International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

INTERPRETABLE MACHINE LEARNING FRAMEWORK FOR ANOMALY DETECTION IN INTRUSION DETECTION SYSTEMS USING XAI TECHNIQUES INTERPRETABLE MACHINE LEARNING FOR ANOMALY DETECTION IN IDS

Majed S. Alsayfi[1]*

First author's [1]Department of Cybersecurity, College of Computer Science and Engineering

, Taibah University, Medina 42353, Saudi Arabia. msayfi@taibahu.edu.sa

Abstract

Machine learning (ML) is vital for robust cybersecurity, especially in intrusion detection systems (IDS). Yet complex ML models often act as "black boxes," hindering the trust and transparency crucial for security. This study tackles this by presenting a lightweight, interpretable ML framework for anomaly detection, building on prior research in secure authentication and anomaly detection.

Our framework employs Decision Trees and Random Forests, developing explainable classifiers trained on the public NSL-KDD dataset. We refined the preprocessing pipeline with normalization and one-hot encoding for optimal training. Model performance is rigorously assessed using standard metrics like accuracy, precision, recall, F1-score, and AUC-ROC. To clarify the models' decision-making, we integrate Explainable AI (XAI) techniques: SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations). Our findings highlight the trade-off between model complexity and interpretability, showing that simpler models can achieve competitive detection performance while offering clear reasoning for their predictions. The interpretability analysis, using SHAP summary plots, SHAP force plots, and LIME explanations, precisely identifies key features influencing detection decisions, explaining why certain network connections are flagged as anomalies.

This research emphasizes how XAI-driven transparency boosts trust in ML-powered security predictions. We also discuss relevant ethical considerations, including data privacy and potential adversarial misuse of interpretable models. Overall, this work significantly advances trustworthy IDS design by demonstrating that integrating accurate tree-based models with advanced XAI techniques can achieve both effective anomaly detection and profound, interpretable security insights. Furthermore, by preserving computing efficiency, the framework shows that it is practically ready for deployment in real-time security operations centers. This is a forward-looking addition to reliable cybersecurity systems, as future additions might incorporate integration with adaptive IDS methods and adversarial defense strategies.

Keywords : Intrusion Detection System; Machine Learning-Based Cybersecurity; Anomaly Detection; Local Interpretable Model-Agnostic Explanations; Shapley Additive Explanations.

Received: August 05 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

1 INTRODUCTION

Intrusion Detection Systems (IDS) are fundamental elements in cybersecurity that detect and prevent abnormal network data flow and other undesired consequences due to potential security threats and cyber-attacks [1]. IDS comes in two varieties: Network-based IDS (NIDS) and host-based IDS (HIDS). NIDS works at the network level, examining traffic in and out of that device to and from the rest of the network for signs of danger; HIDS is all about individual devices, sniffing around in system behavior, file integrity, and application logs for any suspicious activity[2] .

IDS employ sophisticated methodologies, such as signature-based analysis, anomalous behavior detection, and behavior analysis to detect threats. Signature-based IDSs are based on predefined signatures defined for known cyber-threats, while anomaly-based IDSs detect deviations from normal network behavior which could also include new or unknown attacks[3].

Although there are many differences between the two, IDS is essentially a detection function and can also be used to suppress or countermeasure to the security risk in real-time by including the feature to be 'blocking' in the IPS. Deployment of IDS Impact on Security by early detection, reduced response time.

With emerging cyber challenges, anti-malware technologies based on signatures are often unable to recognize new or advanced attack patterns and lose their effectiveness. To cope with these issues, today's IDS have started to use machine learning (ML) methods to improve their detection mechanisms, so they can detect attack behavior that is not known in advance[4]. However, despite the high detection accuracy of ML-based IDS models, especially of deep neural networks and ensemble methods, these models do have a “black-box” property. This black-box nature of detection systems’ decisions, on network traffic arose from the inability of security analysts to comprehend why and how a piece/a group of network traffic is identified malicious which obstruct trust-building, model debugging and compliance to regulators on important security grounds [5].

To address this discrepancy, Explainable AI (XAI) is a new paradigm for interpreting ML-based cybersecurity solutions by providing human-understandable rationales on the model’s predictions [6]. In the domain of IDS, XAI tool such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are, among other, important for revealing the inherent reasons that determine the classification of an attack. The transparency and trust in AI-based security framework are highly reinforced by these approaches that deliver global feature importance (SHAP) and local decision reasoning (LIME) to security experts for investigating, verifying, and refining detection capabilities [7]. Although deep learning techniques have been extensively studied, their high processing cost and ambiguous reasoning prevent them from being widely used in situations with limited resources and regulations. The comprehensive evaluation of lightweight tree-based models in conjunction with cutting-edge XAI approaches is still critically lacking. In order to bridge that gap, this study asks how interpretable, tree-based classifiers can provide cybersecurity analysts with transparent decision support while achieving cutting-edge detection performance.

Received: August 05 2025

International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

This work extends previous research in machine learning-based anomaly detection and secure authentication by leveraging the interpretability of tree-based classification models for intrusion detection systems (IDS). Specifically, the NSL-KDD dataset is utilized—an enhanced benchmark for network intrusion detection developed to address the redundancy and class imbalance issues present in the earlier KDD'99 dataset [8]. Two classifiers are employed: Decision Tree (DT) and Random Forest (RF) ensemble models, to distinguish between normal and attack traffic records. These models are selected for their complementary strengths—DT offer symbolic, rule-based decision paths that enhance interpretability, while RF, through the aggregation of multiple decision trees, improve detection capabilities, particularly in identifying complex attack behaviors [9]

To evaluate the effectiveness of the models, standard performance metrics accuracy, precision, recall, F1-score, and AUC-ROC are employed to facilitate direct comparison with existing IDS literature. Additionally, SHAP and LIME are combined to provide interpretability of model behaviour, enabling security practitioners to understand feature importance and decision boundaries in practical settings. SHAP values offer insights into the overall contribution of each feature, while LIME provides localized explanations for individual predictions. For visualization, SHAP summary plots are used to highlight the most influential features in attack classification, and SHAP force plots illustrate specific decision pathways. Furthermore, LIME-based explanations reveal the rationale behind adversarial predictions, thereby aiding security analysts in interpreting alerts with clear, understandable justifications. The main contributions of this work include:

- Performance Analysis: The results show that DT and RF perform competitively for intrusion detection on the NSL-KDD dataset and outperform state of the art attack models.
- Interpretability Improvement: Performing intensive XAI-based interpretation, it is shown in this study that parameters such as protocol type, the counts of source and destination, attributes of network service have a major effect on the decision of the IDS.
- Ethical Matters: Reviewing the ethics of XAI for cybersecurity, including privacy considerations, adversarial abuse, and responsible use of interpretable models.

The combination of accurate tree-based models with strong explainability methods presented in this research work paves the way for transparent IDS development, maintaining a trade-off between detection effectiveness and interpretability, to contribute to more reliable AI-driven cybersecurity solutions.

