

Enhancement Of Intrusion Detection System In Explainable Artificial Intelligence

¹Anupama P M*, ²Somnath Sinha

¹Research Scholar, Christ University, Bengaluru, India

²Associate Professor Christ University, Bengaluru, India

*Corresponding Author: Anupama P M

Abstract— Intrusion Detection Systems (IDS) remain vital to the security of contemporary networks; however, conventional techniques find it difficult to combat progressively sophisticated cyber-attacks. In order to overcome such difficulty, XAI has become the approach of choice by which the issues of transparency and trust in automated threat detection are solved. The present work introduces a conceptualization of XAI-based IDS, coupling various machine learning classifiers with SHAP and LIME to generate unambiguous, human-readable rationale for every inference. It has methodology such as dataset preprocessing, RFE and correlation analysis based feature selection and CatBoost, MLP and Gaussian Naïve Bayes based model development. XAI methods are used to find out the most significant contributing features and clarify the ways of attack detection on the class-based level. The issue that the presented work is focused on is the lack of interpretability of high-performance IDS models, which restricts their application in the real world. The suggested solution would improve transparency, yet it is very accurate, which allows optimists to have confidence in, authenticate, and refine model choices. Experimentation indicates that CatBoost has validation accuracy of around 97 per cent with almost perfect model accuracy and recall on key attack types. MLP does the same, that is, the SHAP analysis underlines the strongest influential features that make predictions. On balance, the XAI-enhanced IDS are much better in terms of its detection performance as well as interpretability.

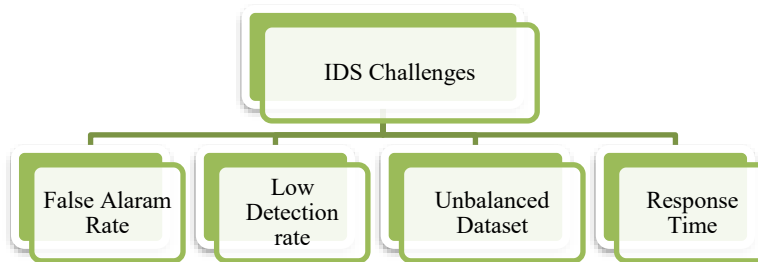
Keywords— Intrusion Detection System, Explainable AI, CatBoost, SHAP, MLP.

1. INTRODUCTION

The widespread use of the internet has led to a sharp rise in cyberattacks against government agencies, energy industries, and financial institutions in recent years. Large companies' websites are easy targets for hackers and attackers [1]. They can attack systems using a variety of vectors, including adware, worms, malware, and fraudulent logins. Security apps are essential for organizations to protect their networks from harmful assaults and improper usage [2, 3]. Intrusion Detection Systems (IDSs) are essential to protecting modern digital infrastructures against the ever-increasing complexity of cyber-attacks [4]. The ongoing rapid growth of cloud infrastructure, Internet of Things (IoT), and 5G networks has increased the attack surface and made traditional detection methods less and less useful [5]. Signature-based or anomaly-based traditional IDS methodologies are hardly useful extracted from new, unknown, or quickly changing attacks and have high false positive rate, scale issues, and poor adaptability in non-static environments [6]. With challenges such as "Advanced Persistent Threats (APTs)", zero-day exploits, and insider attacks, smart and adaptive security solutions are becoming more and more attractive. Machine Learning (ML) and Artificial Intelligence (AI) have become powerful tools to improve Intrusion Detection Systems (IDS) so that they can process complex datasets, recognize latent patterns, and improve detection with more data more accurately [7]. AI-based IDS can also be more efficient than classical systems because they can recognize anomalies and malicious activities using deep learning, ensemble learning, and hybrid models [8]. In some cases, they are effective but, typically, they are ineffective as they are considered black-box models, where the characteristics of their internal decision-making are entirely unclear

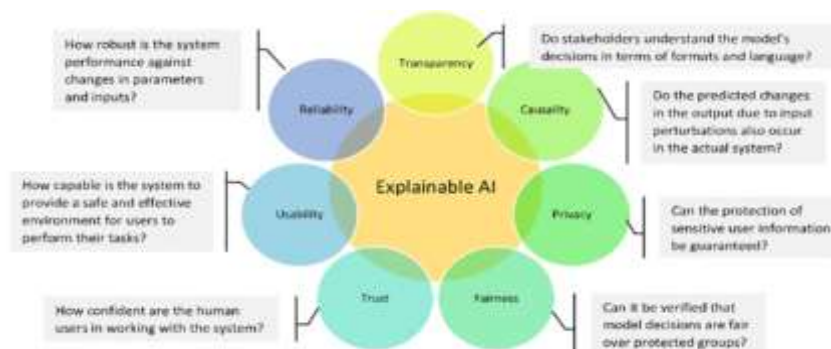
and uninterpretable [9]. This is a major problem with respect to transparency in the sensitive security areas like finance, health, and defense where human analysts must have explicit reasoning of automated decisions. The IDS challenges are depicted in figure 1. In the absence of interpretability, it can be hard for users to trust or take action on AI-based IDS outputs, restricting their use in the real world [10].

Figure 1: IDS Challenges [11]



To address this challenge, implementations of Explainable Artificial Intelligence (XAI) into IDS have been identified as a potential solution [12]. XAI is designed to make complex AI models comprehensive, interpretable, and trustworthy based on giving valid explanations for their predictions (Figure 2). In the context of intrusion detection, XAI allows security analysts to understand why a particular activity is flagged as malicious or benign, highlighting the most influential features and decision pathways. As an example, "SHAP (Shapley Additive Explanations) [13], LIME (Local Interpretable Model agnostic Explanations)" are two of the methods being used to improve the readability of the model's predictions in IDS. Rule-based post-hoc XAI, attention models. These XAI methods provide human comprehensible information about the previous decision process, improving the understanding of the reliability of the detection, and decreasing uncertainty, leading to a more rapid and informed incident response process [15]. The enhancement of IDS through XAI offers several additional advantages [17]. By providing interpretable decision-making, it fosters trust and collaboration between human experts and AI systems, bridging the gap between technical complexity and human reasoning [18]. Transparent IDS frameworks also support regulatory compliance, auditing, and forensic investigations by generating explanation trails that validate detection outcomes. Furthermore, explainability facilitates continuous learning and adaptation, as the insights gained from model explanations can be used to refine datasets, improve feature engineering, and strengthen the overall robustness of IDS. [19].

Figure 2: Goals of explainable AI (XAI) [16].



Thus, explainable AI IDS is a disruptive innovation in the development of a reliable, adaptive, and efficient defense against the modern cyber-attacks. This research seeks to discuss XAI methods and incorporation in IDS and how transparency and interpretability can enhance detection ability, build user trust, and promote the overall goal of trustworthy AI-based cyber-security systems. Some of the possible research objectives of the study are as follows:

- To conduct a comprehensive review of existing IDS techniques (signature-based, anomaly-based, and AI-driven) to highlight their strengths and limitations.
- To select and preprocess benchmark IDS datasets for training and testing AI/XAI-based intrusion detection frameworks.
- To investigate the role of AI and ML in improving the accuracy, adaptability, and efficiency of IDS compared to conventional approaches.
- To evaluate and integrate XAI techniques such as SHAP and LIME explanations into IDS frameworks for enhancing interpretability.

II. LITERATURE SURVEY

The increasing sophistication of cyber threats and the widespread adoption of vulnerable cyber-networks have made IDS an indispensable requirement for ensuring network security. Traditional ML-based IDS have demonstrated promising results in detecting conventional cyber-attacks; however, their lack of robustness against adversarial attacks and limited interpretability reduce their trustworthiness in real-world scenarios.

A few articles have considered the use of XAI with ML and DL-based IDS. Wali et al. (2025) [20] suggested a SHAP-based identification device, where low-credibility outputs are reassessed by secondary classifiers, and adverse robustness against adversarial examples was recorded on the CICIDS dataset. In a similar manner, Alabbadi et al. (2025) [21] used deep learning models (CNN, DNN, and TabNet) with XAI, to improve detection in IoT settings, with CNN reaching 100 percent accuracy on TON_IOT datasets. XI2S-IDS (Mahmoud et al., 2025) [22], a two-step model based on SHAP explanations to enhance transparency and low-frequency attack detection, and Random Forest and Attention-Based Neural Networks (Almolhis et al., 2025) [23] with SHAP and LIME were both proposed to achieve both interpretability and high detection accuracy. Similarly, Arreche et al. (2024) [24] came up with a SHAP and LIME evaluation framework, which established the strengths and limitations of the two XAI methods in IDS tasks. Similarly, Hooshmand et al. (2024) [25] implemented a hybrid approach to imbalance handling (SMOTE with K-means clustering) and SHAP explanations to ensemble-based IDS, and their results were much better on both NSL-KDD and UNSW-NB15 datasets.

