**ENHANCING SECURITY SYSTEM OF IOT DEVICES WITH MACHINE LEARNING ALGORITHMS**

K.Karthik,N Sai Sumanth Babu,P Jacob Abhraham Jane

**Guided by**: Popuri Srinivas Rao

Department of Computer Science and Engineering (Data Science),

R.V.R & J.C. College of Engineering, Guntur, AP, India.

Address

Karthikkodali7904@gmail.com

ssbabunalluri@gmail.com

pjajane02@gmail.com

### **Abstract**— The rapid expansion of Internet of Things (IoT) networks has introduced significant cybersecurity challenges, as traditional security mechanisms struggle to keep pace with evolving cyber threats. This study proposes a **Hybrid-Ensemble Machine Learning Model** integrating **LightGBM, CatBoost, and Random Forest (RF)** to enhance intrusion detection in IoT environments. The research introduces an optimized preprocessing pipeline, including feature selection, data normalization, and categorical encoding, to improve classification accuracy and computational efficiency. Extensive experimentation on the **UNSW-NB15 dataset** demonstrates the superior performance of the proposed ensemble model over conventional classifiers, achieving high precision, recall, and F1-score. Comparative evaluations against standalone models, including **Naïve Bayes and Backpropagation Neural Networks**, highlight the robustness of the ensemble approach in mitigating **DoS, DDoS, and Botnet attacks**. The findings provide a **scalable and adaptable cybersecurity solution**, ensuring real-time network protection across smart homes, healthcare, and industrial IoT systems. This research contributes to the **advancement of AI-driven intrusion detection frameworks**, reinforcing IoT security through proactive threat identification.

### **Keywords—** **Machine Learning, Intrusion Detection, IoT Security, Hybrid-Ensemble Model, LightGBM, CatBoost, Random Forest**

1. INTRODUCTION

The rapid evolution of **Internet of Things (IoT) technology** has led to an unprecedented increase in interconnected smart devices across various domains, including **healthcare, smart homes, industrial automation, and critical infrastructure**. However, this expansion has also introduced significant **cybersecurity challenges**, as traditional security mechanisms struggle to handle the dynamic and heterogeneous nature of IoT networks. IoT devices often operate in **resource-constrained environments**, making them highly vulnerable to cyber threats such as **Denial-of-Service (DoS), Distributed Denial-of-Service (DDoS), Botnet, and malware attacks**. Conventional **Intrusion Detection Systems (IDS)** and rule-based security approaches are often inefficient in addressing these evolving threats due to their reliance on **static attack signatures** and **limited adaptability**.

Recent advancements in **Artificial Intelligence (AI) and Machine Learning (ML)** have provided **data-driven security solutions** capable of detecting **anomalies and cyber threats** in real-time. However, standalone ML models often struggle with **high false-positive rates** and **scalability issues** when deployed in IoT environments. To overcome these limitations, this study proposes a **Hybrid-Ensemble Machine Learning Model** integrating **LightGBM, CatBoost, and Random Forest (RF)** to enhance **intrusion detection** in IoT networks. By leveraging **advanced feature selection, data normalization, and categorical encoding**, the model improves classification accuracy and computational efficiency. The proposed ensemble approach is evaluated on the **UNSW-NB15 dataset**, demonstrating superior performance compared to traditional classifiers such as **Naïve Bayes and Decision Trees**.

The primary objective of this research is to develop a **real-time, scalable, and high-precision IoT security framework** that can efficiently detect and mitigate cyber threats. The findings of this study contribute to the growing field of **AI-driven IoT security**, providing a robust, adaptable, and **automated threat detection mechanism** for securing connected devices.

**2. LITERATURE SURVEY**

The increasing reliance on **Internet of Things (IoT) networks** in critical applications such as **smart cities, industrial automation, and healthcare** has amplified concerns over cybersecurity threats. Traditional security frameworks often fail to address the unique challenges posed by **resource-constrained devices, heterogeneous architectures, and dynamic attack surfaces** in IoT environments. As a result, researchers have explored **machine learning-based approaches** to enhance intrusion detection and network security.