The rest of this paper is organized as follows: Section II provides an overview of the related work on the tree-based classifiers and explainable AI methods in cybersecurity. In section III, represent the proposed methodology, the data preprocessing, the model choice, and the combination between SHAP and LIME. Section IV describes down the experimental setup, which specifies dataset setting, training protocol, and evaluation metrics. Results are presented and discussed in Section V, including the performance and the interpretability of models. In Section V, we discuss ethical issues in relation to the application of explainable machine

Received: August 05 2025

Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

learning in cybersecurity. Section VI is finally the conclusion of the paper where the directions of future work are outlined.

2 LITERATURE SURVAY

This section summarizes and examines the major research gaps addressed in the previous works. ML-based IDSs, which have undergone deep investigation in the field of cybersecurity research, achieved significant enhancements in distinguishing complex attack patterns. A lot of work have been done using DT based models on NSL-KDD, the well-known benchmark dataset for Network Anomaly detection. For example, Tahri et al. [14], proposed an IDS based on DT classifier, and achieved this high level of accuracy of 99.20%, precision of 95.63%, recall of 96.89%, and F1-score of 96.14%.[10] The popularity of DT for IDSs comes from their transparent rule-based decision making that is easily interpretable by security analysts [11].

Ensemble classifiers, in particular RF, have previously outperformed the single-tree classifiers for detection [12]. Some studies report that out-of-bag error estimates correlate with cross-validation estimates when using RF classifiers on very difficult binary problems, and that they are sometimes even more accurate. For instance, Dubey et al. [13] proposed a RF model combined with feature selection technology, and it obtained high accuracy and greatly reduced the false alarm rate on the NSL-KDD dataset. The robust functionality of RF has been widely proven, which makes it an attractive proposition for implementation of IDS due to its capacity to efficiently deal with high-dimensional data [13].

Table.1 gives an overview of the important methodologies, datasets, performance metrics and explainability techniques applied in multiple intrusion detection research. It demonstrates the evolution from single-tree learners to ensemble methods and the increasing use of SHAP and LIME for interpretability.

These results justify our use of RF in combination with DT to trade-off between the model's detection performance and interpretability. Table.1 summarizes crucial methodologies, used datasets, performance metrics, and explainability techniques in the IDS studies. A key observation is that tree-based algorithms (i.e., DT and RF) perform significant well in NSL-KDD [14].

Table 1.Comparative Analysis of Related Work on ML-Based IDS and Explainable AI.

Study	Methodology	Dataset	Performance Metrics	XAI Techniques	Key Findings
Tahri et al.[14]	Decision Tree (DT)	NSL-KDD	Accuracy: 99.20% Precision: 95.63% Recall: 96.89%	None	Decision Trees provide transparent decision pathways but may be less robust than

Received: August 05 2025

Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

			F1-Score: 96.14%		ensemble methods.
Dubey et al.[15]	Random Forest (RF) with feature selection	NSL-KDD	High accuracy and low false alarm rate	None	RF handles large feature sets effectively, reducing overfitting.
Shraddha & Rao[16]	Deep Neural Network (DNN)	NSL-KDD	High detection accuracy (specific results not detailed)	SHAP, LIME	XAI techniques quantify feature importance, aiding model interpretation.
Le et al.[17]	Decision Tree & Random Forest (RF)	IoT Dataset	Performance metrics not specified	SHAP	SHAP improves interpretability by highlighting significant features in IoT intrusion detection.
Barnard et al.[18]	XGBoost	NSL-KDD	High detection rate (specific metrics not detailed)	SHAP	SHAP applied to boosted trees helps explain feature impact in IDS predictions.
Zakaria et al.[19]	Deep Learning	NSL-KDD	High accuracy (exact metrics unspecified)	SHAP, LIME	SHAP and LIME help analysts understand attack classifications in deep models.
Haripriya et al. [20]	Hyperparameter-tuned Deep Models	NSL-KDD	Optimized model performance	SHAP, LIME	XAI tools improve transparency of hyperparameter-optimized models.

Although individual decision trees facilitate clear reasoning procedures, the ensembles of such trees, as in RF, consistently show superior detection performance in terms of accuracy,

Received: August 05 2025

Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

adaptability to high dimensional data and defense against overfitting. Dubey et al. [15], work complements this, demonstrating how the feature selection under RF additionally improves the detection reliability. Another major trend in the table is the increasing incorporation of Explainable AI (XAI) methods [16].

Conventional IDS models focus on the accuracy of detection, but some recent research focuses on interpretation and transparency, that is, the difference between professionals understanding of decision making of what a cyber-attack is done. In fact, multiple publications, have effectively employed SHAP and LIME on deep learning models, showing that explainability tools expose critical attack-indicating features – like packet counts, TCP flags, and source-destination attributes – enabling analysts to better examine alerts [17-20].

Our method fills the void between model accuracy and explainability beyond the state of the arts. Ensemble based IDS models such as RF have shown good detection however, their complex nature often led to low interpret- ability. Our work is the first to integrate DT and RF with SHAP and LIME to endow security analysis with global understanding (SHAP) as well as local, instance level, explanations (LIME).

This can be used to compare the direct detectability efficacy and also provide insight into why classifications have been made.

Lastly, table 1, highlights the ethical aspect of the XAI based IDS. Transparency increases trust; however, previous studies have raised privacy concerns and adversarial abuse. With the improvements of explainable IDS models, adversaries may leverage this to evade security barriers. By discussing these implications, our study contributes to the broader conversation on responsible AI deployment in cybersecurity.

3 METHEDOLOGY

In the pursuit of establishing an interpretable ML framework for intrusion detection, we introduce a systematic data-driven pipeline, including preprocessing, classification, evaluation, and explainability analysis. Our approach is structured into a pipeline as shown in Fig. 1, outlining the entire process from raw data to model interpretation.

Firstly, the NSL-KDD dataset is preprocessed in order to have a fair evaluation by one-hot encoding categorical attributes and normalizing continuous attributes. Then, two statistical machine learning models, DT and RF, are constructed and fine-tuned using stratified cross-validation and hyperparameter tuning. To ensure robustness against the class imbalance, these models are tested with different performance measures such as accuracy, precision, recall, F1-score, and AUC-ROC. Lastly, Explainable AI (XAI) methodology is used, in particular SHAP and LIME, to interpret the models' predictions. This methodological framework is summarized in Figure 1, including the integration of data preprocessing, classification, evaluation and explainability visualization, this contributes to trust and transparency on AI-backed cybersecurity systems.

