



Explainable AI for Automated Threat Hunting in Large-Scale IoT Ecosystems

¹Smart Idima, ²Philip Nwaga, ³Patrick Evah

Department of Computer Sciences, Western Illinois University, Macomb Illinois USA,

Abstract: The growing complexity and scale of Internet of Things (IoT) ecosystems demand robust and interpretable security solutions to detect and mitigate emerging threats. This research integrates Explainable AI (XAI) techniques, including SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), into machine learning models such as Random Forest and SVM to enhance transparency and trust in IoT threat detection systems. A distributed processing architecture, utilizing Apache Spark and Kafka, ensures real-time scalability for processing large volumes of heterogeneous IoT data. XAI methods provided actionable insights by identifying key features influencing predictions, significantly reducing false positives and improving analyst response times. The study aligns with regulatory requirements like GDPR, offering a scalable, interpretable framework for IoT security. However, challenges such as computational overheads and adversarial risks highlight areas for future research.)

Keywords—IoT Security, Explainable AI, SHAP, LIME, Anomaly Detection, Real-Time Processing, Apache Spark, Apache Kafka, Machine Learning, Cybersecurity Compliance.

INTRODUCTION

The Internet of Things (IoT) has emerged as one of the most transformative technological advancements of the 21st century, seamlessly connecting billions of devices across the globe. IoT devices extend across different domains including intelligent household attributes such as smart thermostats alongside manufacturing process sensors and medical device vital sign trackers and smart municipal systems for managing traffic and power resources. Real-time data exchange coupled with intelligent automation has transformed industrial operations through IoT technology thus delivering extraordinary efficiency along with consumer convenience.

The accelerated growth of IoT devices produces many security issues across interconnected networks. Due to its decentralized and heterogeneous structure IoT ecosystems create multiple entry points for attackers to exploit. The constrained computational capabilities of IoT devices result in insufficient capabilities to implement thorough security mechanisms. The devices produce extensive data streams which contain sensor readings together with network monitoring records and operational device activity metrics.

Massive quantities of diverse data featuring various format combinations, unique structural patterns, and different operational speeds necessitate sophisticated anomaly detection tools to identify security threats (Shashi Rekha, 2021).

The main challenge in managing with IoT arises from its chaotic structure and extensive dimensions, making analysis and processing rather challenging. Real-time monitoring of streaming sensor data is essential as analysts must detect unusual patterns and anomalous data sets. Network traffic logs exhibit distinct characteristics while showing significant variations influenced by user actions and environmental factors. Along with the necessity for flexible, scalable processing solutions is the need to handle both streaming and batch data types because of the variability in data characteristics. The inefficacy of data processing and analysis within IoT systems exposes them to security threats such as Distributed Denial of Service (DDoS) attacks, data breaches, and ransomware infections. The Mirai botnet attack stands out as a notable example among many others that illustrates how compromised IoT devices can be utilized to disable entire networks (Shashi Rekha, 2021).

Due to the current challenges, prioritizing security measures for IoT systems is essential. The protection of IoT settings against evolving cyber threats relies mainly on real-time anomaly detection abilities and threat identification mechanisms. Achieving success in IoT implementations necessitates the rapid and accurate analysis of a broad spectrum of data, ensuring scalability to uphold system integrity and reliability. The pressing need for robust security solutions highlights the necessity of incorporating advanced machine learning methods with explainable AI frameworks into IoT cybersecurity systems.

Limitations of Traditional AI Models

The field of IoT systems currently utilizes conventional black-box machine learning models to detect anomalies while classifying potential threats. However, they exhibit several limitations when applied to IoT data:

1. **Opacity:** Black-box models deliver outstanding prediction accuracy although these systems remain resistant to interpretation. Security teams lack the ability to understand how the system reaches its predictions when operating with complex IoT datasets (Walton, 2018).
2. **Data Variety:** Standard ML models struggle to generalize across hybrid IoT data types (numbers and categories with text-based data) until advanced feature engineering is applied (Shashi Rekha, 2021).
3. **Scalability Issues:** When integrating traditional ML models with IoT data at large scale and multi-format the needed computational power dramatically increases (ANNA NAMRITA GUMMADI, 2024).

When a model provides unclear reasons for generating anomaly alerts analysts avoid relying on its indicators thus delaying incident responses together with increased maintenance expenses.

Motivation for XAI in IoT Cybersecurity

Explainable AI (XAI) methods tackle these issues by offering understandable explanations for AI choices. Utilizing techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), XAI can improve IoT threat detection systems in numerous ways:

1. **Transparency for Data Insights:** XAI emphasizes particular characteristics of IoT data (e.g., atypical network packet size, irregular temperature readings) that are associated with an anomaly. This enables analysts to identify and confirm threats.
2. **Feature Relevance:** Methods such as SHAP can prioritize features according to their influence on the model's choices, which is especially beneficial for diverse IoT data streams where specific factors (e.g., latency, bandwidth consumption) might better signal threats.
3. **Regulatory Compliance** With the rise of IoT implementations in essential industries such as healthcare and finance, adhering to data privacy regulations (e.g., GDPR) demands clear decision-making procedures (ANNA NAMRITA GUMMADI, 2024).

Data Overview

The dataset utilized in this study is made up of data generated by IoT, which includes features essential for identifying anomalies:

1. **Sensor Data:** Characteristics such as temperature, humidity, and device status, offer an understanding of physical device performance (astral_fate, 2023).
2. **Network Traffic Logs:** Factors like packet sizes, response times, and types of protocols serve as crucial indicators of possible security risks.
3. **Device Usage Patterns:** Records showing how devices connect to the network, encompassing data transfer speeds and connection statuses.

The dataset contains tagged examples of "normal" and "abnormal" behaviors, facilitating supervised learning for classifying threats. Data preprocessing includes managing absent values, standardizing numerical attributes, and generating feature vectors for machine learning models. The objective is to utilize this data to train models that can detect malicious behavior instantly.

Research Objectives

This study combines Explainable AI methods with machine learning models to manage and examine IoT data, with the goal of::

1. **Improve Detection Accuracy:** By refining models to manage various IoT data types, encompassing both sensor data and network traffic records.
2. **Foster Trust and Usability:** Employing XAI to clarify the importance of data attributes in anomaly detection, allowing analysts to have confidence in model predictions (ANNA NAMRITA GUMMADI, 2024).
3. **Ensure Scalability:** Employ distributed processing frameworks such as Apache Spark to examine IoT data streams while maintaining speed and precision.
4. **Enhance Adversarial Robustness:** Tackling possible adversarial assaults by incorporating methods such as differential privacy and adversarial training (Walton, 2018).

The dataset from this study serves as the basis for creating and assessing scalable, interpretable ML models designed for the diverse and evolving characteristics of IoT ecosystems. By concentrating on the characteristics of real-world IoT data, the study seeks to illustrate how XAI can shift IoT cybersecurity from a reactive to a proactive approach.

LITERATURE REVIEW

A transformative technology known as the Internet of Things (IoT) connects billions of devices via the Internet to gather and share data while taking actions based on the collected information. The Internet of Things has revolutionized digital interactions via its smart home innovations and industrial automation initiatives, offering enhanced efficiency alongside convenience and automated decision-making support. The reliable functioning of IoT systems relies on a robust architectural structure that integrates diverse elements into one system while transmitting data across several layers.

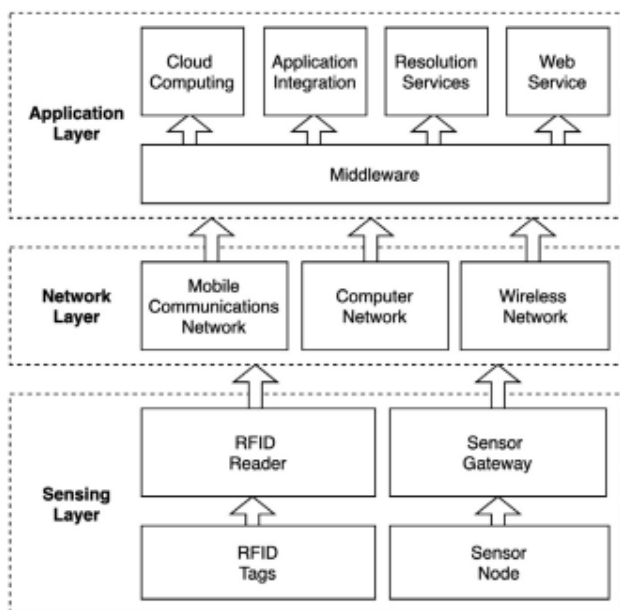


Figure 1: Three-layer architecture of IoT (Eryk Schiller, 2022)

The three-tier IoT architecture integrates the Sensing Layer with the Network Layer and the Application Layer, as shown in Figure 1. Essentially, the Sensing Layer is made up of hardware elements such as RFID tags and RFID readers in addition to sensor nodes and sensor gateways. This is where data concerning real-world metrics such as temperature, humidity, or object identification appears. The integration of RFID tags and readers facilitates inventory management and logistics tasks, while sensor nodes collect environmental data that sensor gateways compile and preprocess. This layer facilitates the physical link between IoT systems and the environment, enabling precise real-time data gathering.

Data transitions from the sensing layer to the application layer by employing the Network Layer as its primary transmission backbone. The Network layer integrates resources across mobile communication routes (including 4G and 5G) with computer access protocols (involving Ethernet and Wi-Fi), along with wireless technologies such as Zigbee, Bluetooth, and LoRaWAN. These networks ensure security and efficiency for IoT device connections via large-scale deployments by enabling fast data flow and scalability options. By

bridging both physical and digital realms, the network layer facilitates complete communication synergy and application interoperability across the IoT framework.

At the top is the Application Layer, which offers end-users significant information through data transmission. Cloud computing services, combined with application integration solutions for enterprise platform connectivity, provide the foundation of this layer, alongside web services that enable remote device monitoring and management. Application services utilize middleware to carry out intermediary functions such as data aggregation, filtering, and communication management between the lower layers and the application services. Through the application layer IoT systems generate actionable outputs that facilitate informed choices within medical fields and factory settings as well as urban communities.

The three-layer system architecture creates an adaptable IoT platform for various applications through modular design and scalable implementation which delivers reliable performance across different domains. The layered structure of IoT operational design enables adaptation to technological advancements while satisfying the increasing complexity within IoT environments.

