Application of Machine Learning Tools and Techniques in Cyber Threat Intelligence

Sarkis Chichkoyan
190425490
Supervisor: Piyajith Wijetunge
MSc Computer Science

Abstract**—In today's rapidly evolving cybersecurity landscape, the rise of sophisticated cyber threats necessitates advanced detection and prevention methodologies. Traditional measures, relying on predefined rules and signatures, are inadequate against zero-day exploits and advanced persistent threats (APTs). This study, part of a Master's project at Queen Mary University of London, explores the integration of machine learning (ML) techniques to enhance Intrusion Detection Systems (IDS) using the HIKARI-2021 dataset. Supervised learning techniques, including Logistic Regression, Random Forest, Gradient Boosting Machine (GBM), and XGBoost, were employed to classify network activities as benign or malicious. Key objectives included enhancing IDS performance, addressing integration challenges, and identifying critical features for threat detection. Data preprocessing involved removing irrelevant columns, frequency encoding for categorical features, and standardisation for feature scaling. The models were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC, with cross-validation ensuring robustness. Results indicate that XGBoost outperforms other models, achieving the highest accuracy, precision, F1-score, and ROC-AUC. The study demonstrates the significant enhancement of IDS capabilities through ML, providing a robust and adaptive security framework.**

Keywords: Cybersecurity, Machine Learning, Intrusion Detection Systems, HIKARI-2021, XGBoost, Data Preprocessing, Threat Detection.

I. INTRODUCTION

In the ever-evolving landscape of cybersecurity, the proliferation of sophisticated cyber threats has underscored the necessity for advanced detection and prevention methodologies. Traditional cybersecurity measures, often reliant on predefined rules and signatures, are increasingly inadequate against zero-day exploits and advanced persistent threats (APTs) that evade conventional detection systems. Intrusion Detection Systems (IDS) have long been a cornerstone of network security, yet their effectiveness is controlled by their dependency on static signatures and inability to adapt to emerging threats. This has paved the way for the integration of machine learning (ML) techniques, which offer dynamic, data-driven approaches to threat detection.

Machine learning in cybersecurity leverages algorithms capable of learning from historical data to identify patterns and anomalies indicative of malicious activity. The application of ML in IDS aims to enhance the detection of novel threats, providing a more robust and adaptive security framework. The HIKARI-2021 dataset, a comprehensive collection of real and synthetic network traffic data, serves as a pivotal resource in this endeavour. By analysing this dataset, researchers can develop and refine machine learning models that significantly improve the accuracy and efficiency of threat detection.

This study, conducted as part of a Master's project at Queen Mary University of London, focuses on the development and integration of ML models to advance cyber threat intelligence. Supervised learning techniques, such as Logistic Regression, Random Forest, Gradient Boosting Machine (GBM), and XGBoost, are employed to classify network activities as benign or malicious. Previous research by Fernandes and Lopes (2022) demonstrated the efficiency of SVMs in reducing feature size while maintaining high accuracy using the HIKARI-2021 dataset. Similarly, Rajak et al. (2022) highlighted the potential of CNN-LSTM models in detecting various attack types within IoT environments, achieving notable accuracy and reliability.

The project's objectives include not only the enhancement of IDS through machine learning but also addressing practical challenges in integrating these models into existing cybersecurity infrastructures. This involves identifying key features for effective threat detection, evaluating model performance using metrics such as accuracy, precision, recall, and F1 score, and ensuring seamless integration into real-time security operations. Ethical considerations, such as data privacy and the implications of deploying AI in cybersecurity, are also meticulously examined.

Furthermore, this project places significant emphasis on the practical implementation of machine learning models using Python and Jupyter Notebook. Using these tools, the study aims to create an accessible and scalable solution for fraud detection within the HIKARI-2021 dataset. The development process involves detailed data preprocessing, model training, and performance evaluation, ensuring that the proposed models are both robust and efficient.

In addition to technical development, this study engages with industry professionals through surveys and interviews to validate the practical relevance of the proposed solutions. Feedback from cybersecurity professionals, data scientists, and machine learning experts will inform the refinement of the models and their applicability in real-world scenarios. By aligning research outcomes with industry needs, this project aims to contribute significantly to the field of cybersecurity, offering advanced tools and strategies to safeguard against the evolving landscape of cyber threats.

This dissertation lays the groundwork for a comprehensive, machine learning-enhanced IDS framework that can be adapted by small to medium-sized enterprises (SMEs) and larger organisations alike. Through the rigorous analysis of the HIKARI-2021 dataset and the development of robust ML models, the study aspires to set a new standard in cyber threat detection and response. The final aim being that cybersecurity measures are as dynamic and adaptive as the threats they aim to counter.

II. LITERATURE REVIEW AND ANALYSIS

The rapid evolution of cyber threats requires advanced methodologies for threat detection and prevention. Traditional cybersecurity measures often fall short in identifying sophisticated attacks, highlighting the need for integrating machine learning (ML) into cyber threat intelligence. Intrusion Detection Systems (IDS) are critical components of cybersecurity infrastructure designed to detect and respond to malicious activities. Over the years, IDS have evolved from simple rule-based systems to complex models leveraging artificial intelligence (AI) and machine learning. Traditional IDS rely heavily on predefined rules and signatures to identify known threats, which are effective against known attacks but struggle with zero-day exploits and advanced persistent threats (APTs) that do not match existing signatures. Modern IDS utilise machine learning to analyse patterns and detect anomalies in network traffic, allowing for the identification of new threats by learning from historical data and continuously improving their detection capabilities.