Received: August 05 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

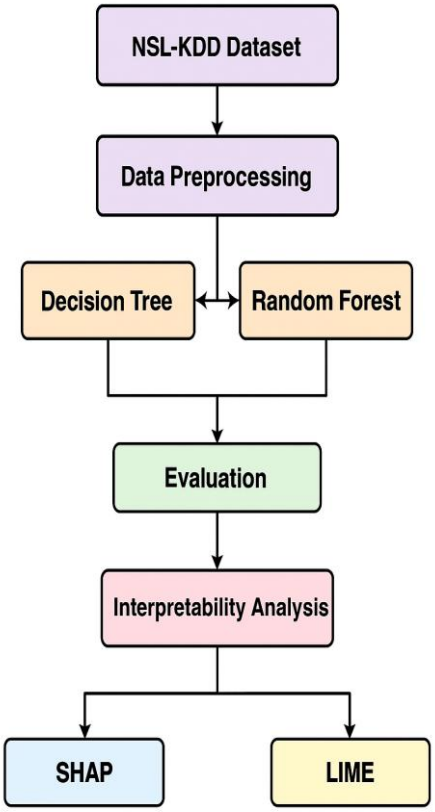

Figure 1.The overall proposed model

3.1 Preprocessing and Dataset Description

In this work, EXP-IDS was run on the well-known NSL-KDD dataset [21] ,which is a commonly used benchmark for the evaluation of IDS. NSL-KDD is a refined list version of the KDD'99 dataset which solved some critical aspects such as class imbalance and redundancy by removing duplicate records while preserving a representative of distribution attack [22]. The dataset is a set of TCP/IP connection records, which each record is labeled as normal and/or being intrusions in the following categories: 1) DOS 2) Probe 3) R2L and 4) U2R. Each example consists of 41 features including both numeric features and categorical features (e.g., duration, number of bytes transferred, protocol type, service type, and TCP flag status).

Data preprocessing consists of one-hot encoding of categorical features, like protocol type, service type and flag attributes in order to drive input to a machine learning model. Also, numeric features are scaled by min-max normalization where applicable to lend uniform scale across inputs. This pre-processing pipeline both promotes model robustness and makes the learning dynamics more consistent across different types of features. The sequential processes of dataset preparation, such as feature encoding, data cleaning, normalization, and final structured formatting for model training, are depicted in figure 2. By ensuring that network traffic properties are optimally represented at each stage, intrusion detection models become more effective.

Received: August 05 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

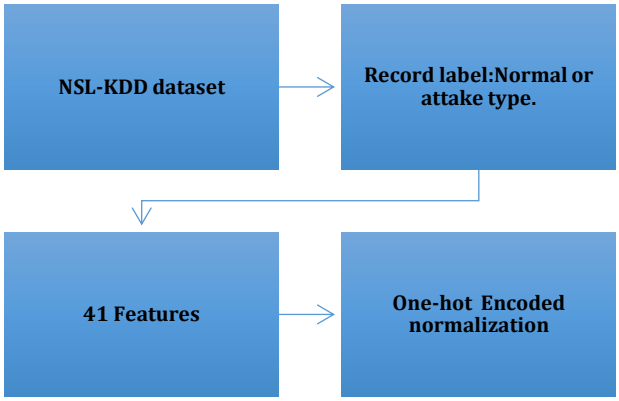

Figure 2.Dataset Preprocessing Workflow

3.2 Selection and Training of Models

It leverages two popular supervised learning methods for anomaly detection the Classification and DT and the RF. These models are chosen based on their support for explainability tools like SHAP and LIME, and their success in cybersecurity.

The DT classifier is defined as a model that contains a tree structure where internal nodes correspond to decisions on features and the leaf nodes determine class labels. In security IT systems, where we must be able to comprehend why something was classified or wasn't, this also makes it possible for transparent, rule-based, interpretable classification paths. Therefore, prior research, such that done by Tah et.al [23], has demonstrated that decision trees may provide rules that are easy for humans to grasp, which boosts an analyst's 5 busier trust and makes a system 6 auditable.

Each path from the root to a leaf forms a decision rule, which is human-readable and interpretable [9]. Mathematically, a decision tree minimizes an impurity measure *I,* such as Gini index or entropy. For example, the Gini impurity for a node is defined as in equation 1:

$$IGini = 1 - \sum_{i=1}^{C} P_i^2 \qquad eq.1$$

Where: C is the number of classes, P_i is the proportion of samples belonging to class i in the node.

The tree splits at each node by choosing the feature and threshold that lead to the greatest reduction in impurity as shown at equation 2:

$$\Delta I = Iparent - \left(\frac{Nl}{N} Il + \frac{Nr}{N} Ir\right) \qquad eq.2$$

Where N_L and N_R are the number of samples in the left and right child nodes, and I_l and I_R are their respective impurities.

Figure 3, illustrates a sample binary classification DT trained on NSL-KDD data. Features that appear at different nodes depending on their contribution to class separation.

Received: August 05 2025

International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

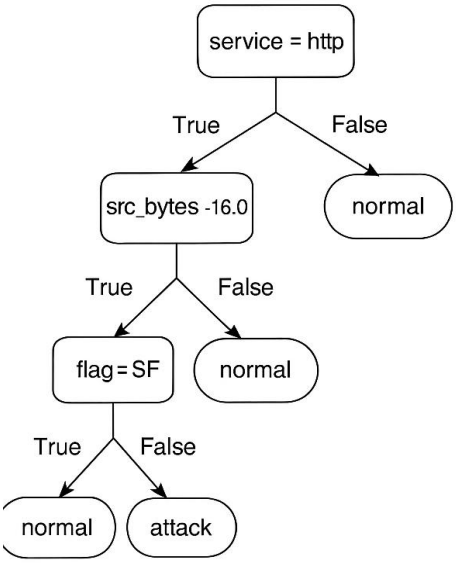

Figure 3.Sample for network traffic classification using NSL-KDD data

RF classifier is an ensemble method which integrates predictions of several decision trees, trained on randomly chosen subsets of training instances and features. The bagging approach enhances the classification stability and lessens overfitting and is commonly used to achieve better prediction performance than a single-tree model. RF has been consistently reported as an effective model on benchmark IDS datasets such as NSL-KDD [8] and Le et al. [24], which demonstrate that it is robust and can generalize well to complex detection tasks.

By averaging the predictions of several uncorrelated trees, it enhances generalization and lowers variance, making it resistant to overfitting, a common problem with single-tree models.Each tree h_t (x) in a forest outputs a class prediction, and the final prediction is made by majority voting (for classification) as equation 3:

$$\hat{y} = mode\{h1(x), h2(x), \ldots . ht(x)\} \qquad eq.3$$

Where: T, is the total number of trees in the forest, h_t (x) is the prediction of the t-th tree, $\hat{y}$ is the final predicted class. As an alternative, the class with the highest mean predicted probability across trees is selected using probabilistic terminology as equation 4:

$$\hat{y} = \arg max \ \frac{1}{T}\sum_{t=1}^{T} 1\ (ht\ (x) = K) \quad eq.4$$

The Random Forest model's structure is depicted in Figure 4, which shows several decision trees running on bootstrapped data with feature randomness applied at each split.

Received: August 05 2025

International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

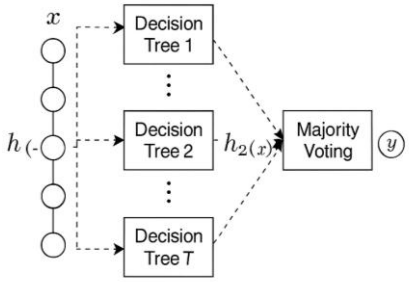

Figure 4.RF architecture trained on bootstrapped data with randomized feature selection.