Other researchers have contributed by expanding XAI-based IDS frameworks across different application domains. In a comparison of LIME and SHAP, Gaspar et al. (2024) [26] ensured better interpretability of an IoT intrusion detection system, and in a different study, Corea et al. (2024) [27] demonstrated the importance of feature engineering in feature engineering in IDS based on the occlusion sensitivity. XAIID-SCPS introduced by Nwakanma et al. (2023) [28] in the context of cyber-physical systems incorporates optimization algorithms with LIME to improve explainability, whereas Sivamohan et al. (2023) [29] used BiLSTM-XAI with SHAP and LIME to explain the industrial networks and achieved a 98.2% accuracy. Likewise, Wardhani et al. (2023) [30] combined classification models with counterfactual explanations and LIME to IoT security through CICIOT2023 and IoTID20 datasets and emphasized their interpretability. Lastly, Bauskar et al. (2023) [31] highlighted the promise of CNN-based IDS which demonstrated 99% accuracy on the NSL-KDD dataset indicating that it can

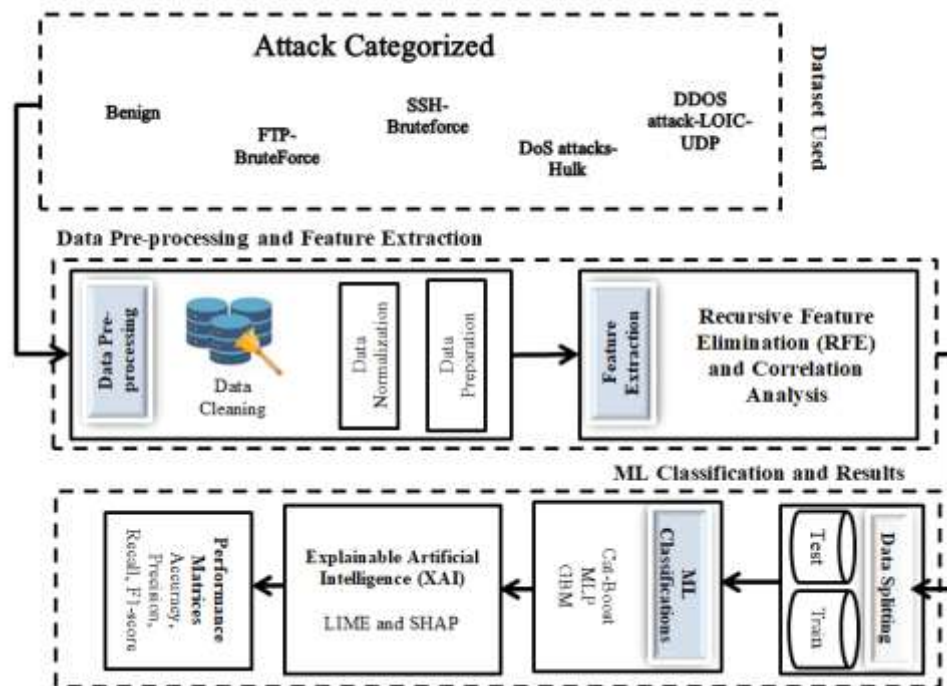
deal better with complex intrusion patterns than more traditional models. Despite progress, major research gaps remain in XAI-based IDS. Many models lack standardized interpretability metrics and rely on offline data, limiting real-world applicability. Real-time deployment in dynamic cyber and IoT environments is still rarely addressed. Explainability methods also remain vulnerable to adversarial attacks, and balancing accuracy, transparency, and efficiency is challenging. These gaps highlight the need for a more robust, transparent, and real-time IDS framework.

III.

RESEARCH METHODOLOGY

The enhanced IDS with XAI integrates machine learning and interpretability to strengthen cybersecurity. After preprocessing and normalizing data, key features are selected using RFE and correlation analysis. Classifiers like CatBoost, Gradient Boosting, and MLP detect intrusions effectively. XAI methods such as “LIME and SHAP” explain model decisions, improving transparency, accuracy, and trust. Figure 3 illustrates the proposed workflow.

Figure 3: Framework of proposed work



A. Dataset Collection

The CSE-CIC IDS 2018 dataset, which was created in 2018 by the “Communication Security Establishment (CSE) and the Canadian Institute for Cybersecurity (CIC)”, comprises a wide range of recent cyberattack vectors and has been extensively used for IDS development [32]. There are 10 distinct CSV files that comprise these attack vectors; each file includes both typical network traffic and various types of intrusions. Consequently, the comprehensive processing of each assault file was impossible because of the constraints of available CPU resources. Using Explainable AI (XAI), this study aimed to create IDS that could evaluate a small number of CSV files from CICIDS 2018 and make choices that were both transparent and credible. Accordingly, three separate files and the six distinct attack routes listed in Table 1 were chosen for this study.

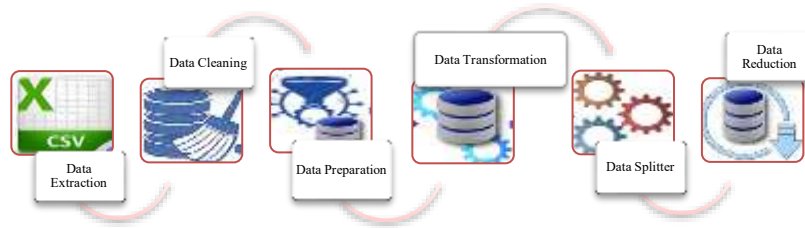
Table 1: Dataset Description

File Name	Network Traffic Type	Total Count
02-14-2018	Benign	667,626
	FTP-BruteForce	193,360
	SSH-Bruteforce	187,589
02-16-2018	DoS attacks-Hulk	461,912
	Benign	446,772
	DoS attacks-SlowHTTPTest	139,890
02-21-2018	DDOS attack-HOIC	686,012
	Benign	360,833
	DDOS attack-LOIC-UDP	1,730

B. Data Preprocessing

It is crucial prior to training any machine learning model because it profoundly influences performance and computing efficiency. As a consequence, the CICIDS 2018 dataset went through numerous data preparation phases, as seen in Figure 4. This section explains each stage of the procedure:

Figure 4: CICIDS 2018 Dataset Preprocessing Phase



Data extraction from CSV files into a Pandas data frame is the initial stage. This process extracts all kinds of network traffic. After that, the data cleaning stage is applied to these data frames, and any incorrect data instances with infinite values or NAN are eliminated. At this level, researchers also deleted any duplicate rows from the data. Before the classifier can make a decision on network traffic, the cleaned data goes through data preparation, where the textual labels are transformed into numerical values and the timestamp column is deleted. Data is normalized in each column after this processing stage, which is known as the transformation stage. Data values are transformed to conform to a specified interval, usually ranging from 0 to 1, by a process called normalization, which is also called min-max scaling. For ML models to work, it is essential that each feature be given equal weight. The case for scale-sensitive models is much more serious in this regard. Given this, they can find the minmax normalization formula in (1):

$$X_{norm} = (X - X_{min}) / (X_{max} - X_{min})$$

The data splitter then divides the normalized data into a 60:40 train-test ratio. After that, methods such as “Recursive Feature Elimination (RFE) and correlation analysis” were used to assess the significance of the dataset’s 36 characteristics (Table 2). In an effort to enhance model performance and decrease overfitting, features with low

variance or high multi-collinearity were eliminated.

