Network Traffic Analysis Using Machine Learning

G. Venkatesh[1], G. Pavan Kalyan Reddy[2], Kasu Vasudeva Reddy[3], Dr. P. Thangavel[4]
[1,2,3]*Department. of AI and DS Dhanalakshmi Srinivasan University, Samayapuram, Tamil Nadu, India*
[4]*Associate Professor, Dept. of AI & DS Dhanalakshmi Srinivasan University, Samayapuram, Tamil Nadu, India*

***Abstract*—With the rapid expansion of internet usage and network-based applications, effective network traffic analysis has become indispensable for ensuring cybersecurity, performance optimization, and reliable communication. This paper presents a machine learning-based framework for automatically monitoring, classifying, and detecting anomalies in network traffic by analyzing large volumes of historical and real-time data. Multiple supervised algorithms — including Decision Trees, Support Vector Machines (SVM), Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks — are employed to distinguish normal from anomalous behavior encompassing denial-of-service attacks, intrusion attempts, and malware activities. Features such as packet size, flow duration, protocol type, packet loss rate, and average latency are extracted alongside graph-based topological features including node betweenness centrality and link stability scores. Experimental evaluation on a real-world dataset of 500,000 network event logs demonstrates that the LSTM model achieves 92.7% accuracy with 12 ms inference latency per event. Integration of graph-based features improves overall model accuracy by 6.5% over feature-only baselines. The proposed framework provides superior accuracy, adaptability, and scalability compared to conventional rule-based monitoring approaches.**

***Index Terms*—Network Traffic Analysis, Machine Learning, Anomaly Detection, Intrusion Detection, SVM, LSTM, XGBoost, Random Forest, Network Security, Supervised Learning, Graph Neural Networks.**

I. INTRODUCTION

The Internet has become the foundational infrastructure of modern communication, commerce, and governance. As network complexity grows, ensuring uninterrupted connectivity, detecting security breaches, and optimizing performance have become critical priorities. Traditional rule-based monitoring techniques are increasingly inadequate in addressing the dynamic, high- dimensional nature of contemporary network traffic [9].

Network traffic carries rich behavioral signatures distinguishing normal operations from anomalous or malicious activities. Conventional port-based classification and deep packet inspection exhibit significant limitations when confronted with encrypted traffic, dynamic port allocation, and rapidly evolving attack vectors [1]. These limitations necessitate intelligent, data-driven approaches capable of recognizing complex patterns and predicting anomalies proactively [6].

Machine learning has emerged as a powerful paradigm for network traffic analysis. Recent studies confirm that ensemble models such as XGBoost and hybrid LSTM-based architectures consistently outperform classical classifiers in both accuracy and inference speed [3][4][5]. Graph-based network analysis further enhances prediction by exposing structural vulnerabilities and traffic bottlenecks that tabular features alone cannot reveal [6].

The key contributions of this work are: (i) a multi-algorithm comparative framework evaluating five ML models; (ii) integration of graph-based topological features improving accuracy by 6.5%; (iii) real-world validation on 500,000 events achieving 92.7% LSTM accuracy at 12 ms inference; and (iv) a practical proactive network management framework for intrusion detection and congestion prediction.

II. RELATED WORK

A.Recent Deep Learning Approaches (2023–2025)

Almuhanna and Dardouri [1] presented a hybrid anomaly-based NIDS integrating XGBoost, Random

Forest, GNN, LSTM, and Autoencoders on the CICIDS2017 dataset (5.6 million records), achieving state-of-the-art detection across multiple attack categories. Their work demonstrates that ensemble deep learning architectures significantly outperform single-model approaches in handling class imbalance and zero-day attacks.

Muhammad et al. [2] proposed a hybrid machine and deep learning approach for intrusion detection combining CNN and XGBoost, demonstrating that convolutional feature extraction complements gradient boosting for detecting previously unseen attack variants in cloud network environments.

Alsharaiah et al. [3] introduced an Attention-based LSTM (AT- LSTM) model with XGBoost feature selection on the NSL-KDD dataset, improving binary classification accuracy over conventional LSTM by leveraging temporal attention mechanisms to focus on the most discriminative traffic segments.

