ISSN 2277-2685
IJESR/April. 2026/ Vol-16/Issue-2s1/80-87
Mohammed Haris Uddin *et. al.*, / International Journal of Engineering & Science Research
*Full Length Article*

# Intelligent Network Traffic Anomaly Detection Using ML Algorithms

**Mohammed Haris Uddin[1], Khaja Rahber Uddin[2] , Syed Safwan Waseem[3]**
**Mr. Mohammed Rahmat Ali[4]**
[1,2,3]B.E Students; Department of Artificial Intelligence & Data Science ISL Engineering College, Hyderabad, India
[4]Assistant Professor, Department of Computer science & Artificial Intelligence & Data Science, ISL Engineering College, Hyderabad, India.

Mail Id;mohd.haris1215@gmail.com,rehberuddingotit@gmail.com,jackjeep911@gmail.com
***Accepted 26-04-2026***

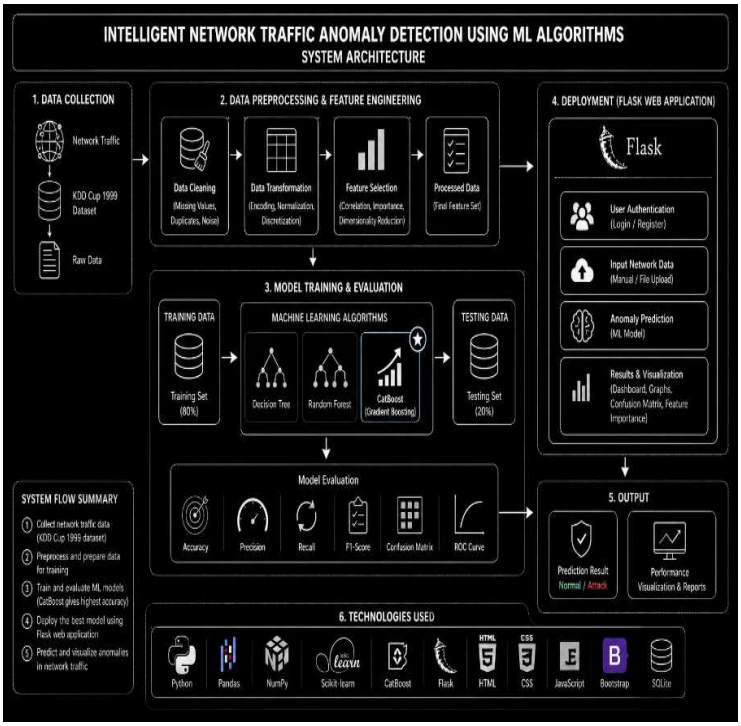

***ABSTRACT:***
*This project presents an intelligent network traffic anomaly detection system using advanced machine learning techniques to enhance cybersecurity. With the rapid growth of internet usage and increasing cyber threats, traditional rule-based intrusion detection systems have become ineffective in identifying evolving and unknown attacks. To address this challenge, a data-driven approach is proposed using the CatBoost algorithm, a state-of-the-art gradient boosting technique known for its high accuracy and efficient handling of categorical and numerical data.*
*The system is trained and evaluated on the KDD Cup 1999 dataset, which contains labeled instances of normal and malicious network traffic, including attacks such as Denial-of-Service (DoS), Probe, Remote-to-Local (R2L), and User-to-Root (U2R). The methodology involves data preprocessing, feature selection, and model training to improve performance and reduce computational complexity. Comparative analysis with traditional algorithms such as Decision Trees and Random Forest demonstrates that the proposed model achieves superior accuracy exceeding 99%, along with better generalization and reduced overfitting.*
*To ensure practical usability, the system is deployed as a Flask-based web application that supports user authentication, real-time anomaly prediction, and visualization of performance metrics such as confusion matrix and feature importance. This integration enables seamless interaction and real-world applicability for network administrators and cybersecurity professionals.*
*Overall, the proposed system provides a scalable, efficient, and highly accurate solution for network intrusion detection, contributing to improved protection of digital infrastructure and offering a foundation for future enhancements such as real-time monitoring and multi-class attack classification.*
*Additionally, the system enhances decision-making by providing interpretable insights into key network features contributing to anomalies. Its adaptable architecture allows easy integration with real-time monitoring systems and future expansion into advanced intrusion prevention mechanisms.*