Table 2: Feature Importance Ranking

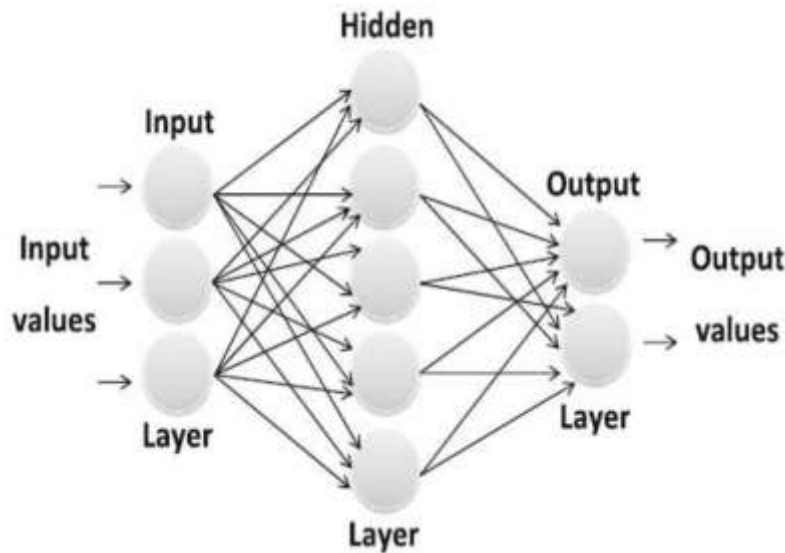
Feature	Weighting	Rank	Feature	Weighting	Rank
Fwd Seg Size Min	0.0777	1	Pkt Size Avg	0.0189	19
Init Fwd Win Byts	0.0746	2	Tot Fwd Pkts	0.0178	20
Fwd Pkt Len Max	0.0503	3	Bwd Seg Size Avg	0.0170	21
TotLen Fwd Pkts	0.0467	4	Fwd IAT Max	0.0153	22
Subflow Fwd Byts	0.0453	5	Fwd Pkt Len Std	0.0149	23
Fwd Header Len	0.0425	6	Bwd Pkt Len Mean	0.0145	24
Flow Pkts/s	0.0393	7	Bwd Header Len	0.0144	25
Fwd Pkts/s	0.0337	8	Pkt Len Max	0.0144	26
Init Bwd Win Byts	0.0318	9	Bwd Pkt Len Std	0.0135	27
Fwd Seg Size Avg	0.0309	10	Subflow Bwd Byts	0.0131	28
Bwd Pkts/s	0.0299	11	Flow IAT Max	0.0130	29
Fwd Pkt Len Mean	0.0279	12	Fwd IAT Mean	0.0130	30
Subflow Fwd Pkts	0.0236	13	Bwd Pkt Len Max	0.0113	31
Flow Duration	0.0208	14	Subflow Bwd Pkts	0.0104	32
Pkt Len Var	0.0203	15	Flow IAT Mea	0.0098	33
Fwd IAT Tot	0.0201	16	Pkt Len Mean	0.0097	34
Pkt Len Std	0.0193	17	TotLen Bwd Pkts	0.0088	35
Tot Bwd Pkts	0.0189	18	PSH Flag Cnt	0.0084	36

C. ML Model Development

- Multilayer Perceptron (MLP)

It is type of feed-forward artificial neural network made up of an input layer, hidden layers (one or more), and the output layer (as observed from Figure 5) [33]. It uses data processing through interconnected neurons with weighted sums and nonlinear activation functions in capturing complex and nonlinear relationships [34]. Because of these strengths, MLP performs exceptionally well in classification problems and has been deployed widely in IDS for the identification of normal and malicious network traffic [35]. While MLP does a great job of detecting things, its opaque structure makes it hard to understand. In order to achieve this goal, it is necessary to combine MLP with XAI methods such as SHAP and LIME. This combination increases IDS adversarial robustness, decreases false alarms, advance analyst trust, and boost transparency [36].

Figure 5: Architecture of MLP [37].



- CatBoost

It is a gradient boosting model and is built to handle categorical and numerical data effectively with minimal preprocessing [38]. It relies on decision trees and uses an ordered boosting algorithm that reduces overfitting and improves generalization and is hence quite useful for large and complex data. The fact that it handles categorical features directly provides it with an advantage over other boosting algorithms such as XGBoost and LightGBM [39]. CatBoost has shown high performance in terms of classifying normal and malicious network packets in correct classification in the IDS scenario [40]. Moreover, its integration with XAI models like SHAP enhances model explainability by illuminating the contribution of features to predictions. This improves the analysts' credibility, reduces false positives, and improves the effectiveness of the IDS against the emergent cyber threats [41].

- Gaussian Naive Bayes (GNB)

It is a Bayesian-based probabilistic classification system that models characteristics with normal (Gaussian) distributions and assumes that they are independent [42]. It is ideal for real-time applications, has low training data requirements, and performs better with high-dimensional datasets [43]. In the context of IDS, GNB has been used widely to classify network traffic as normal or malicious based on its fast learning and prediction capabilities [44]. Moreover, even though the assumption of feature independence is

probably too simple for complex network data, GNB still achieves competitive accuracy performance for intrusion detection. Furthermore, GNB increases interpretability with its integration of XAI, to help increase analysts comprehension of the decision making processes and build trust in IDS conclusions [45].

• Explainability Integration (XAI Framework)

All models employed in this work are black-box models; hence, the rationale behind a certain prediction remains unclear. As a consequence, explaining predictions and the attributes responsible for each class's prediction is crucial for gaining the network administrators' faith in the suggested solution. In order to clarify the models' predictions, this study employs two XAI methods, namely LIME and SHAP.

The LIME approach [46] locally approximates the model surrounding the instance requiring explanation to clarify individual predictions. An interpretable model, such as linear regression, is fitted to perturbed samples surrounding the instance to find out which attributes are most essential for a certain prediction. On the other hand, SHAP [47] determines SHAPley values for each model feature by utilizing cooperative game theory. This approach assures consistent and theoretically grounded explanations by consider all potential feature combinations to equalize the contribution of each characteristic to a prediction.

Among structured datasets, LIME and SHAP are the two most used methods for describing predictions [48,49]. When it comes to cybersecurity, these techniques are best [50]. Researchers chose LIME and SHAP because, as a whole, the datasets utilized in this study are organized. According to Algorithm 1, the LIME findings were computed by following these steps: perturbing the input instance, using the model to predict outputs, matching a surrogate linear model to those predictions, and then assessing feature weights to evaluate significance. In contrast, according to Algorithm 2, the SHAP results were computed initially in order to create feature subsets, forecast outputs for each subset, calculate SHAP values that indicate the contribution of each feature, and determine the relevance of features based on their SHAP magnitudes.

Algorithm 1: LIME

1. Assume that the instance from the dataset that requires explanation is x , which belongs to $\mathbb{R}^d$.
2. Create iterations $\{x_i\}$ for each value of i from 1 to m such that $x_i = x + \epsilon_i$, where ϵ_i is the delta that is changed to x for variations.
3. Use the algorithm as $\hat{y}_i = f(x_i)$, where $i = 1, 2, \dots, m$, to forecast the altered instances' output.
4. The surrogate model g should be fitted to the predictions using the formula $g(x) = \sum_{j=1}^d \omega_j x_j + b$
5. Use the formula $\min_g \sum_{i=1}^m (g(x_i) - \hat{y}_i)^2$ to find the minimum value of the difference between $g(x_i)$ and $\hat{y}_i$.
6. Produce the justifications by examining the weights ω_j to ascertain the significance of the attributes, with features with a greater $|\omega_j|$ regarded as more significant.

Algorithm 2: SHAP

1. The particular example from the dataset that requires explanation is denoted by $x \in \mathbb{R}^d$.
2. The trained ML model is denoted as $f(x)$. It provides the forecast, for example x .

3. Assign all remaining features to the baseline value and create a subset S of features that includes certain x features.
4. Classify each subset S using the model $f(x_s)$, whereby x_s represents an instance of the features in that subset.
5. For feature j , find the SHAPley value (ϕ_j) by using the formula.
 ϕ_j = The mean value of j 's input into the model's forecast.
6. The evaluation of the final forecast $f(x)$ is performed as

$$f(x) = \text{Baseline prediction} + \sum_{j=1}^d \phi_j$$
7. Determine the significance of features: characteristics that have a high ϕ_j value is more crucial for the model's prediction.