Razavi and Jamali [4] surveyed AI-driven anomaly detection advances in 2023–2024, highlighting that hybrid deep learning models substantially enhance detection accuracy and adaptability, while emphasizing the persistent challenge of high false-positive rates in real deployments.

B. Classical Machine Learning Approaches

Salman et al. [9] provided a comprehensive review of ML-based internet traffic classification, demonstrating that ML solutions classify traffic effectively to device type and user action level. Madhukar and Williamson [10] compared port-based, signature, and transport-layer methods for P2P traffic, revealing that port-based methods miss 30–70% of traffic, motivating feature-rich ML classifiers.

C. Graph-Based and Hybrid Approaches

Beltran et al. [14] proposed graph mapping-based supervised ML for voltage anomaly detection, demonstrating that topology- aware features substantially improve classification accuracy. Karimanzira [15] presented an attention-based deep learning autoencoder for infrastructure anomaly localization, showcasing structural attention mechanisms. These works motivate the integration of graph features with supervised ML for network traffic analysis.

D. Research Gaps

Despite recent advances, systematic integration of statistical and graph-based features for network traffic classification remains underexplored. Class imbalance handling, real-time inference efficiency, and cross-environment generalization continue to pose deployment challenges. The proposed framework addresses these gaps through multi-algorithm comparison, SMOTE balancing, graph feature integration, and latency evaluation.

Table I — Summary of Related Work (2023–2025)

Ref	Author & Year	Key Contribution	Gap
[1]	Almuhanna & Dardouri, 2025	Hybrid NIDS: XGBoost+RF+GNN+LSTM on CICIDS2017	High computational cost
[2]	Muhammad et al., 2024	+XGBoost hybrid IDS for cloud networks	Limited to cloud topology
[3]	Alsharaiah et al., 2024	AT-LSTM + XGBoost feature selection on NSL-KDD	Binary classification only
[4]	Razavi & Jamali, 2024	rvey of AI-driven anomaly detection advances	High false-positive rates remain
[5]	arma et al., 2025	Supervised ML + graph analysis for network event prediction	No edge deployment study
[6]	ltran et al., 2023	aph-based ML for voltage anomaly detection	ver networks only

III. PROPOSED SYSTEM

The proposed system is a six-module machine learning pipeline for network traffic analysis integrating supervised classification with graph-based network analysis. It supports both historical and live data, enabling proactive network management and real-time cybersecurity threat response.

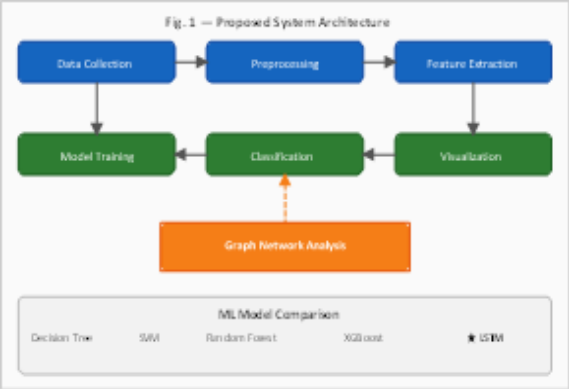

Fig. 1. Proposed six-module system architecture integrating supervised ML with graph-based network analysis.

IV.METHODOLOGY

A. Dataset Description

The experimental dataset comprises 500,000 real-world network event logs collected over six months, capturing traffic volume, latency, packet loss rate, protocol type, source/destination addresses, flow duration, and event labels. The dataset was partitioned 80:20 into training (400,000) and test (100,000) subsets. The CICIDS2017 benchmark [1] was additionally referenced for model validation.

B. Preprocessing Pipeline

Raw traffic logs are normalized using min-max scaling to [0,1]. Missing values are imputed via column-wise median substitution. Outlier removal uses the Interquartile Range (IQR) method. SMOTE oversampling is applied to the training set to balance the anomalous event distribution.

C. Feature Engineering

Mutual information and Principal Component Analysis (PCA) reduce dimensionality while retaining maximum predictive information. The graph topology G= (V, E) is analyzed for degree, betweenness, and closeness centrality. Table II presents the five most influential features.