Machine learning techniques in IDS can be broadly categorised into supervised, unsupervised, and hybrid approaches. Supervised learning techniques involve training models on labeled datasets to classify network activities as benign or malicious. Random Forest and XGBoost are prominent methods used in IDS for their effectiveness and robustness. Random Forest, an ensemble learning method, constructs multiple decision trees and merges them to obtain a more accurate and stable prediction. This technique is highly effective in handling large datasets and dealing with overfitting. Fernandes and Lopes (2022) demonstrated the effectiveness of Random Forest using the HIKARI-2021 dataset, achieving high accuracy by leveraging the model’s ability to handle a large number of input features. XGBoost, or Extreme Gradient Boosting, is another powerful ML algorithm used in IDS. XGBoost enhances the performance of gradient boosting models by implementing a more regularised model formalisation to control overfitting. The efficiency and scalability of XGBoost have made it a popular choice for large-scale machine learning applications. Arfaoui et al. (2023) explored the application of XGBoost for smart home threat detection using the HIKARI-2021 dataset, demonstrating the model's versatility and high performance in identifying various cyber threats.

Unsupervised learning techniques, such as clustering and anomaly detection, identify unusual patterns in network traffic without prior knowledge of what constitutes an attack. K-Means clustering is often used for anomaly detection by grouping similar data points and identifying outliers. For example, the study by Ferriyan et al. (2021) on generating network intrusion detection datasets based on real and encrypted synthetic attack traffic highlighted the effectiveness of clustering techniques in identifying anomalies within vast datasets. However, unsupervised methods can suffer from high false positive rates and may struggle to differentiate between benign anomalies and actual threats. Another approach is the use of auto-encoders, neural network models trained to reconstruct input data and detect anomalies by measuring reconstruction errors. Khan et al. (2019) demonstrated the use of auto-encoders in improving the detection rates of IDS by identifying deviations from normal traffic patterns. Auto-encoders are powerful for anomaly detection but require careful tuning of the reconstruction threshold to balance between false positives and missed detections. These methods underscore the versatility and power of unsupervised learning in scenarios where labeled data is scarce or unavailable, but they also require careful tuning and validation to ensure efficacy and reliability.

Hybrid approaches combine supervised and unsupervised learning to leverage the strengths of both methods, aiming to improve the robustness and accuracy of IDS. A study by Sreelatha et al. (2022) demonstrated the combination of anomaly detection with supervised classification, using extended equilibrium deep transfer learning-based intrusion detection to enhance cloud security. By integrating these techniques, the study achieved lower false positive rates and higher detection accuracy. The integration of multiple techniques can provide a more comprehensive defence mechanism but also increases the complexity and potential for integration issues. These hybrid models demonstrate the potential for creating more resilient and adaptive IDS by incorporating the complementary strengths of various ML techniques.

The HIKARI-2021 dataset is a comprehensive dataset created to address the limitations of existing IDS datasets by including both real and encrypted synthetic attack traffic, providing a robust foundation for training and evaluating ML models. This dataset combines real network traffic with synthetic attacks, ensuring a diverse set of scenarios for model training. Studies by Fernandes and Lopes and Rajak have utilised the HIKARI-2021 dataset to validate their models, demonstrating its suitability for network intrusion detection research. The deployment of ML models in cybersecurity raises ethical and privacy concerns, making it crucial to ensure these models do not inadvertently compromise user privacy or operate without appropriate oversight. Future research should focus on enhancing the capability of IDS to process and analyse data in real-time, developing models that generalise well across different network environments and attack types, and exploring seamless integration of ML models into existing cybersecurity frameworks to enhance real-time threat detection and response. The integration of machine learning into intrusion detection systems represents a significant advancement in cybersecurity.

Arfaoui et al. (2023) explored the application of machine learning for smart home threat detection using the HIKARI-2021 dataset, demonstrating the dataset's versatility and potential in various cybersecurity contexts. Their study highlighted the challenges of deploying ML in IoT environments and underscored the importance of addressing privacy and ethical concerns in such applications. Another significant contribution comes from the work of Fernandes and Lopes (2022), who used the HIKARI-2021 dataset to test various ML algorithms, including Random Forest and K-Nearest Neighbors (KNN), achieving high accuracy and demonstrating the dataset's effectiveness in evaluating different ML models. Similarly, the study by Arfaoui, Mrabet, and Jemai (2023) focused on the impact of identifiable features in ML classification algorithms with the HIKARI-2021 dataset, providing insights into the dataset's role in advancing ML-based cybersecurity solutions.

III. ANALYSIS OF EXISTING SOFTWARE

As discussed in previous sections, the integration of machine learning for cyber threat intelligence is crucial in modern cybersecurity. Several software solutions currently address fraud detection using machine learning and advanced analytics. Understanding these existing solutions can help identify gaps and opportunities for improvement in our project.

