Al-Azhar Bulletin of Science

Volume 35 | Issue 3 Article 7

2024
Section: Computer Science

Enhancing IoT Data Balance Based on Feature Selection and Dataset Integration

Khaled Abdelrahman Abdelhakim
Department of Mathematics, Faculty of Science, Al-Azhar University, Cairo, Egypt.,
khaledabdo@azhar.edu.eg

Assad Ahmed Gad-Elrab
Department of Mathematics, Faculty of Science, Al-Azhar University, Cairo, Egypt., ahmad4@kau.edu.sa

Mohamed Sayed Farag
Department of Mathematics, Faculty of Science, Al-Azhar University, Cairo, Egypt.,
mohamed.s.farag@azhar.edu.eg

Shaban Ebrahim Abu-Youssef
Department of Mathematics, Faculty of Science, Al-Azhar University, Cairo, Egypt.,
abuyousf@hotmail.com

How to Cite This Article

Abdelhakim, Khaled Abdelrahman; Gad-Elrab, Assad Ahmed; Farag, Mohamed Sayed; and Abu-Youssef, Shaban Ebrahim (2024) "Enhancing IoT Data Balance Based on Feature Selection and Dataset Integration," *Al-Azhar Bulletin of Science*: Vol. 35: Iss. 3, Article 7.
DOI: https://doi.org/10.58675/2636-3305.1687

ORIGINAL ARTICLE

Enhancing Internet of Things Data Balance Based on Feature Selection and Dataset Integration

Khaled Abdelrahman Abdelhakim [a,*], Asaad Ahmed Gad-Elrab [a,b], Mohamed Sayed Farag [a,c], Shaban Ebrahim Abu-Youssef [a]

[a] Department of Mathematics, Faculty of Science, Al-Azhar University, Cairo, Egypt
[b] Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[c] Computer Science Department, Obour High Institute for Informatics, Cairo, Egypt

Abstract

The Internet of Things (IoT) has significantly advanced since its creation, revolutionizing both business processes and social interactions by connecting devices and data. However, this progress introduces security challenges. To tackle these issues, this paper presents an Imbalance Reduction Model in IoT Data Sets Based on Feature Selection architecture. This model employs machine learning methods to classify IoT traffic, emphasizing flow and Transmission Control Protocol data from datasets like UNSW-NB15 and Bot-IoT. The introduced model is proficient at distinguishing between normal, Denial of service, and distributed Denial of service traffic, addressing problems like data imbalance and overfitting. It achieves classification accuracy between 98.38% and 100%, significantly improving IoT security by effectively identifying and countering malicious traffic. Utilizing machine learning shows resilience to emerging threats, underscoring its potential as a strong intrusion detection system for IoT settings.

Keywords: Bot-IoT and UNSW-NB15 datasets, Cyber security, Data science, Feature selection, Internet of things, Intrusion detection system, Machine learning

1. Introduction

The Internet of Things (IoT) has been transformative, connecting a vast array of devices that communicate to enhance efficiency and decision-making. These devices, ranging from simple sensors to complex systems, have been widely adopted in various sectors, including smart homes, healthcare, and industrial automation. The global IoT ecosystem has rapidly expanded, with projections suggesting that up to 75 billion devices could be connected by 2025 [1]. However, the widespread adoption of IoT has introduced significant security challenges. The decentralized and heterogeneous nature of IoT networks makes them particularly vulnerable to cyberattacks. These networks consist of diverse devices with varying capabilities and security standards, which cybercriminals often exploit [2]. Traditional security methods, although effective in conventional IT systems, often fall short in addressing the unique demands of IoT environments [3].

To address these security concerns, Machine Learning (ML) and Deep Learning (DL) techniques have been increasingly applied. These methods have shown potential in detecting and classifying cyber threats by analyzing large volumes of data generated by IoT devices. Nonetheless, the effectiveness of these models heavily depends on the quality of the data and the features selected for training. A significant challenge faced in this area is data imbalance, where certain types of network traffic or attack scenarios are underrepresented in the datasets. This imbalance can result in biased models that struggle to detect less frequent or emerging threats [4]. Furthermore, existing studies have primarily focused on detecting specific attack types, such as distributed Denial of service (DDoS), while other IoT-specific

Received 8 August 2024; revised 24 August 2024; accepted 27 August 2024.
Available online 8 November 2024

* Corresponding author at: Department of Mathematics, Faculty of Science, Al-Azhar University, Cairo, 11837, Egypt.
E-mail address: khaledabdo@azhar.edu.eg (K.A. Abdelhakim).

https://doi.org/10.58675/2636-3305.1687

threats, like firmware attacks and sensor spoofing, remain underexplored [5].

IoT's integration with stochastic computing paradigms has facilitated the numerical simulation of complex systems, such as heterogeneous mosquito models. By incorporating randomness and uncertainty, IoT-enabled stochastic computing offers a more accurate simulation of biological systems, enhancing reliability in modeling mosquito populations [6]. Similarly, IoT-enabled heuristic models using Morlet wavelet neural networks have improved the numerical treatment of heterogeneous mosquito release ecosystems, capturing nonlinear dynamics and optimizing intervention strategies [7].

IoT technology has also been applied to solve complex differential equations in engineering and science. For example, a neural network procedure has been developed for solving the sixth-order nonlinear singular pantograph differential model, addressing the challenges posed by its high-order and singular nature [8]. Additionally, Gudermannian neural networks, optimized through IoT, have been used to solve two-point nonlinear singular models in thermal-explosion theory, providing accurate predictions by addressing singularities [9].

With the rapid increase in diverse devices designed to simplify daily life, improve health management, and boost efficiency, the IoT has become essential for enhancing human experiences. This rise in device use has generated vast amounts of data and requests, but it has also introduced vulnerabilities, including attacks that transmit false data to obtain sensitive information. Detecting such erroneous data remains a significant challenge. Although well-known attacks can be managed with established methods, new threats require innovative techniques, including ML and DL, alongside existing hybrid strategies [10].