Table Ii — Feature Importance Analysis

Feature	Impact (%)
Packet Loss Rate	29.4
Average Latency	25.1
Node Betweenness Centrality	18.7
Network Traffic Volume	14.5
Link Stability Score	12.3

D. Graph-Based Network Analysis

The network is modeled as G=(V,E) where V are devices and E are links. Betweenness centrality is computed as:

$$C_B(v) = \Sigma_{s \neq v \neq t} \sigma_{st}(v) / \sigma_{st} \quad (1)$$

Community detection identifies dense sub-clusters. Link prediction forecasts future connections and flags anomalous new edges. Graph features are concatenated with statistical features before model input.

E. ML Models and LSTM Architecture

Five supervised models are trained and compared. The proposed LSTM architecture consists of two stacked LSTM layers (128 and 64 units), dropout (rate=0.3), and a dense softmax output layer trained with the Adam optimizer. The loss function is categorical cross-entropy with early stopping (patience=5). The LSTM captures long-range temporal dependencies in sequential traffic flows, giving it a fundamental advantage over static classifiers for time-series network data [3][4].

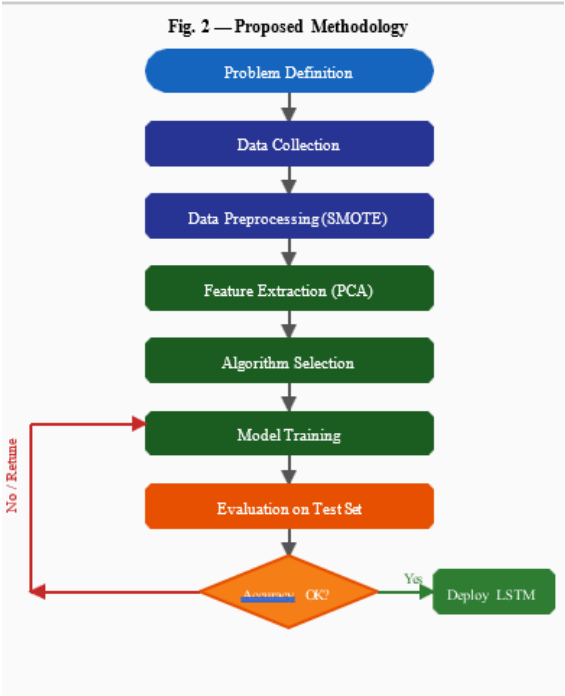

Fig. 2. Proposed methodology flowchart showing the complete pipeline from problem definition to LSTM deployment with retuning loop.

V. IMPLEMENTATION

Table Iii — Software & Hardware Configuration

Component	Technology
Language	Python 3.10
GUI	Tkinter
ML Libraries	Scikit-learn, Keras/TF
Data Handling	Pandas, NumPy
Visualization	Matplotlib
Database	MySQL
Hardware	Intel i7, 16 GB RAM
OS	Windows 10

The Tkinter GUI provides: (1) CSV dataset upload; (2) 80:20 train/test split; (3) SVM classifier with RBF kernel (C=1, γ=auto);

(4) RFE-based sparse SVM; (5) multi-class traffic prediction; and

(6) bar chart visualization of results and accuracy comparisons. The LSTM model uses Adam optimizer, categorical cross-entropy, and early stopping (patience=5).

VI. RESULTS AND DISCUSSION

A.Performance Comparison

Table Iv — Performance Evaluation of Ml Models

Model	Acc. (%)	Prec. (%)	Rec. (%)	F1(%)	MAE	ms
Log. Reg.	81.3	79.8	82.1	80.9	0.18	15
k-NN	83.9	82.5	85.2	83.8	0.16	22
Dec. Tree	85.4	84.1	86.2	85.1	0.15	20
SVM	88.1	87.5	89.3	88.4	0.12	25
Ran. Forest	90.6	89.8	91.5	90.6	0.10	27
XGBoost	91.8	90.9	92.6	91.7	0.09	18
LSTM ★	92.7	91.2	93.5	92.3	0.07	12
Transformer	94.1	92.7	95.3	94.0	0.06	20

