Addressing Imbalanced Data in Network Intrusion Detection: A Review and Survey

Elham Abdullah Al-Qarni[1], Ghadah Ahmad Al-Asmari[2]
Department of Computing and Information Technology, University of Bisha, Bisha, Saudi Arabia[1]
Agency for Planning and Digital Transformation, Ministry of Hajj and Umrah, Macca, Saudi Arabia[2]

***Abstract*—The proliferation of internet-connected devices, including smartphones, smartwatches, and computers, has led to an unprecedented surge in data generation. The rapid rise in device connectivity points to an urgent need for robust cybersecurity measures to counter the mounting wave of cyber threats. Among the strategies aimed at establishing efficient network intrusion detection systems, the integration of machine learning techniques is a prominent avenue. However, the application of machine learning models to imbalanced intrusion detection datasets, such as NSL-KDD, CICIDS2017, and UGR'16, presents challenges. In such intricate scenarios, accurately distinguishing network intrusions poses a formidable challenge. The term "imbalance" refers to the imbalanced distribution of data across classes, which adversely affects the precision of machine learning algorithm classifications. This comprehensive survey embarks on a thorough exploration of the spectrum of methodologies proposed to address the challenge of imbalanced data. Simultaneously, it assesses the efficacy of these methodologies within the realm of network intrusion detection. Moreover, by shedding light on the potential consequences of not effectively tackling imbalanced data, this study aims to provide a holistic understanding of the intricate interplay between machine learning and intrusion detection in imbalanced settings.**

Keywords*—Network intrusion detection system; data imbalance; resampling; data level techniques; hybrid techniques

I. Introduction

In parallel with technological advancements and the proliferation of networks, vulnerabilities to diverse attacks have emerged, potentially leading to system damage, network disruptions, data loss, or unauthorized access. The escalation of network intrusions has become a pressing concern, impacting governments, businesses, and essential infrastructure. Network intrusion detection systems (NIDS) have come to the fore as a means of addressing these challenges. These systems employ advanced algorithms to navigate intricate and extensive data landscapes, functioning as vigilant software that enhances the monitoring of network activities. Their primary mission is to identify and categorize attacks [1].

It has become clear in recent times that the ability to identify attack patterns is crucial given the continuous evolution and increasing sophistication of cyber threats. A report by firewall maker SonicWall shows a significant increase in ransomware attacks of 105% in 2021 compared to the previous year. Additionally, there were a staggering 5.4 billion malware attacks in 2022 [2]. Artificial intelligence (AI) has a central role in addressing this pressing issue, leveraging machine learning and deep learning techniques to construct intelligent NIDS. The remarkable capabilities offered by machine learning enable meticulous analysis of vast volumes of network traffic data, tackling intricate classification challenges and automating decision-making processes.

Over the last decade, researchers have introduced a myriad of machine learning and deep learning-based solutions aimed at enhancing the efficacy of NIDS, pinpointing malicious attacks [3] [4] [5]. The architecture of the machine learning network is a core feature of this complex process. This framework encompasses several key stages: data preprocessing to ensure readiness for data analysis, feature selection to identify relevant variables, model selection to determine the optimal algorithm, training to absorb data patterns, evaluation to assess model performance, and prediction to apply trained models to new data, generating actionable outcomes. This framework is visually depicted in Fig. 3.

However, machine learning and deep learning algorithms often face the challenge of imbalanced class distributions, with certain classes significantly more prevalent than others. This imbalance poses a formidable hurdle as learning algorithms tend to gravitate toward the majority class, impacting the accuracy of classification, particularly for specific intrusion types. Several applications, for example, energy forecasting and climate data analysis [6], operate in nonstationary environments. In other words, the process of generating data is changing over time. Branco et al. [7] undertook a negative impact test of class imbalance on classifiers like decision trees, neural networks, and k-nearest neighbor. It is argued that imbalanced domains are caused by a mismatch between the importance assigned by the user to some predictions and the representativeness of those values when they are applied to the available sample data. This misclassification can have dire consequences, necessitating further investigations into intrusions if normal behavior classes are inaccurately categorized. Moreover, inaccurate intrusion categorization has the potential to inflict harm upon systems [8].

From the vantage point of data mining, the minority class often carries heightened significance. To address biases stemming from imbalanced data scenarios, it is essential to create intelligent systems, which constitute the field of "learning from imbalanced data". The essence of the class imbalance problem is often distilled into a ratio reflecting total occurrences in minority classes relative to their majority counterparts. Imbalanced data embody traits such as overlap, minimal distinct density, noisy data, and dataset variance, collectively posing substantial challenges to effective categorization.

