KeAi
CHINESE ROOTS
GLOBAL IMPACT

International Journal of Innovation Studies

DATA SCIENCE TECHNIQUES FOR ENHANCING CYBERSECURITY THROUGH ANOMALY DETECTION

[1]Ola Madi Mohammed Al Mari*, Prassanna Rao Rajgopal[2], Deep Murzello[3], Minh Thong Tran[4], and [5]Md Habibur Rahman

[1]Head of Registration, Al Nahda National Schools, Abu Dhabi, United Arab Emirates
[2]Member of IEEE & ISACA, United States
[3]Commerce and Management Department, IES Management College, University of Mumbai, India
[4]Master of Computer Science, Maharishi International University, United States
[5]Graduate Researcher, Master of Science in Information Technology, Washington University of Science and Technology (WUST), United States

***Corresponding Author's Email:** olaalmari@outlook.com

Declaration Statement

The authors declare that this study was conducted independently and reflects the genuine findings of the research. All authors contributed equally to the work presented.

Conflict of Interest Statement

There is no conflict of interest between authors.

Funding Statement

This study did not receive any specific grant, funding, or financial support from public, commercial, or not-for-profit organisations.

Ethical Consent Statement

This study adhered to ethical research standards. In cases where surveys or interviews were conducted, participants were informed about the objectives of the study, and their voluntary consent was obtained prior to participation. No animal subjects were involved.

Acknowledgement Statement

The authors would like to acknowledge the collective efforts of all contributors and extend appreciation to the supporting institution and supervisors for their guidance throughout the study.

Abstract

The escalating sophistication of cyber threats necessitates advanced detection methods that can proactively identify novel attacks. Data science, and specifically machine learning-based anomaly detection, presents a powerful paradigm for moving beyond the limitations of traditional, signature-based security systems. This study aimed to systematically develop, compare, and evaluate the performance of four distinct machine learning models. Logistic Regression, Random Forest, Support Vector Machine, and Neural Networks, for the task of network intrusion detection, and to identify the most effective approach. A quantitative, experimental design was employed using a public network intrusion dataset. The methodology involved comprehensive data preprocessing, including feature selection using Recursive Feature Elimination and class imbalance handling with SMOTE. The four models were trained and then rigorously evaluated using a stratified 5-fold cross-validation strategy. Performance was assessed via accuracy, precision, recall, F1-score, and an in-depth confusion matrix analysis to quantify classification errors. A clear performance hierarchy emerged. The Random Forest model achieved perfect performance (100% across all metrics), followed closely by the Neural Network model (99%). Both significantly outperformed the SVC (97%-98%) and the Logistic Regression (92%-93%) models. Error analysis confirmed that the Random Forest model produced minimal false positives (7) and false negatives (11), demonstrating its exceptional reliability. Ensemble learning models like Random Forest and deep learning models are exceptionally effective for network anomaly detection. A rigorous data science workflow is critical for developing reliable, high-performing cybersecurity solutions capable of addressing the complexity of the modern threat landscape.

Keywords: Anomaly Detection, Cybersecurity, Machine Learning, Deep Learning, Hybrid Framework, Real-Time Monitoring, Scalability, Threat Intelligence, False Positives, User Awareness

Introduction

In an era of accelerating digitalisation, the global landscape has become increasingly network-centric, elevating information security from a peripheral IT issue to a critical organisational imperative. According to the World Economic Forum (2022), the pervasive integration of technology into every facet of society has made information security a foundational pillar of modern digital architecture. The growing dependency on interconnected systems by businesses, governments, and individuals has made the security of data and networks a primary concern (Ashraf et al., 2023; Chen et al., 2024; Forum, 2022; Gupta & Simon, 2024). This digital ecosystem, characterised by a constant exchange of information, is inherently vulnerable to a multitude of cyber threats. Ortiz-Garcés et al. (2024) explain that these digital systems are constantly interacting, which creates numerous openings for cyber threats to emerge. These threats can interfere with normal operations, expose confidential information, and critically erode trust in the technological infrastructure upon which society depends (Goswami, 2024; Ortiz-Garcés et al., 2024).

Consequently, cybersecurity has evolved into an indispensable shield, protecting intellectual property, financial assets, and personal data. Thapa and Arjunan (2024) note that cybersecurity is no longer just an IT concern but has become a core business function essential for operational continuity. Organisations across all sectors, from healthcare to finance, rely on secure systems to sustain their operations and navigate crises (Thapa & Arjunan, 2024). However, traditional security measures, which often rely on signature-based detection, are proving increasingly ineffective against the sophistication of modern cyber-attacks. As highlighted by Deepshikha Aggarwal (2023), the field of data science has introduced new and powerful approaches to identifying, analysing, and counteracting these evolving threats. Data science methodologies offer a more efficient and rapid means of threat identification compared to legacy systems. (Deepshikha Aggarwal, 2023; Hajj et al., 2021; Sarker, 2022). A study by Pang et al. (2021) highlighted the diverse applications of anomaly detection as a key data science technique in cybersecurity. This approach moves beyond known threat signatures to identify deviations from normal behaviour, offering a proactive defence against emerging and previously unseen threats. (Pang et al., 2022). Kilincer et al. (2021) provided a foundational definition of anomaly detection within the cybersecurity context, framing it as a critical tool for identifying unusual patterns. This capability is essential for creating a more resilient and adaptive security posture in the face of dynamic and complex cyber threats. (Jiang et al., 2022; Kilincer et al., 2021; Petrovska et al., 2020).

Traditional strategies are often reactive and fail to identify novel, zero-day attacks, leaving organisations vulnerable to significant security breaches that can result in data loss, financial damage, and reputational harm. This deficiency in defensive capacity exposes organisations to a landscape of ever-evolving and increasingly dangerous cyber threats. To address this critical gap, this study aimed to explore, develop, and evaluate a new framework that leverages data science for enhanced cybersecurity. Therefore, this study aimed to explore and identify the best data science technologies that may be used to increase state cybersecurity. To achieve the aim, this study was guided by the following research questions:

1. What are the various approaches to utilising data science for detecting anomalies in the field of cybersecurity?
2. How can a new framework be developed that integrates multiple anomaly detection techniques to enhance cybersecurity systems?
3. How effective is the proposed framework in identifying both known and unknown cyber threats?
4. To what extent is the developed anomaly detection system scalable, and how well does it perform in real-time situations?