A. Splunk

Splunk provides powerful analytics and machine learning capabilities to detect anomalies and potential fraud activities across various data sources. It is widely used in the industry for its flexibility and scalability. Technically, Splunk uses a schema-on-read approach, allowing it to index and search large volumes of unstructured data in real-time. Its machine learning toolkit enables the creation of custom models for anomaly detection and predictive analytics. However, the platform operates on a pricing/ subscription model, which can be prohibitively expensive for some organisations. Additionally, while Splunk offers extensive integration capabilities, it can be complex to set up and manage, requiring significant expertise to utilise its full potential effectively. Furthermore, the proprietary nature of Splunk's technology means that it is not an open-source solution, potentially limiting its adaptability for specific needs.

B. IBM QRadar

IBM QRadar is an integrated security information and event management (SIEM) system that provides real-time threat detection and response. It excels in correlating vast amounts of data from multiple sources to identify fraudulent activities. QRadar employs advanced analytics, machine learning, and behavioural analysis to detect anomalies and potential threats. It integrates seamlessly with other IBM security products, enhancing its capability to provide a comprehensive security solution. However, the platform's high cost and complexity can be barriers for smaller organisations. Additionally, the need for specialised skills to configure and maintain QRadar can limit its accessibility for teams without dedicated security personnel.

C. Darktrace

Darktrace uses AI and machine learning to detect and respond to unusual patterns indicative of fraud within network traffic and user behaviour. Known for its Enterprise Immune System, Darktrace provides autonomous response capabilities, making it a robust solution for real-time fraud detection. Technically, Darktrace leverages unsupervised learning models and a proprietary AI algorithm known as the 'Antigena' to identify and neutralise threats autonomously. Despite its advanced features, Darktrace is also a subscription-based service, which may not be affordable for all organisations. Another consideration is the potential for high false positive rates, which can occur with anomaly-based detection systems, requiring constant tuning and adjustment to maintain accuracy.

These analyses highlight the strengths and limitations of existing fraud detection software solutions. While each platform offers powerful capabilities, common challenges include high costs, complexity, and the need for specialised skills to manage and maintain the systems. These factors underscore the importance of developing more accessible, flexible, and cost-effective solutions for fraud detection in cybersecurity.

IV. REQUIREMENT ANALYSIS AND DESIGN

A. Requirements Gathering

A critical fundamental aspect in any software development phase is its requirements collection. This project aims to develop a machine learning model using Python and Jupyter Notebook to analyse fraud in the HIKARI-2021 dataset, achieving outcomes comparable to existing software solutions. The solution must be flexible, scalable, and capable of integrating with various data sources and existing security products, ensuring it meets the needs of different cybersecurity environments.

Requirements specification gathering spanned several months of meetings with participants with extensive experience in cybersecurity and fraud detection. These participants provided valuable insights into the functionalities and features necessary for an effective fraud detection tool.

After ethics approval was granted, the author scheduled meetings with the participants to gather requirements specifications. The methodology adopted included:

a) Basic introduction and explanation of the study.

b) Presentation of the participant information sheet and consent forms.

c) Discussion on participants' experiences with existing fraud detection tools and their effectiveness.

d) Identification of key performance indicators (KPIs) and metrics crucial for fraud detection.

e) Suggestions for additional features and metrics that could enhance the new tool.

f) Evaluation of proposed wireframes and user interface designs.

g) Follow-up questions to clarify responses and gather detailed insights.

h) Documentation of all responses in a datasheet, ensuring data is pseudonymized to protect participant privacy.

Table.1 Key requirements for a fraud detection tool

Critical Aspect	Requirement Description
Integration Capabilities	The tool must be able to integrate seamlessly with various data sources and existing security products.
Key Metrics	Essential metrics identified include anomaly detection, user behaviour analysis, and real-time threat response capabilities.
Additional Features	Participants suggested features such as customisable dashboards, automated report generation, and support for various data formats.

B. Scope and Context

The scope of the project is to create a machine learning model capable of detecting fraudulent activities in the HIKARI-2021 dataset. The model will be designed to work seamlessly within existing cybersecurity frameworks, providing real-time threat detection and analysis. A successful outcome will be measured by the model's accuracy, precision, recall, and its ability to integrate into current systems without causing disruptions.

C. Detailed Design

Table.2 System architecture and Testing Overview

Component	Description
Architecture	Modular architecture to facilitate easy updates and maintenance.
User Interface	Basic wireframes to outline the user interface components, focusing on usability and accessibility.
Quality and testing	Unit Testing: Regular unit tests to ensure individual components work as expected. Integration Testing: Testing the integrated components to ensure they work together seamlessly.

D. Specific Algorithms and Enhancements

In addition to the general design principles, specific algorithms and enhancements will be implemented to improve the model's performance and address limitations found in previous research and existing projects.

1. Enhanced Feature Selection

Utilise advanced feature selection techniques such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) to identify the most relevant features in the HIKARI-2021 dataset. This will improve model efficiency and accuracy by focusing on the most significant data attributes.