D. Performance Measures

Listed below are the commonly used performance measures for assessing the categorization frameworks. Equations (1)–(5) display the precision, accuracy, recall, F1 score, and FPR.

$$\text{Accuracy} = \frac{TP+TN}{FP+TP+TN+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{FP+TP} \quad (2)$$

$$\text{Recall} = \frac{TP}{FN+TP} \quad (3)$$

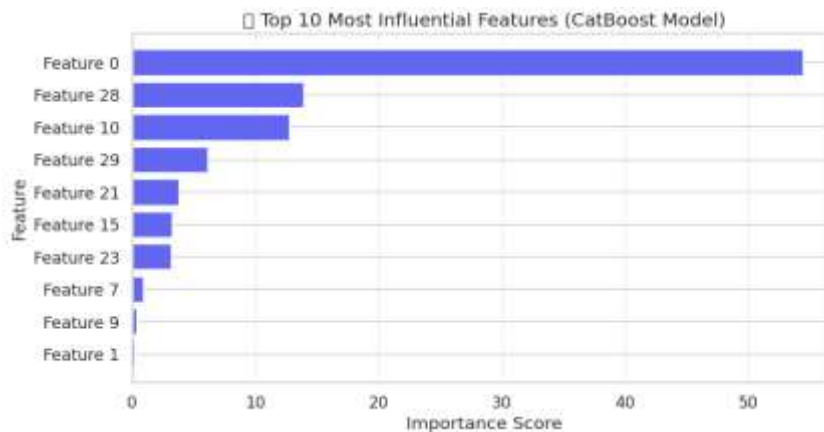
$$\text{F1 Score} = 2 \times \frac{P \times R}{P+R} \quad (4)$$

IV. RESULT AND ANALYSIS

A. Feature Importance Score

The 10 most influential features ranked by the CatBoost model are displayed in figure 6, which also shows how these features contribute to increasing explainability in the proposed intrusion detection system. Feature 0 has a dominant importance score of about 55, thus it can be considered the feature that is most responsible for model decisions. Feature 28 and Feature 10 each have scores close to 15 and 13, correspondingly, while Feature 29 contributes approximately 7. The intermediate-level features-Feature 21, Feature 15, and Feature 23, have importance scores that lie between 3 and 4. The last features (Feature 7, Feature 9, and Feature 1) have a very slight influence, and each of them has a score that is lower than 1.

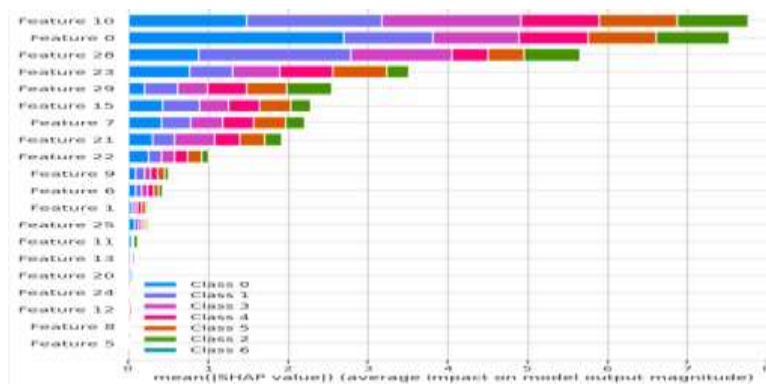
Figure 6: Feature Importance Score



This numerical distribution demonstrates that a small subset of highly influential features drives most predictive power, thereby supporting the explainable AI framework by clearly identifying which attributes most affect intrusion detection outcomes.

Mean SHAP values of the main features are presented in the graph divided into seven classes, and their average influence on the CatBoost model prediction is shown in figure 7. The strongest overall influence is represented by Feature 10, which has a total SHAP value of nearly 8, followed by Feature 0 and Feature 28, each approximately contributing 7. The Features 23, 29, 7, 15, and 21 have a moderate but noticeable effect, with SHAP values varying between 2 and 4 over the different classes. The remaining features, specifically Feature 22, 9, 6, 1, 25, 11, 13, 20, 24, 12, 8, and 2, have very low impact and vast majority of them are below 1. The distribution of numeric values reveals the explainable intrusion detection outcomes' dominant and supporting features.

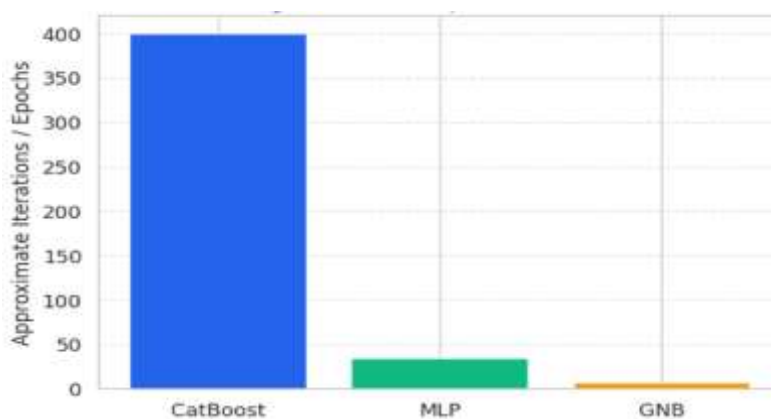
Figure 7: Features mean value using SHAP Value



B. Training Runtime Comparison

In the figure 8, it is evident that the number of iterations or epochs needed by each model are different. The CatBoost also takes the most time to train around 400 iterations, which is iterative boosting. The MLP model has a much higher training rate, as it requires approximately 35 epochs hence mediocre computational cost. On the contrary, GNB is the most effective one, with approximately 5 iterations, as it is a non-iterative probabilistic model. All in all, the fastest computing model is CatBoost and GNB is the fastest model to train.

Figure 8: Models training runtime comparison.

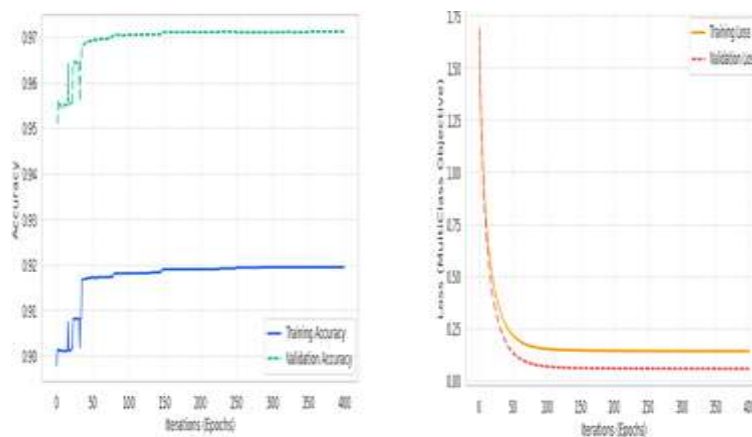


C. Models Performance Evaluation

• CatBoost

Figure 9 shows an accuracy graph showing that across 400 epochs, both training and validation accuracy consistently rise. The training accuracy begins at around 0.90 and rises gradually to around 0.92, indicating steady learning without significant changes. Validation accuracy begins slightly higher, around 0.95, and quickly stabilizes at about 0.97, reflecting strong generalization of the model. The model is not overfitting or underfitting, as seen by the modest difference between training and validation accuracy. The early increase followed by a stable plateau indicates effective optimization and good convergence behavior, which proves that the model can efficiently learn discriminative patterns for intrusion detection.

Figure 9: Training /Validation Accuracy and Loss curve for CatBoost Model



The loss curve reveals that training loss and validation loss have a steep downward trend at the beginning epochs, and training and validation loss decrease to around 1.7 and 0.1, respectively. Both curves level after approximately 100 epochs, which means that the model has optimally learned. Training and validation loss being very close ensures there is minimum generalization error as well as high robustness. The progressively declining component shows successful parameter improvements and robust convergence, which favor the general performance of the explainable intrusion detecting model.

Figure 10 displays the CatBoost model's class performance, which shows that it performs well across most attack categories. Attacks like Benign, HOIC, LOIC-UDP, and Hulk achieve very high detection accuracy with recall, precision, and F1-scores of approximately 1.00. With an F1-score of approximately 0.63 indicating misclassifications, a recall of approximately 0.55, and a precision of approximately 0.75, SlowHTTPTest exhibits modest performance. SSS-Bruteforce has values of 0.94, 0.99, and 0.96, but FTP-BruteForce has a recall of 0.88, a precision of 0.72, and an F1-score of 0.79. A moderately weak category and generally excellent reliability across the board are the model's defining features. The confusion matrix for the CatBoost Model is displayed in Figure 11.