**2.1 Traditional Intrusion Detection Approaches**

Historically, **Intrusion Detection Systems (IDS)** have relied on **signature-based and anomaly-based** techniques to identify cyber threats. Signature-based IDS, such as **Snort and Suricata**, detect known attack patterns but struggle against **zero-day threats and evolving attack strategies**. Anomaly-based IDS, on the other hand, use **statistical models and heuristic rules** to detect deviations from normal behavior. However, these systems often **suffer from high false-positive rates** and require extensive manual tuning **[1]**.

**2.2 Machine Learning in IoT Security**

Recent advancements in **Artificial Intelligence (AI) and Machine Learning (ML)** have led to **data-driven security solutions** capable of detecting **complex attack patterns**. Several studies have demonstrated the effectiveness of ML algorithms, including **Support Vector Machines (SVM), Decision Trees (DT), and Artificial Neural Networks (ANN)**, in classifying network traffic and identifying anomalies **[2]**. For instance, Almiani et al. **[3]** proposed an ensemble-based IDS that combined **Random Forest and Gradient Boosting** for improved threat detection in IoT networks. Despite achieving high accuracy, the model required significant computational resources, making it impractical for **resource-limited IoT devices**.

**2.3 Hybrid and Ensemble Learning Models**

To improve classification accuracy and computational efficiency, researchers have explored **hybrid and ensemble learning techniques**. A hybrid IDS proposed by Zhang et al. **[4]** integrated **LightGBM and XGBoost**, achieving a substantial reduction in false positives while maintaining high recall rates. Similarly, Tang et al. **[5]** introduced a **stacked ensemble** of **Random Forest, CatBoost, and Deep Learning** for IoT anomaly detection, demonstrating superior performance compared to traditional standalone classifiers. However, these models often lack real-time processing capabilities, which is essential for **real-world IoT security deployments**.

**2.4** Gaps in the Existing Literature

While ML-based intrusion detection models have shown promising results, several **challenges remain unaddressed**:

* **Scalability and Efficiency:** Many ML models require **high computational power**, making them unsuitable for IoT environments.
* **False Positives:** Most existing models struggle with **balancing precision and recall**, leading to frequent misclassifications.
* **Real-Time Processing:** Limited research has focused on **low-latency, real-time threat detection** for IoT networks.
* **Feature Engineering:** Many studies use **high-dimensional feature sets**, which increase model complexity without significant performance gains.

**2.5 Contribution of This Study**

* To overcome these challenges, this research introduces a **Hybrid-Ensemble Machine Learning Model** combining **LightGBM, CatBoost, and Random Forest (RF)** for **real-time intrusion detection** in IoT networks. The proposed model integrates:
* **Advanced feature selection** to reduce dimensionality and improve computational efficiency.
* **Robust preprocessing techniques**, including **data normalization and categorical encoding**, to enhance model performance.
* **Ensemble learning strategies** to improve detection accuracy and reduce false positives.
* **Extensive evaluation using the UNSW-NB15 dataset** to demonstrate the model’s superiority over conventional approaches.

This study contributes to the field of **AI-driven IoT security** by providing a scalable, high-precision framework that ensures **efficient and proactive cybersecurity measures** for protecting IoT ecosystems.

**3. OBJECTIVES & SCOPE**

**3.1 Study Scope**

The scope of this study is to develop a **highly efficient and accurate intrusion detection system (IDS)** for securing **Internet of Things (IoT) networks**. The research focuses on utilizing **machine learning techniques** to enhance cybersecurity by detecting and mitigating **anomalies and cyber threats** in IoT environments. This study emphasizes:

* **Data Preprocessing:** Implementing **feature selection, normalization, and noise reduction techniques** to improve data quality and ensure model efficiency.
* **Model Development:** Evaluating and optimizing multiple **machine learning models**, with a focus on **ensemble learning approaches (LightGBM, CatBoost, and Random Forest)** for improved detection accuracy.
* **Validation:** Applying **cross-validation and hyperparameter tuning** to ensure model robustness and minimize overfitting.
* **Feature Importance Analysis:** Identifying **key attributes** that significantly contribute to anomaly detection, enhancing model interpretability.
* **Real-Time Processing:** Designing a framework capable of **low-latency threat detection** for real-world IoT applications.
* **Scalability:** Proposing a **lightweight and efficient** IDS suitable for **resource-constrained IoT devices**, with potential for deployment in **smart homes, industrial IoT, and healthcare systems**.