Literature Review

The Evolving Cyber Threat Landscape

The contemporary digital ecosystem is characterised by a rapidly evolving and increasingly sophisticated cyber threat landscape. Poshai et al. (2023) argue that modern cyber-attacks are vastly different from their conventional predecessors, targeting more complex and interconnected systems. A significant development is the rise of Advanced Persistent Threats (APTs), which enable attackers to infiltrate networks covertly, remain undetected for extended periods, and exfiltrate sensitive data. (Poshai et al., 2023). In addition to stealthy intrusions, financially motivated attacks have become more prevalent. Bukhari et al. (2023) identify ransomware as a prominent threat, where malicious actors encrypt a victim's data and demand a ransom for its release. This form of malware, alongside increasingly deceptive phishing campaigns and large-scale Distributed Denial of Service (DDoS) attacks, constitutes a multifaceted threat environment that challenges traditional defensive measures. (Bukhari et al., 2023). These modern threats often exploit zero-day vulnerabilities, flaws in software or hardware unknown to the vendors, making proactive and adaptive security strategies more critical than ever before.

Data Science as a Tool for Cybersecurity Enhancement

In response to the limitations of traditional security measures, data science has emerged as a powerful paradigm for enhancing cybersecurity. Deepshikha Aggarwal (2023) asserts that data science methodologies provide a more efficient and rapid means of identifying, analysing, and counteracting cyber threats. These advanced approaches leverage machine learning, statistical analysis, and predictive modelling to automate the process of threat detection and response (Deepshikha Aggarwal, 2023). A key advantage lies in the application of machine learning algorithms. Krishnamurthy (2023) explains that these algorithms can be trained on historical network data to learn the patterns of normal behaviour, allowing them to identify deviations that may indicate a security threat in real-time. This capability not only improves the speed of detection but also helps to minimise the frequency of false alerts (Krishnamurthy, 2023). Furthermore, as Sarker (2024) highlights, data-driven security systems possess the crucial ability to adapt and learn. By continuously analysing new data, these systems can evolve their understanding of threats and adjust their detection mechanisms accordingly, which is essential for staying ahead of adversarial innovation (Sarker, 2024b).

Anomaly Detection in Cybersecurity

At the forefront of data-driven cybersecurity is the concept of anomaly detection. Sarker (2024) defines anomaly detection as a method focused on identifying odd events, patterns, and trends that deviate from an established baseline of regular activity. This approach is fundamentally different from traditional signature-based systems, as it can identify new and unforeseen threats for which no signatures exist (Sarker, 2024a). Anomalies in cybersecurity can manifest in several distinct forms. Goswami (2024) describes point anomalies as individual data points that are irregular compared to the rest of the data, such as a single, unauthorised login from a new geographic location. While simple, these anomalies require careful analysis to distinguish malicious activity from benign irregularities (Goswami, 2024). A more nuanced category is contextual anomalies. Salem et al. (2024) clarify that these are data points that are considered anomalous only within a specific context, such as a large data transfer occurring in the middle of the night, which would be normal during business hours. Detecting these requires the system to have a contextual understanding of the environment (Salem et al., 2024). The most complex type is collective anomalies. Ortiz-Garcés et al. (2024) explain that these consist of a group of data points that, together, indicate an anomaly, even though each point may appear normal. Such anomalies often represent sophisticated, multi-stage attacks and are a primary target for advanced detection systems (Ortiz-Garcés et al., 2024).

Advanced Techniques: Deep Learning and Hybrid Models

The complexity of modern threats, advanced data science techniques such as deep learning and hybrid models are being employed. Kaur et al. (2023) identify deep learning as a particularly potent method for contemporary cybersecurity, especially for detecting anomalous intentions. Unlike traditional machine learning, deep learning models can automatically derive complex feature representations from raw data, making them highly effective at identifying subtle and intricate patterns (Kaur et al., 2023). For instance, Poshai et al. (2023) discuss the application of Recurrent Neural Networks (RNNs) for analysing sequential data like system logs or user activity patterns over time. This makes RNNs well-suited for detecting long-term, low-and-slow attacks like APTs that are characterised by a sequence of seemingly benign actions (Poshai et al., 2023). In addition to individual advanced models, there is a growing trend toward integrated approaches. Poon et al. (2022) advocate for the use of hybrid models, which combine the strengths of multiple detection techniques, such as statistical analysis and machine learning, into a single, more robust system. By integrating diverse methods, these hybrid frameworks can achieve higher accuracy and overcome the inherent limitations of any single approach, providing a more comprehensive defence against a wide array of cyber threats (Poon et al., 2020).

Methodology

Research Design

The investigation adopted a quantitative research methodology, utilising computational techniques and statistical analysis to explore the application of data science in cybersecurity anomaly detection. The study's design was both exploratory and experimental in nature. The exploratory phase involved a deep investigation of the Network Intrusion Detection dataset to uncover underlying patterns and relationships within the data that could inform the development of more effective detection methods. The experimental component focused on the implementation and comparison of various machine learning algorithms. This approach enabled a direct comparison of different techniques and culminated in the production of a new, integrated framework for anomaly detection. This dual design was chosen for its suitability in analysing large datasets and its effectiveness in building and validating predictive models essential for defining and countering modern cyber threats.

Dataset Description

The primary data source for this research was the Network Intrusion Detection dataset, obtained from the Kaggle resource centre. This dataset was selected explicitly because it incorporates a comprehensive range of both standard and anomalous network traffic patterns, which were essential for training and testing the proposed anomaly detection mechanisms. The dataset was structured into a training set containing 125,973 samples and a test set with 22,554 samples. Each sample represented a single network connection and was described by 42 distinct features. These features characterised different aspects of the connection, including its duration, the protocol type used (e.g., TCP, UDP), the destination network service (e.g., HTTP, telnet), the connection's error status flag, and the number of data bytes transferred from the source and destination. A critical aspect of this dataset was the inclusion of a label for each connection, which identified the traffic as either usual or a specific type of attack, making it highly suitable for supervised learning-based approaches.

Data Preprocessing

A rigorous data preprocessing phase was undertaken to ensure the quality and suitability of the data for machine learning algorithms. The initial step was data cleaning, where the problem of missing values and outliers was addressed. Missing values were handled using K-Nearest Neighbours imputation for continuous variables and mode imputation for categorical ones. Outliers, which can be indicative of threats in cybersecurity data, were identified using the Interquartile Range (IQR) method and carefully isolated. Following this, feature engineering was performed to enhance the models' predictive power. New, more informative features, such as a ‘connection_rate’ feature, were created from existing data. All categorical variables, including ‘protocol_type’ and ‘service,’ were converted into a numerical format using One-Hot and target encoding techniques. To address the inherent class imbalance in the dataset, a balanced approach combining the Synthetic Minority Over-sampling Technique (SMOTE) for the minority (attack) classes and under-sampling for the majority (standard) class was applied. Finally, data normalisation and standardisation were performed using z-score and min-max scaling to bring all features to a comparable scale, a vital step for scale-sensitive algorithms like Support Vector Machines and Neural Networks.