Recently, an array of cutting-edge learning techniques has emerged, tailored to confronting classification issues embedded within imbalanced datasets [9]. Navigating the reconciliation of class imbalance is closely intertwined with addressing overlap, consistent with the overarching objective of establishing decisive boundaries between classes and facilitating clear differentiation across the spectrum of learning models [10]. This ensemble of techniques enhances accuracy across various strata, spanning inconspicuous elements and random sampling, all without necessitating replacement.

The main objective of this study is to review scientific papers addressing the problem of imbalanced data in the field of NIDS and analyze the methods used to tackle this issue. The analysis showed that oversampling techniques, such as the Synthetic Minority Over-Sampling Technique (SMOTE) and the Adaptive Synthetic (ADASYN) sampling approach, are commonly used to balance datasets.

The remainder of the paper is structured as follows: Section II outlines the survey methodology, Section III presents a comprehensive overview of the datasets, Section IV examines the prevalent techniques employed to address imbalanced data, and finally, Section V provides concluding reflections and highlights potential avenues for future research.

II. Survey Methodology

To conduct the survey concerning imbalanced data within intrusion detection datasets, the study undertook a meticulous analysis of scholarly articles sourced from esteemed publishers of research literature, namely Elsevier, Springer, MDPI, and IEEE. A two-fold approach was employed to select the most pertinent papers. First, we searched specific keywords associated with unbalanced data, such as "class imbalance" and "intrusion detection system", to pinpoint papers likely to address the subject matter. In the second phase, we conducted a meticulous assessment to exclude scientific papers that did not originate from reputable academic journals. This stringent process guaranteed the inclusion of papers that adhered to rigorous academic standards and were founded on robust research methodologies. By applying these dual steps, we identified and curated research papers that offered valuable insights into the intricacies of imbalanced data within intrusion detection datasets.

The holistic workflow of these techniques to address imbalance is succinctly portrayed in Fig. 1.

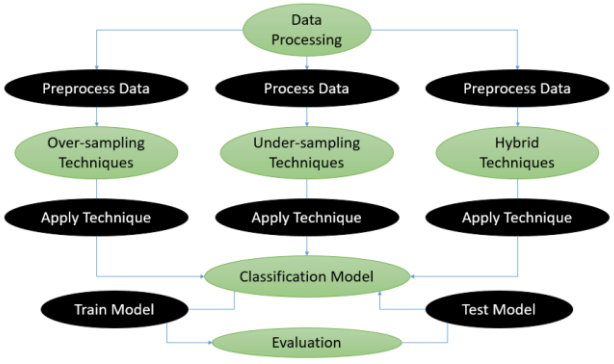

Fig. 1. Flow of imbalance technique approaches.

III. Data Description

Intrusion detection datasets exhibit variations in terms of release dates, sizes, attack classifications, and data collection methods. This section offers a comprehensive overview of prominent datasets utilized in intrusion detection research, providing insights into their key attributes and significance.

A. CICIDS2017

The CICIDS2017 dataset, delivered in 2017, is a significant asset containing roughly 2.8 million records with 83 features. It remains as a demonstration of the developing idea of digital dangers, embodying fourteen unmistakable assault types going from customary Forswearing of Administration (DoS) assaults to additional refined methods like Cross-Site Prearranging (XSS). The expansiveness and profundity of this dataset make it an important resource for concentrating on the complexities of organization intrusion detection [11].

B. CSE-CIC-IDS2018

Presented in 2018 by the Canadian Organization for Network safety (CIC) and Correspondences Security Foundation (CSE), the CSE-CIC-IDS2018 dataset addresses a huge progression in intrusion detection research. With roughly 16.2 million records and 80 features, this dataset gives a rich wellspring of information for examining different sorts of intrusion assaults, including Conveyed Refusal of Administration (DDoS) and beast force web assaults. The sheer volume and variety of assault examples make it an optimal possibility for far reaching examination and assessment of detection systems [12].

C. CIDDS-001

The Coburg Intrusion Detection Data Sets (CIDDS-001), stand apart as a noticeable dataset for network-based intrusion detection, flaunting roughly 32 million records with 14 credits. What sets this dataset separated is its broad inclusion of assault types, incorporating a stunning 92 particular classifications going from Savage Power to Ping Outputs. The wealth and granularity of this dataset make it an important asset for scientists looking to investigate the full range of organization intrusion situations [13].

D. KDD99

The KDD Cup 99 dataset, beginning from 1999 under the support of the Guard Progressed Exploration Ventures Organization (DARPA), addresses a primary asset in the field of intrusion detection. In spite of its age, this dataset remains profoundly significant, containing around five million records with 41 features. Its attention on essential assault types, for example, DoS, test, Client to Root (U2R), and Remote to Nearby (R2L) gives important bits of knowledge into the early scene of digital dangers and the viability of detection procedures. [14].