Researchers have invested significant effort into developing ML-based systems, involving key processes such as gathering data from various sources and devices. This task presents challenges due to the diversity of data and specific attack patterns of different devices. Proper data cleaning and preparation, while maintaining its original characteristics, are essential, followed by feature selection for model training. The system's performance hinges on the quality and relevance of these features, enhancing its speed and robustness [10,11]. Cybercriminals might intercept sensitive communications, disrupt operations, or exploit vulnerabilities to launch DDoS or ransomware attacks, even targeting security devices like cameras.

The research domain in this area covers various aspects, including data collection and the challenges of adapting to new environments. Efforts are made to avoid erroneous data and ensure compatibility with modern systems. Some studies emphasize data cleaning and retaining essential information, while others explore feature selection techniques. These techniques vary from manual selection based on feature descriptions and attack detection impact to mathematical methods (e.g., percentile selection, K-Best, heatmap) and machine learning methods.

However, some studies fail to critically assess feature–class relationships, leading to the inclusion of influential but potentially misleading features. To address this, correlation matrices are often used.

This research addresses these issues by introducing a system that combines two established datasets, Bot-IoT and UNSW-NB15. These datasets were manually cleaned and merged, selecting significant features from a research standpoint. Despite challenges in the manual process, it was crucial to the study. The model employed LSTM with two hidden layers, using the chunking method due to large data volumes. A novel approach was developed to reduce reliance on potentially problematic features: feature selection was performed separately on each dataset, followed by a correlation matrix and merging. This strategy cut the feature count from 13 to 6, speeding up training and minimizing the risk of system failure in identifying zero-day attacks by balancing feature reliance [10,11].

Protective devices, such as cameras, are vulnerable to cyberattacks, with incidents rising more than threefold from the first half of 2019 compared with the latter half of 2018. During this time, IoT devices faced ~2.98 billion cyber-attacks [1]. This increase correlates with the expanding use of IoT devices, which were estimated at 30 billion in 2020 and are expected to reach 75 billion by 2025 [12].

Essential security practices for IoT devices involve encrypting communications to safeguard data, implementing regular updates to fix vulnerabilities, using robust authentication methods, securing networks with firewalls, managing devices securely, conducting regular security audits, and educating users about cybersecurity. With the expanding IoT landscape, focusing on cybersecurity is crucial to maintaining the safety and efficiency of these interconnected systems.

The key contributions of this study include:

(a) Creation of a new, balanced dataset by merging several datasets, covering various attack patterns and addressing data inconsistencies.
(b) Reduction of the total number of features by selecting a subset from the integrated datasets.
(c) Application of feature selection methods to remove features that could impair model accuracy.

(d) Development of an effective model for detecting zero-day attacks and managing real-world. Scenarios, utilizing various machine learning techniques for both binary and multi-class classification with the new dataset.

The paper is structured as follows: Section 2 reviews related research, while Section 3 describes the IRM-BFS architecture in detail. Section 4 covers the experiments conducted with this architecture and their outcomes. Section 5 presents an analysis and comparison with other solutions. Finally, Section 6 concludes with reflections and future research directions.

2. Background and related work

Khraisat *et al*. (2021) [13] offer a comprehensive analysis of IoT intrusion detection systems, examining methods, deployment strategies, datasets, and technologies. They highlight both advantages and challenges related to IoT architecture, summarize recent research, and suggest improvements for IoT IDS performance. The paper also addresses the limitations of traditional systems and outlines future research directions.

Peterson *et al*. (2021) [14] provide an in-depth review of the Bot-IoT dataset, discussing its features, limitations, data cleaning, and previous research applications. Zeeshan *et al*. (2021) [15] propose a Protocol-Based Deep Intrusion Detection (PB-DID) architecture, creating a dataset from IoT traffic by comparing UNSW-NB15 and Bot-IoT features, focusing on flow and Transmission Control Protocol (TCP). The PB-DID model classifies nonanomalous, Denial of service (DoS), and DDoS traffic, addressing issues like imbalance and overfitting, and achieves a classification accuracy of 96.3% using deep learning techniques.

Latif *et al*. (2022) [16] introduce a lightweight dense random neural network (DnRaNN) for IoT intrusion detection, suitable for resource-constrained networks. Their model, tested on the ToN IoT dataset, provides valuable insights and achieves high accuracy in detecting attacks. Addressing imbalanced datasets, oversampling, and undersampling techniques are discussed. Undersampling reduces the majority class to match the minority class, with methods like Random Undersampling and Focused Undersampling. Oversampling increases the minority class representation, using Random Oversampling and the Synthetic Minority Over-sampling Technique, each with its advantages and trade-offs [17,18].

Wang and Liu's study on imbalanced class issues in Bot-IoT achieved 84% accuracy, highlighting bias problems [17]. Wu and Liu explored ensemble methods on UNSW-NB15, with 87% accuracy, but noted computational intensity [19]. Evans and Zhao focused on feature selection's impact, reporting 86% accuracy [20]. Wang and Chen used transfer learning for UNSW-NB15, reaching 92% accuracy, emphasizing the need for relevant pre-trained models [18]. Liu and Wang's reinforcement learning approach for UNSW-NB15 reached 93% accuracy [21]. Larriva *et al*. [22] examined preprocessing techniques, reporting high accuracy for UNSW-NB15, UGR16, and NSL-KDD using normalization methods and Multi-Layer Perceptron classification. Another study [23] applied Synthetic Minority Over-sampling Technique to Bot-IoT, achieving perfect accuracy with a Deep Recurrent Neural Network (DRNN), though normalization might affect realism.