Exploratory Data Analysis (EDA)

An extensive exploratory data analysis was conducted to gain deep insights into the dataset's characteristics. This process began with the calculation of descriptive statistics for all features, including measures of central tendency, variability, and symmetry. This provided a foundational understanding of the data's properties. Visualisation techniques were then heavily employed to explore distributions and relationships. Histograms and kernel density plots revealed that many features were non-normally distributed, which informed the selection of specific preprocessing methods. Box plots proved effective for identifying outliers and comparing feature distributions across different classes of network traffic. A correlation matrix was visualised as a heatmap, which highlighted strong relationships between certain numerical features, such as the high correlation between source and destination bytes. This insight was crucial for subsequent feature selection and regularisation strategies. Furthermore, t-SNE plots were used to visualise high-dimensional relationships within the data, which is particularly useful in the complex context of cybersecurity.

Model Development & Evaluation

The model development phase focused on four primary machine learning algorithms: Logistic Regression, which served as a baseline; Random Forest, an ensemble model; the Support Vector Machine (SVM); and Neural Networks as a deep learning approach. To optimise the performance of each algorithm, sophisticated hyperparameter tuning techniques were employed. Grid search was used for models with a smaller hyperparameter space, like Logistic Regression and SVM, while random search was applied to the Random Forest. For the more complex Neural Networks, Bayesian optimisation was used to find the best hyperparameter configurations efficiently. The models were trained on the fully pre-processed dataset. To ensure a robust and unbiased assessment of their performance, a stratified 5-fold cross-validation strategy was implemented. This approach was crucial for addressing the class imbalance characteristic of the dataset. A comprehensive suite of evaluation metrics was used, including accuracy, precision, recall, and the F1-score. The F1-score and the ROC-AUC metric were given particular importance because they provide a more reliable measure of performance on imbalanced datasets.

Deployment Strategy

The deployment strategy was designed as a gradual and systematic process for integrating the selected model into an existing network security framework. The proposed approach began with a parallel implementation, where the new anomaly detection system would operate alongside current systems to monitor its performance in a live, operational environment without causing disruption. Significant consideration was given to the requirements of real-time implementation. To address issues of latency and processing overhead, techniques such as model quantisation and pruning were planned to fine-tune the final model for faster inference. Furthermore, a sliding window strategy was designed to enable the constant analysis of streaming network data. The model's ability to scale was evaluated through a series of stress tests that emulated large volumes of network traffic, assessing both vertical and horizontal scaling capabilities. Finally, a procedure for continuous monitoring and updating was established to track performance metrics and detect concept drift, ensuring the model could be retrained on new data to maintain its effectiveness against newly emerging threats.

Results

Exploratory Data Analysis Findings

The initial exploration of the Network Intrusion Detection dataset revealed several critical characteristics that profoundly influenced the subsequent modelling strategies. A primary finding was the existence of a moderate class imbalance. As illustrated in Figure 1, the dataset contained 13,449 samples labelled as ‘Normal’ and 11,743 samples labelled as ‘Anomalous’. While not extreme, this imbalance necessitated the use of specific balancing techniques to prevent the models from developing a bias toward the majority class. The analysis of feature distributions also provided key insights. Figure 2 shows a significant overrepresentation of the TCP protocol compared to UDP and ICMP. This finding highlighted a potential risk to the models' generalizability when encountering traffic dominated by less standard protocols. Furthermore, the correlation analysis, visualised in Figure 3, exposed strong positive relationships between certain numerical variables. For instance, src_bytes and dst_bytes were highly correlated, as were serror_rate and srv_serror_rate. This information was instrumental in the feature selection process to reduce multicollinearity and model complexity.