2. Hybrid Model Approach

Combine multiple machine learning models to create a robust hybrid system. For example, use Convolutional Neural Networks (CNNs) for feature extraction from network traffic data, followed by Long Short-Term Memory (LSTM) networks for sequential analysis. This approach leverages the strengths of both models, enhancing detection capabilities for complex fraud patterns.

3. Real-Time Processing Enhancements

Integrate streaming data processing frameworks like Apache Kafka with the machine learning model to handle real-time data streams. This will enable the system to detect and respond to fraudulent activities as they occur, reducing the reaction time and potential damage caused by fraud.

4. Model Explainability

Incorporate explainability techniques such as SHAP (SHapley Additive exPlanations) to provide insights into model predictions. This will help cybersecurity professionals understand why a particular activity was flagged as fraudulent, improving trust in the model and facilitating more informed decision-making.

By integrating these specific algorithms and enhancements, the project aims to develop a cutting-edge fraud detection model that surpasses existing solutions in accuracy, efficiency, and real-time capabilities. This approach will ensure the model is well-equipped to meet the evolving challenges of cybersecurity and fraud detection.

V. IMPLEMENTATION

A. Data Preprocessing

The first step in the implementation phase was data preprocessing, which is critical to ensure the quality and usability of the dataset for machine learning tasks. The HIKARI-2021 dataset was initially loaded and explored to understand its structure and content. Unnecessary columns such as ‘Unnamed: 0.1’, ‘Unnamed: 0’, ‘uid’, and ‘traffic_category’ were removed to streamline the dataset and focus on the most relevant features.

Given the presence of high-cardinality categorical features such as ‘originh’ and ‘responh’, frequency encoding was employed. This method replaced categorical values with their frequency counts, transforming them into numerical values while preserving the underlying distribution.

Feature scaling was performed using standardisation, ensuring that all features contributed equally to the model’s performance. This step was crucial for algorithms like logistic regression and support vector machines, which are sensitive to the scale of input features. The standardised features, along with the target labels, were then split into training and testing sets, maintaining an 80-20 ratio. Additionally, SMOTE (Synthetic Minority Over-sampling Technique) was applied to handle class imbalance, ensuring that the minority class was well-represented in the training set.

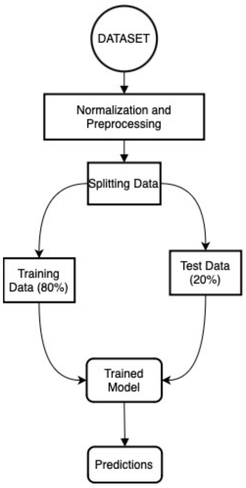

Fig.1 Data Splitting Diagram

Here is some of the code used to pre-process the dataset :

```
# Normalise/scale features
features = data_cleaned.drop(columns=['Label'])
scaler = StandardScaler()
features_scaled = scaler.fit_transform(features)

# Combine scaled features with the label
data_preprocessed = pd.DataFrame(features_scaled,
columns=features.columns)
data_preprocessed['Label'] =
data_cleaned['Label'].values

# Split the data
X = data_preprocessed.drop(columns=['Label'])
y = data_preprocessed['Label']
# Handle class imbalance
smote = SMOTE(random_state=42)
```

```
X_resampled, y_resampled = smote.fit_resample(X,
y)

X_train, X_test, y_train, y_test =
train_test_split(X_resampled, y_resampled,
test_size=0.2, random_state=42)
```

To provide a clear understanding of the dataset composition, the distribution of traffic categories was visualised. The bar plot, displayed below, shows the percentage distribution of each traffic category, highlighting the prevalence of benign traffic compared to various attack types.

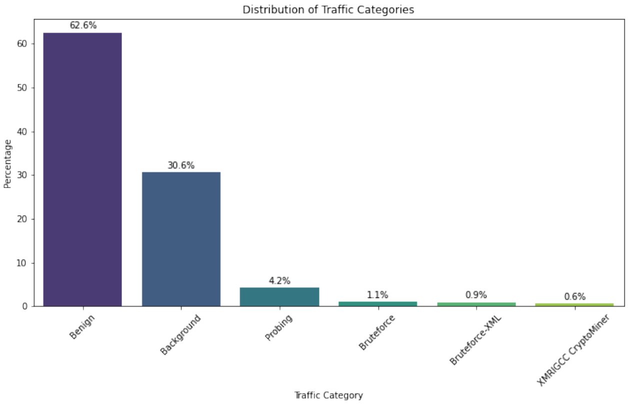

Fig.2 Distribution of Traffic Categories

Feature correlation analysis was conducted to identify the relationships between different features. The correlation matrix, shown below, helps in understanding which features are strongly correlated, allowing for more informed decisions during feature selection and engineering.