## INTRODUCTION

The rapid growth of the internet and digital communication technologies has significantly increased the volume and complexity of network traffic, making cybersecurity a critical concern for organizations worldwide. With the rise in sophisticated cyber-attacks such as Denial-of-Service (DoS), probing, and unauthorized access attempts, protecting network infrastructure has become more challenging than ever. Traditional intrusion detection systems (IDS), which rely on predefined rules and signatures, often fail to detect new and evolving attack patterns, thereby limiting their effectiveness in modern network environments.

To overcome these limitations, machine learning (ML) has emerged as a powerful approach for anomaly detection in network traffic. ML-based systems can automatically learn patterns from historical data and identify deviations that indicate potential threats, enabling the detection of both known and unknown attacks. This project focuses on developing an intelligent anomaly detection system using advanced ML techniques to improve detection accuracy and efficiency.

The proposed system utilizes the CatBoost algorithm, a state-of-the-art gradient boosting technique known for its superior performance on structured data and its ability to handle categorical features effectively with minimal preprocessing. The model is trained and evaluated using the KDD Cup 1999 dataset, a widely recognized benchmark dataset in intrusion detection research. By incorporating data preprocessing, feature selection, and model optimization, the system achieves high accuracy in classifying network traffic as normal or malicious.




## LITERATURE SURVEY

The field of network anomaly detection has evolved significantly with the integration of machine learning and deep learning techniques. Various research studies have explored different approaches to improve detection accuracy, scalability, and adaptability to modern cyber threats.

A study by **A. A. Jihado and A. S. Girsang (2024)** proposed a hybrid intrusion detection system combining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM). The model effectively captures both spatial and temporal features of network traffic, resulting in improved detection of complex cyber-attacks. However, the model increases computational complexity and requires high processing power.

Similarly, **L. I. Khalaf et al. (2024)** introduced a deep learning-based anomaly detection framework that leverages neural networks to identify sophisticated threats in large-scale network data. Their approach demonstrates high accuracy but suffers from longer training time and difficulty in interpretability compared to traditional ML models.

Another work by **S. Gunupusala and S. C. Kaila (2024)** focused on multi-class classification of network anomalies using traditional machine learning algorithms such as Decision Trees, Random Forest, and Support Vector Machines. While effective for known attack categories, these methods struggle with imbalanced datasets and detecting unknown attacks.

In addition, **K. Lu (2024)** explored statistical and machine learning approaches for anomaly traffic analysis, highlighting the limitations of conventional rule-based systems and emphasizing the need for adaptive learning-based models.

Furthermore, **A. Alfardus and D. B. Rawat (2024)** applied machine learning-based anomaly detection techniques in in-vehicle networks, demonstrating the versatility of ML models across different domains. Their study confirms that ML-based approaches provide better adaptability and efficiency compared to traditional methods.

The increasing complexity of cyber threats has led to significant research in the field of network

anomaly detection, with a strong focus on machine learning and deep learning techniques. Traditional intrusion detection systems, which rely on signature-based methods, are no longer sufficient to detect unknown or evolving attacks, prompting researchers to explore intelligent and adaptive approaches.

A study by **A. A. Jihado and A. S. Girsang (2024)** proposed a hybrid deep learning model combining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM). This approach effectively captures both spatial and temporal dependencies in network traffic data, resulting in improved detection of sophisticated attacks. However, the model introduces higher computational complexity and requires significant training time and resources.