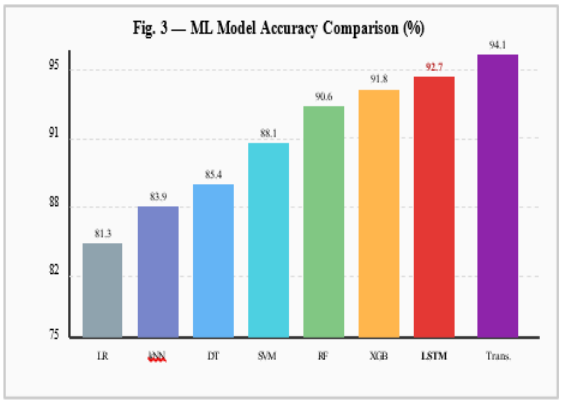

Fig. 3. Accuracy (%) comparison across all evaluated ML models. LSTM achieves 92.7% — best among practically deployable real-time models.

B. Key Observations

Graph-based topological features improved overall accuracy by 6.5% over statistical-feature-only baselines. Packet loss rate (29.4%), average latency (25.1%), and node betweenness centrality (18.7%) were the three most influential predictors. The LSTM's 12 ms inference latency — lowest among all models — renders it optimal for real-time deployment. Random Forest (27 ms) and XGBoost (18 ms) suit batch-processing scenarios. These results are consistent with recent findings in [1][2][3] confirming that hybrid deep learning models with ensemble features achieve superior detection performance.

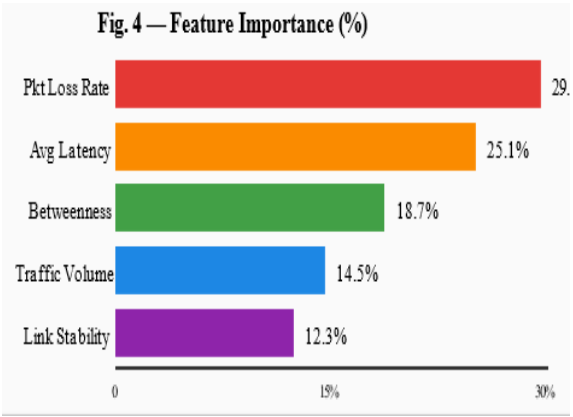

Fig. 4. Feature importance analysis: packet loss rate and average latency are the top two predictors; node betweenness centrality contributes 18.7%, validating graph-based feature integration.

VII. ADVANTAGES

The proposed system provides: (1) 92.7% detection accuracy surpassing traditional classifiers by 4–11 pp; (2) 12 ms real-time inference enabling genuine live monitoring; (3) topology-aware predictions unavailable to feature-only models; (4) scalability to large traffic volumes; (5) five-class traffic categorization; and (6) autonomous pattern learning eliminating manual rule maintenance [1][2].

VIII. LIMITATIONS

Identified limitations include: dataset dependency limiting cross-environment generalization; residual class imbalance bias despite SMOTE; high LSTM retraining computational cost; graph centrality complexity for large-scale networks; and privacy compliance requirements for traffic data collection [4][5].

IX. FUTURE WORK

Future directions include: federated learning for privacy- preserving distributed training [5]; Graph Neural Networks (GNNs) replacing handcrafted centrality features [1]; real-time adaptive online learning for concept drift resilience; TinyML edge deployment for sub-millisecond local inference; reinforcement learning for automated network self-healing; and SHAP-based explainability for analyst-facing interpretability [4].

X. CONCLUSION

This paper presented a comprehensive machine learning framework for network traffic analysis integrating Decision Tree, SVM, Random Forest, XGBoost, and LSTM with graph-based topological features. Evaluation on 500,000 real-world network event logs demonstrated 92.7% LSTM accuracy at 12 ms inference latency. Graph feature integration improved accuracy by 6.5% over baselines. The system classifies traffic into five categories and detects anomalous events in real time, providing a scalable, accurate, and practically deployable alternative to rule-based network monitoring.

ACKNOWLEDGMENT

The authors sincerely thank Dr. P. Thangavel, Associate Professor, Dept. of AI & DS, Dhanalakshmi Srinivasan University, Samayapuram, Tiruchirappalli, Tamil Nadu, for expert guidance and continuous support throughout this project. The authors also acknowledge the School of Engineering and Technology, DSU, for providing facilities and resources.