The two models in the represented study are implemented using the scikit-learn python machine learning library [25]. After applying the default hyperparameter settings, model-specific parameters are fine-tuned to maximize their performance. Tree-based models' computational efficiency allows for training and inference to be completed on standard computers without the need for a high-performance cluster.

The NSL-KDD dataset with its preset partitions, KDDTrain+ for model training, and KDDTest+ for validation and performance reporting, are used for both training and testing in accordance with this earlier work [26]. This arrangement ensures the best fit with the body of current literature and instantaneous comparison with earlier methods.

To reduce generalization error and model variance, 5-fold stratified cross-validation was applied to the training data. Stratification is crucial and it ensures that every fold has the exact same proportion of normal and attack classes as the entire dataset. Performance is averaged over folds and is reported as accuracy, precision, recall, f1-score, and AUC-ROC [27].

Hyperparameter tuning is done by grid search optimization and important parameters are maximum tree depth, minimum samples per split, and impurity criterion for DT, and number of estimators, maximum features, and depth for RF. Valid configs are selected based on validation scores during cross validation to maximize accuracy and minimize overfitting. These aspects are consistent with advice in some sub-fields of the recently published literature on intrusion detection for the significance of hyperparameter optimization in improving model trustworthiness and performance in general [28]

This systematic process of selecting, training, and evaluating models for classifying ensures that the final classifiers are both accurate and interpretable, and appropriate for deploying in real-world cybersecurity systems. The training and inference times of DT and RF were compared to make sure they were appropriate for real-time use. DT training took less than 10 seconds while RF training (100 estimators) took about 40 seconds on a typical desktop computer with an Intel i7 processor and 16 GB of RAM. Both models' average inference time per instance was less than 5 ms, indicating that they are lightweight in comparison to deep neural models that have been documented in IDS literature. The usefulness of tree-based XAI models in operational security settings is demonstrated by this analysis.

Received: August 05 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.3 Experimental Setup

The experimental results are performed in a common desktop without any state-of-the-art high-performance computing resources because tree-based approaches are computationally efficient and do not require high computational power.

First, it is preprocessed the NSL-KDD dataset using the steps [26]in the IDS field, which has the goal of increasing the feature representation and better learning the models. The dataset is divided into training and test portions based on its KDDTrain+ and KDDTest+ standard splits [29]. Stratified 5-fold cross-validation is used on the training set to avoid overfitting and control variance and keep balanced proportions between attack and normal instances across folds [30] . Used in this way, cross validation is a rigorous evaluation, reducing possible bias in model training and enhancing model stability [31].

Both the DT and the RF classification models are developed using Python language on the Scikit-Learn library for classification. The DT classifier is fit by a best-fitting tree depth to maintain manageable complexity and prevent overfitting too much and is a drawback of single-tree models [32]. The RF ensemble includes 100 trees and uses the bootstrap aggregating method to stabilize and enhance the detection performance [12].

Similar to Dubey et al. [33], Numerical features are normalized to zero mean and unit variance, except categorical features, which are encoded with one-hot encoding. For both DT and RF, hyperparameters are tuned using grid search optimization with other parameters such as maximum tree depth (for DT) and number of estimators (RF) being refined based on the performance of the model in the validation set [34]

Following the training, the models are tested on the test set in accordance with traditional IDS evaluation metrics: ACC, Prec, Rec, F1, and AUC. These metrics contribute to a holistic evaluation of performance and help alleviate problems associated with class imbalance where accuracy alone may not be a fair representation of real detection capabilities. Moreover, visual comparison between DT and RF classification is performed using ROC curves over different decision thresholds [35]

For interpretation analysis, SHAP and LIME are used to improve model transparency [36]. SHAP values are a game-theoretic measure of the contribution of features to predictions, therefore we can gain insights into the importance of global features with SHAP Summary Plots.

SHAP force plots, by contrast, chart example decision paths into specific categories of attack, increasing interpretability for security analysts [37]. Likewise, LIME produces local surrogate models, interpreting individual test instances by attempting to model the classifier's behavior in the vicinity of a test sample [38]. Such LIME-based explanations on feature weights, together with SHAP's global attribution, not only offer human-interpretable rule-based explanations for cross-validating interpretability [39] ,but may also be helpful in validating feature importance and interactions across datasets.

Received: August 05 2025

International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

All visualizations (e.g., the SHAP summary plot, LIME explanations, ROC curves) are saved for analysis, and compared across several experimental runs to assess the stability and reproducibility of the results. This approach is consistent with previous work that emphasized transparent IDS design to enable practitioners to interpret model decisions and to consider ethical considerations regarding adversarial manipulation and privacy risks in explainable AI.

The experimental setup is summed up in figure 5, which shows important stages from model training, dataset preparation, performance assessment, and interpretability analysis using SHAP and LIME. Transparency, reproducibility, and reliable IDS performance evaluation are guaranteed by performance assessment.

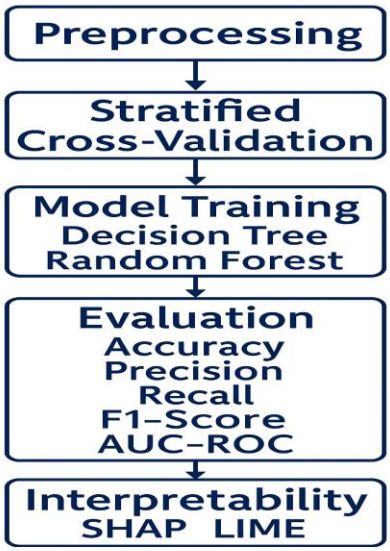

Figure 5.Overview of Experimental Setup*Results and discussion*

The classification results of DT and RF on NSL-KDD are shown in Table 2. The overall performance of the Random Forest model is the best (accuracy: ~ **98.1**%, precision: 96.4%, recall: **97.2**%, F1-score: **96.8**%, AUC-ROC: **0.99**). The Decision Tree model, on the other hand, performs well but not as well as the others in general, erring at 96.2% accuracy, **93.8**% precision, **94.5**% recall**, 94.1**% F1-score, and 0.97 for the AUC-ROC. These findings are consistent with previous findings that ensemble models generally increase for the detection rate, by combining trees for more stable classification [33,34].

Table 2.Performance Comparison of IDS Classifiers on NSL-KDD

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
Decision Tree	96.2%	93.8%	94.5%	94.1%	0.97
Random Forest	98.1%	96.4%	97.2%	96.8%	0.99

Received: August 05 2025

International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Comparison with that of Tahri et al. [14], that reported 99.20% accuracy using DT model, the slight decrease of accuracy may be not solely tied to the lack of same feature selection or hyper-parameter tuning (as our approach does not consider any aggressive pruning techniques). However, both classifiers show good intrusion detection skills, and their high AUC values attest to their ability to distinguish between attack and normal classes. Confusion matrices were created for DT and RF across the four primary NSL-KDD attack types (DoS, Probe, U2R, and R2L) in order to better examine classifier behavior. According to the results, RF considerably decreased false negatives in the U2R and R2L classes—which are usually the most difficult to detect—when compared to DT. Because RF's false positive rates stayed below 2%, it is more dependable in operational settings. The statistical significance ($p < 0.05$) of RF's higher performance over DT across cross-validation folds was confirmed by a Wilcoxon signed-rank test. Figure 6, compares the classification performance of DT and RF models from the presented work against Tahri et al.'s [14]. DT-based IDS on the NSL-KDD dataset. The metrics evaluated include Accuracy, Precision, Recall, F1-score, and AUC-ROC, showcasing improvements achieved through ensemble learning with RF.