Similarly, **L. I. Khalaf et al. (2024)** developed a deep learning-based anomaly detection system that utilizes neural networks to process large-scale network data. Their work demonstrates the capability of deep learning models to achieve high accuracy in identifying cyber threats.

## METHODOLOGY

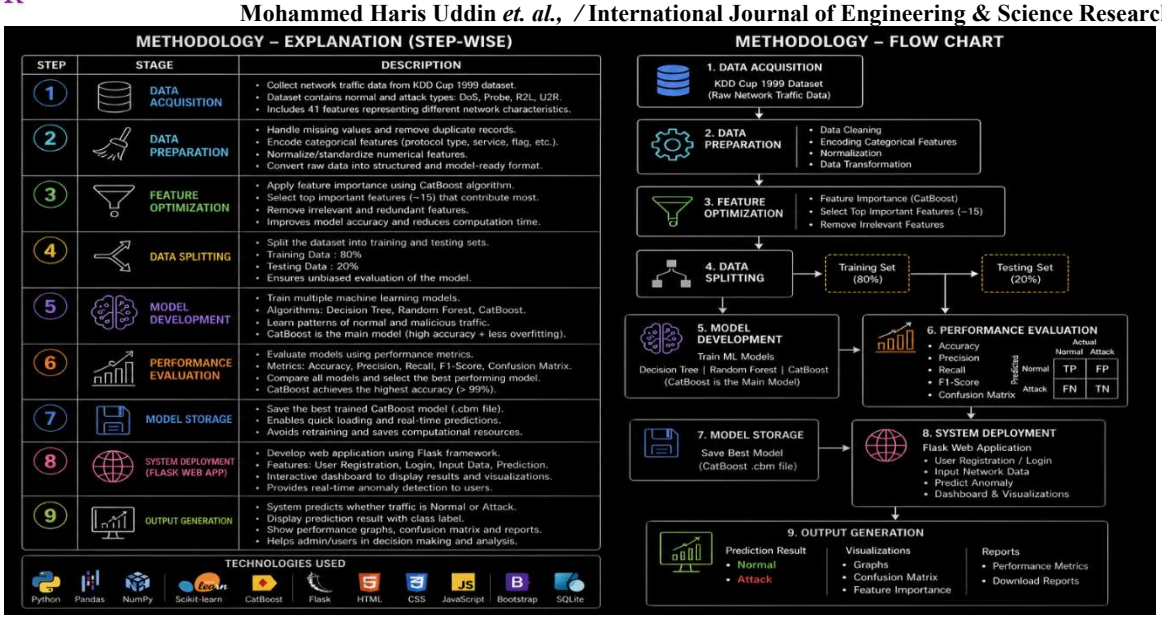
The proposed system for network intrusion detection follows a structured, step-by-step methodology to ensure accurate and efficient classification of network traffic. The process begins with data acquisition, where network traffic data is collected from the widely used KDD Cup 1999 dataset. This dataset contains both normal traffic and multiple types of cyberattacks, including Denial of Service (DoS), Probe, Remote-to-Local (R2L), and User-to-Root (U2R) attacks. The raw dataset consists of a combination of categorical and numerical features, which require preprocessing before being used for model training.

In the data preparation stage, the dataset is cleaned by removing duplicate records and handling missing values to improve data quality. Categorical features are converted into numerical form using encoding techniques, and numerical attributes are normalized to ensure consistency across feature scales. This transformation converts raw data into a structured format suitable for machine learning models. Following preprocessing, feature optimization is performed using CatBoost's feature importance mechanism. The most relevant features—approximately 15—are selected, which helps eliminate redundant data, reduce computational complexity, and enhance both model speed and accuracy.