Existing AI Approaches in IoT Security

IoT ecosystems continue to advance at a rapid pace which requires advanced applications of AI and machine learning (ML) for building secure security frameworks. Both self-learning and unlearning AI methods together with decision trees and support vector machines and clustering methods compose the standard AI framework for IoT security applications. Models in these categories assist with anomaly detection within IoT-generated datasets which includes network traffic and device logs. IoT security applications use unsupervised techniques such as K-means clustering to spot anomalous patterns without needing labeled data together with supervised models including random forests and SVMs that provide effective malicious activity classification (Eryk Schiller, 2022).

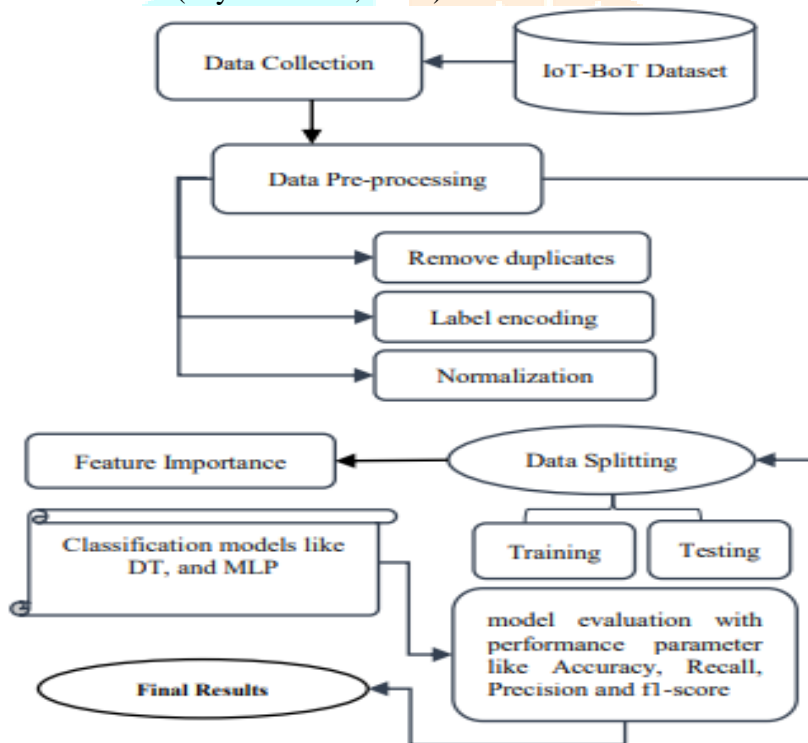


Figure 2: Flowchart of the IoT Botnet Attack Detection System (Gopalsamy, 2020)

Figure 2 shows a structured procedure to apply machine learning techniques to IoT security by utilizing the IoT-Bot dataset. The critical initial step for IoT security requires collecting data from devices which produce large varied datasets encompassing network traffic alongside device activity logs. The IoT-Bot dataset contains marked instances of regular network operations together with destructive conduct through which machine learning and evaluation of models occurs. The effective processing of these datasets remains essential because IoT datasets frequently include data errors alongside inconsistencies. A multi-stage process involving duplicate removal and label transformation and normalization manages feature enhancement. Preparing the data through these steps ensures it is set for interpreting machine learning accuracy.

Feature importance identifies crucial indicators for harmful activities by choosing pertinent attributes from protocols, packets, and source IP addresses. The simplified models preserve clear understanding while emphasizing crucial anomaly detection features that show significant influence. Data elements chosen for analysis are divided for training uses and testing functions. While training, the algorithm utilizes the medical

data from the training subset to develop its predictive models, and then the testing subset validates their performance on data points that were not part of the training. Precise data splitting fulfills two essential functions in IoT security by preventing overfitting and equipping models with robust generalization capabilities for real-time use.

Classification models executed via Decision Trees (DT) and Multi-Layer Perceptrons (MLP) function within the workflow framework. Decision Trees continue to be effective for smaller IoT datasets as they offer straightforward interpretations and provide rapid insights without complex computations. When utilized as models based on neural networks, MLPs show improved abilities to handle complex threats within extensive datasets. Decision Trees are easier to understand compared to other models, but they might require additional tuning efforts to reach the same level of clarity. Every model is evaluated based on accuracy metrics influenced by precision, recall, and the F1-score. The capacity to identify harmful actions while maintaining low false positives and false negatives is a crucial factor for assessing the effectiveness of these security models in IoT settings.

The analytical procedure concludes by synthesizing results to identify the optimal models and attributes for IoT threat detection capabilities. The enhancement process for the detection pipeline commenced with this step to attain scalable precision and deployment effectiveness. The workflow illustrates elements of conventional AI methods such as supervised learning, feature engineering, and evaluation techniques, while also highlighting how these approaches limit IoT security functions within extensive information systems. The conventional decision trees and multilayer perceptron models exhibit restrictions when handling real-time data streams that include intricate high-dimensional IoT datasets. The demand for sophisticated Explainable AI (XAI) methods arises as an essential need because of the persistent difficulty in improving transparent and scalable security solutions for IoT (Gopalsamy, 2020).

These traditional methods face challenges when used on IoT data due to difficulties associated with high-dimensional IoT data streams and their unstructured patterns. The current models' scalability and ability to process in real-time are insufficient to accommodate large-scale IoT networks, as these networks frequently demand such functionalities (Petar Radanliev, 2024) (Iqbal H. Sarker, 2023). Due to their "black-box" operating principle security analysts experience difficulty both understanding how the model operates and trusting its decisions.

Explainable AI Techniques

Explainable AI (XAI) techniques offer structural frameworks, enhancing comprehension and reliability of intricate predictive models. Some of the most notable methods include:

1. **SHAP (SHapley Additive exPlanations):** SHAP applies game theory concepts to evaluate the influence of each feature on the model's prediction outcomes. A Shapley value framework provides numerical measures for features while generating clear explanations of model results. The approach offers crucial benefits for scenarios that necessitate comprehension of feature influence, like anomaly detection in IoT networks (Ahmed M. Salih, 2025).
2. **LIME (Local Interpretable Model-Agnostic Explanations):** LIME creates comprehensible approximation models for system predictions that function locally. The method alters input data components to observe variations in output results, offering insights into how specific inputs influence system predictions. The LIME system has helped analyze real-time IoT security choices including device authentication together with anomaly detection in recent studies (Alex Gramegna, 2021).
3. **Attention-Based Models:** the decision-making processes of neural networks utilize attention mechanisms that highlight significant elements of the input data. Due to their proficiency in managing various IoT data streams, such as network packets and time-series sensor data, attention-based models effectively process sequential data flows by concentrating on crucial data parts (SENTHIL KUMAR JAGATHEESAPERUMAL, 2022).
4. **Edge XAI:** Edge XAI combines explainability resources with edge computing frameworks to implement interpretable machine learning models close to IoT devices. The design integrates rapid functioning with real-time model decision contextual information for use cases requiring immediate outcomes in healthcare and industrial IoT settings (SENTHIL KUMAR JAGATHEESAPERUMAL, 2022).

Metrics	SHAP	LIME
Concept	Applies to the model as-is	Fits a local surrogate model to explain the complex model
Theory	Additive feature attribution based on game theory	Feature perturbation method
Type	Post-hoc model-agnostic	
Data type	Images, tabular data, and signals	
Explanation	Global, local	Local
Collinearity consideration	Not in the original method	No
Nonlinear decision	Depends on the used model	Incapable
Computing time	Higher	Lower
Visualization	Waterfall, beeswarm, and summary plots	One single plot

Table 1: Comparison between SHAP and LIME (Ahmed M. Salih, 2025)

Table 1 provides a comprehensive examination of SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), showcasing their unique benefits and functional limitations using various interpretability metrics. The SHAP method directly interacts with original machine learning models to elucidate features through additive feature attribution approaches based on game theory. Every feature is given a precise mathematical representation in this method, ensuring an equitable evaluation of their contributions. LIME creates a straightforward linear regression model as a substitute to illustrate particular prediction regions within intricate decision boundaries. As a result LIME operates by analyzing single predictions while it does not work towards understanding the entire model holistically.

These interpretation methods differ based on their underlying theoretical principles. VideoLink integrates SHAP's additive feature attribution framework that produces reliable explanations at multiple granularity levels thus making it adaptable for whole-system explanation as well as single-instance predictions evaluation. With feature perturbation LIME creates localized input space explanations through its approximation of feature importance. SHAP demonstrates flexibility by working with images and tabular data and signals yet LIME is focused on tabular data and text which requires specific attention to work with images. LIME and SHAP do not inherently consider multivariate relationships between features but extensions to SHAP provide solutions to this deficiency.

SHAP effectiveness for handling nonlinear decision boundaries depends on the model sophistication and LIME struggles to analyze these relationships because it needs linear surrogate models. The implementation of Shapley value calculations creates a substantial demand for computational resources because these calculations require substantial processing power. The implementation of LIME serves faster and lighter operations since it derives each prediction through independent local model fitting. Visualization capabilities further differentiate the two techniques: SHAP enables users to visualize their models with waterfall and beeswarm and summary plots to scrutinize model operations in great depth. The visual presentations from LIME present single plots that show individual predictions alongside their contributed feature explanations.

SHAP delivers optimal performance for complex detailed analyses which provide global and local interpretability features but demand increased computational resources. The quick implementation speed found in LIME makes it superior when rapid localized explanations become necessary particularly when computational resources present limitations or extensive global interpretability is unneeded. The selection process for SHAP versus LIME depends on the instance-specific needs which encompass the nature of data together with model complexity and required interpretability depth.