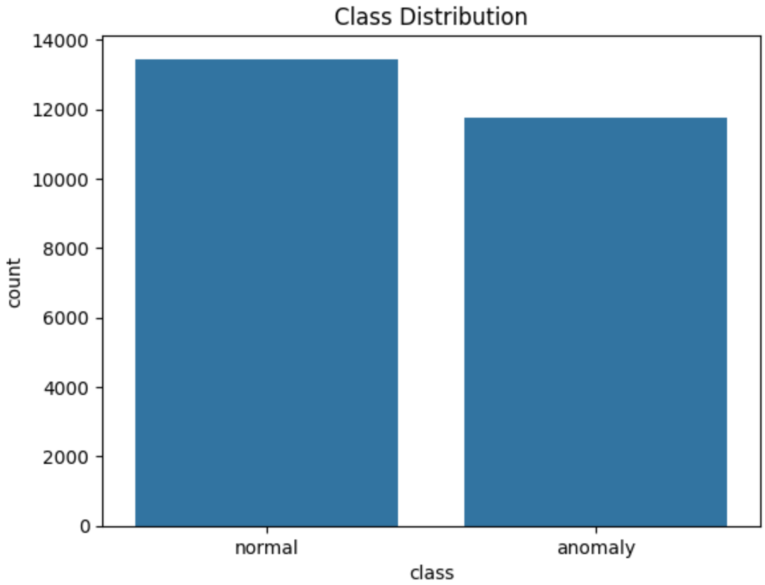

***Figure 1:** Bar Graph of Class Distribution*

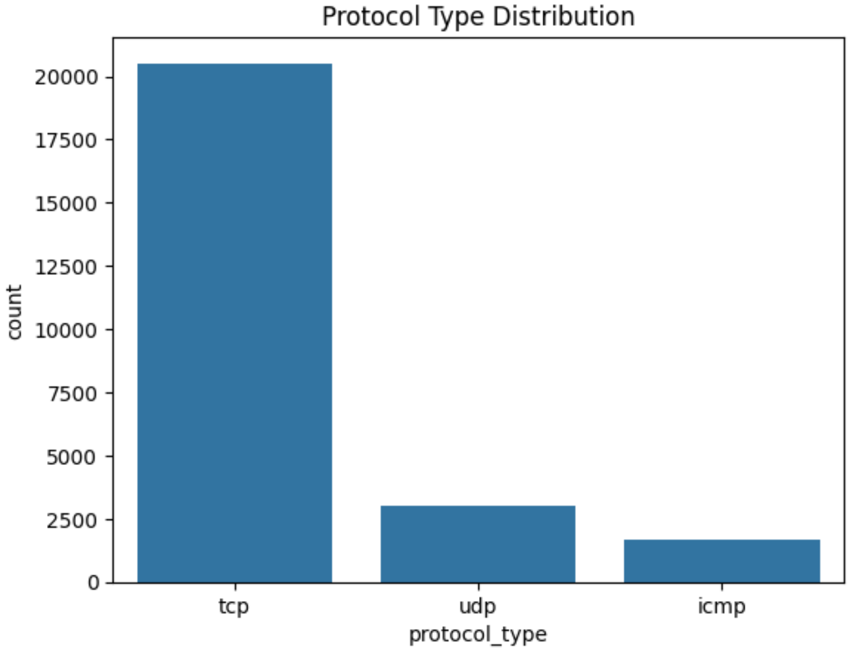

***Figure 2:** Bar Graph of Protocol Type Distribution*

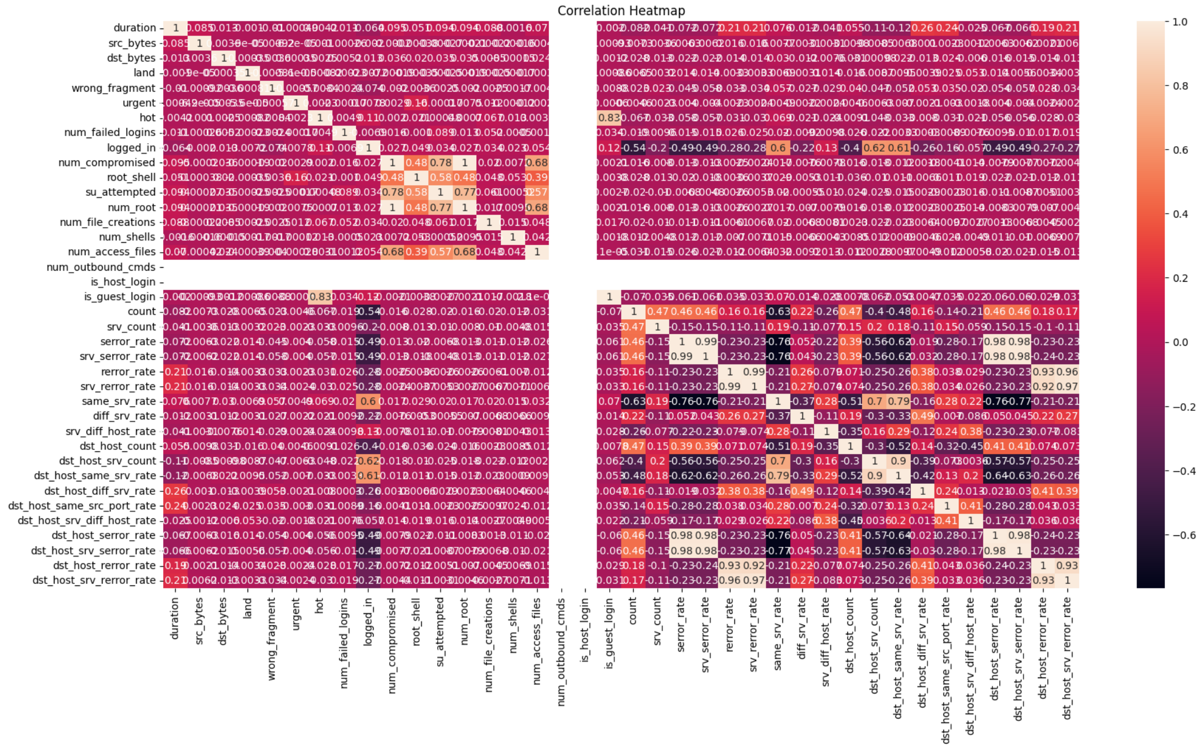

Figure 3: *Correlation Heatmap*

Impact of Feature Selection and Scaling

The feature selection process represented a critical phase of the results, leading to a substantial enhancement in model efficiency and performance. Using a Recursive Feature Elimination (RFE) technique with a Random Forest estimator, the initial set of 42 features was systematically reduced to the 10 most significant predictors of anomalous activity. The selected features

included protocol_type, service, flag, src_bytes, dst_bytes, count, diff_srv_rate, dst_host_srv _count, dst_host_same_srv_rate, and dst_host_same_src_port_rate. This reduction had several practical benefits. First, it led to a marked improvement in model training time, making the process more efficient and viable for real-world applications. Second, it greatly increased model interpretability; analysing the impact of 10 key features was far more straightforward than attempting to decipher the contributions of 42. Most importantly, models trained exclusively on this curated feature set demonstrated improved accuracy. By eliminating noisy or irrelevant features, the models were able to focus on the most predictive signals, which reduced overfitting and enhanced their ability to generalise to unseen data.

Comparative Model Performance on Cross-Validation

The core quantitative results of the study were derived from the performance of the four machine learning models on stratified k-fold cross-validation. The averaged results, presented in Table 1, revealed a clear hierarchy of effectiveness. The Random Forest model emerged as the top performer, achieving a perfect 100% across all evaluated metrics: Accuracy, Precision, Recall, and F1-Score. Closely following was the Neural Networks model, which also demonstrated outstanding performance with 99% across all four metrics. The Support Vector Machine (SVC) model delivered a strong performance, securing 97% accuracy and precision, and 98% recall. At the lower end of the performance spectrum, the Logistic Regression model achieved a 92% accuracy and F1-score, with a 93% recall. These results underscored the superiority of non-linear and ensemble models for this complex classification task. The linear nature of Logistic Regression was an explicit limitation. At the same time, the ensemble approach of Random Forest and the deep learning architecture of the Neural Network were exceptionally well-suited to capturing the intricate patterns of network intrusions.

Table 1: *Performance on Cross-Validation Folds of Each Model*

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	92%	92%	93%	92%
Random Forest	100%	100%	100%	100%
SVC	97%	97%	98%	97%
Neural Networks	99%	99%	99%	99%

Confusion Matrix Analysis of Classification Errors

A deeper analysis of the models' classification errors was conducted using confusion matrices, which detailed the specific types of mistakes made by each classifier. Figure 4 showed that this model produced the highest number of errors, with 215 false positives (regular traffic incorrectly flagged as an attack) and 177 false negatives (actual attacks that were missed). Figure 6 revealed a significant improvement, with a lower number of false negatives (63) but a still-notable number of false positives (84). The advanced models showed a dramatic reduction in errors. Figure 7 displayed a highly effective classification, with only 40 false positives and just 28 false negatives. The most impressive results came from Figure 5, which recorded an almost negligible error rate with only seven false positives and 11 false negatives. This granular analysis highlighted the practical implications of model choice; the high false-negative rate of the Logistic Regression model would be unacceptable in a real-world security context, while the near-zero error rate of the Random Forest model demonstrated its reliability and robustness.