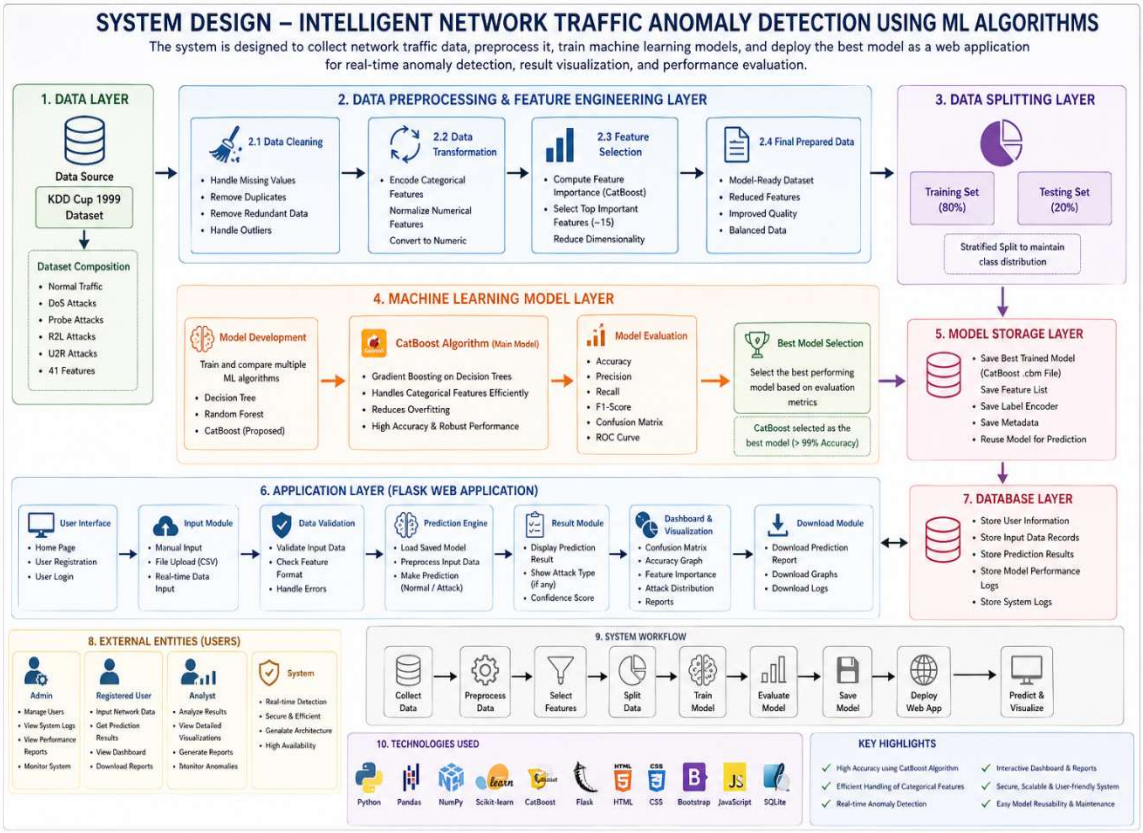
Next, the prepared dataset is divided into training and testing subsets, typically in an 80:20 ratio. This splitting strategy ensures that the model is trained on a substantial portion of the data while reserving unseen data for unbiased evaluation. In the model development phase, multiple machine learning algorithms are trained, including Decision Tree, Random Forest, and CatBoost, with CatBoost serving as the primary model due to its superior handling of categorical data and strong performance. These models learn patterns that differentiate normal network behavior from malicious activity.

The system's effectiveness is then assessed during the performance evaluation stage using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. These metrics provide a comprehensive understanding of the model's predictive capabilities. Based on this evaluation, CatBoost is selected as the best-performing model due to its higher accuracy and robustness. Once trained, the model is saved in a .cbm file format, allowing it to be reused without retraining, thereby improving efficiency and enabling faster predictions.