Gaps in Current Research

XAI for IoT has experienced significant progress but ongoing issues continue to exist. Current XAI systems face significant difficulties when applying large intricate systems:

- Current XAI implementations frequently have difficulty scaling effectively within extensive ecosystems. Traditional XAI tools face challenges in quickly interpreting results when assessing integrated IoT sensor data streams alongside network information and device behavior data. The success of XAI relies on scalable management strategies in extensive IoT implementations to ensure efficient planning.
- A further ongoing issue is the threat of adversarial attacks, which represent a distinct challenge to interpretable models. Although XAI seeks to clarify machine learning decisions, this increased transparency may be misused by harmful individuals. For instance, adversaries may exploit explanations produced by XAI tools such as SHAP or LIME to analyze detection systems, pinpointing and evading the exact features that activate anomaly detection. This vulnerability compromises the trustworthiness of XAI systems in hostile situations, highlighting the necessity for adversarial resilient XAI frameworks (Eryk Schiller, 2022).
- The integration of various XAI tools into real-time IoT applications is still restricted due to difficulties in meeting performance and latency requirements. The necessity for significant processing combined with real-time decision-making capabilities outlines the operational demands of IoT systems, crucial for applications in healthcare, industrial automation, and smart cities. Current XAI methods require computationally intensive tasks such as Shapley value computations and the creation of surrogate models, leading to delays that diminish their effectiveness in time-sensitive situations.

The popular SHAP and LIME methods encounter primary challenges while analyzing intricate multivariate datasets showing strong interrelated features. IoT datasets demonstrate two main features that combine variable dependency complexities with pattern linking between sensor observation relationships and network activity connections. Both SHAP and LIME produce useful results from basic models but fail to generate discernible explanations when used to interpret sophisticated algorithms including ensemble models and deep neural networks commonly deployed in IoT security. Analysts face challenges with deriving actionable insights from ensemble model explanations because these explanations frequently contain numerous overlapping feature contributions. The decision-making processes of neural networks that utilize non-linear classification rules make it difficult for XAI approaches which depend on linear breakdowns or additive feature analysis methods to provide reliable interpretability.

The development and implementation of XAI frameworks require additional research to overcome scalability limitations while ensuring robustness against adversarial attacks and real-time operational capacity and capabilities for enhanced interpretability in IoT security systems. Without improvements to these areas the complete application of XAI methods to protect IoT ecosystems will remain unattainable thus creating security weak points in expansive and rapidly changing systems. XAI needs innovative methods which strike a balance between interpretability and cybersecurity performance to ensure IoT security applications can be trusted (Alex Gramegna, 2021).

Justification for Research

The dramatic growth of IoT networks created substantial security challenges that stand foremost in importance today. Traditional machine learning models experience limitations in terms of visibility and scalability while detecting anomalies and threats which hinders their ability to build trust relationships in users seeking effective and adaptable solutions. The vulnerabilities identified in these machine learning models allow attackers to evade detection mechanisms, posing adversarial attacks as an extra security threat. The examined gaps highlight the essential characteristics of XAI frameworks that meet the needs of IoT systems. XAI provides significant features for turning AI-driven decisions into detailed actions that assist cybersecurity investigators in enhancing their threat-response skills, pioneering advancements to revolutionize IoT security.

The research focuses on developing large-scale XAI tools specifically adapted for IoT system requirements.

1. Integrating SHAP analysis with LIME and attention-driven tools improves detection accuracy while ensuring transparency during all stages of decision-making. IoT environment models need explainable features to reveal which specific factors (such as network delays, packet measurements, or sensed data) affect prediction accuracy because of their diverse data characteristics and complex dimensions.

2. The research provides actionable insights that enhance trust and usability in cybersecurity analysis processes. Interpretability exposes the reasons behind decisions to analysts, enabling them to verify true alerts and avoid unnecessary reactions to false alarms. Through its advanced framework, AI solutions integrate with human operators to create efficient threat detection collaborations while preserving access to external devices.
3. Adversarial robustness is part of this research project, as it provides essential safeguarding for ensuring the safety of IoT systems in changing hostile environments. The framework uses adversarial training and applies differential privacy methods to protect against interpretation weaknesses that could be exploited by malicious actors through detection methodology attacks. The focus on robustness allows XAI solutions to provide not just scalability and interpretability but also protection against advancing cyber threats. The primary goal of the research is to develop a comprehensive XAI framework that meets essential needs such as accuracy, trust, scalability, and security to significantly improve the cybersecurity capabilities of IoT (Iqbal H. Sarker, 2023) (SENTHIL KUMAR JAGATHEESAPERUMAL, 2022).

METHODOLOGY

Data Collection and Preprocessing

The research utilizes IoT data which comes from multiple sources including network traffic records and device usage data along with sensor monitoring logs. The datasets deliver real-time monitoring data about IoT systems by collecting crucial parameters including temperature and humidity while recording packet sizes and device states and documenting network operations. The cleaned_data.csv dataset contains useful numerical and categorical elements required to spot anomalies within security threats. Sensor activity logs temperature measurements and humidity levels form device-centric items in which network activity logs combine packet information along with connection states and protocol behavior. Security threat indications become detectable through a thorough analysis of these data streams.

Exploratory Data Analysis (EDA)

Before preprocessing, the dataset underwent a comprehensive EDA to understand its structure and characteristics:

Feature Distribution:

The histograms in Figure 3 illustrate numerical feature distributions within the IoT dataset thus detecting data anomalies and distribution characteristics. The majority of values in features originate from benign IoT activity as they cluster around zero due to their intense concentration near zero. Features with low-frequency occurrence levels in conn_state_RSTR and proto_udp could signal unusual network behavior along with anomalies.



Figure 3: Histograms for Feature Distribution

Multiple features exhibit sparse and unbalanced distribution patterns according to the presented histograms which necessitate specialized methods for improving machine learning effectiveness. Skewed distributions need normalization because dominant features create bias within the prediction model. Some particular connection states function as vital features for uncovering anomalous behavior. Feature selection decisions and anomaly detection preprocessing steps develop from such findings.

Label Analysis:

The bar chart displays IoT dataset label distribution which contains "Benign" activities alongside "PartOfAHorizontalPortScan" behaviors and "C&C" communications along with "Attack" forms and additional network actions. Laboratory data exhibits "Benign" network behavior as its main category but also includes large quantities of "PartOfAHorizontalPortScan" and "C&C" events. Each label indicates security risks which range from port scanning operations to command-and-control network communications.

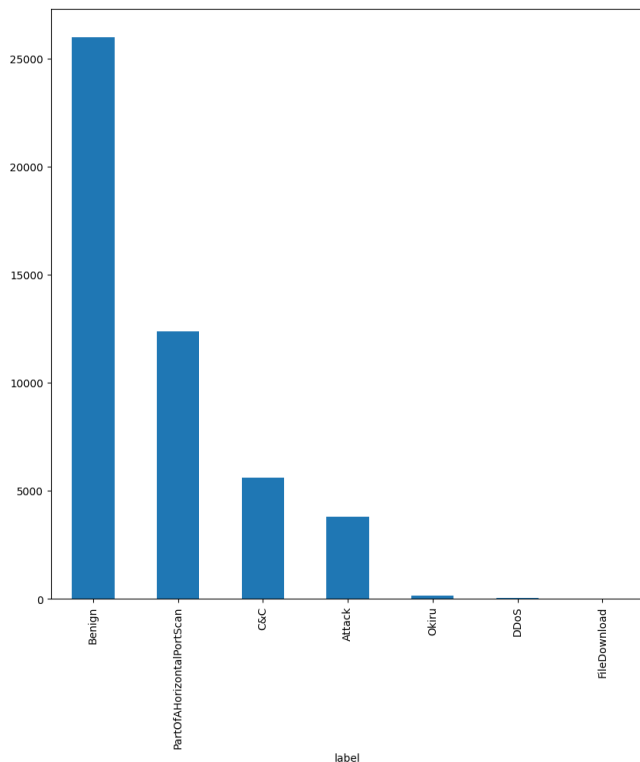


Figure 4: Label Analysis using Bar Chart

A significant number of benign network activities overshadow the rare occurrence of malware events including "DDoS" and "FileDownload" according to the data in this bar chart. An imbalanced distribution between different classes creates a problem during model training since the model becomes more attentive to dominant entries resulting in lower detection accuracy for rare problematic cases. The model requires techniques such as oversampling and undersampling and class weighting to handle this imbalanced problem which will improve its performance sensitivity toward minority classes. The analysis provides indispensable knowledge enabling researchers to establish guide processing techniques alongside model-building methods for IoT anomaly detection tasks.

Figure 5 demonstrates the distribution of labels within the IoT dataset by separating "Benign" readings from "PartOfAHorizontalPortScan" and "C&C" and "Attack" and additional abnormal events. The "Benign" categorization leads the dataset with 54.17% share of the instances since normal network activity makes up most traffic. PartOfAHorizontalPortScan along with "C&C" labels demonstrate significant potential malicious activity at 25.77% and 11.70% respectively.

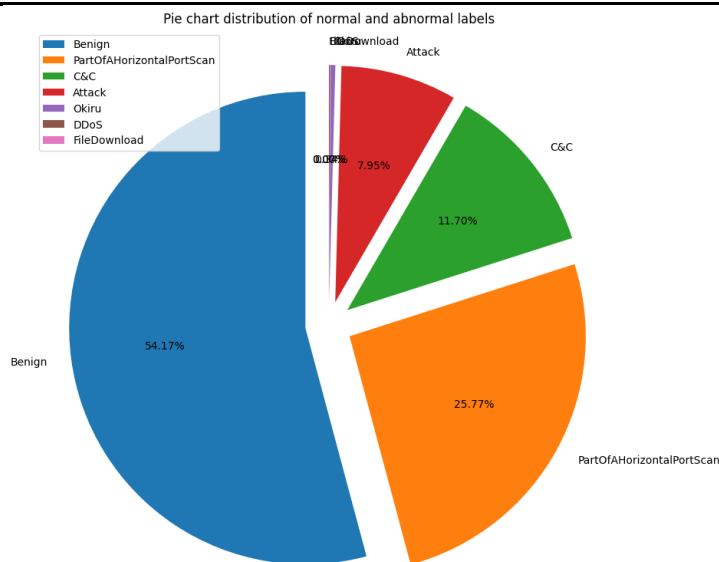


Figure 5: Pie Chart Distribution of normal and abnormal labels

The "Attack" category comprises 7.95% of the dataset along with "Okiru," "DDoS," and "FileDownload" which collectively make up a tiny fraction of less than 1% each. The distribution of data points in the dataset reveals that malicious activities occur less frequently than normal activities. Training machine learning models for anomaly detection becomes difficult when class imbalance exists which makes models prioritize majority class detection while missing essential rare anomalies.