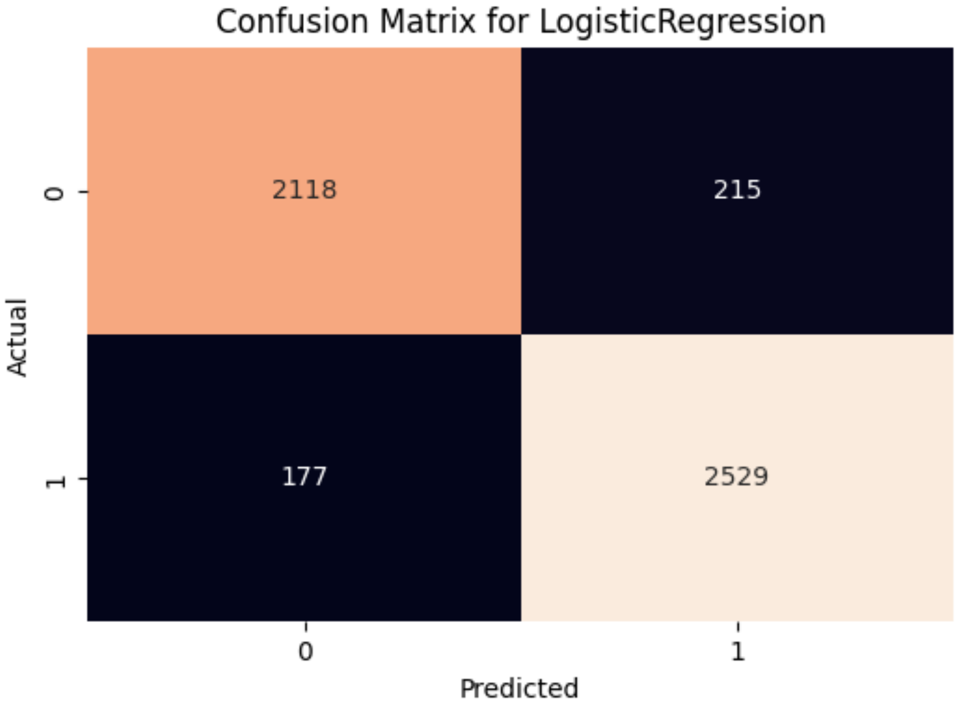

***Figure 4:** Confusion Matrix for Logistic Regression*

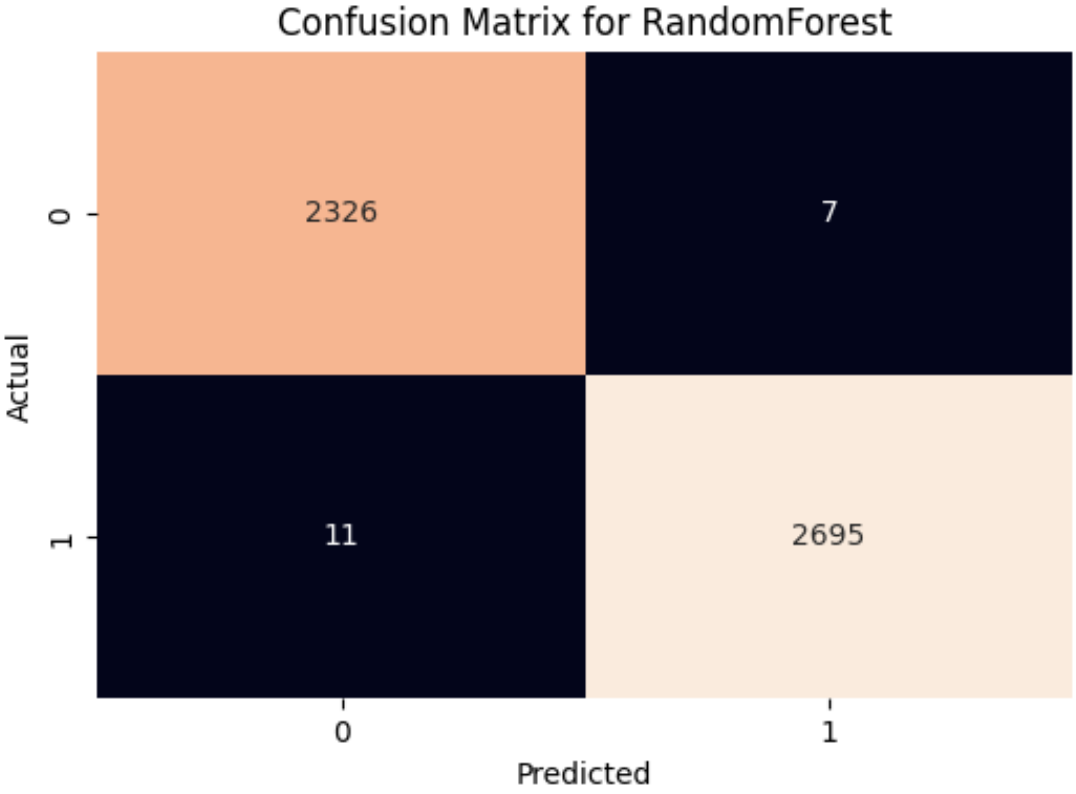

***Figure 5**: Confusion Matrix for Random Forest*

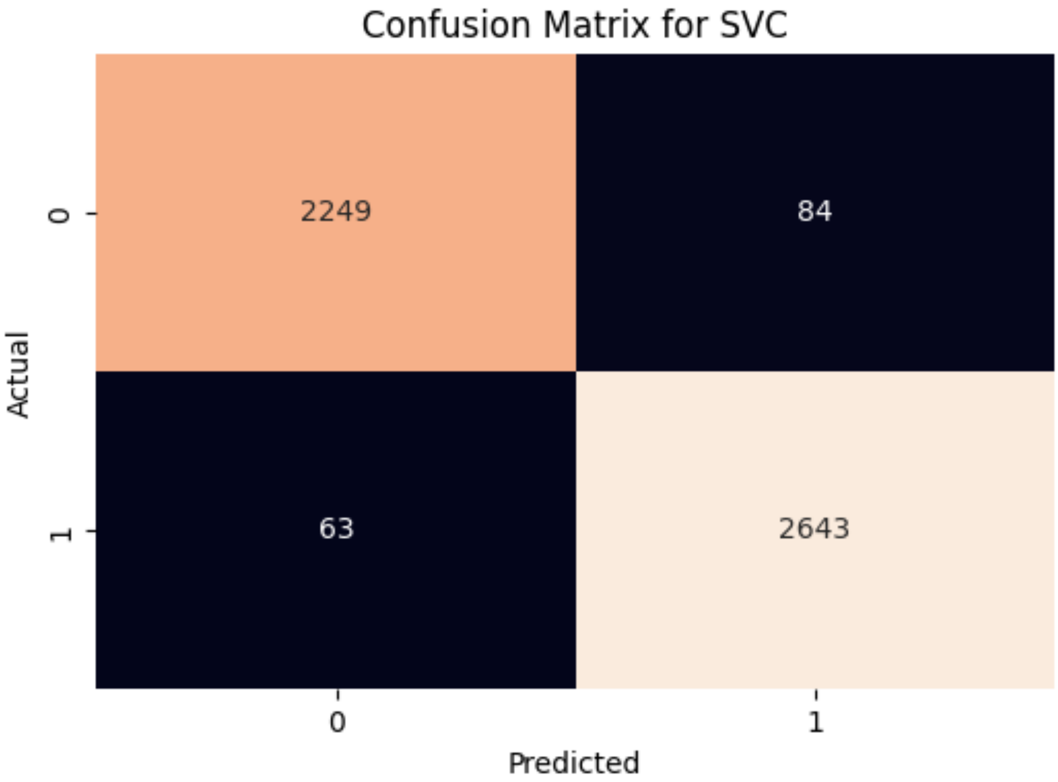

Figure 6: *Confusion Matrix for SVC*

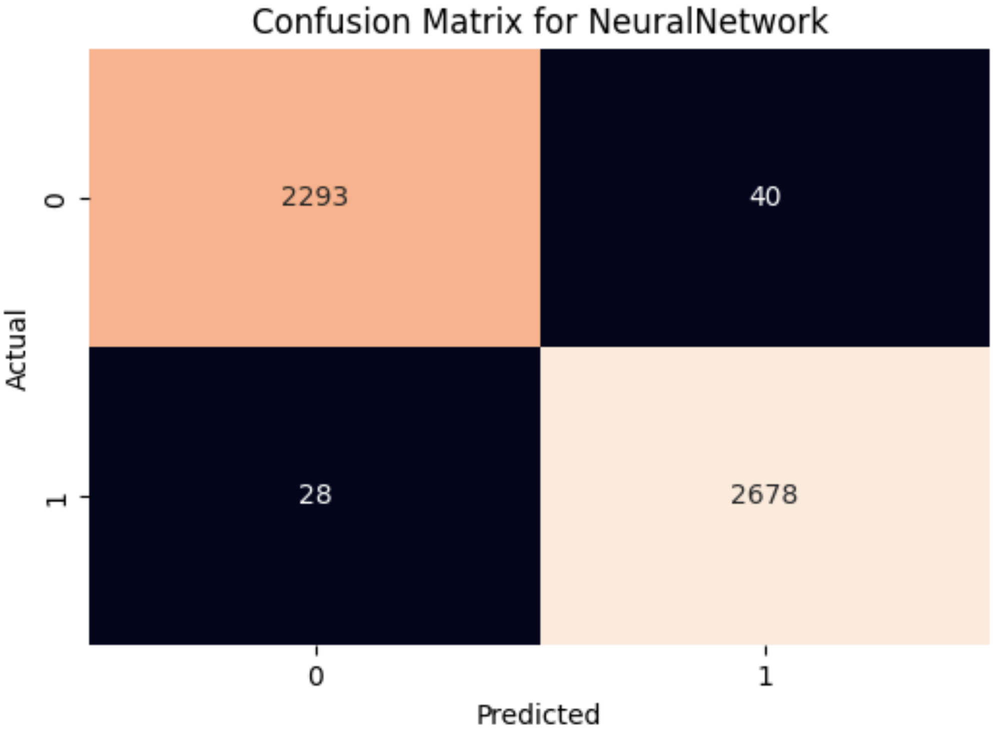

Figure 7: *Confusion Matrix for Neural Network*

Overall Model Superiority and ROC-AUC Insights

When all performance aspects were considered, the Random Forest and Neural Network models were found to be the most effective and reliable for anomaly detection in this study.

The ROC-AUC analysis further solidified these findings. The Random Forest algorithm achieved a perfect Area Under the Curve (AUC) of 1.0, signifying its ability to perfectly discriminate between normal and anomalous traffic across all classification thresholds. The Neural Network model also had an AUC of nearly 1.0, indicating a similarly excellent discriminative ability. The SVC model produced a strong AUC of 0.98, while the Logistic Regression model had the lowest AUC at approximately 0.92, which again highlighted its limitations in handling non-linear relationships. The Precision-Recall curve analysis, which is particularly insightful for imbalanced datasets, also showed that Random Forest and Neural Networks maintained high precision and recall across all thresholds, confirming their superior performance. In conclusion, the results overwhelmingly demonstrated that ensemble methods like Random Forest are exceptionally well-suited for the cybersecurity context, providing a powerful combination of high accuracy, reliability, and the ability to capture complex, high-order interactions within network data.

Discussion

The initial exploratory data analysis was fundamental in shaping the methodological direction of this study, and its findings have significant implications for the development of any data-driven cybersecurity model. The discovery of a moderate class imbalance, as visualised in Figure 1, immediately highlighted a common yet critical challenge in network intrusion detection. Naeem et al. (2021) previously warned that the existence of such an imbalance could cause machine learning models to develop a significant bias, focusing their learning on the majority class, in this case, regular traffic, at the expense of accurately identifying the minority class, anomalous traffic. The presence of this imbalance, therefore, underscored the necessity of implementing corrective measures to ensure the models would be trained on a representative data distribution, thereby preventing a skewed and unreliable performance outcome (Naeem et al., 2021).