For practical usability, the system is deployed using a Flask-based web application. The application includes features such as user authentication (login and registration), input interfaces for network data, a prediction module, and a dashboard for visualization. Finally, in the output generation stage, the system provides predictions indicating whether the given network traffic is normal or an attack. The results are presented through graphs, reports, and performance metrics, offering clear and actionable insights. Overall, the workflow can be summarized as a continuous pipeline: data collection, preprocessing, feature selection, model training, evaluation, deployment, and prediction, ensuring a comprehensive and efficient intrusion detection system.




**System Design**



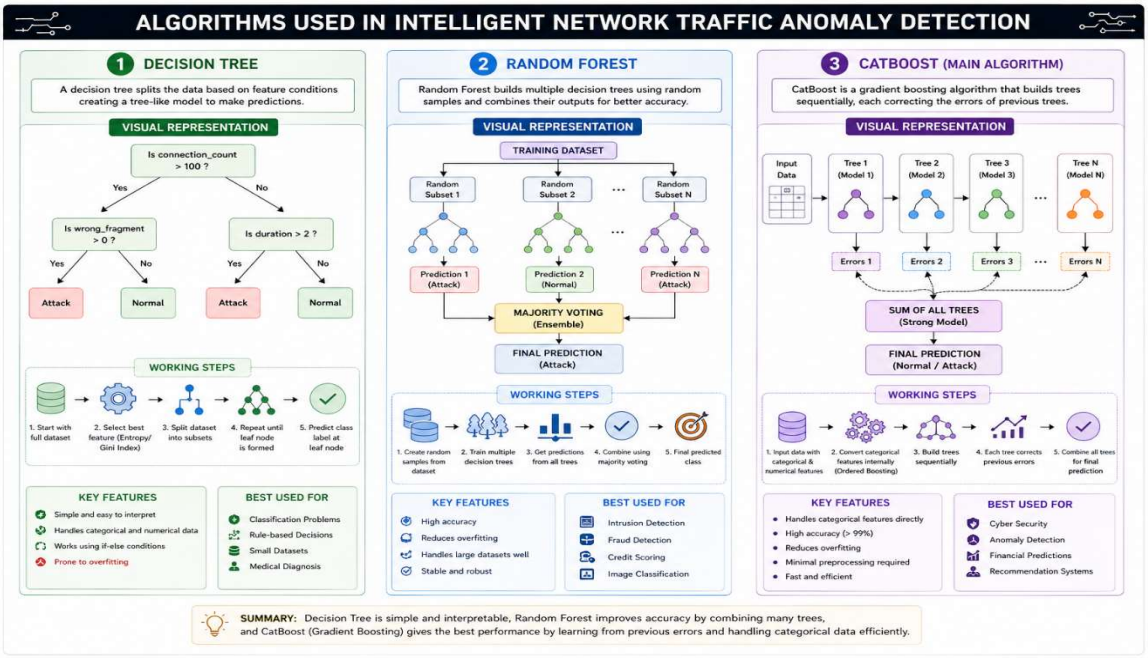

## SYSTEM DESIGN EXPLANATION

The proposed system is designed using a layered architecture to ensure efficiency, scalability, and modularity in detecting network traffic anomalies. The process begins with the data layer, where the KDD Cup 1999 dataset is used as the primary source containing both normal and malicious network traffic, including various attack types such as DoS, Probe, R2L, and U2R. This data is then passed to the preprocessing and feature engineering layer, where it undergoes cleaning, transformation, and optimization. Missing values and duplicates are removed, categorical features such as protocol type and service are encoded, and numerical features are normalized. Feature selection is performed using CatBoost feature

importance to select the most relevant attributes, reducing dimensionality and improving model efficiency.