Figure 10: CatBoost Model Performance Comparison

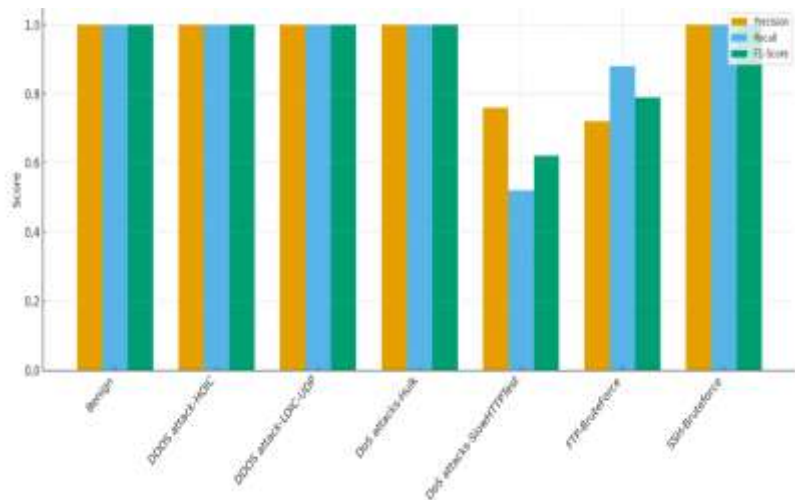
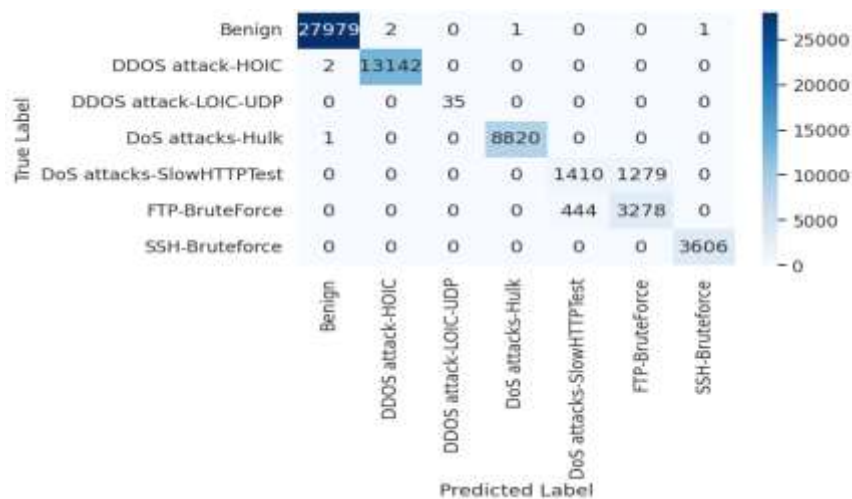
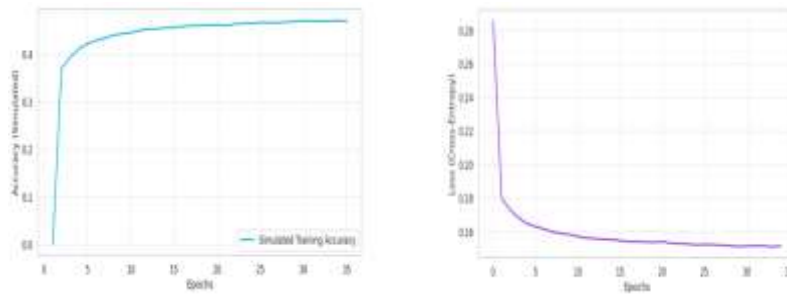


Figure 11: Confusion Matrix for CatBoost Model

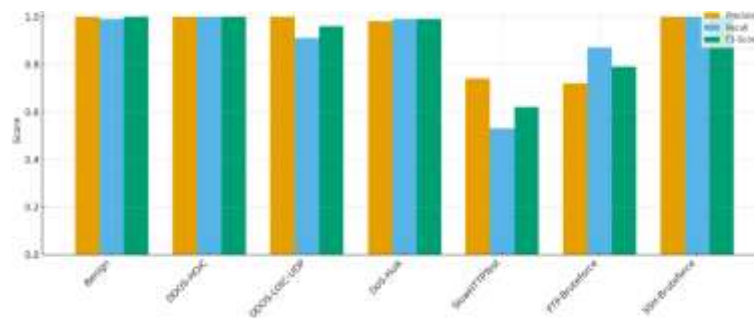
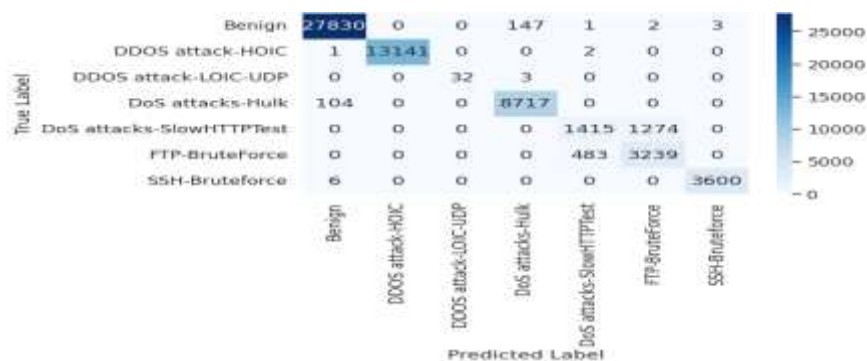


• MLP

The MLP model's training accuracy curve reveals a regular and uniform development throughout 35 epochs. The accuracy starts off very low, but is raised dramatically within the first few epochs, reaching approximately 0.38 by epoch 5 shown in figure 12. As the training goes on, the accuracy keeps increasing slowly, getting close to 0.53 around epoch 30 and making a tiny jump to about 0.54 by epoch 35. Such a behavior is a clear indication that the model is an effective learner from the given data, although it achieves fewer improvements after the initial learning phase. The training loss commences at about 0.28 and decreases drastically during the initial epochs until about epoch 5 where it steps almost to 0.17. During the further development of the epochs, the loss is reduced more gradually, getting even at 0.155 between epochs 20 and 30. At epoch 35, the loss is almost 0.152 with little variation. This is a smooth decreasing curve that represents efficient optimization and desirable convergence pattern of MLP model.

Figure 12: MLP model training accuracy and loss curve

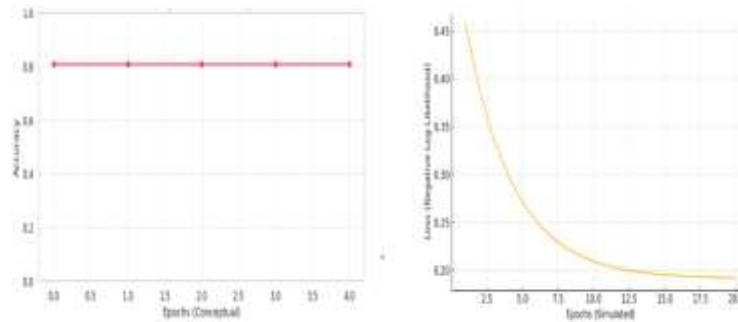
The figure 13 presented the accuracy, recollection, and the F1-score of the MLP model on eight traffic classes. Benign, DDOS-HOIC, DDOS-LOIC-UDP, and DoS-Hulk are near-perfect at 0.99100, that is, there is good classification. In the case of DoS-SlowHTTPTest, the precision is estimated to be approximately 0.75, the recall is estimated to be almost 0.60, and F1-score is estimated to be almost 0.65, their performance is moderate. FTP-BruteForce is precise with about 0.72, recalls recall with about 0.90 and F1-score with 0.80, that is, recall domination is prominent. The SSH-BruteForce has a high score of close to 0.98 to 1.00 in all metrics. On the whole, MLP model is good, except there are slight flaws in SlowHTTPTest and FTP-BruteForce classes. The confusion matrix in figure 14 shows accurate classification for major classes, with Benign (27830), HOIC (13141), Hulk (8717), and SSH-Bruteforce (3600) correctly identified. Misclassifications occur mainly in SlowHTTPTest, where 1415 samples are correctly predicted but 1274 are misclassified as FTP-Bruteforce. Minor errors appear in Benign classification, where 147 samples are misclassified as Hulk and a few scattered errors across other classes.