E. UNSW-NB15

Delivered in 2015 by the Digital Reach Lab, the UNSW-NB15 dataset offers a cutting edge viewpoint on intrusion detection, highlighting 49 features and enveloping nine assault types. Its consideration of assorted assault classes, including Conventional, Exploits, and Observation, mirrors the developing idea of digital dangers in contemporary

organizations. Additionally, its generally late delivery guarantees its importance in tending to ebb and flow difficulties in intrusion detection research [14].

F. UNSW-NB18

The UNSW-NB18 BoT-IoT dataset, an expansion of the UNSW-NB15 dataset, addresses a significant extension regarding information volume and assault groupings. With more than 72 million records and assault classes like Keylogging, operating system, and Information exfiltration, this dataset offers remarkable bits of knowledge into the complicated interaction between IoT gadgets and organization security. Its accessibility in different renditions, incorporating a consolidated variant with roughly three million records, gives adaptability for scientists fluctuating computational assets [15].

G. NSL-KDD

The NSL-KDD dataset fills in as an improvement to the KDD Cup 99 dataset, tending to weaknesses like information overt repetitiveness and copies. While its emphasis stays on essential assault classes steady with KDD Cup 99, its smoothed out construction and end of superfluous information make it a more productive and open asset for intrusion detection research [6].

H. UWF-ZeekData22

Arising in 2022, the UWF-ZeekData22 dataset addresses a spearheading exertion in network checking, utilizing imaginative information assortment procedures and examination strategies. With roughly 18 million records and 14 sorts of assaults, this dataset offers new bits of knowledge into arising dangers and weaknesses in present day organizations. Its joining with the open-source Zeek instrument further upgrades its utility for specialists and experts the same [16].

I. UGR'16

The UGR'16 dataset, custom-made for recognizing network security peculiarities, includes two unmistakable sets: Alignment and TEST. Laid out in 2016, this dataset catches a different exhibit of malware classes, including secure shell (SSH), spam, and port filtering. Its attention on abnormality detection highlights the developing significance of proactive safety efforts in alleviating arising dangers [17]. Table I provides a summary of the characteristics of these datasets.

TABLE I. SUMMARIZES THE OVERALL CHARACTERISTICS OF ALL DATASETS

Dataset	Dataset Type	Records	Features	Number of attacks
CICIDS2017	Multi class	2830540	83	14
CIDDS-001	Multi class	31.959.175	14	92
CSE-CIC IDS2018	Multi class	16.232.943	80	6
KDD99	Binary class	4.898.430	41	4
NSL-KDD	Binary class	N/A	41	4
UGR'16	N/A	16.900.000	12	7
UNSW-NB15	Multi class	2.540.044	49	9
UNSW-NB18	Multi class	3.668.522	42	6

In spite of the lavishness and variety of these datasets, a common worry across each of the nine is information lopsidedness. The lopsided conveyance of assault occasion classes and the striking inconsistency between ordinary traffic cases and those addressing different assault classifications present huge difficulties for intrusion detection research. Tending to these irregular characteristics requires cautious thought of examining procedures, highlight determination, and algorithmic ways to deal with guarantee vigorous and dependable detection capacities.

IV. COMMON STRATEGIES FOR ADDRESSING IMBALANCED DATA

Dealing with imbalanced data has emerged as one of the most formidable challenges in the field of machine learning. Studies have proposed various approaches to mitigate this issue, encompassing data sampling, cost-sensitive analyses, ensemble learning, algorithmic methodologies, and more. These strategies can be categorized into three main types, as shown in Fig. 2.

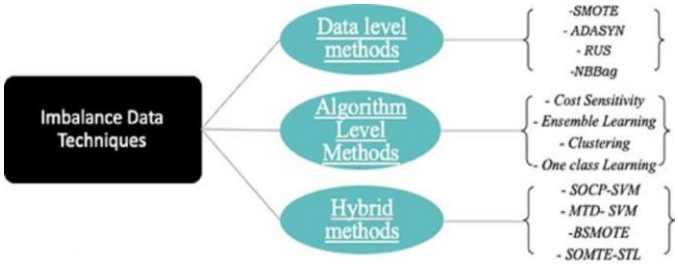

Fig. 2. Handling imbalanced data methods.

Extensive research has been undertaken to tackle the problem of data imbalance, particularly within network intrusion detection systems. This section provides an overview of select studies that examine these techniques. Table II presents a compilation of studies that have employed different approaches across the most widely used intrusion detection datasets to provide a comprehensive understanding of the diverse methodologies aimed at combating imbalanced data issues,

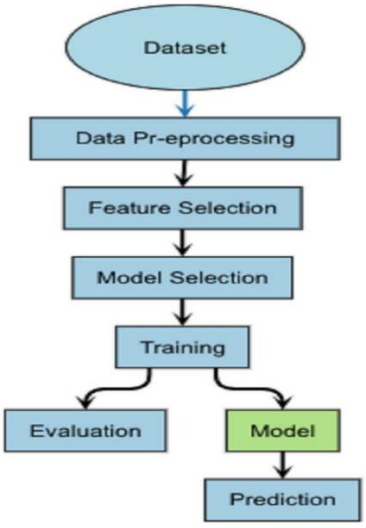

Fig. 3. Machine learning framework.