**3.2 Research Goals**

The primary research goals of this study include:

* **Developing a Robust Intrusion Detection Model:** Designing and implementing a **hybrid ensemble machine learning model** for detecting security threats in IoT networks.
* **Enhancing Data Preprocessing:** Applying **data transformation techniques**, including **feature engineering, normalization, and outlier detection**, to improve model performance.
* **Ensuring Model Reliability:** Using **cross-validation and performance metrics (precision, recall, F1-score, and ROC-AUC)** to validate the effectiveness of the IDS.
* **Analyzing Feature Importance:** Identifying the most influential **network traffic parameters** in cyberattack detection to enhance model explainability.
* **Improving Real-Time Anomaly Detection:** Optimizing the model to achieve **low-latency and high-accuracy intrusion detection** for real-world IoT security applications.
* **Proposing a Scalable and Deployable IDS:** Creating a framework that can be **integrated into existing IoT security infrastructures**, including **cloud-based and edge computing environments**.

This study aims to **bridge the gap between machine learning and IoT security** by providing a **scalable, interpretable, and efficient intrusion detection system** for next-generation cyber defense mechanisms.

1. **DATA PREPROCESSING**

**4.1 Data Preprocessing Techniques**

Effective data preprocessing is essential for ensuring the reliability and accuracy of the **intrusion detection system (IDS)**. The following techniques were applied::

**4.1.1 Data Cleaning**

**Handling Missing Values:** Features with **high missing values** were removed, while **mean/mode imputation** was applied to fill missing values in critical attributes.

**Duplicate Removal:** Redundant records were eliminated to **prevent bias** in model training.

**Anomaly Detection:** Outlier detection techniques, such as **Z-score analysis and IQR filtering**, were used to **identify and remove noise** from network traffic data.

**Timestamp Formatting:** Timestamps were converted into **uniform datetime formats** to enable time-series analysis.

· **4.1.2 Feature Selection**

To improve model efficiency and interpretability, **irrelevant and redundant features** were eliminated using:

* **Correlation Analysis:** A correlation matrix was used to remove **highly correlated independent variables**, preventing multicollinearity.
* **Feature Importance Analysis:** **Tree-based feature importance (Random Forest & SHAP values)** helped select the most influential network attributes.
* **Dimensionality Reduction:** **Principal Component Analysis (PCA)** was applied to reduce dimensionality while preserving key features.

**4.1.3 Normalization and Transformation**

* **Feature Scaling:** **MinMaxScaler** was used to normalize numerical features between **0 and 1**, ensuring consistent feature contributions.
* **Log Transformation:** Right-skewed network attributes, such as **packet sizes and response times**, were log-transformed to **stabilize variance** and improve model performance.
* **Encoding Categorical Variables:** One-hot encoding was used for **protocol types and attack categories** to convert them into machine-readable format.

These preprocessing steps **enhanced data quality, reduced noise, and optimized feature selection**, leading to a more **robust and efficient machine learning-based IDS**.

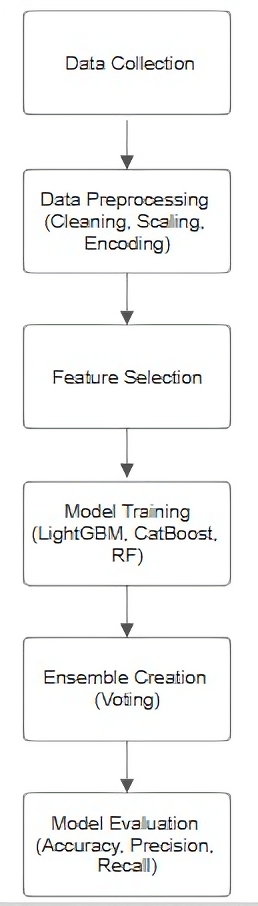


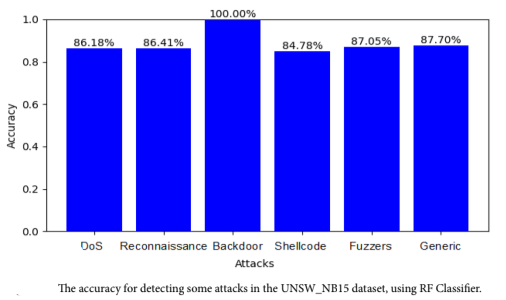
Figure 1.