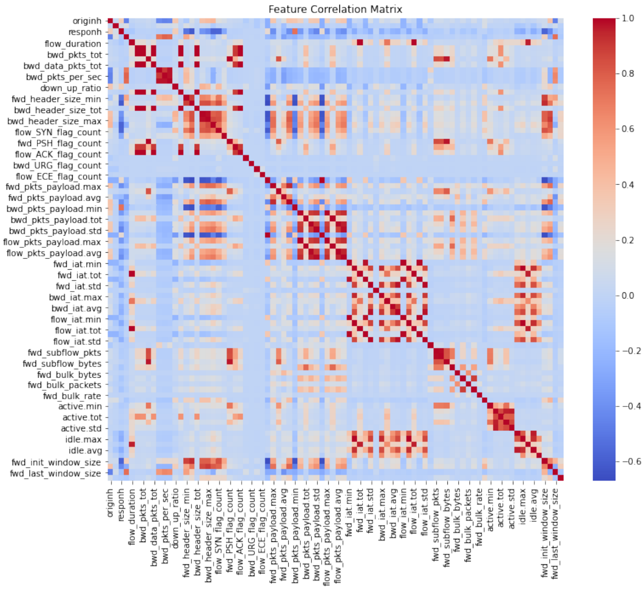

Fig.3 Feature Correlation Matrix

B. Model Training and Evaluation

Four machine learning algorithms were chosen for this project: Logistic Regression, Random Forest, Gradient Boosting Machine (GBM), and XGBoost. These models were selected based on their diverse approaches and strengths in classification tasks.

Logistic Regression was selected for its simplicity and effectiveness in binary classification problems. Despite its straightforward approach, logistic regression provides a solid baseline for comparison against more complex models.

Random Forest was implemented due to its robustness and ability to handle a large number of features efficiently. It is an ensemble method that combines multiple decision trees to improve predictive accuracy and control overfitting.

Gradient Boosting Machine (GBM) was chosen for its ability to build models sequentially, each new model correcting the errors made by the previous ones. This iterative approach makes GBM highly effective for classification tasks.

XGBoost was included to leverage its powerful classification capabilities, particularly in high-dimensional spaces. XGBoost is known for its effectiveness in finding the optimal hyperplane that best separates the classes.

Each model was trained on the preprocessed training set and evaluated on the test set using the following metrics:

- Accuracy: This metric measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total instances. It is a general indicator of model performance but may not be sufficient when dealing with imbalanced datasets.

- Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positives. High precision indicates a low false positive rate, making it crucial for tasks where false alarms are costly.

- Recall: Also known as sensitivity, recall is the ratio of correctly predicted positive observations to all actual positives. High recall is important for identifying as many true positive instances as possible, crucial in scenarios where missing a positive instance is critical.

- F1-Score: The F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. It is particularly useful when dealing with imbalanced datasets, as it considers both false positives and false negatives.

- ROC-AUC: The Receiver Operating Characteristic - Area Under Curve (ROC-AUC) measures the model's ability to distinguish between classes. A higher AUC indicates better performance in distinguishing between positive and negative classes.

Cross-Validation was performed to assess the generalisability of the models and ensure robust performance across different subsets of the data. This technique involves dividing the data into multiple folds and training the model on different combinations of these folds, providing a more comprehensive evaluation of model performance.

C. Feature Importance

Feature importance analysis was conducted for Random Forest and GBM models to identify the most influential features in the dataset. This analysis helps in understanding which features contribute the most to the model's predictions. Features such as 'responh', 'fwd_subflow_bytes', and 'originp' were consistently highlighted as significant contributors to the models' performance, providing insights into the key indicators of fraudulent activities. Visualising the feature importances can also guide future feature engineering efforts and model improvements.