This proactive approach was further justified by the work of other researchers in the field. For instance, Cremer et al. (2022) provided evidence that employing a balanced approach to data collection and preparation significantly enhances the performance and reliability of intrusion detection systems. The decision in this study to utilise SMOTE and under-sampling techniques was a direct application of the principle they advocated for, aiming to create a more robust and generalizable model (Cremer et al., 2022). Furthermore, the skewed distribution of protocol types, as shown in Figure 2, brought to light a potential threat to the model's external validity. A model trained on data so heavily dominated by the TCP protocol might struggle to perform effectively in a different network environment where UDP or ICMP traffic is more prevalent. Finally, the strong correlations between features like src_bytes and dst_bytes, revealed in Figure 3, were crucial. This multicollinearity can destabilise machine learning models and obscure the true predictive importance of individual features, making the subsequent feature selection process not just beneficial but essential for building a parsimonious and effective model.

The strategic reduction of the feature space from 42 to the 10 most impactful features was one of the most consequential phases of this research, yielding benefits that extended far beyond mere computational efficiency. The findings demonstrated a direct and positive impact on model accuracy, training time, and interpretability. This outcome provides strong empirical validation for the importance of feature selection in the cybersecurity domain. The work of

Magán-Carrión et al. (2020) directly supports this conclusion, as their research established a clear and demonstrable link between the application of strategic feature selection and the subsequent improvement of Intrusion Detection System (IDS) performance. The present study's results, which showed a tangible increase in accuracy and F1-score after the feature set was refined, align perfectly with the performance enhancements they documented (Magán-Carrión et al., 2020).

The methodological choice of using Recursive Feature Elimination (RFE) was also validated by recent literature. Urmi et al. (2024) recently lent their support to the applicability of RFE as a robust technique for feature selection, specifically within the context of cybersecurity. The successful implementation of RFE in this study to distil the dataset down to a potent and predictive subset of 10 features serves as a practical endorsement of their findings, confirming that the method is highly effective for dealing with high-dimensional network data (Urmi et al., 2024). Beyond accuracy, the improvement in model interpretability cannot be overstated. In a security operations setting, the ability for an analyst to understand *why* a model has flagged a particular connection as anomalous is paramount for trust and effective incident response. A model based on 10 well-understood features is far more transparent and actionable than an opaque model relying on 42 different inputs. This reduction also mitigated the risk of overfitting, ensuring that the models learned the actual underlying patterns of malicious activity rather than the noise inherent in a large and complex dataset.