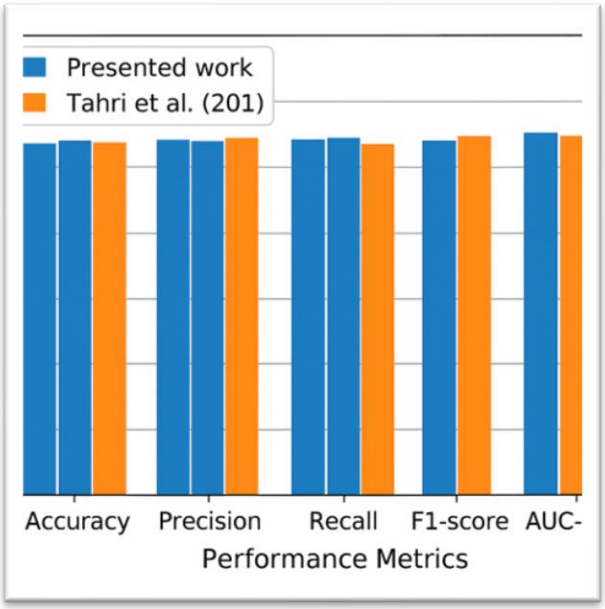

Figure 6.Performance Comparison of IDS Classifiers Between the Presented Work and Tahri et al. [14]

Interpretability Analysis using SHAP and LIME: Knowing the reason behind a prediction is critical in high stakes domains such as cybersecurity where accuracy and interpretability of a decision are required. Interpretability techniques such as SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are applied to the trained Random Forest classifier in order to achieve this result. These tools offer global as well as local perspectives on feature importance and allowing analysts to follow the reasoning process of the model.

SHAP Interpretability, SHAP adopts a unified approach to explain individual predictions, rooted in cooperative game theory, but under the assumption that each feature value can be

Received: August 05 2025

International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

considered a "player," all working together to make the final prediction. SHAP value of a feature is the average contribution of that feature, over all possible combinations of features. This is what makes SHAP particularly attractive in the space of cybersecurity, as this community requires a transparent model for supporting analysts' decisions and policy compliance.

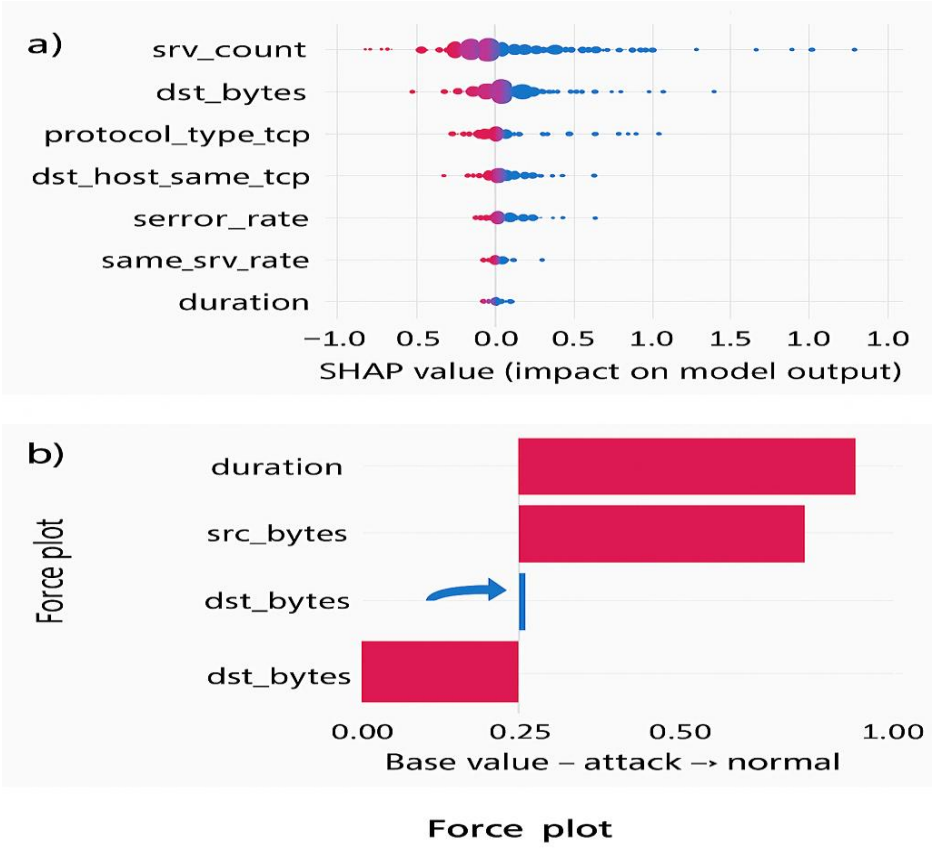

Figure 7.SHAP-Based Interpretability of IDS Models

Figure 7 (a) shows the SHAP summary plot for the Random Forest model built on the NSL-KDD dataset. In the plot above, all of the input features are ranked by their mean absolute SHAP values, which in effect represents how much influence the feature has on the model's output. Features with great SHAP values make a greater impact on classification. The following some are the prominent features immeasurable:

srv_count: Number of connections to the same service. High scores indicate suspicious repetitive activity and are usually indicative of probe or DoS attacks. As an example, in the summary plot red dots indicate that when srv_ount has a high value there is increased probability of being classified with an "attack". dst_bytes: Number of bytes from destination to source. Lower values can indicate failed or blocked responses, which are possibly the result of probing. duration and src_bytes: Also show significant effect, especially in separating normal session from highclass or bad connections.

Received: August 05 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

protocol_type_tcp: This is in terms of one hot encoding, it normally has lower SHAP value indicating that gaining most decision relies on surrounding other parameter value for that context.

The color shift in the SHAP plot also indicates which value of each feature is affecting the prediction. Red points (high feature values) on the right side (Positive SHAP values) push the model to be more confident in classifying the observation as an attack, whereas blue points (lower feature values) pull predictions towards normal. Figure 7(b) presents the SHAP force plot for one test instance, which predicted as an attack. The force plot decodes the model's baseline prediction (average value across the dataset) and updates it with each SHAP value of a feature moving it closer to the final prediction. For the chosen attack example:

The positive SHAP contributions are driven by a few features such as duration, src_bytes, and srv_count which strongly push the prediction score upward the decision threshold indicating that the record is more likely to be assigned to the attack type. Negative 'features' such as low dst_bytes value provides a weak counter- weight against the overall classification of the attack. This finer-grained view lets us check which exact features and values "tipped" the model's decision. SHAP based interpretability like this is in line with the results obtained by Mohale et al. [40], who also showed the effectiveness of SHAP for diagnosing the behavior of the IDS model on the NSL-KDD.