A. Data-Level Techniques for Addressing Imbalanced Data

- Data-level procedures include preprocessing steps pointed toward amending awkward nature inside datasets. These procedures, otherwise called outside strategies, try to accomplish information proportionality by either decreasing greater part class tests or increasing minority-class tests. Normal information level procedures include:
- SMOTE: SMOTE involves generating synthetic instances for the minority class by interpolating existing minority class instances [18]. A new specimen is created by selecting a random k-nearest neighbor (KNN) of an underrepresented instance and generating a value from a random combination of both interpolated instances. This method aids in spreading minority classes into the space occupied by majority classes, resulting in better defined decision boundaries.
- ADASYN: Sampling methods such as ADASYN enhance learning from data distributions by reducing the bias caused by class imbalances and reshaping classification boundaries toward challenging examples [31].

1) Used considerable amount of research has been conducted in the field of NIDS employing data-level techniques, as shown in Table II.

TABLE II. RECENT RESAMPLE TECHNICAL BASED NIDS STUDIES

Dataset	Technique	AI-based approaches		Year	Reference
		ML	DL		
NSL−KDD	SMOTE−ENN	No	Yes	2019	Zhang et al. [19]
CICIDS2017	Uniform Distribution Based Balancing (UDBB)	Yes	No	2019	Abdulhammed et al. [20]
CICIDS2017	SMOTE	Yes	No	2019	Yulianto et al. [21]
CSE−CIC−IDS2018	SMOTE	Yes	No	2020	Karatas et al. [22]
UGR'16	GAN	Yes	No	2020	Yilmaz et al. [23]
NSL−KDD UNSW−NB15	OSS and SMOTE	No	Yes	2020	Jiang et al. [24]
CIDDS-001 UNSW-NB15	SMOTE-STL	Yes	Yes	2021	Al and Dener [13]
NSL−KDD UNSW−NB15 CICIDS2017	ADASYN	Yes	No	2021	Liu et al. [25]
UNSW-NB15	SMOTE	Yes	No	2022	Ahmed et al. [26]
KDD99 NSL−KDD UNSW-NB15	SMOTE	No	Yes	2022	Meliboev et al. [27]
NSL-KDD	ADASYN	No	Yes	2022	Fu et al. [28]
UNSW-NB15	SMOTE	No	Yes	2023	Almarshdi et al. [29]
CICIDS2017 KDD99 UNSW-NB15	Ensemble method	Yes	No	2023	Thockchom et al. [30]
UWF-ZeekData22 UNSW-NB15	Random under sampling before splitting. Random under sampling after splitting. (B−SMOTE)	Yes	No	2023	Bagui et al. [16]

a) Ahmed et al. [26] proposed a NIDS framework using various machine learning schemes to detect network attack categories. The framework includes techniques such as data standardization, normalization, and SMOTE. This model achieved 95.1% accuracy on the UNSW-NB15 dataset.

b) Yulianto et al. [21] applied a similar technique to address data imbalance in their proposed IDS. They employed principal component analysis (PCA), ensemble feature selection (EFS), and SMOTE to enhance AdaBoost-based IDS performance on the CICIDS2017 dataset, achieving accuracy, precision, recall, and an F1 score of 81.83%, 81.83%, 100%, and 90.01%, respectively.

c) Karatas et al. [22] employed the CSE-CIC-IDS2018 dataset to build an efficient IDS using the SMOTE technique, resulting in an average increase in accuracy of 4.01% to attain 30.59% accuracy across different machine learning models.

d) Meliboev et al. [27] applied machine learning and deep learning techniques to detect security attacks. They used SMOTE to enhance model performance on the UNSW-NB15, KDD99, and NSL-KDD datasets, achieving accuracy scores of 91.2%, 95.2%, and 82.6%, respectively.

e) Almarshdi et al. [29] developed a hybrid deep learning IDS using convolution neural network (CNN) and long short-term memory (LSTM) algorithms, combined with the SMOTE technique. This model achieved 92.10% accuracy on the UNSW-NB15 dataset compared to 89.90% for the basic CNN model.

f) Fu et al. [28] introduced the Deep Learning Network Intrusion Detection (DL-NID) model using bidirectional LSTM (Bi-LSTM) and attention mechanisms, incorporating the ADASYN technique. This model achieved an accuracy of 90.73% on the NSL-KDD dataset.

g) Liu et al. [25] also employed the ADASYN technique in their proposed IDS, achieving accuracy scores of 92.57%, 85.89%, and an impressive 99.91% on the NSL-KDD, UNSW-NB15, and CICIDS2017 datasets, respectively.