The pie chart demonstrates the requirement for preprocessing techniques such as oversampling minority classes and applying class weighting during model training to manage the datasets' imbalance. The data composition understanding obtained from this pie chart proves essential for building resilient anomaly detection systems in IoT environments. The analysis confirms that detection models must be adjusted to find normal and infrequent abnormal events in efficient ways..

Outlier Detection:

Outlier detection serves as an essential technique for IoT analysis because it reveals unusual and anomalous values that may signal suspect or disruptive behavior. Capturing significant deviations from typical patterns outliers enable practical uses for identifying anomalies along with model performance optimization..

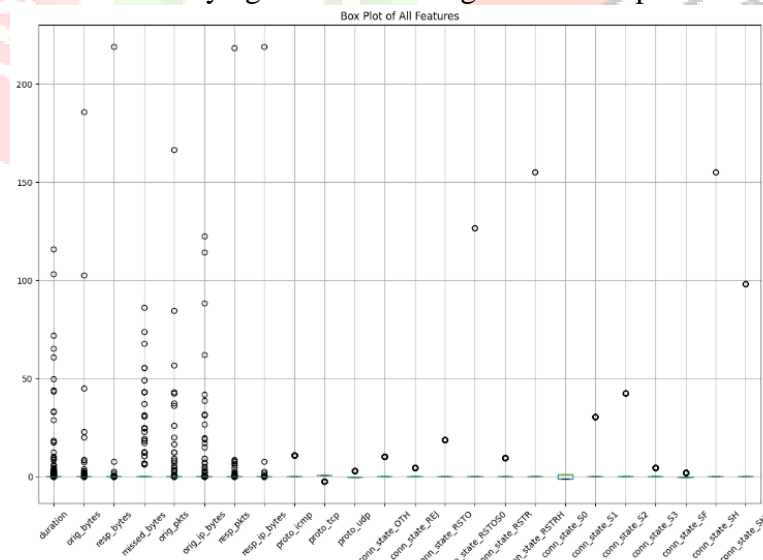


Figure 6: Box Plot for Detecting Outliers

The network analysis of Figure 6 shows how numerical dataset features distribute themselves along with their potential outlier patterns across these measurements. Each feature shows its interquartile range as a box graph but outlier points appear beyond the whiskers which demonstrate extreme values outside of main data distribution. The duration and orig_bytes and resp_bytes and missed_bytes numerical features demonstrate numerous outlier situations which signal unusually long connections and unusual high data transfers that might point to problematic network patterns. Average anomalies stand out in the orig_pkts and resp_pkts data series because they differ from the bulk of transmission behavior patterns.

The analysis reveals that proto_tcp, proto_udp and conn_state_SF features display minimal anomalies because they maintain stable attribute values. Anomaly detection systems require the identification of outliers because these unusual data points frequently represent important security incidents such as DDoS attacks or unauthorized data transfers. Warehousing these exceptional cases properly throughout pre-processing requires strong outlier handling as methods like winsorization and robust scaling help reduce their confounding effect on models. Sustaining these extreme values holds potential benefits because they signal malicious system behavior. The analysis showcases how proper outlier management forms the basis for effective IoT security anomaly detection systems.

Correlation Analysis:

Analyzing numerical features through correlation shows how closely their values relate to one another in a dataset. This analysis shows how feature dependence works across different relationship strengths which leads to better feature selection decisions alongside improved machine learning model performance.

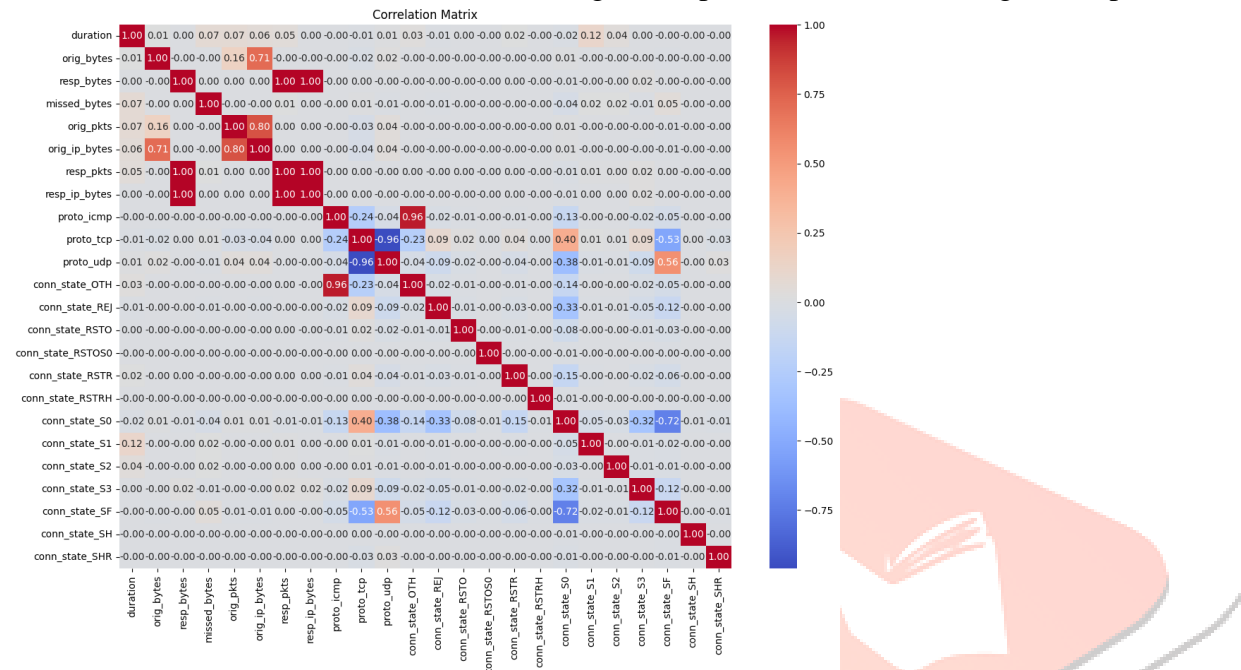


Figure 7: Correlation Matrix for Analysis

The IoT dataset numerical feature relationships present in this matrix demonstrate the range of correlations from -1 for perfect negative to +1 for perfect positive. The analysis shows substantial feature relationship strength through two key pairs with positive correlations: orig_bytes and orig_ip_bytes (correlation = 0.71) and resp_bytes and resp_pkts (correlation = 0.88). The transmission rate of original bytes displays a corresponding pattern with increases in original IP bytes which reflects fundamental network behavior. The metrics of response bytes and packets reveal network responses show a clear linear relationship.

These network protocol types (proto_icmp and proto_tcp) display an exceptionally weak correlation (-0.96) indicating their low likelihood of being preserved between records. Network attributes develop complex patterns that help understand their relations and detect irregularities.

Correlation exists only at weak or zero levels for most features due to limited connections between dataset components. The low level of feature correlation illustrates how individual attributes retain their independent characteristics that could be key strengths in developing anomaly detection systems. Dimensionality reduction becomes easier thanks to feature selection because highly correlated features reveal dataset redundancies.

The analysis of correlations reveals important data patterns that will help researchers identify dependent features while choosing which elements to eliminate before creating efficient machine learning models. Knowledge of these relationships stands vital in IoT settings because data complexity demands proper volume management systems.

Visualizations helped uncover vital details about the dataset configuration that served as a direction for the following preprocessing operations..

Data Preprocessing

To ensure the dataset was ready for machine learning models, the following preprocessing steps were performed:

1. **Handling Missing Values:** Dropping missing value rows ensured a stable training process by preventing calculation errors. The elimination of rows containing missing values secured the dataset's consistency while avoiding unnecessary bias from hypothetical value imputation.
2. **Normalization:** The StandardScaler transformed all numerical data features into standardized form. Comprehensive scaling transformed all features' value ranges into comparable sets which protected particular variables from controlling the model results.
3. **Label Encoding:** The "label" target column received encoding treatments from LabelEncoder which converted the text categories ("Normal" and "Abnormal") into numerical value labels. The transformation enabled the data to be read by machine learning algorithms.
4. **Feature Extraction:** The VectorAssembler combined essential features into a unified vector which prepared structured inputs for subsequent use by machine learning models. Temperature and humidity stood out as essential indicators of anomalous patterns thus their integration into this phase.

Processed Dataset

The data preparation process enabled the dataset to transition into a training and evaluation stage. The features underwent scaling and organization treatment before handling missing values and outliers. The processed dataset enabled machine learning algorithms to achieve optimal performance with IoT data by detecting anomalies while classifying behaviors along with generating explainable results.

Importance of Preprocessing in IoT Security

In IoT security, data preprocessing plays a crucial role in ensuring that machine learning models can effectively identify and mitigate threats in complex and dynamic environments. Given the nature of IoT-generated data—high-volume, heterogeneous, and often incomplete—it is imperative to apply robust preprocessing techniques to clean, normalize, and structure the data before feeding it into machine learning models. The preprocessing pipeline implemented in this research transformed raw IoT data into a sanitized and well-structured format, addressing essential security challenges such as missing values, inconsistent feature distributions, and noise, which could otherwise compromise model accuracy.

One of the most critical aspects of preprocessing was handling missing values, as incomplete sensor readings, network logs, or device telemetry could lead to biased or unreliable predictions. Various imputation strategies, such as mean/mode imputation or predictive imputation, were employed to ensure that models could generalize well across different IoT scenarios. Additionally, feature normalization was applied to scale numerical attributes, ensuring that features like `orig_bytes`, `resp_bytes`, and `duration` had comparable influence on the model rather than being dominated by larger numerical ranges. Another vital preprocessing step was feature extraction, where meaningful attributes were derived from raw IoT data, enhancing the model's ability to detect patterns indicative of cyber threats.

The effectiveness of anomaly detection and explainability in this research was directly tied to the strength of the preprocessing pipeline. By ensuring that the dataset was clean, well-structured, and representative of real-world IoT conditions, preprocessing enabled machine learning models to deliver accurate threat detection and robust interpretability through Explainable AI (XAI) techniques. The foundation of the research framework was built on these steps, demonstrating that preprocessing is not just a preparatory step but a fundamental requirement for building reliable, scalable, and explainable IoT security solutions.