Churcher *et al*. [24] compared machine learning techniques on Bot-IoT, achieving 99% accuracy with K-Nearest Neighbors (KNN) by selecting features with high relevance. The Improved Conditional Variational AutoEncoder (ICAVE) method balanced datasets like NSLKDD and UNSW-NB15, with varying accuracy across dataset variants [25]. Shafiq *et al*. [26] used various machine learning techniques on Bot-IoT, focusing on top features for high accuracy. Guizani *et al*. [27] and Alkadi *et al*. [28] applied LSTM and BiLSTM techniques on UNSW-NB15 and Bot-IoT, achieving notable accuracy with BiLSTM reaching over 98%.

The research highlights the importance of feature selection and dataset integration, but many studies rely on individual datasets, potentially overlooking vital information. Challenges arise in creating balanced datasets from existing ones, such as manual merging and feature selection. This study will explore methods for constructing reliable datasets and models, emphasizing feature selection and the use of various ML techniques for classification.

A common approach in these studies is to use feature importance or information gain for feature selection. However, adding new data might necessitate feature replacement, requiring the identification of crucial features that remain relevant with additional data. While many studies use a single dataset for model training and validation, few compare and integrate features from multiple datasets to form a new dataset. Utilizing the full dataset can lead to more comprehensive learning and avoid missing critical details.

Previous research indicates that most studies apply ML or DL methods to develop models for detecting cyber-attacks, relying on existing datasets. This often results in a gap between model performance and real-world data, which includes various attacks and

inconsistencies. Moreover, current studies have not compared datasets to identify overlapping features, which is crucial for training effective ML or DL models. Creating models with real data is preferred, but acquiring such data is challenging due to the time, cost, and effort required.

One study proposed generating a balanced dataset from multiple existing datasets but faced issues such as manual dataset merging and feature selection. Many studies aiming for high model efficiency by focusing on class-specific features may find their models less effective in real-world scenarios. This study will address these challenges by developing methods for building a new dataset and constructing a model that selects reliable features. The study will also examine the proposed model using various classification methods and ML techniques, detailed in the following sections.

3. The proposed imbalance reduction model in IOT data sets based on feature selection (IRM-BFS)

3.1. Basic idea

Given the difficulties faced in the research, a new and effective approach was introduced to tackle the issues of combining two distinct datasets. The Imbalance Reduction Model in IoT Data Sets Based on Feature Selection (IRM-BFS) is proposed to address these challenges. This model incorporates several key elements: (1) Data preprocessing, (2) Feature selection (FS), (3) Integration of datasets using shared features, (4) Feature elimination, and (5) Application of standard machine learning methods to the combined datasets. The FS technique was applied separately to each dataset to ensure optimal integration and efficiency. The common features selected were crucial for merging the datasets, resulting in a more cohesive and refined dataset that highlights the essential attributes for the analysis.

3.2. Proposed model

The amalgamated dataset, now refined to encompass these nine selected features, underwent training and testing via four distinct machine learning techniques: Decision Trees (DT), k-Nearest Neighbors (KNN), Random Forest (RF), and XGBoost. The focus of the analysis lay in the classification of data into three categories of attacks: Normal, DoS, and DDoS.

However, the outcomes of the model testing phase presented a perplexing scenario. Strikingly similar results were yielded by all four machine learning methods, implying a seemingly illogical correspondence between them. This called for a comprehensive investigation to discern the root causes of this unexpected uniformity.

To delve into this issue further and illuminate the reasons behind the congruent results, a comprehensive examination was initiated. A correlation matrix was employed to explore the relationships between each feature and their impact on classification. This analytical approach served as a critical step in elucidating the intricacies of the model's behavior Fig. 1.

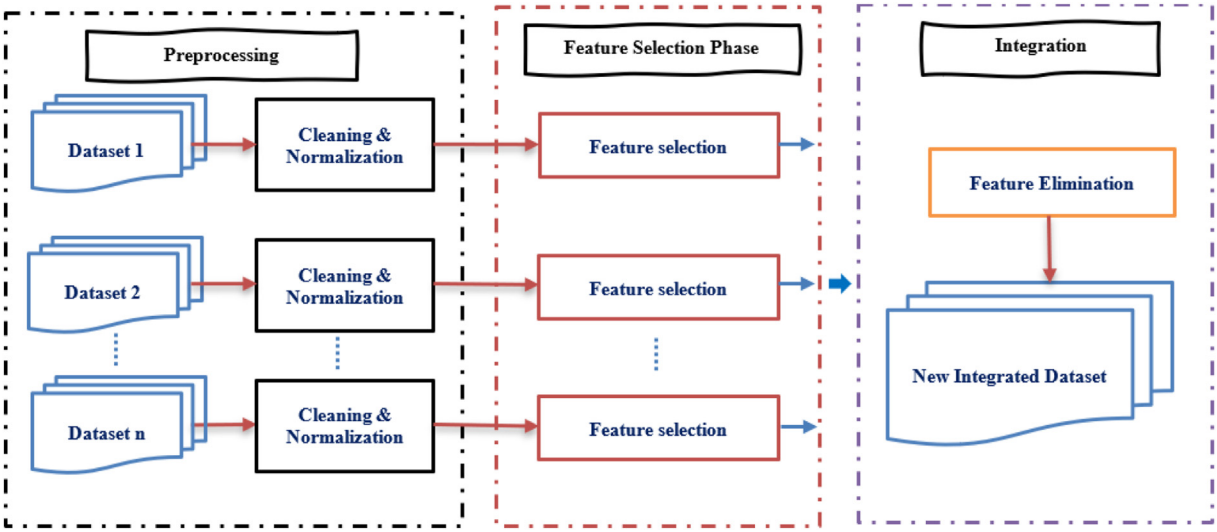

Fig. 1. The proposed imbalance reduction model based on feature selection Model.