LIME Interpretability, to complement the SHAP insights, we make use of the LIME algorithm on the same instances. SHAP offers a global and local explanation of the model behavior by attributing importance to each feature both locally and globally, while LIME concentrates only on local fidelity by computing a sparse linear approximation of the model locally around the single prediction. A sample LIME explanation result for one typical test instance is listed in Table 3. LIME identifies the most important features that drove the model's prediction for that instance:

Table 3.LIME feature weights

FEATURE	WEIGHT	IMPACT ON DECISION
SRV_COUNT	+0.95	Pushes towards attack
DST_HOST_COUNT	+0.87	Pushes towards attack
SERVICE=HTTP	-0.76	Pushes towards normal
FLAG=SF	-0.65	Pushes towards normal

The negative weights (such as service=http, flag=SF) indicate that these terms detracted from predicting the "anomaly" class. Positive weights (such as "srv_count", "dst_host_count") contributed to the possibility to the sample to be classified as "attack". The intuition-based outputs of LIME make it easier for analysts to quickly understand why a decision was made,

Received: August 05 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

e.g., in terms of pseudo-rules such asif srv_count is high and dst_host_count is also elevated then the session is likely an attack unless service=http and flag=SF.

While SHAP (with its grounding in additive game theory) and LIME (with its focus on local linear surrogates) are dissimilar methods, both of the interpretability tools suggested the importance of components such as srv_count, providing more trust in the model's learned behaviour. SHAP offered uniform global trends together with fine details on the local level, as LIME produced simplified rules.

Taken together, SHAP and LIME represent a powerful pair of complementary view of the model. SHAP explains what are the important features globally and why predictions are made in a specific way locally. LIME also returns another human-interpretable explanation for a specific instance. Security operations can particularly benefit from these insights as analysts need to comprehend alerts in a timely manner to respond. Being able to link an alert back to arguments such as high connection counts or questionable traffic patterns provides an operational trust – and results in our being able to qualify or disqualify false positives.

In summary, our results are consistent with the previous work that the trust of IDS models can be improved with the introduction of interpretability methods. Using that knowledge, analysts can triage the alerts more efficiently— for instance, if Offensive SHAP indicates high connection counts as a sign of an attack, they might investigate potential port scanning efforts or denial of service attacks [41]

This increased transparency builds more trust in AI-based cyber systems, and overcomes the problem of black-box models in the delicate balance between detecting correctly and understanding what gets detected.

Moreover, these results also support prior work that highlights the importance of transparency in cybersecurity AI systems . Through revealing black-box decision-making, SHAP and LIME enhance explainability and enforcement of ethical deployment and monitoring in cybersecurity infrastructures.

4 ETHICAL CONSIDERATIONS IN CYBERSECURITY

The use of interpretable ML in cybersecurity brings up the very important ethical concerns. Firstly, users' privacy is a problem. Inherently, XAI methods take into account the input of a model, and in case of IDS, network traffic features. As Olasehinde [42] explains, this approach may inadvertently leak private data about the users. For instance, if an XAI tool reports that “our explanation flagged this because the user’s email domain is unusual,” this could entail a leak of confidential metadata [42]. Likewise, anomaly detection via network flows can be used to capture personal information without consent. Striking a balance between transparency and privacy is crucial: we need to avoid explanations that expose raw sensitive content. If not, one may need to use feature refinement or compliance with data protection laws (GDPR) as proposed in existing work in the literature.

And second, there’s always a trade-off between transparency and security. Although explanations inspire trust for defenders, they might be abused if acquired by attackers.

Received: August 05 2025

International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Olasehinde[42] and other researchers [43] warn that lifting the veil from how a model arrives at a decision also can provide attackers with details that could allow them to sidestep detection.

For example, if an insider infers alarms are raised for high srv_count, they may choose to blend into the normal relationship of some services. Consequently, access to XAI-system capabilities should be carefully regulated: only technical reports might be made available for security professionals, while summarized presentations would be exposed to the operators. Implementing an ethical deployment is getting gauge how much detail and to whom, as [42] suggested.

Third is the potential for bias and fairness. Even understandable models can obfuscate the actual bias in the training data (e.g., an overrepresentation of specific IP ranges or services). XAI could potentially pinpoint and such biases (e.g., always classifying a service as malicious) [44] .We need to also ask whether the models disproportionately frame particular users or behaviors. Transparency can assist in this area, but designers need to actively verify and address bias.

And finally, the regulatory and societal effects count. There are so many workloads that need to be auditable: IDS decisions should be understandable for compliance. Conversely, an overbearing amount of transparency could compromise an institution's security by making the internal cogs visible. Normative guidelines recommend a control (an ethics board as XAI user oversight) be put into place [42]. In short: Explainable models are one of the best ways to inspire trust and enforce accountability, but we must delicately balance all privacy (keeping user data private), no adversarial assistance (no giving attackers' mechanism to design adversarial inputs), reasonable oversight.

Restricted-access explanation dashboards, which guarantee that only trusted analysts have access to important feature attributions, can be used to reduce the dangers of adversarial usage of XAI. Furthermore, using privacy-preserving strategies like federated IDS training or differential privacy could lessen the amount of sensitive traffic data that leaks while preserving explainability. These guidelines are in line with new suggestions made by regulatory frameworks like the NIST AI Risk Management Framework and the GDPR.

5 CONCLUSION

The escalating volume and sophistication of cyber threats underscore the critical need for Intrusion Detection Systems (IDS) that are not only accurate but also transparent and interpretable. This research directly addresses this imperative by proposing a robust framework that integrates Decision Tree (DT) and Random Forest (RF) classifiers with SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations). This synergy creates a powerful and comprehensible system for intrusion detection, bridging the gap between precise threat identification and actionable, human-interpretable decision support.

Our experimental validation confirms the efficacy of the proposed models. The Random Forest Classifier (RFC) achieved an impressive detection accuracy exceeding 98%, coupled with strong precision and recall rates, indicating its proficiency in accurately identifying both normal and anomalous network traffic. While the Decision Tree (DT) classifier demonstrated slightly

Received: August 05 2025

Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

lower but still competitive accuracy, its inherent ability to provide clear and interpretable rules offers significant practical utility in real-world scenarios. These results highlight that a judicious balance between accuracy and interpretability can be achieved by carefully controlling the depth of decision trees.

The integration of SHAP and LIME substantially enhanced model interpretability. SHAP provided both global insights, revealing the most influential features across the dataset, and local, per-prediction explanations, attributing individual outcomes to specific input features. LIME complemented this by offering simplified, localized explanations for single decisions. This combined explainability empowered analysts to understand not just *what* the model was predicting, but *why*, fostering greater trust, facilitating auditing processes, and improving response effectiveness.