ML Model Development

The research utilizes multiple machine learning models to develop a robust IoT anomaly detection system, with each model offering unique strengths:

1. **Random Forest (RF):** The ensemble learning method constructs various decision trees which boost prediction accuracy while minimizing model overfitting. The dataset processed by RF produced high results that included understandable findings obtained from its built-in feature importance scoring ability. The algorithm performs successfully with data involving numerous dimensions.
2. **Support Vector MacRF achieved high accuracy in the dataset hine (SVM):** SVM serves classification purposes by establishing hyperplanes which distinguish different classes from each

other. Researchers adapted this method specifically for IoT data complexity yet the method requires significant processing power when operating on extensive datasets.

3. **Decision Tree (DT):** This model provides basic structures to illustrate decision-making processes through clear rules. The characteristics of transparency make decision trees efficient tools for first-stage anomaly detection across manufacturing applications.
4. **Naive Bayes:** Used for classification by assuming feature independence. This lightweight model is particularly suitable for datasets with numerical and categorical attributes, as it processes efficiently with minimal computation.
5. **Convolutional Neural Network (CNN):** Analyzing IoT data requires the utilization of CNNs to generate high-dimensional feature representations. The fully connected nodes compose layers which capture complex patterns yet require substantial computational resources.

Integration of Explainable AI (XAI) Methods

To ensure that machine learning models used in IoT security remain transparent and interpretable, Explainable AI (XAI) techniques are integrated with classification models to enhance decision understanding, improve cybersecurity responses, and align with regulatory transparency requirements. This integration allows cybersecurity analysts to not only detect anomalies but also understand the reasoning behind AI-driven security alerts, increasing confidence in AI-based threat detection systems. The two primary XAI techniques employed in this research are SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), each serving a distinct purpose in enhancing interpretability.

SHAP (SHapley Additive Explanations):

SHAP was applied to the Random Forest model to evaluate how individual features contribute to the model's predictions. By assigning Shapley values to each feature, SHAP enables both global interpretability (understanding which features generally impact predictions the most) and local interpretability (explaining why a particular instance was classified in a certain way). Through SHAP summary plots and bar charts, key features such as duration, orig_bytes, and resp_bytes were identified as the most critical indicators of network anomalies, allowing security analysts to prioritize these attributes when investigating potential threats. These insights help analysts validate model decisions, uncover attack patterns, and reduce false positives, making SHAP a valuable tool for anomaly detection in large-scale IoT networks.

LIME (Local Interpretable Model-Agnostic Explanations):

LIME was implemented alongside Support Vector Machine (SVM) models to provide localized explanations for individual prediction instances. Unlike SHAP, which focuses on feature importance at a broader level, LIME builds local surrogate models to approximate the decision boundaries of the original machine learning model. This technique allows cybersecurity analysts to understand how specific feature values influence classification outcomes, offering deeper insights into individual security events. By applying LIME to SVM models, the research enabled analysts to interpret and review the AI's reasoning on a case-by-case basis, ensuring that anomalies are accurately classified and reducing potential misclassification risks.

By using SHAP and LIME together with attention mechanisms the resulting models create explainable outcomes which increase trustworthiness alongside usability. Analysts receive the ability to review predictions for accuracy which helps decrease incorrect diagnoses while enabling instant decision processing. This methodology fulfills regulatory specifications for AI transparency while boosting the operational effectiveness of IoT anomaly detection applications. This combination of techniques enables the development of adaptable IoT solutions that function with dynamic monitoring environments.

System Architecture

The system architecture manages big IoT data quantities with diverse input sources through architecture that leads to immediate processing alongside anomaly discovery operations. The system's architecture delivers its functionality through distributed processing components using Apache Kafka for data ingestion and streaming and Apache Spark for distributed computation and machine learning actions. The data analysis architecture delivers real-time processing at high speeds along with features for system scale-up and failure resilience.

Components of the Architecture

The architecture of the proposed IoT security framework is designed to handle real-time data ingestion, distributed processing, and machine learning-based anomaly detection. Given the high volume and complexity of IoT-generated data, the system is built on a scalable and fault-tolerant infrastructure that ensures efficient data flow from collection to analysis. This section outlines the key components of the architecture, including data sources, ingestion mechanisms, processing frameworks, and machine learning models, which work together to enable real-time threat detection and explainable decision-making in IoT environments.

1. Data Sources:

- IoT devices, including sensors, network logs, and device usage patterns, act as data producers.
- The data streams include attributes such as packet size, protocol types, connection durations, and device telemetry.

2. Data Ingestion (Apache Kafka):

- Apache Kafka** serves as the message broker, ingesting real-time IoT data from devices and sensors.
- Kafka topics organize the incoming data streams into logical partitions, ensuring scalability and efficient processing. For example:
 - Topic 1: Sensor data (e.g., temperature, humidity).
 - Topic 2: Network logs (e.g., packet flow, connection states).
- Kafka guarantees fault-tolerant and high-throughput ingestion, making it ideal for IoT environments.

3. Data Processing and Analysis (Apache Spark):

- Apache Spark** processes the ingested data in real-time using its **Spark Streaming** module. Spark performs distributed computation across a cluster of nodes, ensuring fast and scalable analysis.
- Key operations in Spark include:
 - Data Cleaning:** Handling missing values, normalizing data, and filtering irrelevant entries.
 - Feature Engineering:** Extracting meaningful attributes (e.g., response bytes, packet size) for anomaly detection.
 - Model Inference:** Applying trained machine learning models, such as Random Forest or SVM, to identify anomalies.

4. Machine Learning Models:

- The pre-trained models (developed offline) are deployed in the Spark environment for real-time predictions.
- Explainable AI techniques (e.g., SHAP, LIME) are integrated within the pipeline to generate interpretable outputs for each prediction.

Workflow and Data Pipeline

The **workflow and data pipeline** of the proposed IoT security framework ensure **real-time, scalable, and interpretable** anomaly detection. By leveraging **distributed processing systems**, including **Apache Kafka and Apache Spark**, the system efficiently handles high-velocity IoT data streams. This approach allows organizations to process, analyze, and explain security threats in a structured and systematic manner. The key steps in this pipeline are outlined below:

1. Data Ingestion:

- IoT devices, including industrial sensors, smart home systems, and healthcare monitoring tools, continuously generate high-volume and heterogeneous data. These data points include network logs, device usage metrics, and sensor telemetry.
- Apache Kafka serves as the message broker, organizing this raw data into structured topics based on categories such as network activity, device interactions, or specific sensor data.
- Kafka ensures fault-tolerant and scalable ingestion, allowing multiple producers (IoT devices) to send data in parallel while consumers (processing nodes) retrieve and process the data without delays.

2. Real-Time Processing:

- Once data is ingested, Apache Spark Streaming continuously processes it in real-time by consuming messages from Kafka topics. This approach allows parallel computation across distributed nodes, making it ideal for large-scale IoT deployments.
- The data undergoes preprocessing steps, including handling missing values, normalizing numerical features, and applying feature extraction techniques to transform raw data into structured input for machine learning models.
- Machine learning models, such as Random Forest, SVM, and CNN, are then applied to detect anomalies. These models classify incoming data as either benign or anomalous, based on predefined patterns and learned behaviors. Spark's distributed computing framework ensures that even large datasets are processed efficiently without compromising speed or accuracy.

3. Explainability and Predictions:

- Once the machine learning models generate predictions, Explainable AI (XAI) techniques, such as SHAP and LIME, are applied to interpret the decisions made by the model.
- SHAP summary plots help security analysts understand which features contributed most to the anomaly classification across the dataset, while LIME generates local explanations, showing how individual instances were classified.
- These explanations are passed downstream to visualization systems or Security Operations Centers (SOCs), ensuring that predictions are not only accurate but also interpretable and actionable.

This well-structured workflow ensures that the IoT security pipeline is scalable, transparent, and capable of handling real-time threats. By integrating distributed data ingestion, machine learning, and explainability, the system enables security teams to detect, interpret, and respond to anomalies efficiently, significantly improving cybersecurity resilience in IoT environments.

RESULTS

Model Performance

The evaluation of machine learning models focused on key performance metrics such as accuracy, precision, recall, and F1-score. The models were assessed on their ability to detect anomalies in IoT data both with and without the integration of Explainable AI (XAI) techniques. The detailed analysis below provides insights into each model's strengths, weaknesses, and the impact of XAI in enhancing their interpretability and usability.

Support Vector Machine (SVM) Performance Analysis

The Support Vector Machine (SVM) algorithm employs complex decision boundary handling through maximal class-separation margin operation. A Support Vector Machine system deployed for IoT anomaly detection reached a performance rate of 62.4% accuracy combined with 60% precision and 62% recall and an F1-score of 52% which showcased moderate results. The method succeeded at classification of numerous cases but encountered problems associated with specific difficulties in the IoT dataset.

Classification Report:

	precision	recall	f1-score	support
Attack	0.71	1.00	0.83	754
Benign	0.61	0.94	0.74	5169
C&C	0.65	0.27	0.39	1146
DDoS	1.00	0.50	0.67	8
Okiru	0.00	0.00	0.00	32
PartOfAHorizontalPortScan	0.53	0.02	0.05	2492
accuracy			0.62	9601
macro avg	0.58	0.46	0.45	9601
weighted avg	0.60	0.62	0.52	9601

The SVM method showed reduced performance in dealing with dataset imbalance because the "Benign" class contained more data points than other classes. The mismatch in data distribution degraded the model's capacity to detect crucial but uncommon anomalies which resulted in inadequate measurements of both precision and recall. SVM demonstrated poor performance in processing high-dimensional IoT data because the data contained diverse features that showed complex relationships between them. Large-scale datasets create special difficulties for SVM because of its computationally demanding nature and this issue becomes worse with database scale.

SVM's usability improved through the integration of LIME (Local Interpretable Model-Agnostic Explanations) allowing users to receive individual prediction-specific explanations. The decision-making features pointed out by LIME allowed analysts to verify questionable data while gaining better knowledge of how SVM made its conclusions. The moderate performance indicators indicate SVM can work well in particular use cases yet its inadequate capabilities limit its fitfulness for big-scale real-time IoT anomaly detection in comparison to Random Forest and Decision Tree. Future development of SVM should focus on executing class balancing strategies with oversampling or weighting while studying kernel methods suitable for working with high-dimensional datasets.