4. Preprocessing

Dataset preprocessing involves several essential steps performed to guarantee the data's quality and appropriateness for further analysis. These steps include cleaning data to address inconsistencies and errors, integrating data from various sources, transforming data to standardize and scale features, and reducing data dimensions while preserving crucial information. When these preprocessing operations are conducted methodically, they improve the reliability and efficiency of intrusion detection systems within IoT networks.

Handling missing values is a common aspect of data cleaning. If we represent our dataset as $X = \{x_1, x_2, \ldots, x_n\}$, then replacing each missing value xi with the mean μ can be represented as:

$$x_i' = \begin{cases} x_i \text{ if } x_i \text{ is not missing} \\ \mu \text{ if } x_i \text{ is missing} \end{cases} \quad (1)$$

Normalization techniques, like Min-Max normalization, adjust the data to a defined range, often between 0 and 1. The Min-Max normalization formula for a value xi in a dataset is:

$$x_i' = \frac{x_i - \min(X)}{\max(X) - \min(X)} \quad (2)$$

where:

xi is the original value,

$\min(X)$ and $\max(X)$ are the minimum and maximum values in the dataset X,

$x'\ i$ is the normalized value.

4.1. Feature selection

At this stage, each dataset is independently subjected to four distinct feature selection techniques. The primary goal is to achieve optimal alignment of features between the two datasets before merging them into a single, comprehensive dataset. This process aims to reduce the number of features, thereby decreasing the data volume for subsequent phases. Feature selection methods are employed to accomplish this objective. The essence of this phase lies in the meticulous selection and refinement of features to harmonize the characteristics of both datasets. This feature alignment is critical before dataset integration, ensuring that only the most pertinent and informative features are retained, thus streamlining subsequent data analysis and modeling. By strategically reducing the feature count, this phase enhances resource efficiency, model interpretability, and performance in the later stages of the research or project.

In the context of feature selection for each cleaned dataset, C_i, a crucial procedure is involved. This involves meticulously selecting features to acquire a subset that best serves the desired objectives. The variable S_i is introduced, signifying the collection of carefully chosen features.

For every cleaned dataset, C_i, a diligent and systematic feature selection process is executed. This process seeks to sift through the available features and determine the most relevant and valuable subset, represented by the variable S_i, which encapsulates the chosen features Eq. (3).

$$S_i = SelectFeatures(C_i) \quad (3)$$

4.2. Integrating common features

To integrate common features from the datasets, an intersection operation is performed on the selected features. This involves identifying and extracting the features that are common across all datasets, resulting in a unified set of shared characteristics. This process ensures that only the features present in every dataset are retained, leading to a consistent and integrated set of attributes. The result of this intersection is a unified collection of common features in Eq. (4).

$$CommonFeatures = S_1 \cap S_2 \cap \ldots \cap S_i \quad (4)$$

4.3. Features elimination

Feature elimination is essential for boosting the performance and effectiveness of machine learning models. By removing correlated features related to the classes, redundancy is minimized, which enhances model interpretability and generalization. This technique decreases dimensionality, reduces overfitting, and streamlines the model, resulting in improved computational efficiency. The Pearson correlation coefficient between two variables, X and Y, is defined in Eq. (5):

$$r_{xy} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}} \quad (5)$$

Where, r_{xy} represents the correlation coefficient between two variables, denoted as X and Y. The subscripts i refer to individual sample points, where X_i and Y_i represent specific values in the datasets. X and Y signify the mean (average) of the X and Y samples, respectively. The numerator, $\sum (X_i - \overline{X})(Y_i - \overline{Y})$ computes the sum of the product of deviations of paired scores from their respective means. The denominator

involves the product of the square root of each sum of squares of the deviations. The correlation matrix R for a set of variables $(X_1, X_2, \ldots, X_n)$ is constructed in Matrix (6):

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \ldots & r_{1n} \\ r_{21} & 1 & r_{23} & \ldots & r_{2n} \\ r_{31} & r_{32} & 1 & \ldots & r_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \ldots & 1 \end{pmatrix} \tag{6}$$

In this matrix: Each element r_{ij} is the correlation coefficient between X_i and X_j. The diagonal elements are all 1, as the correlation of a variable with itself is always perfect, and the matrix is symmetric, meaning $(r_{ij} = r_{ji})$.

4.4. Machine learning methods

DT: is a flowchart-like structure in which each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. In mathematical terms, the decision at each node is made using a function like Eq. (9).

$$f(x) = \begin{cases} \mathrm{f}_1(x) \; if \; \mathrm{condtion}_1 \\ \mathrm{f}_2(x) \; if \; \mathrm{condtion}_2 \\ \quad \vdots \\ \quad \vdots \\ \mathrm{f}_3(x) \; if \; \mathrm{condtion}_3 \end{cases} \tag{9}$$

Where x represents an input sample, $f(x)$ represents the decision function, $f_i(x)$ represents the decision function at the i-th node, and $\mathrm{condition}_i$ represents the decision rule derived from the data at the i-th node.

Gini Impurity: is a measure of how often a randomly chosen element would be incorrectly identified. An attribute with a lower Gini impurity is preferred. The Gini Impurity of a set S is defined as:

$$I_{\mathrm{Gini}}(S) = 1 - \sum_{i=1}^{c} p_i^2 \tag{10}$$

where p_i is the proportion of the elements in class i in the set S, and c is the number of classes.

Entropy (Information Gain): Entropy measures disorder or uncertainty. The goal of machine learning models is to reduce this uncertainty. The entropy of a set S is defined as shown in Eq. (11):

$$I_{\mathrm{Entropy}}(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{11}$$

where p_i is the proportion of the elements in class i in the set S, and c is the number of classes.