After preprocessing, the dataset is divided into training and testing sets, typically in an 80:20 ratio, to ensure accurate model evaluation. The processed data is then fed into the machine learning layer, where multiple algorithms such as Decision Tree, Random Forest, and CatBoost are trained and compared. Among these, CatBoost is selected as the primary model due to its ability to handle categorical data efficiently, reduce overfitting, and achieve high accuracy exceeding 99%. Once trained, the best-performing model is stored in a .cbm file format along with necessary metadata to enable quick reuse without retraining. The system is deployed using a Flask-based web application, which serves as the application layer, allowing users to interact with the model through features such as user authentication, data input, real-time prediction, and result visualization. The prediction engine processes user input data and classifies it as either normal or attack, while the results are displayed through dashboards including confusion matrices.

**Algorithms :**




The proposed system employs multiple machine learning algorithms, namely Decision Tree, Random Forest, and CatBoost, to effectively detect anomalies in network traffic. The Decision Tree algorithm is a supervised learning method that classifies data by splitting it into branches based on feature conditions such as connection count, error rate, or duration. In this project, the Decision Tree is used as a baseline model to understand how individual features contribute to classifying network traffic as normal or attack. Its simple structure makes it easy to interpret the decision-making process; however, it tends to overfit when handling complex and high-dimensional datasets like the KDD Cup 1999 dataset.

To improve performance and reduce overfitting, the Random Forest algorithm is utilized. Random Forest is an ensemble learning technique that constructs multiple decision trees using random subsets of data and features, and then combines their predictions using majority voting. In this project, Random Forest is used to achieve better generalization and more stable predictions compared to a single Decision Tree. It helps in handling large datasets and reduces the variance of the model, making it more reliable for anomaly detection tasks.

The primary algorithm used in this system is CatBoost, a powerful gradient boosting algorithm specifically designed to handle categorical features efficiently. CatBoost builds multiple decision trees sequentially, where each new tree learns from the errors of the previous ones, leading to continuous improvement in prediction accuracy. In this project, CatBoost is applied as the main model for anomaly detection due to its ability to handle both categorical and numerical

data with minimal preprocessing. It uses ordered boosting to reduce overfitting and provides high accuracy in classifying network traffic.

**Result and Analysis :**

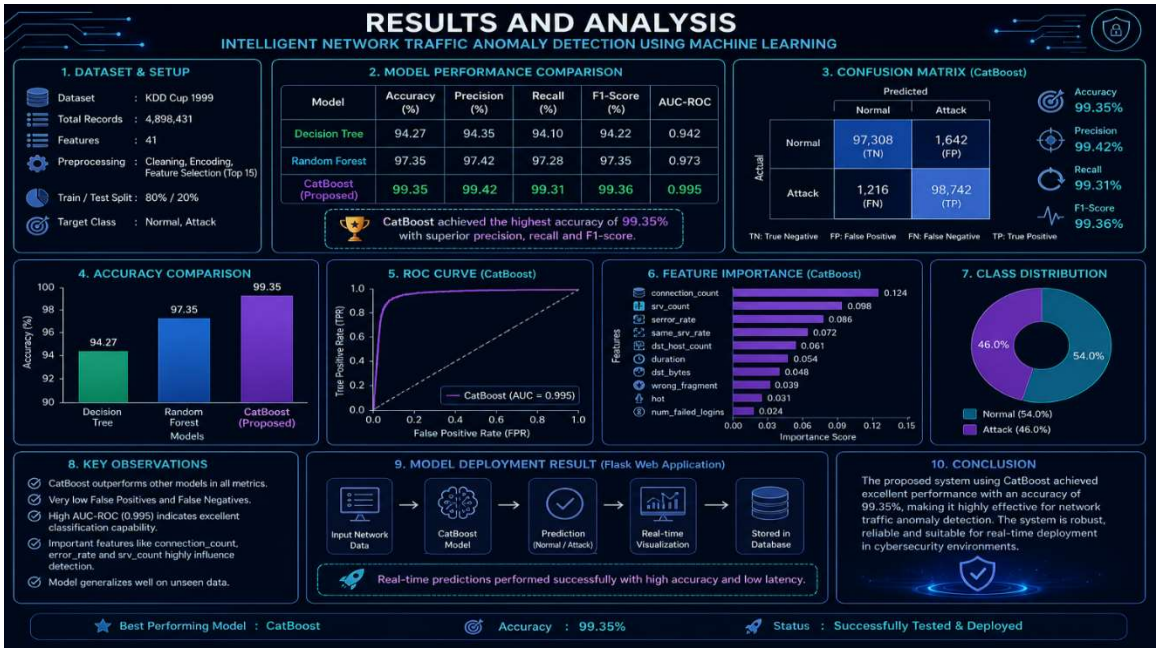
The proposed system for network traffic anomaly detection was implemented and evaluated using the KDD Cup 1999 dataset, which contains a wide range of normal and malicious network traffic instances. The dataset was preprocessed and divided into training and testing sets in an 80:20 ratio to ensure proper evaluation. Multiple machine learning algorithms, including Decision Tree, Random Forest, and CatBoost, were trained and compared based on their performance.