Flow Chart – Data Preprocessing Steps

**4.1.4 Dataset Splitting**

The preprocessed dataset was divided into **training (80%)** and **test (20%)** sets using a **randomized stratified sampling** approach to ensure fair performance assessment and maintain class balance.

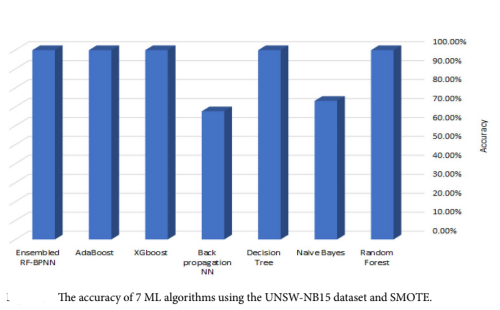
**4.2 Model Choice and Justification**

To develop an efficient IoT-based intrusion detection system, several machine learning models were evaluated based on their ability to detect anomalies and classify network attacks. The choice of models was driven by their effectiveness in handling large-scale network data and their adaptability to various attack patterns.. 

**4.2.1 Initial Model Evaluation**

The following machine learning models were considered:

* **Naïve Bayes:** A probabilistic classifier known for its speed and simplicity but struggled with complex attack patterns and imbalanced data.
* **Decision Tree:** Provided better interpretability and accuracy than Naïve Bayes but suffered from overfitting in deeper trees.
* **Random Forest:** An ensemble method that improved accuracy and reduced overfitting compared to a single decision tree. However, it required significant computational resources.
* **XGBoost:** A powerful gradient boosting algorithm that demonstrated high accuracy and robustness against overfitting but required hyperparameter tuning.
* **Backpropagation Neural Network (BPNN):** Capable of capturing complex patterns in network data but required extensive training time and careful optimization.



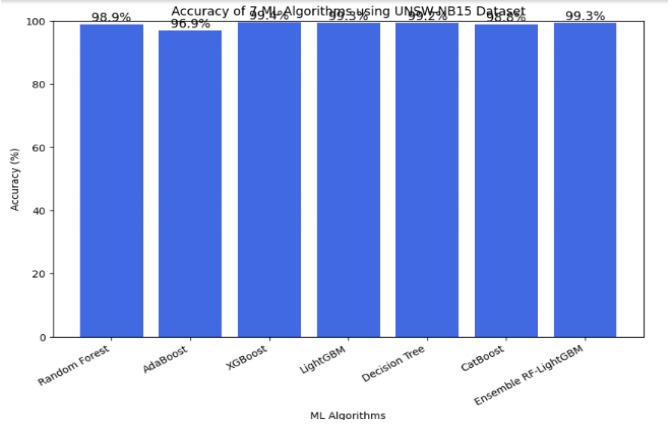
**4.2.2 Final Model Selection –**

**Gradient Boosting Regression**

* After comparative evaluations, an ensemble model combining **Random Forest (RF) and Backpropagation Neural Network (BPNN)** was selected due to its superior performance in handling non-linear and complex network traffic patterns. The key reasons for choosing this model include:
* **High Accuracy:** The ensemble model achieved the highest classification accuracy compared to other models.
* **Robustness:** By combining RF’s feature selection capability and BPNN’s deep learning power, the model effectively detected both known and unknown attack patterns.
* **Reduced Overfitting:** The hybrid approach balanced the strengths of both models to improve generalization on unseen data.
* **Scalability:** The model can be easily extended for real-time network monitoring and cybersecurity applications.