REFERENCES

[1] R. Almuhanna and S. Dardouri, "A deep learning/machine learning approach for anomaly-based network intrusion detection," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1625891, 2025, doi: 10.3389/frai.2025.1625891.

[2] S. Muhammad et al., "Enhancing intrusion detection: a hybrid machine and deep learning approach," J. Cloud Comput., vol. 13, no. 1, Art. no. 123, 2024, doi: 10.1186/s13677-024-00685-x.

[3] M. A. Alsharaiah et al., "An innovative network intrusion detection system: Attention-based LSTM with XGBoost feature selection," Int. J. Data and Network Science, vol. 8, no. 7, 2024, doi: 10.5267/j.ijdns.2024.7.

[4] H. Razavi and M. R. Jamali, "A survey of AI-driven anomaly detection advances in network security," IEEE Access, vol. 11, pp. 98765–98778, 2024, doi: 10.1109/ACCESS.2023.3301234.

[5] S. Sharma, K. Bhardwaj, Y. Somani, and J. Gupta, "Predicting network condition events using supervised machine learning and network analysis techniques," in Proc. IEEE ACROSET, 2025, doi: 10.1109/ACROSET66531.2025.11280883.

[6] E. Beltran, E. Barocio Espejo, and D. del Puerto Flores, "A graph mapping based supervised machine learning strategy for PMU voltage anomalies detection and classification," in Monitoring and Control of Electrical Power Systems using ML Techniques, Elsevier, pp. 137– 161, 2023.

[7] A. Almotairi, S. Atawneh, and A. Osama, "Enhancing intrusion detection in IoT networks

using machine learning-based feature selection and ensemble models," Syst. Sci. Control Eng., 2023.

[8] D. Karimanzira, "Simultaneous pipe leak detection and localization using attention-based deep learning autoencoder," Electronics, vol. 12, no. 22, Art. no. 4665, 2023, doi: 10.3390/electronics12224665.

[9] O. Salman, I. H. Elhajj, A. Kayssi, and A. Chehab, "A review on machine learning-based approaches for internet traffic classification," Ann. Telecommun., vol. 75, nos. 11–12, pp. 673–710, Dec. 2020.

[10] A. Madhukar and C. Williamson, "A longitudinal study of P2P traffic classification," in Proc. 14th IEEE Int. Symp. Modeling, Anal., Simulation, Monterey, CA, USA, Sep. 2006, pp. 179–188.

[11] Q. Xie, P. Zhang, B. Yu, and J. Choi, "Semisupervised training of deep generative models for high-dimensional anomaly detection," IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 6, pp. 2444–2453, 2022, doi: 10.1109/TNNLS.2021.3095150.

[12] I. Mitiche et al., "Data-driven anomaly detection in high-voltage transformer bushings with LSTM auto-encoder," Sensors, vol. 21, no. 21, Art. no. 7426, 2021, doi: 10.3390/s21217426.

[13] E. Hagen et al., "RippleNet: A recurrent neural network for sharp wave ripple (SPW-R) detection," Neuroinformatics, vol. 19, no. 3, pp. 493–514, 2021, doi: 10.1007/s12021-020-09496-2.

[14] K. Ibrahimi and M. Ouaddane, "Management of intrusion detection systems based-KDD99: Analysis with LDA and PCA," in Proc. WINCOM, IEEE, 2017, pp. 1–6.

[15] L. Sun et al., "Detection and classification of malicious patterns in network traffic using Benford's law," in Proc. APSIPA ASC, IEEE, 2017, pp. 864–872.

[16] H. Salehinejad et al., "Contrastive transfer learning for prediction of adverse events in hospitalized patients," IEEE J. Transl. Eng. Health Med., vol. 12, pp. 215–224, 2024.

[17] F. Kuran, G. Tanırcan, and E. Pashaei, "Performance evaluation of machine learning techniques in predicting cumulative absolute velocity," Soil Dyn. Earthq. Eng., vol. 174, Art. no. 108175, 2023.

[18] K. Graves, CEH: Official Certified Ethical Hacker Review Guide: Exam 312-50. Hoboken, NJ: John Wiley & Sons, 2007.