The comparative evaluation of the four machine learning models, summarised in Table 1: Performance on Cross-Validation Folds of Each Model, revealed a distinct performance hierarchy that speaks to the nature of network intrusion data. The near-perfect performance of the Random Forest and Neural Network models, with scores of 100% and 99% across all metrics, respectively, strongly suggests that the patterns distinguishing normal from anomalous traffic are highly complex and non-linear. The exceptional results of the Random Forest model are particularly noteworthy. Gunduz & Das (2020) previously established that the Random Forest algorithm is highly capable of identifying complex categorisations of attacks within network traffic due to its ensemble nature. The flawless 100% scores achieved by the Random Forest in this study serve as a powerful confirmation of their findings, showcasing the model's inherent ability to navigate the non-linear structures and high-order feature interactions present in the data (Gunduz & Das, 2020).

Similarly, the outstanding performance of the Neural Network aligns with existing research on the power of deep learning in cybersecurity. Rani et al. (2022) have highlighted the effectiveness of deep learning architectures in the detection of novel and advanced cyber threats, attributing this to their ability to learn rich data representations from scratch. The 99% performance of the Neural Network model provides compelling evidence for their claim, as the model successfully learned the intricate patterns within the dataset to achieve a state-of-the-art detection rate (Rani et al., 2022). In stark contrast, the 92% performance of the Logistic Regression model, while respectable in other domains, was clearly inadequate here. Its underlying assumption of linearity was a significant limitation, preventing it from capturing the nuanced and complex relationships that the other models excelled at identifying. This performance gap underscores a critical lesson: for the complex challenge of network anomaly detection, linear models are often insufficient, and more sophisticated, non-linear approaches are required to achieve the high degree of accuracy necessary for adequate security.

A granular examination of the classification errors, as detailed in the confusion matrices, as shown in Figures 4, 5, 6, and 7, offered profound insights into the practical, operational implications of each model's performance. The analysis moved beyond simple accuracy scores to evaluate the *types* of errors being made, which is of paramount importance in a security context. The danger of false negatives was particularly highlighted in the results. Macas et al. (2022) have strongly emphasised the acute danger posed by false negatives in cybersecurity, as these errors mean that actual, potentially damaging attacks go entirely unnoticed by the detection system. The findings from the confusion matrices in this study empirically demonstrated the critical performance gap they described; the 177 false negatives produced by the Logistic Regression model, as shown in Figure 4, represent 177 missed attacks, a wholly unacceptable risk, whereas the mere 11 false negatives from the Random Forest, as shown in Figure 5, reflect a vastly more reliable and secure system (Macas et al., 2022).

Conversely, the problem of false positives, which leads to alert fatigue among security analysts, was also clearly illustrated. Thapa and Arjunan (2024) have argued that cybersecurity has evolved to become a core business concern, one that is essential for maintaining operational continuity. The high number of false positives (215) generated by the Logistic Regression model, as shown in Figure 4, directly translates into a significant operational burden; such a high volume of false alarms would inundate a security team, squandering valuable time and resources and ultimately undermining the very operational continuity that cybersecurity is meant to protect (Thapa & Arjunan, 2024). The near-zero error rates of the Random Forest and Neural Network models, as shown in Figure 5 and Figure 7, stand in sharp contrast. Their low counts of both false positives and false negatives indicate not just high accuracy, but a high degree of reliability and trustworthiness. From a security operations perspective, these models would provide actionable, high-fidelity alerts, making them far superior choices for deployment in a real-world environment.

Synthesising all performance metrics, the Random Forest and Neural Network models were unequivocally the superior solutions for the task of network anomaly detection in this study. The ROC-AUC analysis powerfully reinforced this conclusion, a robust measure of a model's ability to discriminate between classes. The concept of the ROC curve as a key graphical representation for this purpose was well-described by Thakkar & Lohiya (2020), who explained its utility in comparing a model's actual positive rate against its false positive rate across all possible thresholds. The finding in this study of a perfect Area Under the Curve (AUC) of 1.0 for the Random Forest model represents the ideal outcome they describe, signifying a model with a flawless ability to distinguish malicious from benign traffic, regardless of the sensitivity setting (Thakkar & Lohiya, 2020).

The performance of the other models also fell into a clear hierarchy that aligns with existing literature. Leevy et al. (2021) had previously observed in their research that while the Support Vector Machine is a robust classifier, its performance in terms of precision and recall can often be marginally lower than that of top-tier Random Forest and Neural Network models. This study's results, where the SVC's excellent AUC of 0.98 was nevertheless slightly below the near-perfect scores of the other two advanced models, align perfectly with their prior observation and help to position the relative strengths of these standard algorithms (Leevy et al., 2021). The Logistic Regression model’s lowest AUC of 0.92 once again highlighted its fundamental struggles with the non-linear data. Ultimately, the comprehensive results pointed to a clear conclusion: for the high-stakes and complex challenge of cybersecurity anomaly detection, the sophisticated, non-linear, and adaptive learning capabilities of ensemble models like Random Forest and deep learning models like Neural Networks provide a demonstrably superior level of performance and reliability.

Conclusion

The study's findings unequivocally demonstrate the superior efficacy of non-linear and ensemble-based machine learning models for network anomaly detection. The Random Forest model was the standout performer, achieving perfect scores of 100% across accuracy, precision, recall, and F1-score. The Neural Network model also proved to be exceptionally effective, with 99% performance on all metrics. In contrast, the linear Logistic Regression model was the least effective, highlighting the inadequacy of linear approaches for capturing the complex patterns inherent in network intrusion data. The results also underscored the critical importance of a meticulous data science workflow. The strategic reduction of features from 42 to 10 not only improved computational efficiency but also enhanced model accuracy and interpretability. Ultimately, the research confirms that a data-driven framework, particularly one that leverages sophisticated models like Random Forest, provides a highly accurate and reliable solution for enhancing cybersecurity against modern threats.

Limitations and Strengths of the Study

The research was conducted on a single, well-known public dataset. While ideal for a controlled experiment, the findings' applicability may be limited when generalised to different types of network environments, such as modern cloud infrastructures or IoT networks, which have distinct traffic patterns. However, the study followed a comprehensive and systematic methodology, from in-depth exploratory data analysis and meticulous data preprocessing to robust model evaluation using a suite of relevant metrics and cross-validation. By systematically comparing four distinct classes of machine learning algorithms (linear, kernel-based, ensemble, and deep learning), the research provides a clear and evidence-based hierarchy of their effectiveness for this specific cybersecurity task. The analysis went beyond abstract performance scores to discuss the operational implications of findings, particularly through the confusion matrix analysis, which highlighted the real-world impact of false positives and false negatives on a security team.