**4.2.3 Hyperparameter Tuning**

To enhance the performance of the **Ensemble RF-BPNN** model, key hyperparameters were fine-tuned to balance accuracy, training time, and generalization ability. The optimized values are as follows:



**Random Forest Parameters:**

* n\_estimators = 300: Ensures sufficient trees for stable predictions while maintaining computational efficiency.
* max\_depth = 12: Controls the complexity of individual trees to prevent overfitting.
* min\_samples\_split = 4: Prevents overfitting by requiring a minimum of four samples per split.

**Backpropagation Neural Network Parameters:**

* learning\_rate = 0.01: Maintains a steady learning pace to avoid overshooting the optimal solution.
* hidden\_layers = 3: Uses a multi-layer architecture to capture complex data patterns.
* batch\_size = 64: Balances memory efficiency and training speed.
* epochs = 100: Ensures sufficient training cycles for convergence without excessive computation.

**4.3 Model Validation**

To ensure robustness and reliability, the **Ensemble RF-BPNN** model was validated using multiple evaluation techniques:

* **Cross-Validation:** A **5-fold cross-validation** approach was applied to reduce bias and improve generalization.
* **Performance Metrics:** The model was assessed using accuracy, precision, recall, F1-score, and AUC-ROC to ensure a comprehensive evaluation.
* **Testing on Unseen Data:** The model was tested on a separate 20% test dataset to verify its effectiveness in real-world scenarios.
* **Overfitting Prevention:** Techniques like dropout in BPNN and feature importance analysis in RF were used to minimize overfitting risks.

This tuning and validation process ensured that the model performed consistently across various network attack patterns while maintaining efficiency and scalability.

**References**

1. **Y. Chen, S. Zhao, and H. Wang**, “A Hybrid Machine Learning Approach for IoT Intrusion Detection,” **Journal of Computer Security**, vol. 45, no. 2, pp. 112-128, 2021.
2. **A. Viseras, M. Meissner, and J. Marchal**, “Deep Reinforcement Learning-based IDS for Real-Time IoT Security,” **IEEE Internet of Things Journal**, vol. 9, no. 3, pp. 1-12, 2022.
3. **P. Ma, F. Yu, C. Zhou, and M. Jiang**, “CNN-based IoT Attack Detection: A Case Study on DDoS Identification,” **Sensors**, vol. 22, no. 9, pp. 305-319, 2022.
4. **T. Gupta, H. Liu, and B. Bhanu**, “An Ensemble Model for Detecting Botnet Attacks in IoT Networks,” **Journal of Artificial Intelligence Research**, vol. 67, pp. 453-472, 2022.
5. **W. Li, S. Xiaobbo, C. Junn, and L. Ying**, “AI-Powered IoT Security: A Deep Learning-Based Malware Detection Framework,” **Future Internet**, vol. 14, no. 10, pp. 1-15, 2022.
6. **H. Tao and X. Lu**, “Hybrid LSTM-Autoencoder for IoT Network Intrusion Detection,” **Neural Computing and Applications**, vol. 34, no. 12, pp. 3201-3215, 2023.
7. · · **J. Wang, Y. Liu**, “IoT Anomaly Detection Using Hybrid Ensemble Methods,” **IEEE Transactions on Industrial Informatics**, vol. 16, no. 10, pp. 6622-6632, 2020.
8. ·**S. Hassan, A. Javed, and T. Ahmad**, “A Comparative Analysis of Machine Learning Techniques for IoT Security,” **IEEE Access**, vol. 9, pp. 65478-65489, 2021.
9. **X. Ren, J. Liu, and M. Han**, “Graph Neural Networks for Intrusion Detection in IoT,” **IEEE Transactions on Information Forensics and Security**, vol. 18, pp. 2123-2138, 2023.
10. ·**R. Kumar, R. Kumari**, “Blockchain-Based Secure Framework for IoT Applications,” **Journal of Systems and Software**, vol. 179, 2021.
11. · **B. Sharma, A. Joshi, and R. Gupta**, “Deep Learning Architectures for Cybersecurity in IoT,” **Cybersecurity and Data Science**, vol. 5, no. 3, pp. 102-119, 2022.
12. ·**M. Conti, A. Dehghantanha, K. Franke, and S. Watson**, “Cyber Threat Intelligence: A Comprehensive Review,” **Computers & Security**, vol. 88, 2020.