Table III illustrates the ratios of measures that were attained by researchers in each study before addressing the issue of data imbalance using data-level techniques. As can be seen, Liu et al. [25] attained the highest accuracy rate of 99.86% on the CICIDS2017 dataset. In contrast, the lowest accuracy score achieved was 55% in the study conducted by Meliboev et al. [27] on the UNSW-NB15 dataset.

Table IV illustrates the ratios of measures that were attained by researchers in each study after addressing the issue of data imbalance using data-level techniques. Liu et al. achieved the highest accuracy rate of 99.91% when applying the ADASYN technique in their study. Several NIDS models demonstrated improved accuracy after implementing various data-level techniques. For instance, Meliboev et al. conducted a study on the UNSW-NB15 dataset using the recurrent neural network (RNN) algorithm. Initially, they obtained an accuracy rate of 55%, but after applying the SMOTE technique, the accuracy increased significantly to 71.90%. These findings highlight the effect and importance of data-level techniques in enhancing NIDS performance.

TABLE III. RESULTS OF MODELS BEFORE HANDLING IMBALANCED DATA

Algorithm / framework	Dataset	Accuracy	F1-score	Recall	Precision	Reference
RF	UNSW-NB15	89.5%	73.7%	72.3%	77.3%	Ahmed et al. [26]
DT		88.5%	70.7%	72%	70.9%	
LR		82.2%	41.9%	42.3%	51.3%	
KNN		84%	53.3%	51.3%	57.8%	
ANN		85.2%	54.4%	54.6%	61.2%	
AdaBoost	CICIDS2017	-	-	-	-	Yulianto et al. [21]
KNN	CSE-CIC-IDS2018	98.52%	98.89%	98.52%	99.28%	Karatas et al. [22]
RF		99.21%	99.25%	99.2%	99.30%	
Gradient Boosting		99.11%	99.29%	99.11%	99.51%	
AdaBoost		99.69%	99.7%	99.69%	99.7%	
DT		99.66%	99.60%	99.66%	99.66%	
Linear Discriminant Analysis		90.80%	99%	99.11%	98.90%	
CNN	UNSW-NB15	85.8%	87.8%	99.4%	80.9%	Meliboev et al. [27]
LSTM		84.9%	87.7%	98.3%	79.2%	
GRU		57%	71.3%	97.3%	56.3%	
RNN		55%	71%	100%	55.1%	
CNN + LSTM		80.8%	85%	99.3%	74.4%	
CNN	KDD99	92.3%	95.2%	91%	99.8%	
LSTM		91.8%	94.7%	91.1%	98.6%	
GRU		90.7%	93.8%	88.70%	99.7%	
RNN		91.7%	94.6%	90.2%	99.4%	
CNN + LSTM		92.7%	95.2%	91%	99.8%	
CNN	NSL-KDD	78.8%	77.7%	65%	96.7%	
LSTM		76.2%	74.2%	60.2%	96.9%	
GRU		72.5%	68.5%	52.4%	98.7%	

RNN		63.2%	70.5%	56.2%	94.6%	
CNN + LSTM		85.5%	85.9%	77.1%	96.1%	
CNN + LSTM	UNSW-NB15	91.86%	91.7%	90.91%	91.8%	Almarshdi et al. [29]
Bi-LSTM+ attention mechanisms	NSL-KDD	-	-	-	-	Fu et al. [28]
LightGBM	NSL-KDD	89.79%	-	-	-	Liu et al. [25]
	UNSW-NB15	83.98%	-	-	-	
	CICIDS2017	99.86%	-	-	-	