Figure 13: MLP Model Performance Comparison**Figure 14: Confusion matrix for MLP model**

- **GNB**

The accuracy curve for the GNB model in figure 15 stays pretty steady throughout the epochs, leveling off at around 0.80. This suggests that the model quickly hits its peak classification performance and doesn't show much improvement with additional iterations. Since GNB is a non-iterative probabilistic classifier, its accuracy generally remains consistent across epochs, which is why we see that flat and stable performance curve.

Figure 15: Accuracy and Loss curve for GNB Model



The loss curve is steeply downward trend at the very beginning and then it shows a gradual stabilization as the epochs go on. It is the negative log-likelihood loss that drops sharply in the very first stages, so in fact the model quickly learns the underlying data distribution. When the curve becomes level, it is a sign that the GNB model has achieved convergence, and thus, there are very few, if any, further improvements in prediction uncertainty reduction.

Figure 16: GNB Model Performance Comparison

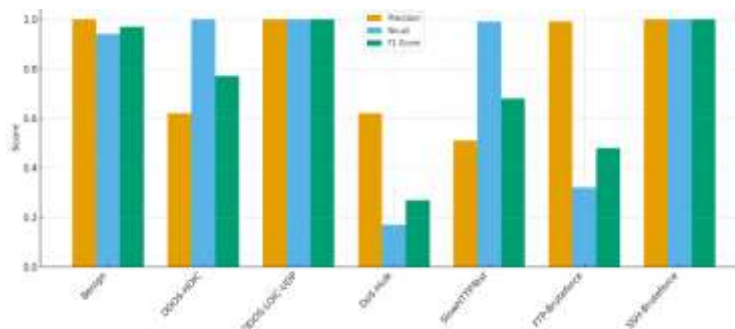
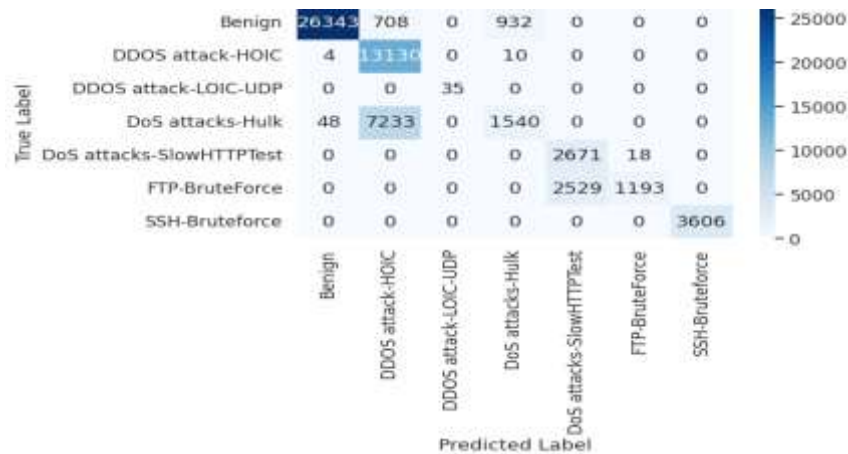


Figure 16 represents precision, recall and the F1-score of Gaussian Naive Bayes (GNB) model in seven classes of traffic. The model works well on DDOS-LOIC-UDP and SSH-Bruteforce with a 1.00 score on the three metrics. It also shows high levels of precision in Benign (1.00) and FTP-Bruteforce (0.99), but the overall recall of these classes is quite wide, ranging 0.94 and only 0.32, respectively. It is distinguished by a significant drop in recall (0.17) and F1-score (0.27) in the DoS-Hulk, which means that the model experiences some problems in the recognition of the examples of this kind of attack. On the same note, DDOS-HOIC and SlowHTTPTest exhibit an imbalanced performance; the former has low precision (0.62) and recall (1.00) whereas the latter has moderate precision (0.51) and recall (0.99). In general, the numeric trends indicate that GNB is

able to classify certain types of attacks outstandingly well but is vulnerable to misclassification in certain ones as the distribution of the features intersect with one another, causing asymmetrical trends in precision–recall. Figure 17 shows the confusion matrix for GNB model.

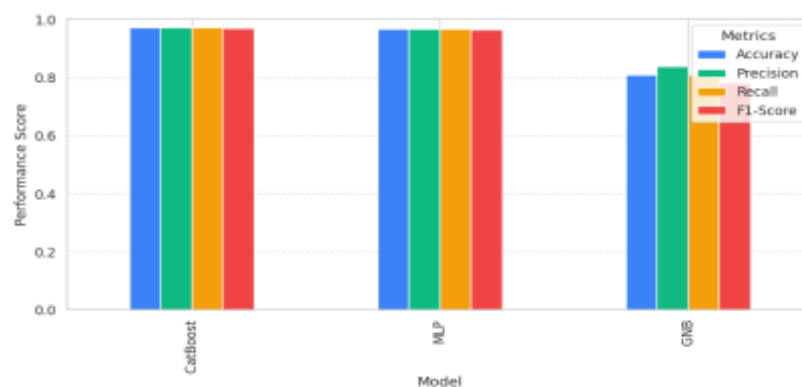
Figure 17: Confusion Matrix for GNB Model



D. Models Performance Comparison

Figure 18 illustrates how CatBoost, MLP, and GNB models stack up against each other in terms of performance metrics. Both CatBoost and MLP shine with nearly perfect scores, hovering around 0.97 to 0.98 across the board, which shows they deliver highly reliable and consistent classification results. On the other hand, the GNB model lags behind, with an accuracy of about 0.81, precision around 0.84, recall close to 0.78, and an F1-score of roughly 0.80. These figures clearly indicate that while GNB does a decent job, it simply can't compete with the superior performance of the more advanced CatBoost and MLP models in every evaluation metric.

Figure 18: Comparison of Models Performance

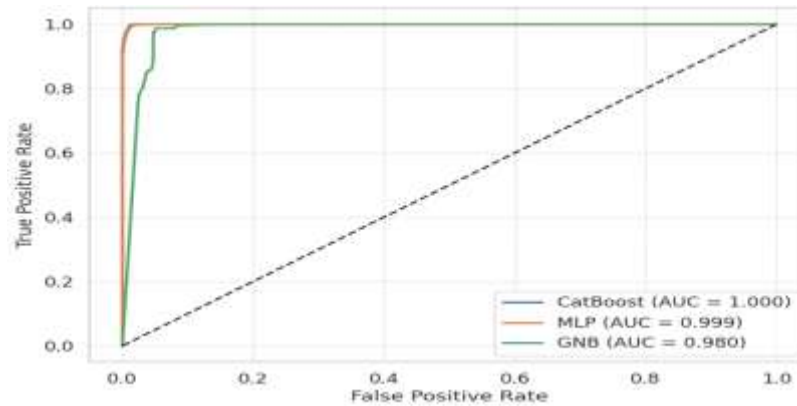


E. ROC Curves

The AUC scores representing how well CatBoost, MLP, and GNB are able to separate the two classes shown through the use of a ROC curve in figure 19. The AUC score of CatBoost indicates an ability to completely separate the two classes in this example (perfect score of 1.000). The AUC score for MLP was just barely below Catboost's score but still shows an almost perfect score (0.999). The GNB model did not perform as well,

with an AUC score of 0.980. However, it still met the criteria for a very good predictive model with a strong ability to differentiate between the two classes. In terms of all three AUC curves were quite near to or surpassed their maximum values; however, CatBoost had the highest AUC score of the three models.

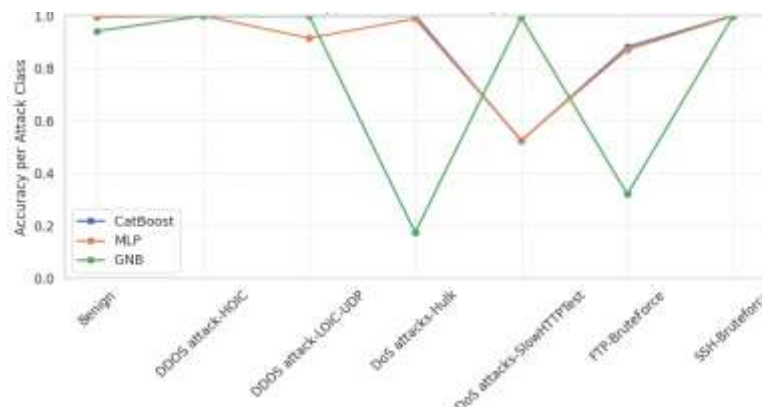
Figure 19: ROC-AUV Curve



F. Class-wise Detection Accuracy

The Figure 20 shows the accuracy of the per-class detection for CatBoost, MLP, and GNB, broken down into seven different attack categories. Both CatBoost and MLP are able to keep their accuracy at a high level almost all the time, which is mostly within the range of 0.90 to 1.00, thus, they are able to correctly identify almost all the attack types including Benign, HOIC, LOIC-UDP, and SSH-Bruteforce. Their minimum accuracy for SlowHTTPTest is estimated to be around 0.88-0.90. On the other hand, GNB performance is very inconsistent and variable, resulting in good accuracy (0.95-1.00) for Benign, HOIC, LOIC-UDP, and SSH-Bruteforce and a very low accuracy for the rest of the attacks i.e., Hulk (~0.18), SlowHTTPTest (~0.55), and FTP-Bruteforce (~0.33). This emphasizes that GNB has a hard time differentiating between certain complicated attack patterns.