The objective function for a DT is to maximize the information gained at each node, which is defined as the difference in impurity before and after the split. Information gain IG for a split on dataset D with subsets $(D_1, D_2, \ldots, D_k)$ based on feature A can be computed as shown in Eq. (12).

$$IG(D, A) = I_{\mathrm{Entropy}}(D) - \sum_{j=1}^{k} \frac{|D_j|}{|D|} I_{\mathrm{Entropy}}(D_j) \tag{12}$$

where $|\mathrm{D_j}|$ is the number of elements in subset $\mathrm{D_j}$, and $|\mathrm{D}|$ is the total number of elements in dataset D.

XGBoost: is a gradient boosting framework that optimizes the following objective function, which comprises a loss function and a regularization term as shown in Eq. (13)

$$\mathrm{Obj} = \sum_{i=1}^{n} l(y_i, \widehat{y_i}) + \sum_{k} \Omega(f_k) \tag{13}$$

Where $l(y_i, \widehat{y_i})$ is the loss function, and $\Omega(f_k)$ is the regularization term.

Random Forest (RF): RF algorithm creates a set of decision trees from randomly selected subsets of the training set and aggregates their predictions. For classification and regression, the final prediction $\widehat{y}$ for a sample x shown in Eq. (14):

$$\widehat{y} = \mathrm{mode}\{T_1(x), T_2(x), \ldots, T_n(x)\} \tag{14}$$

k-Nearest Neighbors (KNN): Given a training set and a new sample x, KNN finds the kNN and predicts the output based on these neighbors. The Euclidean distance is calculated as shown in Eq. (15)

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{p} (x_{il} - x_{jl})^2} \tag{15}$$

where p is the number of features. The final prediction is made by majority voting or averaging of the labels of these k neighbors.

5. Experiments and results

5.1. Datasets

Bot-IoT Dataset: The Bot-IoT dataset is a large and detailed collection of data related to IoT network traffic, featuring approximately 46 attributes. These attributes cover various aspects of network activity,

including source and destination IP addresses, port numbers, packet lengths, and flow durations. The dataset supports thorough analysis and the creation of advanced machine-learning models for detecting anomalies and intrusions. Its extensive and varied features are ideal for researchers aiming to develop robust security measures for IoT environments.

UNSW-NB15 Dataset: The UNSW-NB15 dataset is a significant resource frequently used in network intrusion detection research. It contains 1,753,342 records and encompasses nine types of attacks, making it suitable for in-depth analysis. With a total of 49 features, the dataset provides a broad range of data for developing and accessing intrusion detection systems, benefiting from its size and diversity to tackle security issues effectively.

5.2. Performance evaluation

In this study, the classifiers' performance has been evaluated using critical metrics: Precision, Recall, and F-measure. These metrics are obtained from the confusion matrix analysis shown in Table 1. The definitions and explanations of these key evaluation criteria are detailed below.

Confusion matrix: A confusion matrix is a table used to evaluate the performance of a classification model. It is defined as follows in Table 1:

Precision: Precision measures how well the model identifies true positive (TP) cases, reflecting the ratio of correctly identified intrusions to all instances labeled as such. It is calculated as the ratio of TP to the sum of TP and false positive (FP), mathematically expressed as:

$$P = \frac{TP}{(TP + FP)} \tag{16}$$

Recall: Recall, also referred to as Sensitivity or TP rate, gauges how well the model identifies all genuine positive cases. It is determined by the ratio of TP to the total of TP and FN, and is expressed as:

$$R = \frac{TP}{(TP + FN)} \tag{17}$$

F1-measure: The F1-measure, also known as the F1-score, integrates Precision and Recall into a unified evaluation metric. It is calculated as follows:

Table 1. Confusion matrix.

Actual	Predicted	
	Positive	Negative
Positive	*True Positive (TP)*	*False Negative (FN)*
Negative	*False Positive (FP)*	*True Negative (TN)*

$$F1 = \frac{2 * P * R}{(P + R)} \tag{18}$$

This metric, together with the confusion matrix, provides a thorough assessment of the model's intrusion detection efficacy, balancing precision and recall. A high F1-score reflects a robust model that effectively detects intrusions while reducing false alarms.

Figures 2 and 3 illustrate the confusion matrices for binary classification where the task is to differentiate between DDoS attacks and Normal traffic. Specifically, Fig. 2 presents the results of RF, DT, KNN, and XGBoost classifiers in distinguishing DoS from Normal traffic, while Fig. 3 focuses on the same classifiers applied to DDoS versus Normal traffic. The matrices demonstrate how each model

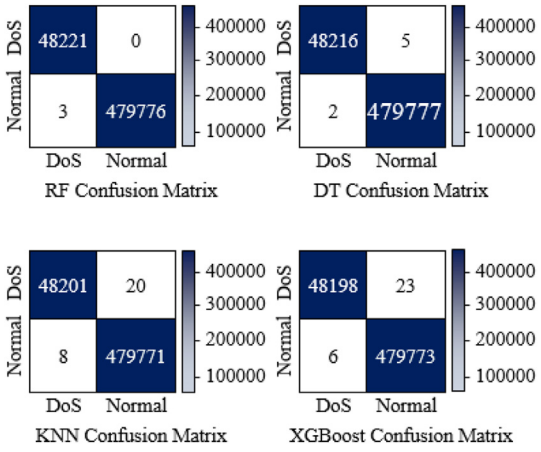

Fig. 2. Confusion matrix for binary classification ML techniques DDoS/Normal.