TABLE IV. RESULTS AFTER HANDLING IMBALANCED DATA

Technique	Algorithm / framework	Dataset	Accuracy	F1-score	Recall	Precision	Ref.
SMOTE	RF	UNSW-NB15	95.1%	95.1%	95.7%	94.8%	Ahmed et al. [26]
	DT		94.7%	94.8%	95.4%	94.4%	
	LR		69.4%	56.2%	59.4%	61%	
	KNN		84.7%	83.1%	85.1%	82.2%	
	ANN		77.6%	71.5%	70.6%	76.2%	
SMOTE + EFS	AdaBoost	CICIDS2017	81.83%	90.01%	100%	81.83%	Yulianto et al. [21]
SMOTE	KNN	CSE-CIC-IDS2018	98.8%	98%	98.08%	97.92%	Karatas et al. [22]
	RF		99.35%	99.35%	99.34%	99.35%	
	Gradient Boosting		99.29%	99.3%	99.29%	99.3%	
	AdaBoost		99.6%	99.6%	99.61%	99.6%	
	DT		99.57%	99.56%	99.57%	99.56%	
	Linear Discriminant Analysis		91.18%	91.57%	91.18%	91.96%	
SMOTE	CNN	UNSW-NB15	91.2%	91.5%	96.1%	87.5%	Meliboev et al. [27]
	LSTM		88.9%	89.5%	94.8%	84.8%	
	GRU		77.9%	79%	83.2%	75.3%	
	RNN		71.9%	76.5%	91.3%	65.8%	
	CNN + LSTM		87.6%	88%	90.6%	85.5%	
	CNN	KDD99	95.2%	94.9%	90.7%	99.5%	
	LSTM		95.4%	95.1%	91.4%	99.4%	
	GRU		94.1%	93.8%	88.9%	99.1%	
	RNN		94.1%	93.6%	90%	98%	
	CNN + LSTM		95.2%	94.9%	90.8%	99.5%	
	CNN	NSL-KDD	79.3%	74.8%	61.4%	95.5%	
	LSTM		75.8%	69.2%	54.2%	95.4%	
	GRU		79.1%	74.5%	61.2%	95.4%	
	RNN		76.1%	71.7%	60.5%	88%	
	CNN + LSTM		82.6%	79.8%	68.9%	99.5%	
SMOTE	CNN + LSTM	UNSW-NB15	92.10%	90.11%	91.75%	92.85%	Almarshdi et al. [29]
ADASYN	Bi-LSTM+ attention mechanisms	NSL-KDD	90.73%	89.65%	93.17%	86.38%	Fu et al. [28]
ADASYN	LightGBM	NSL-KDD	92.57%	-	-	-	Liu et al. [25]
		UNSW-NB15	85.89%	-	-	-	
		CICIDS2017	99.91%	-	-	-	

B. Algorithm-Level Approaches

Algorithm-level approaches focus on enhancing the learning capacity of classifier algorithms with regard to minority classes. These methods are often referred to as internal approaches. Techniques such as adjusting the probability estimation or modifying class-specific costs can be employed to benefit minority classes [31].

1) Cost-Sensitive Learning: The cost-sensitive learning framework lies between internal and external approaches. This technique integrates both algorithmic and data-level modifications in a unified approach by altering the learning process and assigning costs to samples accordingly [13].

2) Ensemble Learning: Ensemble learning combines various methodologies to address imbalanced classes. Ensembles based on techniques such as bagging and boosting are commonly used to tackle class imbalance issues.

a) Thockchom et al. [30] employed ensemble methods, specifically the stacking ensemble technique, on the KDD99, CIC-IDS2017, and UNSW-NB15 datasets. The stacking ensemble combined Gaussian naïve Bayes (GNB), decision tree (DT), and logistic regression (LR) classifiers. The proposed model achieved high accuracy levels: 99.80% on CIC-IDS2017, 93.88% on UNSW-NB15, and 99.84% on KDD99.

b) Yilmaz et al. [23] proposed a model for intrusion detection using the UGR'16 dataset. They used generative adversarial networks (GANs) to address the issue of unbalanced data. The degree of balance in the training dataset was determined using multilayer perceptron.

c) Abdulhammed et al. [20] presented an anomaly-based IDS applied to the CICIDS2017 dataset. The imbalanced distribution in the dataset was handled using a uniform distribution-based balancing method. Performance metrics were calculated based on five classifiers: the random forest (RF) algorithm, a Bayesian network, linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA). The model achieved the highest accuracy of 98.80%.

C. Hybrid Approaches

Hybrid strategies amalgamate methods from both algorithmic and information levels in ideal extents. These methodologies consolidate the qualities of algorithmic and information level strategies while relieving their particular shortcomings, at last further developing order accuracy. Normal hybrid calculations include:

1) SOCP-SVM: Support Vector Machines with Second-Order Cone Programming.

2) MTD-SVM: Multi-Threshold Decision-Support Vector Machines.

3) B-SMOTE: Borderline SMOTE.

4) SOMTE-STL: Synthetic Minority Over-sampling Technique-SMOTE with Tomek Links.

Application of Hybrid Techniques for Handling Imbalanced Data

1) Jiang et al. [24] employed two methods to address data imbalance on the NSL-KDD and UNSW- NB15 datasets. They combined one-side selection (OSS) to reduce majority samples and SMOTE to increase minority sample sizes. A deep hierarchical network model integrating CNN with BiLSTM achieved accuracy rates of 83.58% and 77.16%, respectively. Zhang et al. [19] also employed a hybrid sampling method combining SMOTE with edited nearest neighbors (SMOTE-ENN) to achieve an accuracy of 83.31% on the NSL-KDD dataset using CNN.

2) Al and Dener [13] utilized hybrid sampling with SMOTE and Tomek-Links Sampling (STL) to address imbalance in the CIDDS-001 and UNSW-NB15 datasets. Their Hybrid Deep Learning Approach combined CNN and LSTM algorithms, outperforming other deep learning and machine learning algorithms.