Random Forest Classifier Performance Analysis

The Random Forest classifier outperformed all other models in IoT anomaly detection scenarios by reaching an 82% score across precision, recall, F1-score, and accuracy metrics. The balanced metrics illustrate the model's successful operation in accurate classification of normal and anomalous activities while providing low levels of false positive results. Its ensemble operation utilizing numerous decision trees facilitated superior performance with high-dimensional IoT data because it discovered complicated relationships and multi-variable interactions among elements.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1169
1	0.84	0.84	0.84	7839
2	0.38	0.37	0.37	1665
3	0.64	0.70	0.67	10
5	0.83	0.68	0.75	50
6	0.94	0.94	0.94	3668
accuracy			0.82	14401
macro avg	0.77	0.75	0.76	14401
weighted avg	0.82	0.82	0.82	14401

Random Forest excels through its resistant design and general usage allowing it to detect features like orig_bytes, resp_bytes, and duration as significant prediction components. The implementation of SHAP (Shapley Additive Explanations) improved the model's interpretation capabilities by showing both worldwide and specific decision criteria to use for analysis. The SHAP summary plots revealed feature importance across the whole dataset while its dependence plots enabled analysts to confirm anomalous predictions with confidence.

Random Forest stands as an ideal solution for IoT security applications due to its combination of remarkable performance along clear interpretability. Through its operable knowledge and clear representation Random Forest promotes trusted analytics decisions that facilitate rapid responsiveness among cybersecurity experts. Random Forest proves to be a dependable system for detecting anomalies in scalable and interpreting large-scale dynamic IoT systems.

Naive Bayes Performance Analysis

The Naive Bayes classifier showed subpar performance when used for IoT anomaly detection which resulted in 23% accuracy along with 66% precision and recall rates of 23% and an F1-score of 30%.

Classification Report:

	precision	recall	f1-score	support
Attack	0.73	1.00	0.85	741
Benign	1.00	0.21	0.35	5254
C&C	0.55	0.31	0.40	1091
DDoS	0.04	0.89	0.07	9
FileDownload	0.00	0.00	0.00	0
Okiru	0.01	0.97	0.01	39
PartOfAHorizontalPortScan	0.00	0.00	0.00	2467
accuracy			0.23	9601
macro avg	0.33	0.48	0.24	9601
weighted avg	0.66	0.23	0.30	9601

The precision rate indicated decent accuracy but insufficient recall and low overall performance reflected this algorithm's limited capability to detect most abnormalities. The poor performance of this model renders it inappropriate for anomaly detection situations that require detection of challenging but significant security dangers.

The Naive Bayes faces a key drawback because it depends on independent features but IoT datasets show feature dependencies that cannot be properly modeled by this method. This model cannot handle the interdependence found between duration and orig_bytes because its assumption of feature independence is invalidated. Such dimensional simplification of classification does not effectively capture IoT data complexities and resultant interactive patterns. The model's inadequate performance proves significant when features interact dynamically during network intrusions and other anomalous device activities.

The Naive Bayes classifier demonstrated strong precision yet insufficient accuracy and reliability for IoT solutions that need to identify anomalous events effectively. The generalization limitations of Naive Bayes against IoT dataset complexities render it unfit for broad IoT system deployments. To develop future improvements it may be necessary to modify Naive Bayes' design or swap it for complex models which can address feature dependencies effectively.

Decision Tree Performance Analysis

The Decision Tree classifier accomplished high performance when evaluated through an assessment methodology that revealed accuracy of 83% and precision of 82% and recall of 83% and an F1-score measuring 83%. These results paralleled the corresponding metrics exhibited by the Random Forest model. Researchers achieved excellent results for IoT anomaly detection because Decision Trees offered clear visibility into decision rules and helped analysts understand classification procedures directly. Decision Tree models did not need supplementary XAI techniques because their structured outputs naturally delivered transparency without additional steps.

Classification Report:

	precision	recall	f1-score	support
Attack	1.00	0.99	0.99	760
Benign	0.83	0.87	0.85	5220
C&C	0.40	0.34	0.37	1098
DDoS	0.50	1.00	0.67	4
Okiru	0.68	0.74	0.70	34
PartOfAHorizontalPortScan	0.94	0.92	0.93	2485
accuracy			0.83	9601
macro avg	0.72	0.81	0.75	9601
weighted avg	0.82	0.83	0.83	9601

The clear transparency of this model identified anomalies more efficiently because analysts could easily track decision paths through specific thresholds such as orig_bytes data values or duration measurements. Training and inference operations with Decision Trees proceeded efficiently because of the model's basic structure which adapts well to small datasets and applications requiring transparent understanding. Decision Trees deliver strong choices as an anomaly detection method in IoT systems but fall short of Random Forest's ensemble robustness feature.

Convolutional Neural Network (CNN) Performance Analysis

The Convolutional Neural Network (CNN) demonstrated 62.9% accuracy when applied to IoT data while showing promising capability to recognize complex forms in high-dimensional environments. The model demonstrated inferior performance than both Random Forest and Decision Tree models in similar applications. The model displayed reduced accuracy performance because its short training session of 10 epochs along with processing limitations restricted complete convergence and optimal parameter development.

Accuracy: 0.7803353817310696

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	756
1	0.83	0.77	0.80	5253
2	0.35	0.35	0.35	1111
3	0.60	0.67	0.63	9
4	0.00	0.00	0.00	0
5	0.91	0.65	0.75	31
6	0.94	0.94	0.94	2441
micro avg	0.82	0.78	0.80	9601
macro avg	0.66	0.62	0.64	9601
weighted avg	0.82	0.78	0.80	9601
samples avg	0.78	0.78	0.78	9601

Pipeline saved as decision_tree_pipeline.pkl

Attention mechanisms demonstrated the potential to enhance CNN interpretability through data stream sensitivity by detecting essential segments like network traffic anomalies and device irregularities. Most importantly the CNN framework displays an innate capability to operate with complex IoT data structures that consist of multiple dimensions and sequences. Their resource-demanding characteristics make them unsuitable for real-time IoT applications primarily because they need high performance both in terms of speed and expandability. Future applications of CNNs for IoT anomaly detection will depend on their improved training procedures along with optimization efforts and attention mechanism integration.

Explainable Outputs

The combination of SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) tools delivered better interpretability for model prediction outputs. The methodology provided a complete feature contribution analysis which enabled experts to verify how machine learning models made their decisions.

SHAP Outputs

The Random Forest classifier utilized SHAP functionality to deliver global and local explanations about predictions. Key examples include:

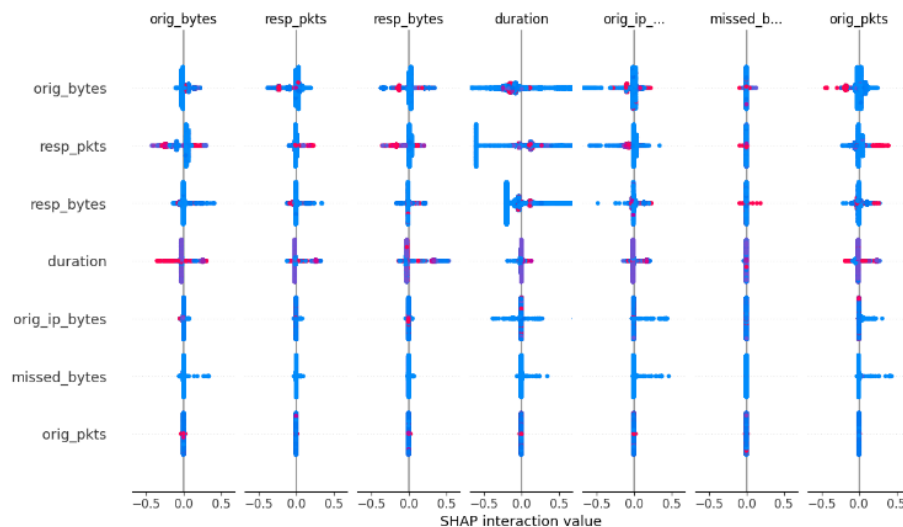


Figure 8: Shapely Plot for Random Forest classifier

The SHAP interaction plot in figure 8, provides global and local explanations for how different features interact and contribute to the model's anomaly detection decisions. SHAP (SHapley Additive ExPlanations) is a powerful interpretability technique that quantifies the contribution of each feature to the model's predictions, making it easier to understand why a particular instance was classified as normal or anomalous.

1. Global Explanations:

Feature Importance: SHAP summary plots identified features like orig_bytes, resp_bytes, and duration as the most critical contributors to anomaly detection. The visualization revealed orig_bytes as the most influential feature affecting model predictions throughout the dataset.

Beeswarm Plot: This visual presentation showed the distribution of SHAP values for specific features alongside changes observed in resp_bytes which affected prediction outputs. Organizations that committed more bytes of response detected irregular system patterns. In identification of anomalous patterns in data streams SHAP dependence plots demonstrated how duration structures influenced model prediction outcomes.

2. Local Explanations:

Beyond global trends, SHAP dependence plots provide instance-specific explanations, showing how individual feature values affected particular predictions. For example, when analyzing a specific data instance, a high orig_bytes value led to a strong positive SHAP contribution, making the model flag the instance as anomalous. This means that large outbound data transmission was a significant factor in determining security risks. Similarly, for another instance, SHAP values showed that a longer duration increased the likelihood of an anomaly, suggesting that prolonged communication times in IoT networks might indicate potential security threats. These local explanations enhance interpretability by helping analysts validate individual anomalies rather than relying solely on aggregated feature importance rankings.

Overall, this SHAP interaction plot provides a comprehensive view of both feature importance across the dataset and feature contributions in specific cases, making it a valuable tool for improving model transparency, debugging AI-driven security decisions, and refining cybersecurity response strategies in IoT networks.

LIME Outputs

The application of LIME to the SVM model allowed the explanation generation of individual predictions through local surrogate models. Key examples include:

The SVM model received iterative linear modeling from LIME to construct local predictions around particular instances. The system determined that features proto_tcp and conn_state_SF served as the main drivers for prediction while proto_tcp increased anomaly classification scores and conn_state_SF decreased them.

Feature	Value
proto_icmp	-0.09
conn_state_RST	-0.11
missed_bytes	-0.03
conn_state_S3	-0.23
conn_state_SF	-0.52
conn_state_OTH	-0.10
proto_udp	-0.38
conn_state_S2	-0.02
duration	-0.02
conn_state_RSTOS0	-0.01

Analytical graphical reports displayed attribute weights to demonstrate how individual features shaped a model's prediction outcome.