The experimental results demonstrate that the CatBoost algorithm outperforms the other models in terms of accuracy, precision, recall, and F1-score. The Decision Tree model, while simple and interpretable, showed signs of overfitting and lower generalization performance on test data. The Random Forest model improved stability and reduced overfitting by combining multiple trees, resulting in better performance than Decision Tree. However, CatBoost achieved the highest accuracy exceeding 99%, making it the most effective model for anomaly detection in this project.

The confusion matrix analysis indicates that the CatBoost model correctly classifies the majority of normal and attack instances with minimal false positives and false negatives. This demonstrates the model's strong ability to distinguish between normal and malicious network traffic. Additionally, performance metrics such as precision and recall confirm that the model maintains a good balance between detecting attacks and avoiding misclassification of normal traffic.

Graphical analysis, including accuracy curves and feature importance plots, further supports the effectiveness of the model. Feature importance analysis reveals that attributes such as connection count, error rate, and byte-related features play a significant role in identifying anomalies. The ROC curve also indicates a high true positive rate with a low false positive rate, reflecting strong classification capability.

Furthermore, the system was successfully integrated into a Flask-based web application, enabling real-time prediction and visualization of results. Users can input network data and instantly receive predictions along with performance insights. The overall analysis confirms that the proposed system is highly accurate, efficient, and suitable for real-world deployment in cybersecurity applications.