Trial review accentuate the power of information driven oversampling calculations in reinforcing base classifier execution, really tending to irregularity issues across different models, including AI and profound learning. The SMOTE has demonstrated success in diverse domains by creating new minority instances, circumventing overfitting and promoting classifier generalization [32]. This approach effectively addresses imbalance issues across various models, including machine learning and deep learning.

V. Conclusions

All in all, this review has embraced an exhaustive examination of strategies for tending to class irregularity in interruption identification datasets, with an emphasis on the viability of different procedures. Through our examination, we have assessed the presentation of oversampling techniques, for example, Destroyed and ADASYN, revealing insight into their adequacy in moderating the difficulties presented by imbalanced information.

Our examination has added to the comprehension of how these strategies can be applied with regards to interruption location, giving bits of knowledge into their assets and impediments. We have emphasized ADASYN's notable effectiveness in rebalancing datasets and increasing classification accuracy in particular.

While our review takes care of many systems, it's fundamental to perceive the developing idea of interruption location research. While we zeroed in basically on oversampling procedures, there are different methodologies, for example, bunch based under-examining that warrant further investigation. This features the continuous quest for creative procedures to handle class unevenness in interruption location situations.

In synopsis, our review fills in as an important asset for scientists and professionals in the field, offering experiences into the present status of the workmanship and making ready for future progressions in tending to the difficulties of imbalanced information in network interruption location.

1) Statistical Analysis: To measure the exhibition of intrusion detection frameworks when applying SMOTE and

ADASYN oversampling methods, we look at the exactness, F1-score, review, and accuracy measurements straightforwardly as shown in Table V and Fig. 4:

TABLE V. AVERAGE TECHNIQUE LIST

Technique	Accuracy	F1-score	Recall	Precision
SMOTE	88.13%	86.96%	85.68%	90.68%
ADASYN	92.28%	89.65%	93.17%	86.38%

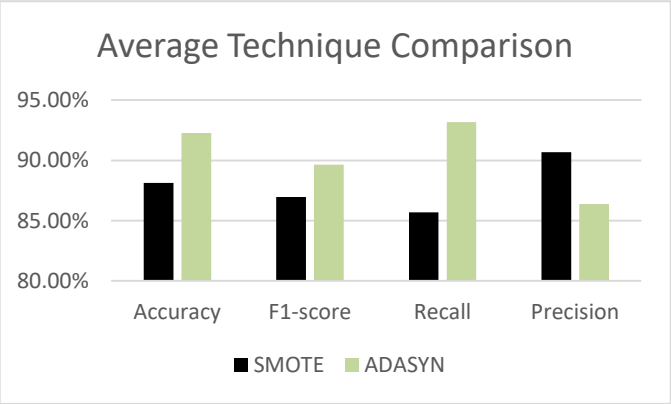

Fig. 4. Average technique comparison.

These discoveries exhibit that while the two strategies work on the general execution of intrusion detection frameworks, ADASYN gives better exactness, F1-score, and review, while SMOTE might be ideal for keeping up with higher accuracy.

REFERENCES

[1] T. Saranya, S. Sridevi, C. Deisy, T. D. Chung, and M. K. A. A. Khan, "Performance analysis of machine learning algorithms in intrusion detection dystem: a review," 2020. doi: 10.1016/j.procs.2020.04.133.

[2] S. Inc., "2022 SonicWall cyber threat report," 2022. https://www.sonicwall.com/resources/white-papers/2022-sonicwall-cyber-threat-report/ (accessed Jan. 01, 2024).

[3] H. Alqahtani, I. H. Sarker, A. Kalim, S. M. Minhaz Hossain, S. Ikhlaq, and S. Hossain, "Cyber intrusion detection using machine learning classification techniques," 2020. doi: 10.1007/978-981-15-6648-6_10.

[4] A. Halimaa and K. Sundarakantham, "Machine learning based intrusion detection system," in 2019 3rd International conference on trends in electronics and informatics (ICOEI), 2019, pp. 916–920.

[5] L. Ashiku and C. Dagli, "Network intrusion detection system using deep learning," 2021. doi: 10.1016/j.procs.2021.05.025.

[6] Z. Wang, Z. Li, J. Wang, and D. Li, "Network intrusion detection model based on improved BYOL self-supervised learning," Secur. Commun. Networks, 2021, doi: 10.1155/2021/9486949.

[7] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," ACM Computing Surveys. 2016. doi: 10.1145/2907070.

[8] A. A. Alqarni and E. S. M. El-Alfy, "Improving intrusion detection for imbalanced network traffic using generative deep learning," Int. J. Adv. Comput. Sci. Appl., 2022, doi: 10.14569/IJACSA.2022.01304109.

[9] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of imbalanced data: review of methods and applications," IOP Conf. Ser. Mater. Sci. Eng., 2021, doi: 10.1088/1757-899x/1099/1/012077.