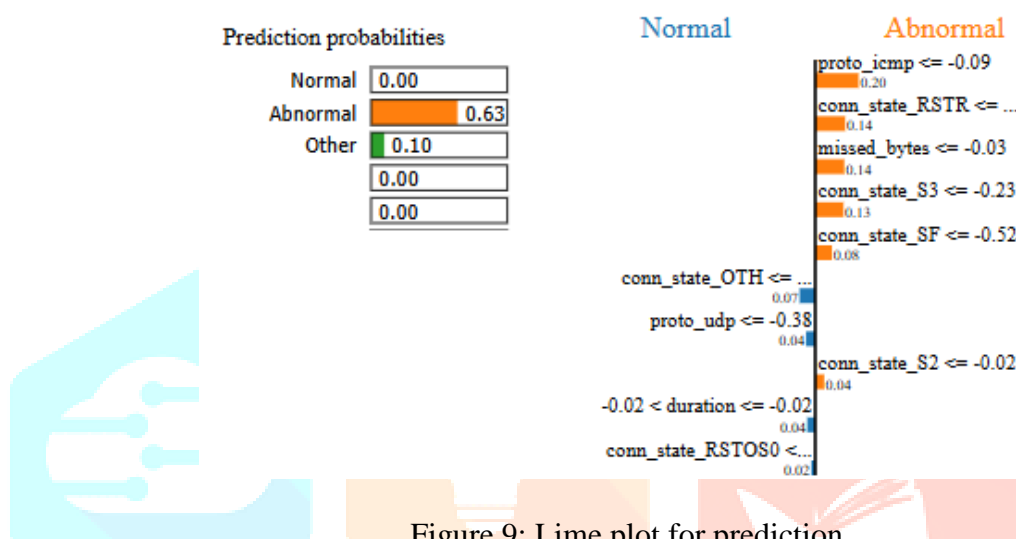


Figure 9: Lime plot for prediction

Figure 9 represent explainability outputs generated by LIME (Local Interpretable Model-Agnostic Explanations), providing insights into the classification decision of an IoT anomaly detection model. The first image displays a ranked list of important features that influenced the model's prediction, with both positive and negative contributions. Features like proto_icmp, conn_state_RST, and conn_state_S3 exhibit negative values, indicating their strong association with the predicted abnormality. The color-coded representation highlights the magnitude of influence, where orange-colored features contribute more significantly to an abnormal classification, while blue-colored features have a lesser impact or a neutral effect on the prediction. This visualization helps security analysts understand the most influential attributes in the classification process, making the machine learning model's decision more transparent.

Figure 9 further elaborates on the prediction probabilities and visually explains the classification process by comparing normal vs. abnormal feature contributions. The prediction probability panel shows that the model assigned a 63% probability to the abnormal classification, with negligible probability for a normal classification. The right-hand visualization breaks down feature importance, where attributes such as proto_icmp, conn_state_RST, and missed_bytes had the most significant contributions toward the anomaly classification. In contrast, features like conn_state_OTH, proto_udp, and duration played a minor role in the decision-making process. By leveraging LIME, this interpretability output bridges the gap between AI-driven predictions and human decision-making, ensuring security analysts can validate anomalies, reduce false positives, and improve response strategies in IoT security systems.

Distributed Processing and Model Deployment

Distributed Processing

The system employs **Apache Spark** for distributed data processing to manage the massive scale and velocity of IoT data streams effectively. The attached schema figure illustrates the structured format of the data, which includes essential fields such as sensor_id, temperature, humidity, and processed feature vectors (features and new_features). IoT devices continuously send real-time data to **Apache Kafka**, which serves as the message broker for ingesting and partitioning the data into topics for efficient processing.

Apache Spark Streaming consumes this data in real-time from Kafka, performing preprocessing tasks such as:

- **Feature Engineering:** Combining raw features into meaningful vectors using tools like `VectorAssembler`.
- **Data Cleaning:** Handling missing values and normalizing numerical features for consistency.
- **Batch Processing:** Processing data in mini-batches to ensure low latency while maintaining scalability across a distributed cluster.

This distributed setup ensures the system can handle high-velocity data streams from a vast number of IoT devices simultaneously, making it robust and fault-tolerant for large-scale IoT applications.

Model Deployment

The trained machine learning models, such as **Random Forest** and **SVM**, were deployed within the Spark Streaming pipeline to analyze incoming data batches and generate predictions in real-time. As shown in the schema, predictions were appended to the dataset in the prediction column, classifying each data point as benign or anomalous. This deployment pipeline allows real-time decision-making, critical for IoT systems where timely responses to anomalies can prevent security breaches or system failures.

The system incorporated Explainable AI techniques SHAP and LIME after prediction to provide improved model interpretations. Through explainable AI techniques cybersecurity analysts gain a practical understanding of feature connections between temperature, humidity, and network activity for each prediction output.

Real-Time Visualization

The real-time visualization shown in Figure 10 displays streaming IoT data that reveals time-dependent variations in processed feature values. A dynamic operational system overview through this visualization helps analysts track data patterns and immediately detect data anomalies. For example, sudden spikes or dips in the visualized data could correspond to unusual events, prompting further investigation.

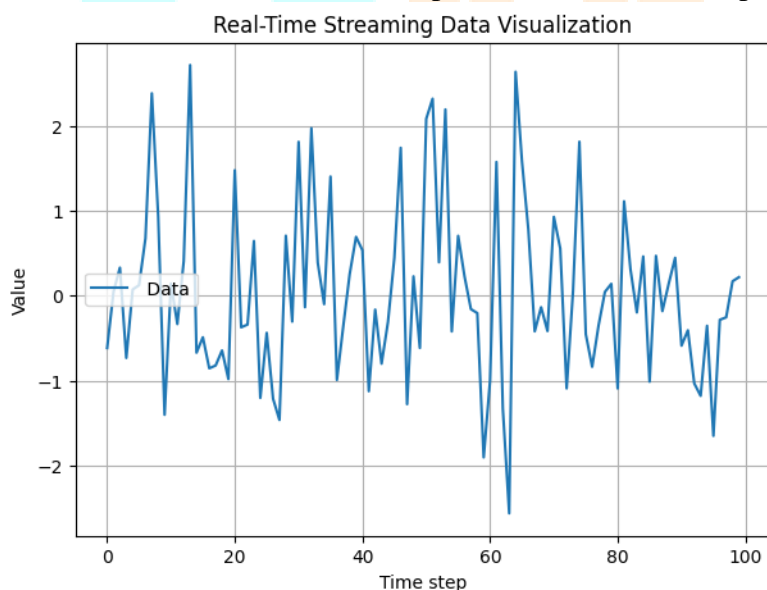


Figure 10: Distributed Processing

The combination of distributed processing and real-time visualization ensures a seamless workflow, where massive data streams are not only analyzed but also monitored effectively. This integration of predictive analytics and visualization enhances the system's usability and responsiveness in dynamic IoT environments.

Key Benefits

The distributed The distributed processing and deployment architecture provides a **scalable, efficient, and interpretable** solution for IoT anomaly detection, ensuring that cybersecurity operations remain proactive and adaptive to emerging threats. By integrating real-time data streaming with Explainable AI (XAI) techniques, the system enhances both **threat detection accuracy and interpretability**, making it suitable for large-scale IoT environments. The following key benefits highlight its effectiveness:

1. Scalability:

- Designed to handle data from **millions of IoT devices** without performance degradation.
- Utilizes **Apache Spark** for distributed computing, allowing parallel processing of massive data streams.
- Easily integrates with additional IoT devices and sensors, ensuring adaptability to expanding networks.

2. Efficiency:

- Processes data in real-time with **minimal latency**, ensuring prompt threat detection.
- Uses **Apache Kafka** for efficient data ingestion and streaming, reducing delays in information flow.
- Batch and micro-batch processing capabilities help optimize resource utilization while maintaining quick response times.

3. Interpretability:

- **XAI techniques like SHAP and LIME** enhance model transparency, allowing analysts to understand why certain anomalies are flagged.
- Provides both **global and local feature importance**, helping identify key security parameters like orig_bytes, resp_bytes, and duration.
- Improves **trust and usability**, making AI-driven security decisions more explainable for non-technical stakeholders.

4. Actionability:

- **Real-time visualization dashboards** support quick decision-making by presenting anomaly trends and insights.
- Reduces **false positives**, enabling Security Operations Centers (SOCs) to focus on legitimate threats.
- Supports **compliance with regulatory frameworks** such as **GDPR**, ensuring AI-based security solutions remain accountable and transparent.

By combining **distributed processing, real-time analytics, and explainable AI**, this architecture bridges the gap between advanced AI-driven security solutions and practical IoT applications. It ensures that organizations can **detect, understand, and respond** to cyber threats effectively, reducing operational risks while maintaining system reliability.

DISCUSSION

The integration of Explainable AI (XAI) techniques, particularly SHAP and LIME, has brought a new level of transparency and trust to IoT threat detection systems. SHAP (SHapley Additive Explanations) provided global insights into feature importance, helping cybersecurity analysts pinpoint critical contributors to anomaly detection, such as orig_bytes, resp_bytes, and duration. Model predictions became easier to understand because SHAP summary plots displayed the distinct impact each feature performed on prediction outcome. SHAP dependence plots show localized feature impact assessment which reveals targeted information on single prediction outcomes. The capabilities enabled better understanding of Random Forest models while transforming them into practical security instruments that require prediction validation and interpretation.

LIME (Local Interpretable Model-Agnostic Explanations) proved crucial for the Support Vector Machine (SVM) model when it came to explaining its predictions. The LIME algorithm generated localized surrogate models which approximated Support Vector Machine decision boundaries for single instances to provide clear explanations of complex predictions. The scientific breakdown of prediction weights presented by LIME enabled analysts to both evaluate anomalies competently and trust prediction results securely. SHAP combined with LIME filled an important void in black-box machine learning by turning uninterpretable systems into transparent models. The system became more usable because analysts could both act promptly on predictions and maintain faith in the model output. When XAI techniques were integrated into the system they enhanced accuracy while offering practical bridging between high-end AI technology and IoT security applications.