The implications of these findings are profound for cybersecurity operations. Enhanced transparency leads to more rapid navigation of alarms, minimization of false positives, and improved handling of data access anomalies. Crucially, this framework fosters a stronger human-AI partnership, enabling domain experts without extensive machine learning backgrounds to better comprehend model outputs. This interpretability facilitates more informed and agile responses to cyber threats.

Future work will focus on integrating this approach into real-time operational environments, evaluating its performance on noisy datasets, and incorporating adaptive learning techniques. Furthermore, we plan to delve deeper into the ethics of transparency, particularly the intricate trade-offs between explainability, data privacy, and the adversarial robustness of AI systems. In conclusion, this paper makes three contributions: In order to provide both global and local transparency, (i) DT and RF were used to develop a lightweight yet interpretable IDS framework; (ii) complementary XAI techniques (SHAP and LIME) were integrated; and (iii) operational and ethical considerations were incorporated for responsible deployment. This study is among the first to systematically evaluate tree-based interpretable IDS on the NSL-KDD dataset because of these contributions. Ultimately, this study offers a pragmatic, forward-looking, and responsible contribution to the development of reliable, efficient, and explainable cybersecurity systems.

Statements and Declarations

Competing Interests:

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Author Contributions:

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Majed. S. and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Received: August 05 2025

Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

REFERENCES

[1] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," *IEEE Symposium on Computational Intelligence for Security and Defense Applications, CISDA 2009*, Dec. 2009, doi: 10.1109/CISDA.2009.5356528.

[2] A. EFE and İ. N. ABACI, "Comparison of the Host Based Intrusion Detection Systems and Network Based Intrusion Detection Systems," *Celal Bayar Üniversitesi Fen Bilimleri Dergisi*, vol. 18, no. 1, pp. 23–32, Mar. 2022, doi: 10.18466/CBAYARFBE.832533.

[3] "[2306.09451] Host-Based Network Intrusion Detection via Feature Flattening and Two-stage Collaborative Classifier." Accessed: Jun. 12, 2025. [Online]. Available: https://arxiv.org/abs/2306.09451

[4] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv Neural Inf Process Syst*, vol. 2017-December, pp. 4766–4775, May 2017, Accessed: Jun. 11, 2025. [Online]. Available: https://arxiv.org/pdf/1705.07874

[5] V. Z. Mohale and I. C. Obagbuwa, "Evaluating machine learning-based intrusion detection systems with explainable AI: enhancing transparency and interpretability," *Front Comput Sci*, vol. 7, p. 1520741, May 2025, doi: 10.3389/FCOMP.2025.1520741.

[6] F. Charmet *et al.*, "Explainable artificial intelligence for cybersecurity: a literature survey," *Annales des Telecommunications/Annals of Telecommunications*, vol. 77, no. 11–12, pp. 789–812, Dec. 2022, doi: 10.1007/S12243-022-00926-7/TABLES/3.

[7] V. Z. Mohale and I. C. Obagbuwa, "Evaluating machine learning-based intrusion detection systems with explainable AI: enhancing transparency and interpretability," *Front Comput Sci*, vol. 7, p. 1520741, May 2025, doi: 10.3389/FCOMP.2025.1520741.

[8] "A novel machine learning-based artificial intelligence approach for log analysis using blockchain technology ," Sigma Journal of Engineering and Natural Sciences. Accessed: Jun. 12, 2025. [Online]. Available: https://sigma.yildiz.edu.tr/storage/upload/pdfs/1726036942-en.pdf?utm_source=perplexity

[9] V. Z. Mohale and I. C. Obagbuwa, "Evaluating machine learning-based intrusion detection systems with explainable AI: enhancing transparency and interpretability," *Front Comput Sci*, vol. 7, p. 1520741, May 2025, doi: 10.3389/FCOMP.2025.1520741.

[10] R. Tahri, A. Lasbahani, A. Jarrar, and Y. Balouki, "Intelligent Intrusion Detection Using Decision Trees and the NSL-KDD Dataset: An All-Inclusive Method for Cyber Attack Detection," *Journal of Southwest Jiaotong University*, vol. 59, no. 5, 2024, doi: 10.35741/ISSN.0258-2724.59.5.13.

[11] T. T. H. Le, H. Kim, H. Kang, and H. Kim, "Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method," *Sensors 2022, Vol. 22, Page 1154*, vol. 22, no. 3, p. 1154, Feb. 2022, doi: 10.3390/S22031154.

Received: August 05 2025

Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[12] D. A. Salem, M. H. Moharam, and E. M. Hashem, "Development of Machine Learning Regression Models for Predicting the Performance of Nanofibrous Scaffolds for Skin Tissue Engineering," *J BioX Res*, vol. 7, 2024, doi: 10.34133/JBIOXRESEARCH.0008.

[13] Haripriya C1 and P. J. M. P. LastName, "An Explainable and Optimized Network Intrusion Detection Model using Deep Learning." Accessed: Jun. 12, 2025. [Online]. Available: https://thesai.org/Downloads/Volume15No1/Paper_45-An_Explainable_and_Optimized_Network_Intrusion_Detection.pdf

[14] B. Ingre, A. Yadav, and A. K. Soni, "Decision tree based intrusion detection system for NSL-KDD dataset," *Smart Innovation, Systems and Technologies*, vol. 84, pp. 207–218, 2018, doi: 10.1007/978-3-319-63645-0_23.

[15] P. A. A. Resende and A. C. Drummond, "A survey of random forest based methods for intrusion detection systems," *ACM Comput Surv*, vol. 51, no. 3, Jul. 2018, doi: 10.1145/3178582.

[16] S. Mane and D. Rao, "Explaining Network Intrusion Detection System Using Explainable AI Framework," Mar. 2021, Accessed: Jun. 12, 2025. [Online]. Available: https://arxiv.org/pdf/2103.07110

[17] T. T. H. Le, H. Kim, H. Kang, and H. Kim, "Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method," *Sensors 2022, Vol. 22, Page 1154*, vol. 22, no. 3, p. 1154, Feb. 2022, doi: 10.3390/S22031154.

[18] S. W. Sajid, K. M. Rashid Anjum, M. Al-Shaharia, and M. Hasan, "Investigating Machine Learning Algorithms with Model Explainability for Network Intrusion Detection," *Cyber Security and Business Intelligence: Innovations and Machine Learning for Cyber Risk Management*, pp. 121–1236, Jan. 2023, doi: 10.4324/9781003285854-8/INVESTIGATING-MACHINE-LEARNING-ALGORITHMS-MODEL-EXPLAINABILITY-NETWORK-INTRUSION-DETECTION-SAD-WADI-SAJID-RASHID-ANJUM-MD-AL-SHAHARIA-MAHMUDUL-HASAN.

[19] "IEEE Xplore Full-Text PDF:" Accessed: Jun. 12, 2025. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9875264

[20] I. García-Magariño, J. Bravo-Agapito, and R. Lacuesta, "Cybersecure XAI Algorithm for Generating Recommendations Based on Financial Fundamentals Using DeepSeek," *AI 2025, Vol. 6, Page 95*, vol. 6, no. 5, p. 95, May 2025, doi: 10.3390/AI6050095.