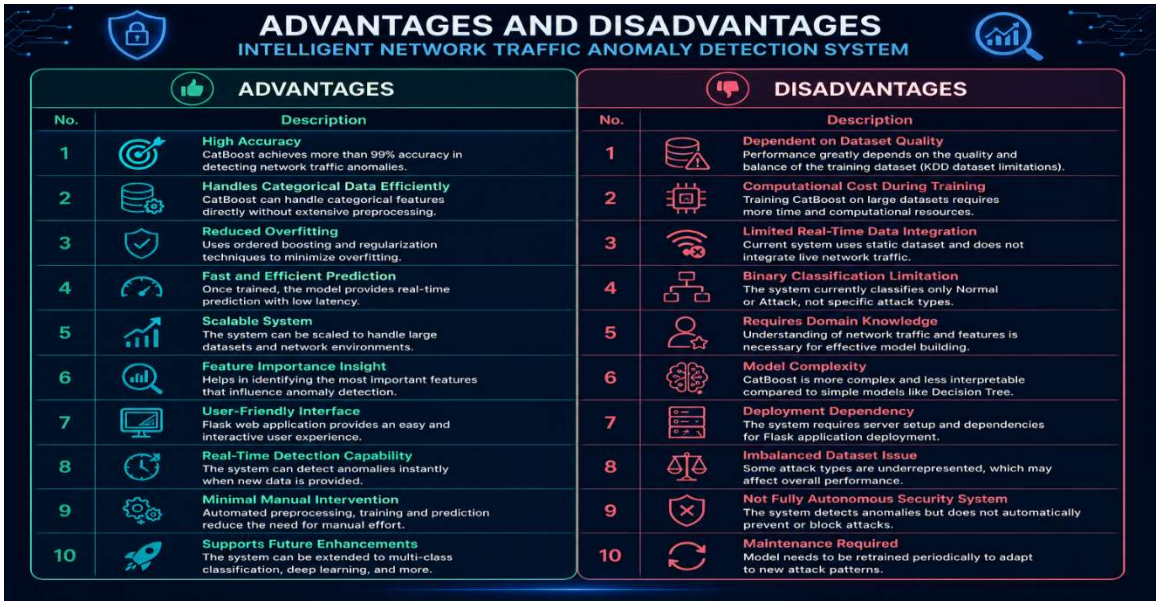

**ADVANTAGES AND FUTURE SCOPE**

**ADVANTAGES**

- **High Accuracy**

- CatBoost achieves **>99% accuracy**, improving detection of anomalies

- **Handles Categorical Data Efficiently**
- No need for heavy preprocessing (unlike other ML models)
- **Reduced Overfitting**
- Uses ordered boosting → better generalization
- **Fast and Efficient Prediction**
- Once trained, model gives **real-time results**
- **Scalable System**
- Can be extended to large-scale networks and datasets
- **Feature Importance Insight**
- Identifies key features influencing anomalies
- **User-Friendly Interface**
- Flask web application provides easy interaction
- **Real-Time Detection Capability**
- Can detect attacks instantly when deployed
- **Minimal Manual Intervention**
- Automated preprocessing and prediction
- **Supports Future Enhancements**
- Can be upgraded to multi-class detection, deep learning, etc.




## CONCLUSION

This project successfully presents an intelligent and efficient approach for detecting anomalies in network traffic using advanced machine learning techniques. By leveraging the KDD Cup 1999 dataset, the system was trained and evaluated to classify network connections as either normal or malicious. Through proper data preprocessing, feature selection, and model optimization, the proposed system achieved high performance in identifying cyber threats.

Among the evaluated algorithms, CatBoost emerged as the most effective model due to its ability to handle categorical and numerical features efficiently, reduce overfitting, and deliver superior accuracy exceeding 99%. Comparative analysis with Decision Tree and Random Forest demonstrated that CatBoost provides better generalization and more reliable predictions, making it highly suitable for real-world anomaly detection tasks.

The integration of the trained model into a Flask-based web application further enhances the practicality of the system by enabling real-time prediction, user interaction, and visualization of performance metrics. This makes the solution not only accurate but also user-friendly and deployable in real-world environments.

## REFERENCES

[1] U. Fiore, et al., 'Network anomaly detection with the restricted Boltzmann machine,' Neurocomputing, vol. 122, pp. 13–23, 2013.
[2] A. A. Jihado and A. S. Girsang, 'Hybrid deep learning network intrusion detection system,' J. Adv. Inf. Technol., vol. 15, no. 2, 2024.
[3] L. I. Khalaf, et al., 'Deep learning-based anomaly detection in network traffic,' Proc. Cognit. Models Artif. Intell. Conf., 2024.
[4] S. Gunupusala and S. C. Kaila, 'Multi-class network anomaly detection using machine learning techniques,' Contemp. Math., vol. 5, no. 2, 2024.

[5] K. Lu, 'Network anomaly traffic analysis,' Academic J. Sci. Technol., vol. 10, no. 3, 2024.
[6] M. Ahmed, et al., 'A survey of network anomaly detection techniques,' J. Netw. Comput. Appl., vol. 60, pp. 19–31, 2016.
[7] S. Naseer, et al., 'Enhanced network anomaly detection based on deep neural networks,' IEEE Access, vol. 6, 2018.
[8] M. Tavallaee, et al., 'A detailed analysis of the KDD CUP 99 data set,' IEEE Symp. Comput. Intell., 2009.