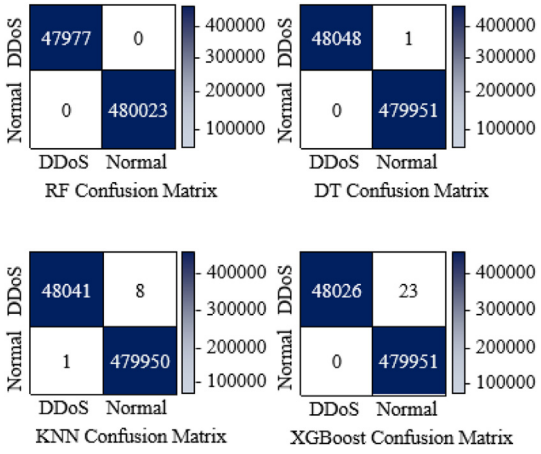

Fig. 3. Confusion matrix for binary classification ML techniques DDoS/Normal.

performs in terms of true positives, false positives, false negatives, and true negatives, with the majority of instances being correctly classified, as indicated by the high values along the diagonal. Fig. 4 extends the analysis to multi-class classification, where the task is to distinguish among three classes: Normal, DoS, and DDoS traffic. The confusion matrices for RF, DT, KNN, and XGBoost classifiers in this more complex scenario reveal that while the classifiers generally perform well, the presence of additional classes introduces more misclassification, particularly between DoS and DDoS attacks. The matrices in Fig. 4 indicate that some models have difficulty distinguishing between the two types of attacks, as evidenced by the nonzero values in the off-diagonal elements, particularly in the RF and XGBoost models. These figures collectively illustrate the effectiveness and challenges of applying different ML techniques to both binary and multi-class classification tasks in the context of IoT network security.

5.3. Feature selection methods

In the process of feature selection for each cleaned dataset, referred to as C_i, a key procedure is implemented. This involves the careful selection of features to obtain a specific subset that meets the objectives. During this stage, the variable S_i is introduced, representing the set of features that have been meticulously chosen.

During this stage, four distinct feature selection techniques are independently applied to each dataset. The goal is to achieve optimal feature alignment between the datasets before combining

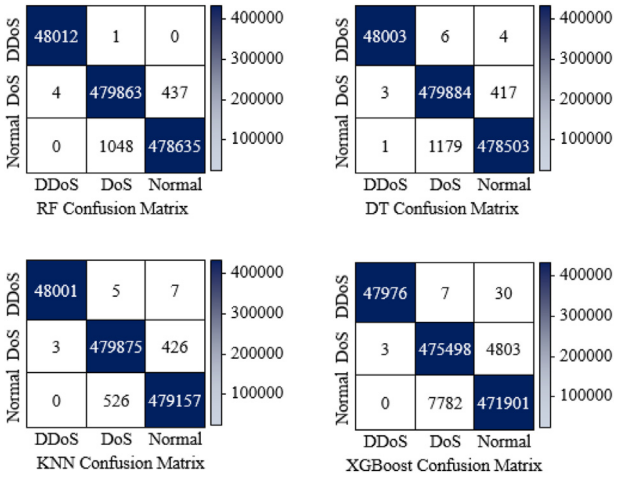

Fig. 4. Confusion matrix for Multi Classification ML.

Table 2. IRM-BFS binary classification results for DoS and normal classes.

ML	Precision				Recall				F1-score				Accuracy			
	KNN	DT	XGBoost	RFs	KNN	DT	XGBoost	RFs	KNN	DT	XGBoost	RFs	KNN	DT	XGBoost	RFs
DoS	99.99%	99.96%	99.98%	99.94%	99.95%	99.99%	99.96%	100%	99.97%	99.93%	99.97%	99.97%	99.95%	99.99%	99.95%	99.99%
Normal	99.95%	99.99%	99.96%	100%	99.99%	100%	99.98%	99.99%	99.97%	99.99%	99.97%	100%				

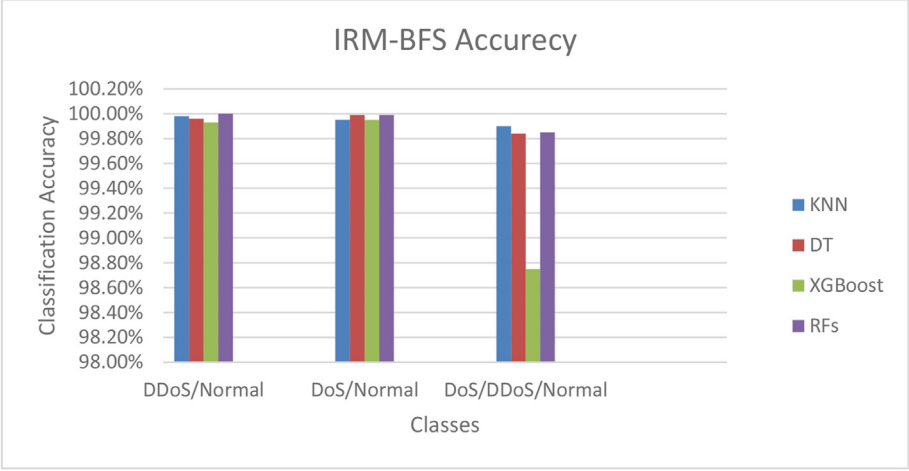

Fig. 5. Imbalance reduction model based on feature selection Classification Accuracy.

them into a comprehensive dataset. This process aims to reduce the number of features, thereby decreasing the volume of data for later stages. The feature selection methods are employed with this main objective in mind.

The results indicated that the Select Percentile method excelled compared to others, identifying a total of 8 matching features, as shown in the accompanying table. The success of the Select Percentile method highlights its efficiency in feature selection. These 8 features are crucial as they align well between datasets, facilitating smooth data integration. This alignment aids subsequent analysis phases, ensuring the dataset focuses on key attributes. The effectiveness of Select Percentile demonstrates its ability to identify and prioritize the most informative features, contributing to the project's success. This achievement reflects the precision of the feature selection process, ensuring that subsequent analyses are conducted on a refined dataset.