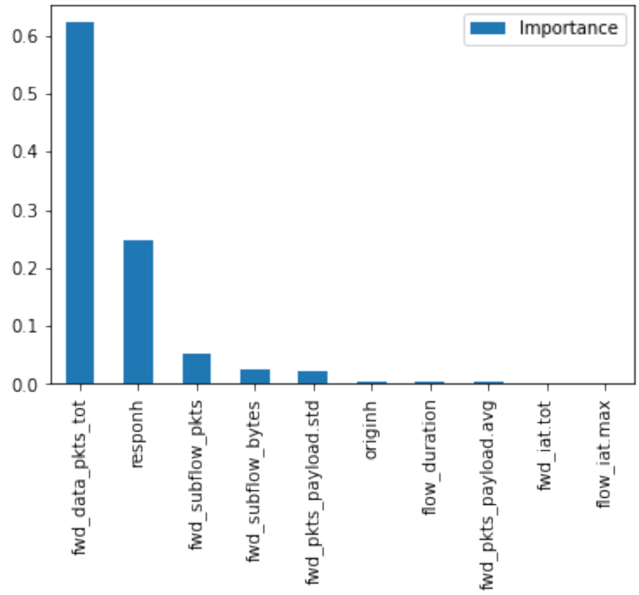

Fig.4 Top 10 Feature Importances from GBM

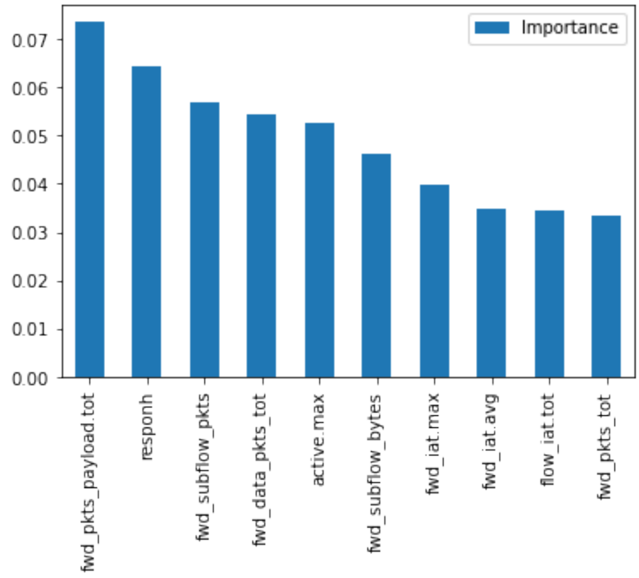

Fig.5 Top 10 Feature Importances from Random Forest

VI. EVALUATION AND TESTING

The evaluation phase focuses on analysing the performance of four different machine learning models: Logistic Regression, Random Forest, Gradient Boosting Machine (GBM), and XGBoost. The evaluation metrics used include accuracy, precision, recall, F1-score, and ROC-AUC, which provide a comprehensive assessment of each model's performance in detecting fraudulent activities.

A. Performance Metrics Analysis

To ensure a thorough evaluation, each model's performance was measured using several key metrics. The results are summarised in the following table:

Table.3 Model results

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.9221	0.8669	0.9957	0.9274	0.9221
Random Forest	0.9220	0.9083	0.9387	0.9232	0.9220
Gradient Boosting Machine	0.9281	0.8747	0.9994	0.9329	0.9282
XGBoost	0.9573	0.9529	0.9620	0.9574	0.9573

- Accuracy: XGBoost achieved the highest accuracy of 95.73%, indicating its superior ability to correctly classify both fraudulent and non-fraudulent instances.
- Precision: XGBoost also led in precision with 95.29%, which is critical in minimising false positives. This high precision means that the model is effective at identifying true fraud cases without incorrectly flagging legitimate transactions.
- Recall: GBM exhibited the highest recall of 99.94%, highlighting its effectiveness in identifying the majority of fraud cases. However, this comes with a trade-off as it also flagged more false positives compared to XGBoost.
- F1-Score: XGBoost had the highest F1-Score of 95.74%, indicating a balanced performance between precision and recall. This metric is particularly important in fraud detection, where both false positives and false negatives have significant implications.
- ROC-AUC: The ROC-AUC score for XGBoost was 95.73%, which is the highest among the models, demonstrating its excellent performance in distinguishing between fraudulent and non-fraudulent transactions.

B. Confusion Matrix Analysis

The confusion matrices provide a detailed breakdown of each model's classification performance, showing the number of true positives, false positives, true negatives, and false negatives. These matrices are crucial for understanding the types of errors each model makes.

- Logistic Regression: The confusion matrix for Logistic Regression indicates a high number of true positives but also a significant number of false negatives, which impacts the recall.

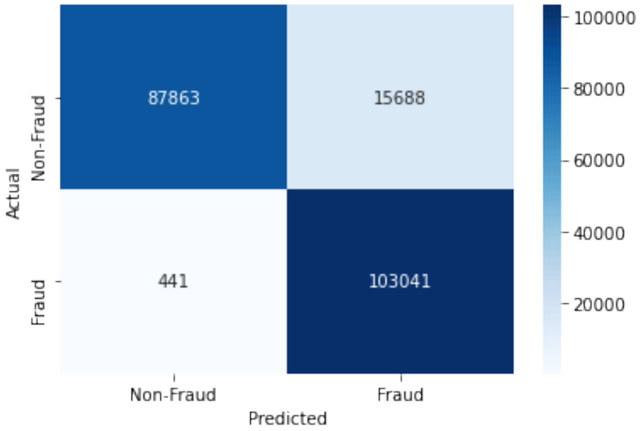

Fig.6 Confusion Matrix for Logistic Regression

- Random Forest: The Random Forest model shows a better balance between true positives and false positives, but with some false negatives still present.

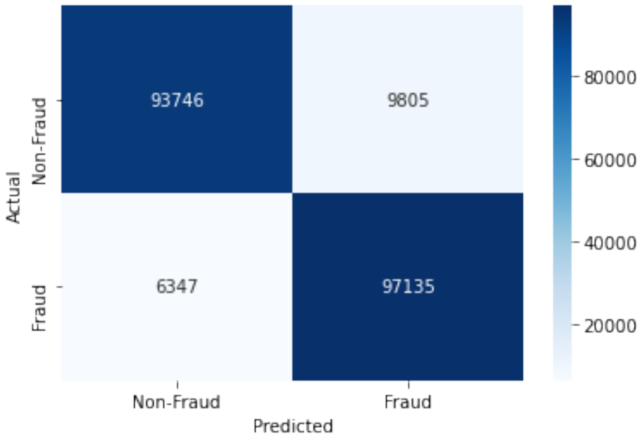

Fig.7 Confusion Matrix for Random Forest

- Gradient Boosting Machine: GBM's confusion matrix highlights its strong recall, with very few false negatives, but this comes with an increased number of false positives.