[21] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," *IEEE Symposium on Computational Intelligence for Security and Defense Applications, CISDA 2009*, Dec. 2009, doi: 10.1109/CISDA.2009.5356528.

[22] E. M. Hashem and D. A. Salem, "Transfer Learning and Machine Learning With MRI Radiomics For Alzheimer's Disease Diagnosis," *National Radio Science Conference, NRSC, Proceedings*, pp. 243–251, 2024, doi: 10.1109/NRSC61581.2024.10510543.

Received: August 05 2025

Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[23] R. Tahri, A. Lasbahani, A. Jarrar, and Y. Balouki, "Intelligent Intrusion Detection Using Decision Trees and the NSL-KDD Dataset: An All-Inclusive Method for Cyber Attack Detection," *Journal of Southwest Jiaotong University*, vol. 59, no. 5, 2024, doi: 10.35741/ISSN.0258-2724.59.5.13.

[24] T. T. H. Le, H. Kim, H. Kang, and H. Kim, "Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method," *Sensors 2022, Vol. 22, Page 1154*, vol. 22, no. 3, p. 1154, Feb. 2022, doi: 10.3390/S22031154.

[25] F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, Accessed: Jun. 12, 2025. [Online]. Available: http://scikit-learn.sourceforge.net.

[26] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," *IEEE Symposium on Computational Intelligence for Security and Defense Applications, CISDA 2009*, Dec. 2009, doi: 10.1109/CISDA.2009.5356528.

[27] K. Pal and B. V. Patel, "Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques," *Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020*, pp. 83–87, Mar. 2020, doi: 10.1109/ICCMC48092.2020.ICCMC-00016.

[28] C. Haripriya and J. M. P. Prabhudev, "An Explainable and Optimized Network Intrusion Detection Model using Deep Learning," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 1, pp. 482–488, 2024, doi: 10.14569/IJACSA.2024.0150145.

[29] T. Su, H. Sun, J. Zhu, S. Wang, and Y. Li, "BAT: Deep Learning Methods on Network Intrusion Detection Using NSL-KDD Dataset," *IEEE Access*, vol. 8, pp. 29575–29585, 2020, doi: 10.1109/ACCESS.2020.2972627.

[30] S. Widodo, H. Brawijaya, and S. Samudi, "Stratified K-fold cross validation optimization on machine learning for prediction," *Sinkron*, vol. 7, no. 4, pp. 2407–2414, Oct. 2022, doi: 10.33395/SINKRON.V7I4.11792.

[31] C. Kumar, G. Walton, P. Santi, and C. Luza, "Random Cross-Validation Produces Biased Assessment of Machine Learning Performance in Regional Landslide Susceptibility Prediction," *Remote Sens (Basel)*, vol. 17, no. 2, p. 213, Jan. 2025, doi: 10.3390/RS17020213/S1.

[32] "(PDF) A Forest of Possibilities: Decision Trees and Beyond." Accessed: Jun. 13, 2025. [Online]. Available: https://www.researchgate.net/publication/374118383_A_Forest_of_Possibilities_Decision_Trees_and_Beyond

Received: August 05 2025

Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[33] "(PDF) A review of technical factors to consider when designing neural networks for semantic segmentation of Earth Observation imagery." Accessed: Jun. 13, 2025. [Online]. Available: https://www.researchgate.net/publication/373245500_A_review_of_technical_factors_to_consider_when_designing_neural_networks_for_semantic_segmentation_of_Earth_Observation_imagery

[34] A. M. Shetty, M. F. Aljunid, D. H. Manjaiah, and A. M. S. Shaik Afzal, "Hyperparameter Optimization of Machine Learning Models Using Grid Search for Amazon Review Sentiment Analysis," *Lecture Notes in Networks and Systems*, vol. 821, pp. 451–474, 2024, doi: 10.1007/978-981-99-7814-4_36.

[35] "(PDF) COMPARISON OF SUPPORT VECTOR MACHINES, RANDOM FOREST AND DECISION TREE METHODS FOR CLASSIFICATION OF SENTINEL - 2A IMAGE USING DIFFERENT BAND COMBINATIONS." Accessed: Jun. 13, 2025. [Online]. Available: https://www.researchgate.net/publication/346776010_COMPARISON_OF_SUPPORT_VECTOR_MACHINES_RANDOM_FOREST_AND_DECISION_TREE_METHODS_FOR_CLASSIFICATION_OF_SENTINEL_-_2A_IMAGE_USING_DIFFERENT_BAND_COMBINATIONS

[36] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," *Brain Inform*, vol. 11, no. 1, p. 10, Dec. 2024, doi: 10.1186/S40708-024-00222-1.

[37] P. Yasin *et al.*, "The Potential of a CT-Based Machine Learning Radiomics Analysis to Differentiate Brucella and Pyogenic Spondylitis," *J Inflamm Res*, vol. 16, pp. 5585–5600, 2023, doi: 10.2147/JIR.S429593.

[38] D. Gaspar, P. Silva, and C. Silva, "Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron," *IEEE Access*, vol. 12, pp. 30164–30175, 2024, doi: 10.1109/ACCESS.2024.3368377.

[39] W. Yang *et al.*, "Survey on Explainable AI: From Approaches, Limitations and Applications Aspects," *Human-Centric Intelligent Systems 2023 3:3*, vol. 3, no. 3, pp. 161–188, Aug. 2023, doi: 10.1007/S44230-023-00038-Y.

[40] V. Z. Mohale and I. C. Obagbuwa, "Evaluating machine learning-based intrusion detection systems with explainable AI: enhancing transparency and interpretability," *Front Comput Sci*, vol. 7, 2025, doi: 10.3389/FCOMP.2025.1520741.

[41] V. Z. Mohale and I. C. Obagbuwa, "A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity," *Front Artif Intell*, vol. 8, p. 1526221, Jan. 2025, doi: 10.3389/FRAI.2025.1526221/BIBTEX.

Received: August 05 2025

Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[42] "(PDF) Ethical Considerations in Explainable AI for Cybersecurity." Accessed: Jun. 13, 2025. [Online]. Available: https://www.researchgate.net/publication/387224493_Ethical_Considerations_in_Explainable_AI_for_Cybersecurity

[43] D. Mhlanga, "Industry 4.0 in Finance: The Impact of Artificial Intelligence (AI) on Digital Financial Inclusion," *International Journal of Financial Studies 2020, Vol. 8, Page 45*, vol. 8, no. 3, p. 45, Jul. 2020, doi: 10.3390/IJFS8030045.

[44] V. Z. Mohale and I. C. Obagbuwa, "Evaluating machine learning-based intrusion detection systems with explainable AI: enhancing transparency and interpretability," *Front Comput Sci*, vol. 7, p. 1520741, May 2025, doi: 10.3389/FCOMP.2025.1520741/BIBTEX.

Received: August 05 2025