In the study, a Binary and Multiclassification experiment was conducted to analyze and compare the model's performance under various conditions. This experiment involved classifying data into three specific categories:

(a) Binary Classification: Normal, and DoS.
(b) Binary Classification: Normal, and DDoS.
(c) Multi-Classification: Normal, Dos, and DDoS.

Figure 5.

These features had a significant impact on binary classification, serving as major factors in decision-making. Conversely, these features had a lesser impact in multi-classification scenarios. In response, the model was strategically adjusted. Features critical to binary classification were deliberately excluded from both the training and testing phases to mitigate bias and reduce undue focus, thereby aiming to enhance overall accuracy and reliability. The IRM-BFS achieved an accuracy between 98.38 and 100%, as detailed in Tables 2–4, reflecting a notable performance improvement. The following section meticulously documents the refinement process and outcomes of the model (see Table 3).

Table 5 provides a comparative analysis of IRM-BFS with other studies, covering aspects such as datasets, data sampling methods, algorithms, feature counts, accuracy, and classification types. This comparison highlights IRM-BFS's strengths and distinguishes it from similar research, offering insights into the field of intrusion detection and the effectiveness of various methods.

In conclusion, the model effectively integrated two diverse datasets and optimized feature selection, resulting in a refined dataset suitable for analysis. The systematic evaluation of machine learning techniques, classification categories, and feature impact contributed to the model's accuracy and reliability, addressing the initial research challenges.

Table 3. Imbalance reduction model based on feature selection binary classification results for DDoS and Normal classes.

ML	Precision				Recall				F1-score				Accuracy			
	KNN	DT	XGBoost	RFs	KNN	DT	XGBoost	RFs	KNN	DT	XGBoost	RFs	KNN	DT	XGBoost	RFs
DDoS	99.94%	100%	100%	100%	99.99%	99.96%	99.92%	100%	99.91%	99.98%	99.96%	100%	99.98%	99.96%	99.93%	100%
Normal	99.99%	99.96%	99.92%	100%	99.94%	100%	100%	100%	99.91%	99.98%	99.96%	100%				

Table 4. Imbalance reduction model based on feature selection multi classification accuracy.

ML	Precision				Recall				F1-score				Accuracy			
	KNN	DT	XGBoost	RFs	KNN	DT	XGBoost	RFs	KNN	DT	XGBoost	RFs	KNN	DT	XGBoost	RFs
DDoS	99.99%	99.99%	99.99%	99.99%	99.98%	99.98%	99.92%	100%	99.98%	99.99%	99.96%	99.99%	99.90%	99.84%	98.75%	99.85%
DoS	99.89%	99.75%	98.39%	99.78%	99.91%	99.91%	99.00%	99.91%	99.90%	99.83%	98.69%	99.84%				
Normal	99.91%	99.91%	98.99%	99.91%	99.89%	99.75%	98.38%	99.78%	99.90%	99.83%	98.68%	99.85%				

Table 5. Imbalance reduction model based on feature selection comparison with other related techniques.

Technique	Algorithm	Data Sets	Data Sample	Features	*Accuracy*			Classes
Alkadi *et al.*, [29]	*BiLSTM*	UNSW-NB15 Bot-IoT	14,000 pkts	Full	***99.41 98.91***			16
Ibitoye *et al.*, [30]	*FNN SNN*	Bot-IoT	20%	10 bests	***95 91***			5
Larriva *et al.*, [22]	*MLP*	UNSW-NB15	10%	Full	***99.2***			NA
PB-DID [15]	*LSTM*	Bot-IoT UNSW-NB15	96% 87%	26	***96.3***			3
					DoS/Normal	***DDoS/Normal***	***Multi***	Binary and Multi
	KNN				***99.95***	***99.98***	***99.90***	
Proposed IRM-BFS	*DT*	Merged Bot-IoT and UNSW-NB15	96% 87%	6	***99.99***	***99.96***	***99.84***	
	XGBoost				***99.95***	***99.93***	***98.75***	
	RF				***99.99***	***100***	***99.85***	

5.4. Conclusion

This study introduces a novel approach to enhancing the security of IoT networks through the development of an IRM-BFS. By integrating and refining features from two prominent datasets, Bot-IoT and UNSW-NB15, the model successfully addresses the challenges of data imbalance and feature relevance in intrusion detection systems. The application of machine learning techniques, including DT, RF, KNN, and XGBoost, demonstrated high accuracy in detecting and classifying various types of cyberattacks, particularly DoS and DDoS attacks.

The key contribution of this research lies in its ability to create a more balanced and representative dataset by merging multiple datasets and applying feature selection methods that reduce the potential for overfitting and improve model robustness. The model achieved classification accuracy between 98.38 and 100%, underscoring its effectiveness in real-world scenarios. This work not only highlights the importance of feature selection and dataset integration in improving intrusion detection but also sets the stage for future advancements, including the incorporation of deep learning techniques and real-time data processing to further enhance IoT network security.

Ethics information

Compliance with ethical standards.

Funding

There are no funding sources.

Author contributions

The author's contributions for this work encompassed various aspects, including conceptualization, methodology, software development, validation, formal analysis, investigation, resource management, data curation, and the initial drafting of the manuscript. These tasks were collectively undertaken by Ahmed. A. A. Gad-Elrab and Khaled A. A. Khalaf-Allah.

Ahmed. A. A. Gad-Elrab also assumed the role of reviewing and editing the manuscript to ensure its quality and coherence. In addition, supervision of the research project was carried out by Ahmed. A. A. Gad-Elrab, in collaboration with Mohamed S. Farag and S. E. Abo-Youssef who collectively provided oversight and guidance to ensure the successful execution of the study.

Conflict of interest

There are no conflicts of interest.