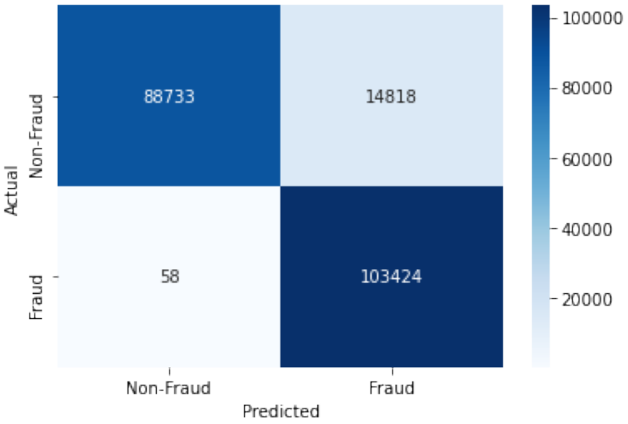

Fig.8 Confusion Matrix for Gradient Boosting Machine (GBM)

- XGBoost: XGBoost exhibits the best overall performance with a balanced confusion matrix, showing high true positives and minimal false negatives and false positives.

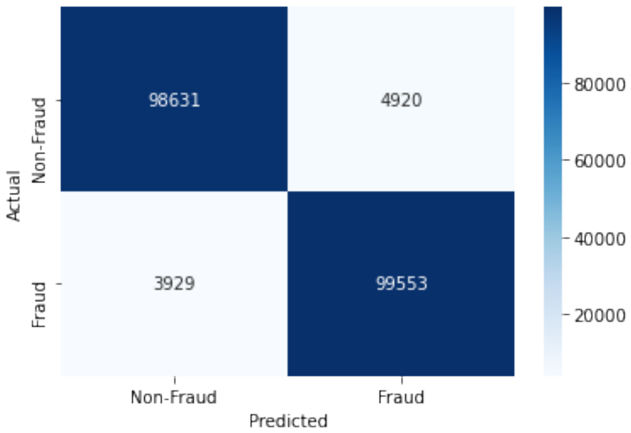

Fig.9 Confusion Matrix for XGBoost

C. Cross-Validation Results

To assess the generalisability of the models, cross-validation was performed. The cross-validation scores provide an insight into how well the models perform across different subsets of the data, ensuring the robustness of the results.

Table.4 Cross Validation Scores

Model	Mean CV Accuracy	Std Dev CV Accuracy
Logistic Regression	0.9017	0.0180
Random Forest	0.8658	0.0742
Gradient Boosting Machine	0.8900	0.0341
XGBoost	0.8761	0.0574

- Logistic Regression: Achieved a mean cross-validation accuracy of 90.17% with a low standard deviation, indicating consistent performance.
- Random Forest: Showed more variability in performance with a higher standard deviation of 7.40%.
- Gradient Boosting Machine (GBM): Demonstrated a balanced performance with a mean accuracy of 89.00% and a standard deviation of 3.41%.
- XGBoost: Had a mean accuracy of 87.61% with a standard deviation of 5.74%, indicating good performance with moderate variability.

Our approach demonstrated the effectiveness of the chosen methodologies. The study demonstrates that XGBoost is the most effective model for fraud detection on the HIKARI-2021 dataset. The detailed analysis of confusion matrices and cross-validation results ensures that the findings are robust and applicable to real-world scenarios.

D. Comparative Assessment

Table.5 Comparaison results

Model	Accuracy
Logistic Regression	0.9221
Random Forest	0.9220
Gradient Boosting Machine	0.9281
XGBoost	0.9573
Random Forest (Razi Sohail's Project)	0.9377
XGBoost (Razi Sohail's Project)	0.9302
Random Forest (Arfaoui's Project)	0.9228
Random Forest (Fernandes's Project)	0.99

As illustrated in Table 5, our XGBoost model achieved the highest accuracy at 95.73%, outperforming not only our other models but also those from Razi Sohail's and Arfaoui's projects. In contrast, our Logistic Regression, Random Forest, and Gradient Boosting Machine (GBM) models, while performing well, did not reach the accuracy levels reported by Fernandes, whose Random Forest model achieved a near-perfect accuracy of 99%.

Razi Sohail's Project demonstrated notable results with their Random Forest model achieving an accuracy of 93.77% and their XGBoost model reaching 93.02%. These results highlight the strength of ensemble methods like Random Forest and XGBoost in handling the complexities of the HIKARI dataset, which features a variety of network traffic scenarios, including both real and synthetic attack traffic.

Arfaoui's Project presented results that were comparable to ours, with their Random Forest model achieving an accuracy of 92.28%. This similarity suggests that our approach to implementing Random Forest is consistent with best practices in the field, although the slight difference in accuracy may be attributed to variations in preprocessing, hyperparameter tuning, or dataset partitioning.

The most striking difference is observed when comparing our results with those of Fernandes's Project. Their Random Forest model achieved an accuracy of 99%, significantly higher than ours. This discrepancy could be due to a more tailored feature selection process or the integration of additional preprocessing steps that enhanced the model's ability to generalise across the dataset.

In conclusion, while our models, particularly XGBoost, show strong performance and are competitive with those from other studies, there is room for improvement. Future work could explore more advanced hyperparameter tuning, the use of more complex ensemble methods, or even the incorporation of additional features to bridge the gap with the top-performing models from other research efforts.