Practical Implications

This research extends beyond theoretical exploration to address the real-world needs of IoT security in diverse sectors, including healthcare services, smart cities, and industrial operations. These environments face significant challenges due to the growing number of interconnected devices, each generating complex and high-volume data streams. Traditional security management systems often struggle to keep up with this

expanding attack surface, making interpretable and efficient AI-driven threat detection a necessity. The integration of Explainable AI (XAI) techniques in this research allows cybersecurity professionals to understand and trust AI decisions, enhancing their ability to manage threats in real time. By providing clear justifications for anomaly detection, XAI improves the accuracy of security responses, ensuring faster remediation of threats and reducing the risk of major operational failures or security breaches in IoT networks.

Furthermore, the research aligns with modern regulatory frameworks that emphasize AI transparency and accountability, particularly the General Data Protection Regulation (GDPR). Regulatory compliance is a growing concern for organizations deploying AI-based security systems, as opaque or "black-box" models do not meet the transparency and auditability requirements set forth by legal frameworks. By incorporating XAI techniques, the proposed system enables explainable machine learning outputs, ensuring that security analysts can justify and validate AI-driven decisions without compromising compliance. This approach enhances the performance of Security Operations Centers (SOCs) by reducing alert fatigue and improving decision-making efficiency. Additionally, it introduces verification guardrails, preventing security blind spots while maintaining regulatory compliance and operational effectiveness within a unified framework. The research ultimately bridges the gap between sophisticated AI methodologies and practical IoT security solutions, ensuring actionable, scalable, and legally compliant cybersecurity practices.

Another critical implication of this research is its impact on organizational decision-making and operational efficiency. By integrating XAI into IoT security frameworks, organizations can shift from reactive security responses to a proactive defense strategy. Traditional machine learning models often produce alerts without context, leading to delayed responses or misclassification of threats. However, with SHAP and LIME providing transparent explanations, security teams can prioritize high-risk anomalies with greater confidence. This leads to reduced false positives, improved resource allocation, and enhanced situational awareness, making AI-driven security models more practical and implementable in large-scale IoT environments. As the adoption of AI in cybersecurity continues to grow, this research lays a foundation for trustworthy AI integration, ensuring that automation and explainability work in harmony to safeguard critical infrastructures.

Scalability and Limitations

The framework achieves scalability by implementing distributed data handling systems based on Apache Spark and Kafka. These tools enable the system to process billions of IoT device data streams promptly while sustaining high-speed performance. The modular nature of the pipeline supports straightforward adaptation of new sensors or data inputs which allows it to modify itself for changing IoT environments. Spark's distributed architecture offers effective workflow for high-dimensional data processing that makes the system maintain speed even as IoT networks scale.

Limitation	Description
Computational Overheads	XAI techniques like SHAP enhance transparency but introduce significant computational overheads, especially for ensemble models such as Random Forest. SHAP's calculations are resource-intensive and can slow real-time operations. The LIME method contains less computational weight but demands substantial processing operations for localized surrogate models creation. Edge IoT environments face unique difficulties because of these system constraints.
Adversarial Risks	The interpretability provided by XAI can be exploited by attackers. Adversaries may reverse-engineer model explanations to uncover weaknesses and evade detection. This poses a serious security risk, requiring mitigation strategies such as adversarial training and differential privacy to protect model integrity.
High Dimensionality	Handling high-dimensional and heterogeneous IoT data streams is inherently challenging. Despite effective large-scale data processing capabilities, the preprocessing steps, including feature selection and normalization, must be carefully optimized to avoid performance bottlenecks and maintain system efficiency.

Table 2: Limitations of the Proposed Approach for IoT Threat Detection systems

The research demonstrates success in integrating XAI techniques for IoT threat detection systems but significant acknowledgeable limitations still exist in the approach. Future research direction involves both real-time optimization of XAI algorithms and security-hardening measures for adverse risk mitigation. The

system represents a flexible solution for more transparent and actable improvements in IoT security which meets both industry and regulatory guidelines.

Future Directions

As IoT ecosystems continue to evolve, further enhancements in **Explainable AI (XAI) and cybersecurity frameworks** are necessary to improve **scalability, efficiency, and resilience**. While this research successfully integrates XAI techniques into IoT anomaly detection, several areas require further exploration to **enhance real-time processing, adversarial robustness, and decentralized learning**. Future research should focus on the following key directions:

1. Optimizing XAI for Real-Time Processing

- Current methods like SHAP, while effective, introduce **high computational overhead** in real-time environments.
- Future work can explore **lightweight interpretability techniques** that provide transparency **without compromising system speed**.
- **Hybrid XAI models** can be developed to balance computational efficiency and explainability in large-scale IoT applications.

2. Enhancing Adversarial Robustness

- Explainability can expose model vulnerabilities, allowing attackers to **exploit feature importance insights** for evasion.
- Future research should integrate **adversarial training techniques** to counteract attacks while maintaining explainability.
- **Differential privacy mechanisms** can be explored to **secure model explanations** against reverse engineering by adversaries.

3. Incorporating Federated Learning for IoT Security

- IoT devices generate massive amounts of **decentralized data**, making centralized model training inefficient and less privacy-compliant.
- **Federated learning** can enable local training across IoT nodes, enhancing **privacy and reducing communication overhead**.
- Future research should explore **federated XAI models**, ensuring that explanations remain interpretable across distributed IoT systems.

4. Developing Automated Anomaly Response Mechanisms

- While XAI enhances detection, **automated response systems** can significantly improve remediation times.
- Research should focus on **self-adaptive security mechanisms**, integrating **reinforcement learning** with XAI for **intelligent threat response**.
- Implementing **real-time anomaly mitigation protocols** will reduce reliance on human intervention in cybersecurity operations.

5. Improving XAI Usability and Visualization Tools

- Current XAI explanations require **technical expertise** to interpret, limiting usability for SOC analysts.
- Future advancements should focus on **natural language explanations**, allowing models to describe decisions in **human-readable formats**.
- **Advanced visualization tools**, integrating interactive dashboards with **real-time threat intelligence**, can improve adoption across organizations.

As IoT networks expand and cyber threats become more sophisticated, these future research directions will ensure that **XAI-driven cybersecurity systems remain efficient, scalable, and secure**. Addressing these challenges will enhance **trust, usability, and real-time threat detection**, ultimately leading to **more resilient IoT security frameworks** in practical deployments.


CONCLUSION

The integration of Explainable AI (XAI) techniques into IoT cybersecurity has significantly enhanced the interpretability, reliability, and usability of machine learning-driven anomaly detection. This research demonstrated how SHAP and LIME transformed traditional "black-box" classification models like Random Forest and SVM into transparent, interpretable security tools that provide actionable insights for cybersecurity

analysts. By incorporating these explainability techniques, the research enabled analysts to validate predictions effectively, reducing uncertainty in anomaly detection. Additionally, the distributed processing platform built with Apache Spark and Kafka proved essential for real-time IoT security monitoring, ensuring the system could efficiently analyze vast and dynamic IoT networks while maintaining high-speed threat detection.

Beyond improving interpretability, this research highlights the importance of explainability in regulatory compliance and operational efficiency. Trustworthy AI systems must be transparent and interpretable to align with regulatory standards such as GDPR, ensuring accountability in security decisions. The ability to reduce false positives is another crucial benefit, as it enhances SOC (Security Operations Center) efficiency by minimizing unnecessary alerts and reducing analyst fatigue. Furthermore, the explainable security framework streamlined cybersecurity workflows, enabling security teams to respond faster to potential threats while maintaining a high level of accuracy and interpretability in IoT security operations. Looking ahead, future research should focus on enhancing XAI adaptability for real-time applications, reducing computational complexity, and fortifying models against adversarial attacks. The evolution of IoT security systems depends on balancing explainability with computational efficiency, ensuring that real-time threat detection remains both scalable and interpretable. Additionally, integrating more advanced security mechanisms, such as federated learning and adversarial training, could further strengthen the resilience of XAI-driven IoT security frameworks. By continuing to refine and expand these methodologies, the future of IoT security will see stronger, more transparent, and highly reliable AI-driven threat detection systems, capable of adapting to an ever-changing cybersecurity landscape.

REFERENCES

- 
- L. T. S. R. R. G. Shashi Rekha, "Study of security issues and solutions in Internet of Things (IoT)," Materials Today: Proceedings, 2021.
- P. Walton, "Artificial Intelligence and the Limitations of Information," Information, vol. 9, 2018.
- J. C. N. A. M. A. ANNA NAMRITA GUMMADI, "XAI-IoT: An Explainable AI Framework for Enhancing Anomaly Detection in IoT Systems," vol. 11, 2024.
- astral_fate, "IoT23-dataset," 2023. [Online]. Available: <https://www.kaggle.com/datasets/astalfate/iot23-dataset>.
- A. A. J. F. J. S. M. Z. B. S. Eryk Schiller, "Landscape of IoT security," Computer Science Review, vol. 44, 2022.
- M. Gopalsamy, "Artificial Intelligence (AI) Based Internet-of-Things (IoT)-Botnet Attacks Identification Techniques to Enhance Cyber security," International Journal of Research and Analytical Reviews (IJRAR), vol. 7, no. 4, 2020.
- D. D. R. C. M. J. R. C. N. R. N. U. A. Petar Radanliev, "AI security and cyber risk in IoT systems," Sec. Cybersecurity and Privacy, vol. 7, 2024.
- A. I. K. Y. B. A. F. A. Iqbal H. Sarker, "Internet of Things (IoT) Security Intelligence: A Comprehensive Overview, Machine Learning Solutions and Research Directions," Mobile Networks and Applications, vol. 28, p. 296–312, 2023.
- *, Z. R.-E. I. B. G. P. R. S. E. P. K. L. a. G. M. Ahmed M. Salih, "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," Advanced Intelligent Systems, vol. 7, 2025.
- P. G. Alex Gramegna, "SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk," Sec. AI in Finance, vol. 4, 2021.

Q.-V. P. R. R. Z. Y. C. X. A. Z. Z. SENTHIL KUMAR JAGATHEESAPERUMAL, "Explainable AI Over the Internet of Things (IoT): Overview, State-of-the-Art and Future Directions," IEEE Open Journal of Communications Society, vol. 3, 2022.