Acknowledgements

This research received support from the Faculty of Science at Al-Azhar University inCairo, Egypt, and partial support from King Abdul-Aziz University in Jeddah, Saudi Arabia. The contributions from both organizations are greatly appreciated and acknowledged as essential components that enabled the research endeavour.

References

[1] Department SR. Internet of things (IoT) connected devices installed base worldwide from 2015 to 2025. Url:, https://wwwstatista.com/statistics/471264/iot-number-of-connected devices-worldwide. 2016.
[2] Khan MAASK. IoT security: review, blockchain solutions, and open challenges. Future Generat Comput Syst 2018;82:395–411.
[3] Jing QaVAVaWJaLJaQD. Security of the internet of things: perspectives and challenges. Wireless Network 2014;20:2481–501.
[4] Chen SaXHaLDaHBaWH. A vision of IoT: applications, challenges, and opportunities with China perspective. IEEE Internet Things J 2014;1:349–59.
[5] Sicari SaRAaGLAaCPA. Security, privacy and trust in Internet of Things: the road ahead. Comput Network 2015;76:146–64.
[6] Latif SaSZaRMAZaAGCaNRASaGDOaSRaAMR. IoT technology enabled stochastic computing paradigm for numerical simulation of heterogeneous mosquito model. Multimed Tool Appl 2023;82:18851–66.
[7] Sabir ZaNKaRMAZaHMRaUMaIAAAaLDN. IoT technology enabled heuristic model with Morlet wavelet neural network for numerical treatment of heterogeneous mosquito release ecosystem. IEEE Access 2021;9:132897–913.
[8] Sabir ZaUMaSSaST. A reliable neural network procedure for the novel sixth-order nonlinear singular pantograph differential model. Mod Phys Lett B 2024:2450473. https://doi.org/10.1142/S0217984924504736.
[9] Fatima SaSZaBDaASE. Gudermannian neural networks for two-point nonlinear singular model arising in the thermal-explosion theory. Neural Process Lett 2024;56:1–27.
[10] Khraisat A, Gondal I, Vamplew P, Kamruzzaman J. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity 2019;2:1–22.
[11] Peterson G, Singh K, Choo KKR. Review of the bot-IoT dataset and its use in the detection of IoT botnet attacks. Comput Secur 2021;102:102117.
[12] Alsop T. Internet of things security spending worldwide from 2016 to 2021. Url: https://www.statista.com/statistics/543089/iot-security-spending-worldwide. 2020.
[13] Khraisat A, Alazab A. A critical review of intrusion detection systems in the internet of things: techniques, deployment strategy, validation strategy, attacks, public datasets and challenges. Cybersecurity 2021;4:1–27.
[14] Peterson JM, Leevy JL, Khoshgoftaar TM. A review and analysis of the BoT-IoT dataset. In: 2021 IEEE international conference on service-oriented system engineering (SOSE); 2021. p. 20–7.
[15] Zeeshan M, Riaz Q, Bilal MA, Shahzad MK, Jabeen H, Haider SA, et al. Protocol-based deep intrusion detection for dos and ddos attacks using UNSW-NB15 and BoT-IoT datasets. IEEE Access 2021;10:2269–83.

[16] Latif R, Asim M, Qadir J, Al-Fuqaha A. Dense random neural network for IoT intrusion detection. IEEE Access 2022;10: 22311−25.
[17] Wang X, Liu Z. Comprehensive study on imbalanced class issues in BoT-IoT. Journal of IoT Security 2020;14:208−23.
[18] Wang Y, Chen Z. Transfer learning with pre-trained models for UNSW-NB15. Journal of Cybersecurity Studies 2021;24: 89−105.
[19] Wu Q, Liu H. Ensemble methods on UNSW-NB15: a computational intensity perspective. International Journal of Cybersecurity 2020;21:134−50.
[20] Evans J, Zhao M. Impact of feature selection on imbalanced IoT datasets. International Journal of Cybersecurity 2020;21: 212−28.
[21] Liu X, Wang P. Reinforcement learning for dynamic imbalance in UNSW-NB15. Int J Netw Secur 2022;17:32−48.
[22] Larriva-Novo XaVVAaVBMaRDaSRM. An IoT-focused intrusion detection system approach based on preprocessing characterization for cybersecurity datasets. Sensors 2021;21:656.
[23] Popoola SI, Adebisi B, Ande R, Hammoudeh M, Anoh K, Atayero AA. smote-drnn: a deep learning algorithm for botnet detection in the internet-of-things networks. Sensors 2021;21:2985.
[24] Churcher A, Ullah R, Ahmad J, Masood F, Gogate M, Alqahtani F, et al. An experimental analysis of attack classification using machine learning in IoT networks. Sensors 2021;21:446.
[25] Yang Y, Zheng K, Wu C, Yang Y. Improving the classification effectiveness of intrusion detection by using improved conditional variational autoencoder and deep neural network. Sensors 2019;19:2528.
[26] Shafiq M, Tian Z, Bashir AK, Du X, Guizani M. CorrAUC: a malicious bot-IoT traffic detection method in IoT network using machine-learning techniques. IEEE Internet Things J 2020;8:3242−54.
[27] Guizani N, Ghafoor A. A network function virtualization system for detecting malware in large IoT based networks. IEEE J Sel Area Commun 2020;38:1218−28.
[28] Alkadi O, Moustafa N, Turnbull B, Choo KKR. A deep blockchain framework-enabled collaborative intrusion detection for protecting IoT and cloud networks. IEEE Internet Things J 2020;8:9463−72.
[29] Alkadi OaMNaTBaCKKR. A Deep blockchain framework-enabled collaborative intrusion detection for protecting IoT and cloud networks. IEEE Internet Things J 2021;8: 9463−72.
[30] Ibitoye OaSOaMA. Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks. Waikoloa, HI, USA. In: 2019 IEEE global communications conference (GLOBECOM); 2019.