Figure 20: Models accuracy comparison

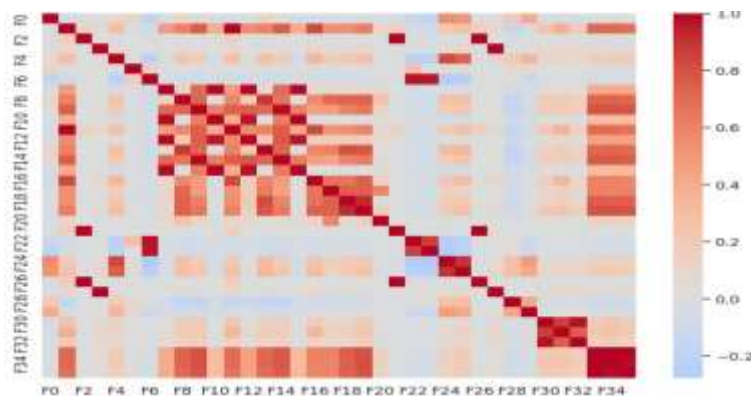


G. Feature Correlation Heatmap

The characteristic association Figure 21 is a heatmap showing the degree of linear correlations between all dataset features. Many of the feature pairs exhibit strong

positive correlations, which are represented by dark red blocks with values close to 1.0, and hence, they are considered as redundancy or similar transactions. A feature cluster around F2–F14 is established, which is the most correlated region meaning these attributes are significantly overlapping in terms of all the information they share. On the other hand, a few feature pairs have weak or negative correlations which are represented in light blue, thus showing that there is either no or inversely related dependency. The heatmap provides a clear view of highly interdependent feature groups along with distinct, weakly related attributes which are beneficial for feature selection and dimensionality reduction.

Figure 21: Feature Correlation Heatmap



V. Conclusion

By integrating Intrusion Detection Systems with Explainable Artificial Intelligence (XAI), the proposed research demonstrates that the technology becomes more transparent, trustworthy, and understandable without sacrificing its ability to detect various cyberattacks. By combining state-of-the-art ML algorithms like CatBoost, MLP, and GNB with interpretability techniques like SHAP and LIME, we are able to identify the important features for each inference and provide strong support for the classification. While GNB demonstrates moderate and inconsistent performance due to the varying degrees of attack complexity, CatBoost and MLP considerably surpass GNB in all metrics measured (>97% across all classes), including performance metrics. Additionally, CatBoost and MLP demonstrate strong detection capabilities for individual classes.

The interpretability of the decisions made by the models is further confirmed by the explainable outputs and at the same time, the data analysts are able to pinpoint the key parameters; hence, the alerts that are falsely flagged can be reduced and the trust in the automated threat detection is also lifted. Furthermore, the study has proved that XAI-based IDS not only offers the advantage of understanding the machine learning models better but also provides support in real-time decision making through conducting proactive analysis of the attack methods. To sum up, this advanced framework for IDS represents a robust, quick, and scalable option for coping with the constantly changing and very demanding situation in cybersecurity, making the detection of complex and sophisticated threats an easy task and being adaptable in different network conditions.

VI. References

1. Wang, Maonan, Kangfeng Zheng, Yanqing Yang, and Xiujuan Wang. "An explainable machine learning framework for intrusion detection systems." *IEEE Access* 8 (2020): 73127-73141.
2. Vigneswaran, Rahul K., R. Vinayakumar, K. P. Soman, and Prabakaran

- Poornachandran. "Evaluating shallow and deep neural networks for network intrusion detection systems in cyber security." In 2018 9th International conference on computing, communication and networking technologies (ICCCNT), pp. 1-6. IEEE, 2018.
3. Patil, Shruti, Vijayakumar Varadarajan, Siddiqui Mohd Mazhar, Abdulwodoood Sahibzada, Nihal Ahmed, Onkar Sinha, Satish Kumar, Kailash Shaw, and Ketan Kotecha. "Explainable artificial intelligence for intrusion detection system." *Electronics* 11, no. 19 (2022): 3079.
 4. Mohale, Vincent Zibi, and Ibidun Christiana Obagbuwa. "A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhance transparency and interpretability in cybersecurity." *Frontiers in Artificial Intelligence* 8 (2025): 1526221.
 5. Birahim, Shaikh Afnan, Avijit Paul, Fahmida Rahman, Yamina Islam, Tonmoy Roy, Mohammad Asif Hasan, Fariha Haque, and Muhammad EH Chowdhury. "Intrusion detection for wireless sensor network using particle swarm optimization based explainable ensemble machine learning approach." *IEEE Access* (2025).
 6. Ahsan, Shakil Ibne, Phil Legg, and SM Iftekharul Alam. "An explainable ensemble-based intrusion detection system for software-defined vehicle ad-hoc networks." *Cyber Security and Applications* 3 (2025): 100090.
 7. Zhang, Hangsheng, Dongqi Han, Shangyuan Zhuang, Zhiliang Wang, Jiyan Sun, Yinlong Liu, Jiqiang Liu, and Jinsong Dong. "Explainable and transferable adversarial attack for ml-based network intrusion detectors." *IEEE Transactions on Dependable and Secure Computing* (2025).
 8. Ali, Daim, Muhammad Kamran Abid, Muhammad Baqer, Yasir Aziz, Naeem Aslam, and Nasir Umer. "IMPROVING THE EXPLAINABILITY AND TRANSPARENCY OF DEEP LEARNING MODELS IN INTRUSION DETECTION SYSTEMS." *Kashf Journal of Multidisciplinary Research* 2, no. 02 (2025): 149-164.
 9. Rane, J., S. K. Mallick, O. Kaya, and N. L. Rane. "Enhancing black-box models: advances in explainable artificial intelligence for ethical decision-making." *Future Research Opportunities for Artificial Intelligence in Industry* 4 (2024): 136-180.
 10. Shen, Max W. "Trust in AI: Interpretability is not necessary or sufficient, while black-box interaction is necessary and sufficient." *arXiv preprint arXiv:2202.05302* (2022).
 11. Al-Janabi, Mohammed, Mohd Arfian Ismail, and Ahmed Hussein Ali. "Intrusion Detection Systems, Issues, Challenges, and Needs." *Int. J. Comput. Intell. Syst.* 14, no. 1 (2021): 560-571.
 12. Kalasampath, Khushi, K. N. Spoorthi, Sreeparvathy Sajeew, Sahil Sarma Kuppa, Kavya Ajay, and M. Angulakshmi. "A Literature review on applications of explainable artificial intelligence (XAI)." *IEEE Access* (2025).
 13. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).
 14. Muhammad, Aoun E., Kin-Choong Yow, Nebojsa Bacanin-Dzakula, and Muhammad Attique Khan. "L-XAIDS: A LIME-based eXplainable AI framework for Intrusion Detection Systems." *arXiv preprint arXiv:2508.17244* (2025).
 15. Manthena, Harikha, Shaghayegh Shajarian, Jeffrey Kimmell, Mahmoud Abdelsalam, Sajad Khorsandroo, and Maanak Gupta. "Explainable artificial intelligence (xai) for malware analysis: A survey of techniques, applications, and open challenges." *IEEE Access* (2025).
 16. Fiok, Krzysztof, Farzad V. Farahani, Waldemar Karwowski, and Tareq Ahram. "Explainable artificial intelligence for education and training." *The Journal of Defense Modeling and Simulation* 19, no. 2 (2022): 133-144.
 17. Nazat, Sazid, Lingxi Li, and Mustafa Abdallah. "XAI-ADS: An explainable artificial