LIMITATIONS AND CONCLUSION

Despite the promising results, this study has certain limitations. The primary limitation is the reliance on the HIKARI-2021 dataset, which, while comprehensive, may not encompass all possible real-world network scenarios. The dataset's synthetic nature could lead to models that perform exceptionally well on this data but may not generalise as effectively to different environments. Additionally, the computational resources required for training advanced models like Gradient Boosting Machine (GBM) and XGBoost are significant, posing challenges for deployment in resource-constrained settings. The project also primarily focuses on supervised learning techniques, potentially overlooking the benefits of unsupervised or semi-supervised methods in detecting novel threats.

This study underscores the critical role of machine learning in enhancing Intrusion Detection Systems (IDS). With the integration Logistic Regression, Random Forest, GBM, and XGBoost models, the research demonstrates substantial improvements in detecting malicious activities using the HIKARI-2021 dataset. The XGBoost model, in particular, showed superior performance across multiple evaluation metrics. The data preprocessing steps, including frequency encoding and standardisation, were crucial in preparing the dataset for effective model training. Despite the limitations, the findings highlight the potential of ML-driven IDS to provide a robust and adaptive cybersecurity framework. Future work should focus on validating these models across diverse datasets and exploring hybrid approaches to further enhance threat detection capabilities.

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my supervisor, Piyajith Wijetunge, for his invaluable guidance, support, and encouragement throughout this project. His expertise and insights were instrumental in shaping the direction and outcomes of this research. I am also grateful to the faculty and staff at Queen Mary University of London for their support and providing the resources necessary to complete this study. Lastly, I would like to thank my family and friends for their support and encouragement throughout my academic journey.

BIBLIOGRAPHY

1. Rajak, P., Lachure, J., & Doriya, R. (2022). 'CNN-LSTM-based IDS on Precision Farming for IIoT data', IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications.
2. Fernandes, R., & Lopes, N. (2022). 'Network Intrusion Detection Packet Classification with the HIKARI-2021 Dataset: a study on ML Algorithms', 10th International Symposium on Digital Forensics and Security.
3. Liu, Z., Thapa, N., Shaver, A., Roy, K., Yuan, X., & Khorsandroo, S. (2020). 'Anomaly detection on IoT network intrusion using machine learning', International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems.
4. Ferriyan, A., Thamrin, A. H., Takeda, K., & Murai, J. (2021). 'Generating Network Intrusion Detection Dataset

Based on Real and Encrypted Synthetic Attack Traffic', Applied Sciences, 11(17), 7868.
5. Khan, R. U., Zhang, X., Alazab, M., & Kumar, R. (2019). 'An improved convolutional neural network model for intrusion detection in networks', Cybersecurity and Cyberforensics Conference.
36. bin Abd Halim, I. H., bin Abdul Azziz, M. H., & bin Mohd Fuzi, M. F. (2021). 'Sinkhole Attack in IDS: Detection and Performance Analysis for Agriculture- based WSN using Cooja Network Simulator'.
7. Ferrag, M. A., Shu, L., Friha, O., & Yang, X. (2021). 'Cyber Security Intrusion Detection for Agriculture 4.0: Machine Learning-Based Solutions, Datasets, and Future Directions', IEEE/CAA Journal of Automatica Sinica.
8. Kumar, P., Gupta, G. P., & Tripathi, R. (2021). 'PEFL: Deep Privacy-Encoding based Federated Learning Framework for Smart Agriculture', IEEE Micro.
9. Sreelatha, G., Babu, A. V., & Midhunchakkaravarthy, D. (2022). 'Improved security in cloud using sandpiper and extended equilibrium deep transfer learning based intrusion detection', Cluster Computing.
10. Alferidah, D. K., & Jhanjhi, N. (2020). 'Cybersecurity impact over bigdata and IoT growth', International Conference on Computational Intelligence.
11. Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H., & Wang, C. (2018). 'Machine learning and deep learning methods for cybersecurity', IEEE Access, 6, 35365-35381.
12. Raghuvanshi, A., Singh, U. K., Sajja, G. S., Pallathadka, H., Asenso, E., & Kamal, M. (2022). 'Intrusion detection using machine learning for risk mitigation in IoT-enabled smart irrigation in smart farming', Journal of Food Quality.
13. Sohail, R. (2024). 'Cloud Enhanced Intrusion Detection: Evaluating Deep Learning and Ensemble Methods Using HIKARI 2021 Dataset', School of Technology.
14. Arfaoui, R., Mrabet, A., & Jemai, A. (2023). 'The impact of identifiable features in ML Classification algorithms with the HIKARI-2021 Dataset', Applied Sciences.
15. Splunk. (n.d.). 'Splunk for Fraud Detection'. Retrieved from (https://www.splunk.com/)
16. IBM QRadar. (n.d.). 'IBM QRadar: Security Intelligence Platform'. Retrieved from (https://www.ibm.com/security/security-intelligence/qradar)
17. Darktrace. (n.d.). 'Darktrace: The Enterprise Immune System'. Retrieved from (https://www.darktrace.com/)
18. Securonix. (n.d.). 'Securonix: Next-Gen SIEM Platform'. Retrieved from (https://www.securonix.com/)