[10] D. Devi, S. K. Biswas, and B. Purkayastha, "Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique," Conn. Sci., 2019, doi: 10.1080/09540091.2018.1560394.

[11] R. Panigrahi and S. Borah, "A detailed analysis of CICIDS2017 dataset for designing intrusion detection dystems," Int. J. Eng. Technol., 2018.

[12] S. Alzughaibi and S. El Khediri, "A cloud intrusion detection dystems based on DNN using backpropagation and PSO on the CSE-CIC-IDS2018 dataset," Appl. Sci., 2023, doi: 10.3390/app13042276.

[13] S. Al and M. Dener, "STL-HDL: a new hybrid network intrusion detection system for imbalanced dataset on big data environment," Comput. Secur., 2021, doi: 10.1016/j.cose.2021.102435.

[14] S. Choudhary and N. Kesswani, "Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 Datasets using deep learning in IoT," 2020. doi: 10.1016/j.procs.2020.03.367.

[15] Y. Pacheco and W. Sun, "Adversarial machine learning: a comparative study on contemporary intrusion detection datasets," 2021.

[16] S. Bagui, D. Mink, S. Bagui, S. Subramaniam, and D. Wallace, "Resampling imbalanced network intrusion datasets to identify rare attacks," Futur. Internet, 2023, doi: 10.3390/fi15040130.

[17] M. Nkongolo, J. P. van Deventer, and S. M. Kasongo, "Ugransome1819: a novel dataset for anomaly detection and zero-day threats," Inf., 2021, doi: 10.3390/info12100405.

[18] H. Ali, M. N. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq, "Imbalance class problems in data mining: a review," Indones. J. Electr. Eng. Comput. Sci., 2019, doi: 10.11591/ijeecs.v14.i3.pp1552-1563.

[19] X. Zhang, J. Ran, and J. Mi, "An intrusion detection system based on convolutional neural network for imbalanced network traffic," 2019. doi: 10.1109/ICCSNT47585.2019.8962490.

[20] R. Abdulhammed, H. Musafer, A. Alessa, M. Faezipour, and A. Abuzneid, "Features dimensionality reduction approaches for machine learning based network intrusion detection," Electron., 2019, doi: 10.3390/electronics8030322.

[21] A. Yulianto, P. Sukarno, and N. A. Suwastika, "Improving AdaBoost-based intrusion detection system (IDS) performance on CIC IDS 2017 dataset," 2019. doi: 10.1088/1742-6596/1192/1/012018.

[22] G. Karatas, O. Demir, and O. K. Sahingoz, "Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset," IEEE Access, 2020, doi: 10.1109/ACCESS.2020.2973219.

[23] I. Yilmaz, R. Masum, and A. Siraj, "Addressing imbalanced data problem with generative adversarial network for intrusion detection," 2020. doi: 10.1109/IRI49571.2020.00012.

[24] K. Jiang, W. Wang, A. Wang, and H. Wu, "Network intrusion detection combined hybrid sampling with deep hierarchical network," IEEE Access, 2020, doi: 10.1109/ACCESS.2020.2973730.

[25] J. Liu, Y. Gao, and F. Hu, "A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM," Comput. Secur., 2021, doi: 10.1016/j.cose.2021.102289.

[26] H. A. Ahmed, A. Hameed, and N. Z. Bawany, "Network intrusion detection using oversampling technique and machine learning algorithms," PeerJ Comput. Sci., 2022, doi: 10.7717/PEERJ-CS.820.

[27] A. Meliboev, J. Alikhanov, and W. Kim, "Performance evaluation of deep learning based network intrusion detection system across multiple balanced and imbalanced datasets," Electron., 2022, doi: 10.3390/electronics11040515.

[28] Y. Fu, Y. Du, Z. Cao, Q. Li, and W. Xiang, "A deep learning model for network intrusion detection with imbalanced data," Electron., 2022, doi: 10.3390/electronics11060898.

[29] R. Almarshdi, L. Nassef, E. Fadel, and N. Alowidi, "Hybrid deep learning based attack detection for imbalanced data classification," Intell. Autom. Soft Comput., 2023, doi: 10.32604/iasc.2023.026799.

[30] N. Thockchom, M. M. Singh, and U. Nandi, "A novel ensemble learning-based model for network intrusion detection," Complex Intell. Syst., 2023, doi: 10.1007/s40747-023-01013-7.

[31] M. O. Miah, S. S. Khan, S. Shatabda, and D. M. Farid, "Improving detection accuracy for imbalanced network intrusion classification using cluster-based under-sampling with random forests," 2019. doi: 10.1109/ICASERT.2019.8934495.

[32] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," Journal of Artificial Intelligence Research. 2018. doi: 10.1613/jair.1.11192.