- intelligence framework for enhancing anomaly detection in autonomous driving systems." *Ieee Access* 12 (2024): 48583-48607.
18. Senevirathna, Thulitha, Bartlomiej Siniarski, Madhusanka Liyanage, and Shen Wang. "Deceiving post-hoc explainable ai (xai) methods in network intrusion detection." In 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC), pp. 107-112. IEEE, 2024.
 19. Muneer, Salman, Umer Farooq, Atifa Athar, Muhammad Ahsan Raza, Taher M. Ghazal, and Shadman Sakib. "A critical review of artificial intelligence based approaches in intrusion detection: A comprehensive analysis." *Journal of Engineering* 2024, no. 1 (2024): 3909173.
 20. Wali, Syed, Yasir Ali Farrukh, and Irfan Khan. "Explainable AI and random forest based reliable intrusion detection system." *Computers & Security* (2025): 104542.
 21. Alabbadi, Adel, and Fuad Bajaber. "An Intrusion Detection System over the IoT Data Streams Using eXplainable Artificial Intelligence (XAI)." *Sensors (Basel, Switzerland)* 25, no. 3 (2025): 847.
 22. Mahmoud, Maiada M., Yasser Omar Youssef, and Ayman A. Abdel-Hamid. "XI2s-IDS: An explainable intelligent 2-stage intrusion detection system." *Future Internet* 17, no. 1 (2025): 25.
 23. Almolhis, Nawaf Abdualaziz. "Intrusion Detection Using Hybrid Random Forest and Attention Models and Explainable AI Visualization." *Journal of Internet Services and Information Security* 15, no. 1 (2025): 371-384.
 24. Arreche, Osvaldo, Tanish R. Guntur, Jack W. Roberts, and Mustafa Abdallah. "E-xai: Evaluating black-box explainable ai frameworks for network intrusion detection." *IEEE Access* 12 (2024): 23954-23988.
 25. Hooshmand, Mohammad Kazim, Manjaiah Doddaghatta Huchaiah, Ahmad Reda Alzighaibi, Hasan Hashim, El-Sayed Atlam, and Ibrahim Gad. "Robust network anomaly detection using ensemble learning approach and explainable artificial intelligence (XAI)." *Alexandria Engineering Journal* 94 (2024): 120-130.
 26. Gaspar, Diogo, Paulo Silva, and Catarina Silva. "Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron." *IEEE Access* 12 (2024): 30164-30175.
 27. Corea, Pap M., Yongxin Liu, Jian Wang, Shuteng Niu, and Houbing Song. "Explainable ai for comparative analysis of intrusion detection models." In 2024 IEEE International Mediterranean Conference on Communications and Networking (MeditCom), pp. 585-590. IEEE, 2024.
 28. Nwakanma, Cosmas Ifeanyi, Love Allen Chijioke Ahakonye, Judith Nkechinyere Njoku, Jacinta Chioma Odirichukwu, Stanley Adiele Okolie, Chinebuli Uzundu, Christiana Chidimma Ndubuisi Nweke, and Dong-Seong Kim. "Explainable artificial intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: A review." *Applied Sciences* 13, no. 3 (2023): 1252.
 29. Sivamohan, Sivanandam, and S. S. Sridhar. "An optimized model for network intrusion detection systems in industry 4.0 using XAI based Bi-LSTM framework." *Neural Computing and Applications* 35, no. 15 (2023): 11459-11475.
 30. Wardhani, Rini Wisnu, Dedy Septono Catur Putranto, Uk Jo, and Howon Kim. "Toward enhanced attack detection and explanation in intrusion detection system-based IoT environment data." *IEEE Access* 11 (2023): 131661-131676.
 31. Bauskar, Sanjay. "Enhancing Cybersecurity In Network Intrusion Detection System Based On Artificial Intelligence (AI) Techniques." Available at SSRN 4987362 (2023).
 32. <https://data.mendeley.com/datasets/29hdbdxx2r/1>
 33. Mohammed, Abdulrahman Jassam, Muhanad Hameed Arif, and Ali Adil Ali. "A multilayer perceptron artificial neural network approach for improving the accuracy

- of intrusion detection systems." *IAES International Journal of Artificial Intelligence* 9, no. 4 (2020): 609.
34. Mahmod, Mahmod S., Zakaria A. Hamed Alnaish, and Ismail Ahmed A. Al-Hadi. "Hybrid intrusion detection system using artificial bee colony algorithm and multi-layer perceptron." *International Journal of Computer Science and Information Security* 13, no. 2 (2015): 1.
35. Esmaily, Jamal, Reza Moradinezhad, and Jamal Ghasemi. "Intrusion detection system based on multi-layer perceptron neural networks and decision tree." In *2015 7th Conference on Information and Knowledge Technology (IKT)*, pp. 1-5. IEEE, 2015.
36. Mahbooba, Basim, Mohan Timilsina, Radhya Sahal, and Martin Serrano. "Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model." *Complexity* 2021, no. 1 (2021): 6634811.
37. Al-Naymat, Ghazi, Mouhammd Al-Kasassbeh, Nosaiba Abu-Samhadanh, and Sherif Sakr. "CLASSIFICATION OF VOIP AND NON-VOIP TRAFFIC USING MACHINE LEARNING APPROACHES." *Journal of Theoretical & Applied Information Technology* (2016).
38. Mutlu, Gökay, and Nese Rihani. "Intrusion Detection System with Explainable AI and Federated Learning."
39. Hajjouz, Abdulkader, and Elena Yurievna Avksentieva. "An approach to configuring CatBoost for advanced detection of DoS and DDoS attacks in network traffic." *Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика* 3 (2024): 64-74.
40. Saleem, Moeed, Muhammad Azam, Zeeshan Mubeen, and Gohar Mumtaz. "Machine learning for improved threat detection: LightGBM vs. CatBoost." *Journal of Computing & Biomedical Informatics* 7, no. 01 (2024): 571-580.
41. Kaya, Kiymet, Elif Ak, Sumeyye Bas, Berk Canberk, and Sule Gunduz Oguducu. "X-CBA: Explainability aided CatBoosted Anomal-E for intrusion detection system." In *ICC 2024-IEEE International Conference on Communications*, pp. 2288-2293. IEEE, 2024.
42. Islam, Riadul, Maloy K. Devnath, Manar D. Samad, and Syed Md Jaffrey Al Kadry. "GGNB: Graph-based Gaussian naive Bayes intrusion detection system for CAN bus." *Vehicular Communications* 33 (2022): 100442.
43. Naiem, Sarah, Ayman E. Khedr, Amira M. Idrees, and Mohamed I. Marie. "Enhancing the efficiency of Gaussian Naïve Bayes machine learning classifier in the detection of DDOS in cloud computing." *IEEE Access* 11 (2023): 124597-124608.
44. Jafer, Shatha H. "Optimize network intrusion detection system based on PCA feature extraction and three naïve bayes classifiers." In *Journal of Physics: Conference Series*, vol. 2322, no. 1, p. 012092. IOP Publishing, 2022.
45. Rafin, Nafiz Imtiaz, Swattic Ghose, Ahsanul Islam, Fahim Ul Islam, Dr Md Golam Rabiul Alam, and Amitabha Chakrabarty. "Enhancing Intrusion Detection: A Robust Ensemble Learning Approach with Explainable AI and Strategic Fused Voting-Stacking Mechanism." In *Proceedings of the 3rd International Conference on Computing Advancements*, pp. 1050-1057. 2024.
46. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135-1144. 2016.
47. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).
48. Salih, Ahmed M., Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E. Petersen, Karim Lekadir, and Gloria Menegaz. "A perspective on explainable artificial intelligence methods: SHAP and LIME." *Advanced Intelligent Systems* 7, no. 1

- (2025): 2400304.
49. Gaspar, Diogo, Paulo Silva, and Catarina Silva. "Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron." *IEEE Access* 12 (2024): 30164-30175.
50. Islam, Md Tohidul, Md Khalid Syfullah, Md Golam Rashed, and Dipankar Das. "Bridging the gap: advancing the transparency and trustworthiness of network intrusion detection with explainable AI." *International Journal of Machine Learning and Cybernetics* 15, no. 11 (2024): 5337-5360.