, 2023.

[32] A. Çayır, U. Ünal, and H. Dağ, "Random capsnet forest model for imbalanced malware type classification task," *Computers & Security*, vol. 102, p. 102133, 2021.

[33] Y.-D. Lin, Z.-Q. Liu, R.-H. Hwang, V.-L. Nguyen, P.-C. Lin, and Y.-C. Lai, "Machine learning with variational autoencoder for imbalanced datasets in intrusion detection," *IEEE Access*, vol. 10, pp. 15 247–15 260, 2022.

[34] A. Vehabovic, H. Zanddizari, N. Ghani, G. Javidi, S. Uluagac, M. Rahouti, E. Bou-Harb, and M. S. Pour, "Ransomware detection using federated learning with imbalanced datasets," in *2023 IEEE 20th International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*. IEEE, 2023, pp. 255–260.

[35] T. Shi, R. A. McCann, Y. Huang, W. Wang, and J. Kong, "Malware detection for internet of things using one-class classification," *Sensors*, vol. 24, no. 13, p. 4122, 2024.

[36] H. Harahsheh, M. Shraideh, and S. Sharaeh, "Performance of malware detection classifier using genetic programming in feature selection," *Informatica*, vol. 45, no. 4, 2021.

[37] H. H. Al-Khshali, M. Ilyas, F. Sohrab, and M. Gabbouj, "Malware detection with subspace learning-based one-class classification," *IEEE Access*, 2024.