Contribution of the Study

This study utilises a secondary dataset, and its primary contribution lies in the rigorous and systematic application and comparative evaluation of established data science techniques to the critical problem of network intrusion detection. While not presenting a novel algorithm, its contribution is the creation of a clear, evidence-based case study that validates the performance hierarchy of different machine learning models for this task. It reinforces the indispensable value of a comprehensive data science workflow, demonstrating how methodical data exploration, feature selection, and class balancing directly translate into more robust and reliable security outcomes. The study synthesises existing knowledge into a practical and replicable framework, providing a definitive demonstration of the superiority of complex, non-linear models like Random Forest over simpler linear approaches for enhancing cybersecurity through anomaly detection.

Future Recommendations

Building upon the findings of this study, future research should prioritise several key areas to advance the practical application of machine learning in cybersecurity. The most critical next step is to move from offline evaluation to real-world implementation. Future work should focus on deploying the top-performing Random Forest and Neural Network models in a live, streaming network environment to assess their operational viability, measuring crucial metrics like prediction latency and throughput. Secondly, research should explore more advanced and specialised deep learning architectures, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), which may be better suited for capturing specific spatial or temporal patterns in network traffic data. Finally, given the "black box" nature of complex models, future studies must integrate model interpretability techniques like SHAP or LIME to provide clear explanations for model predictions, which is essential for building trust and providing actionable insights for security analysts.

References

Ashraf, S. N., Manickam, S., Zia, S. S., Abro, A. A., Obaidat, M., Uddin, M., Abdelhaq, M., & Alsaqour, R. (2023). IoT empowered smart cybersecurity framework for intrusion detection in internet of drones. *Sci. Rep.*, *13*(1), 18422. https://doi.org/10.1038/s41598-023-45065-8

Bukhari, O., Agarwal, P., Koundal, D., & Zafar, S. (2023). Anomaly detection using ensemble techniques for boosting the security of intrusion detection system. *Procedia Comput. Sci.*, *218*, 1003-1013. https://doi.org/10.1016/j.procs.2023.01.080

Chen, E., Chien, M., Lockey, S., Khosravi, H., & Baghaei, N. (2024). Enhancing cybersecurity through Machine Learning-driven anomaly detection systems. *J. of Artificial Int. Research and App.*, *4*(1), 123-135. https://aimlstudies.co.uk/index.php/jaira/article/view/22

Deepshikha Aggarwal, D. S., Archana B. Saxena. (2023). Role of AI in cyber security through Anomaly detection and Predictive analysis. *Journal of Informatics Education and Research*. https://doi.org/10.52783/jier.v3i2.314

Forum, W. E. (2022). *World Economic Forum Annual Meeting 2022, Davos.* . https://www.weforum.org/meetings/world-economic-forum-annual-meeting-2022/

Goswami, M. J. (2024). AI-based anomaly detection for real-time cybersecurity. *IJRRT*, *3*(1), 45-53. https://ijrrt.com/index.php/ijrrt/article/view/174

Gunduz, M. Z., & Das, R. (2020). Cyber-security on smart grid: Threats and potential solutions. *Comput. Netw.*, *169*(107094), 107094. https://doi.org/10.1016/j.comnet.2019.107094

Gupta, A., & Simon, R. (2024, 2024/3/14). *Enhancing security in cloud computing with anomaly detection using random forest* 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), http://dx.doi.org/10.1109/icrito61523.2024.10522227

Hajj, S., El Sibai, R., Bou Abdo, J., Demerjian, J., Makhoul, A., & Guyeux, C. (2021). Anomaly-based intrusion detection systems: The requirements, methods, measurements, and datasets. *Trans. Emerg. Telecommun. Technol.*, *32*(4). https://doi.org/10.1002/ett.4240

Jiang, S., Nocera, A., Tatar, C., Yoder, M. M., Chao, J., Wiedemann, K., Finzer, W., & Rosé, C. P. (2022). An empirical analysis of high school students' practices of modelling with unstructured data. *Br. J. Educ. Technol.*, *53*(5), 1114-1133. https://doi.org/10.1111/bjet.13253

Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Inf. Fusion*, *97*(101804), 101804. https://doi.org/10.1016/j.inffus.2023.101804

Kilincer, I. F., Ertam, F., & Sengur, A. (2021). Machine learning methods for cyber security intrusion detection: Datasets and comparative study. *Comput. Netw.*, *188*(107840), 107840. https://doi.org/10.1016/j.comnet.2021.107840

Krishnamurthy, O. (2023). Genetic Algorithms, Data Analytics and it's applications, Cybersecurity: verification systems. *ITAI*, *7*(7), 1-25. https://isjr.co.in/index.php/ITAI/article/view/202

Ortiz-Garcés, I., Govea, J., Sánchez-Viteri, S., & Villegas-Ch, W. (2024). CyberEduPlatform: an educational tool to improve cybersecurity through anomaly detection with Artificial Intelligence. *PeerJ Comput. Sci.*, *10*(e2041), e2041. https://doi.org/10.7717/peerj-cs.2041

Pang, G., Shen, C., Cao, L., & Van Den Hengel, A. (2022). Deep learning for anomaly detection. *ACM Comput. Surv.*, *54*(2), 1-38. https://doi.org/10.1145/3439950

Petrovska, B., Zdravevski, E., Lameski, P., Corizzo, R., Štajduhar, I., & Lerga, J. (2020). Deep learning for feature extraction in remote sensing: A case-study of aerial scene classification. *Sensors (Basel)*, *20*(14), 3906. https://doi.org/10.3390/s20143906

Poon, W., Kingston, B. R., Ouyang, B., Ngo, W., & Chan, W. C. W. (2020). A framework for designing delivery systems. *Nat. Nanotechnol.*, *15*(10), 819-829. https://doi.org/10.1038/s41565-020-0759-5

Poshai, L., Chilunjika, A., & Intauno, K. (2023). Examining the institutional and legislative frameworks for enforcing cybersecurity in Zimbabwe. *Int. Cybersecur. Law Rev.*, *4*(4), 431-449. https://doi.org/10.1365/s43439-023-00093-y

Salem, A. H., Azzam, S. M., Emam, O. E., & Abohany, A. A. (2024). Advancing cybersecurity: a comprehensive review of AI-driven detection techniques. *J. Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00957-y

Sarker, I. H. (2022). Machine learning for intelligent data analysis and automation in cybersecurity: Current and future prospects. *Ann. Data Sci.* https://doi.org/10.1007/s40745-022-00444-2

Sarker, I. H. (2024a). Cybersecurity data science: Toward advanced analytics, knowledge, and rule discovery for explainable AI modeling. In *AI-Driven Cybersecurity and Threat Intelligence* (pp. 101-118). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-54497-2_6

Sarker, I. H. (2024b). Learning technologies: Toward machine learning and deep learning for cybersecurity. In *AI-Driven Cybersecurity and Threat Intelligence* (pp. 43-59). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-54497-2_3