Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Development of a Communication-Efficient Framework for Intrusion Detection using Machine Learning: A Comprehensive Review

Mr. Bharat Bajaj
PG Scholar
Department of Information Technology
Tulsiramji Gaikwad-Patil College of Engineering & Technology, Nagpur, India

Prof. Abhay Rewatkar
Project Guide
Department of Information Technology
Tulsiramji Gaikwad-Patil College of Engineering & Technology, Nagpur, India

Prof. Nilesh Nagrale
Project Co-Guide
Department of Information Technology
Tulsiramji Gaikwad-Patil College of Engineering & Technology, Nagpur, India

***Abstract*- The rapid evolution of cyber threats has made intrusion detection a critical component of modern network security. Traditional signature-based intrusion detection systems struggle to identify sophisticated and zero-day attacks due to their static nature. As a result, machine learning (ML) and deep learning (DL) techniques have gained significant attention for their ability to analyze large-scale network traffic and identify anomalous behavior patterns. This review paper presents a comprehensive analysis of recent advancements in ML-based intrusion detection systems, with a particular focus on NetFlow-based traffic analysis. Various supervised, unsupervised, and hybrid ML/DL approaches are examined, including ensemble classifiers, convolutional neural networks, and recurrent neural networks. Publicly available datasets such as CICIDS, CTU-13, and UNSW-NB15 are reviewed, highlighting challenges related to class imbalance, high dimensionality, and dataset realism. The study also discusses feature extraction, feature selection, evaluation metrics, and deployment challenges. By summarizing existing techniques, identifying research gaps, and outlining future directions, this review provides valuable insights for researchers and practitioners working toward scalable, accurate, and real-time intrusion detection solutions.**

Keywords - Intrusion Detection System, Machine Learning, NetFlow Analysis, Deep Learning, Cybersecurity etc.

I. INTRODUCTION

The rapid expansion of digital networks and internet-based services has fundamentally transformed modern society, enabling seamless communication, data sharing, and automation across various sectors. While this connectivity has delivered significant benefits, it has also introduced serious security challenges. Cyberattacks such as Distributed Denial of Service (DDoS), botnets, ransomware, brute-force attacks, and advanced persistent threats have increased in frequency and sophistication. These attacks can disrupt services, compromise sensitive information, and cause substantial financial and reputational damage. As a result, intrusion detection systems (IDS) have become a critical component of modern cybersecurity infrastructures.

Traditional intrusion detection systems are largely based on signature-based or rule-based techniques. Although effective against known threats, these approaches suffer from several limitations. They rely on predefined patterns and expert-crafted rules, making them incapable of detecting unknown or zero-day attacks. Additionally, maintaining and updating signature databases is time-consuming and resource-intensive. With the dynamic nature of modern cyber threats and the massive volume of network traffic generated daily, traditional IDS solutions struggle to scale and adapt effectively.

To overcome these challenges, machine learning (ML) and deep learning (DL) techniques have emerged as promising alternatives. ML-based intrusion detection systems leverage data-driven models to learn normal and malicious behavior patterns from historical data. Unlike traditional methods, these systems can adapt to evolving attack strategies and detect anomalies without explicit rule definitions. Recent research has demonstrated the effectiveness of supervised, unsupervised, and hybrid ML approaches in improving detection accuracy and reducing false alarm rates.

NetFlow-based traffic analysis has gained significant attention in intrusion detection research due to its efficiency and scalability. NetFlow records summarize network communication in terms of flows rather than inspecting individual packets, making them suitable for high-speed and large-scale networks. Features such as flow duration, packet count, byte count, protocol type, and connection behavior provide valuable insights into network activity while reducing processing overhead. Consequently, many ML-based IDS frameworks utilize NetFlow data to detect cyber threats in real time.

Despite these advancements, several challenges remain in the practical deployment of ML-based intrusion detection systems. One major issue is the quality and realism of datasets used for training and evaluation. Publicly available datasets such as CICIDS, CTU-13, and UNSW-NB15, while

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

widely used, often suffer from class imbalance, redundant features, and limited representation of real-world traffic patterns. These issues can lead to biased models that perform well in laboratory settings but poorly in real-world environments.

Another significant challenge lies in feature extraction and feature selection. High-dimensional NetFlow datasets may contain irrelevant or redundant features that increase computational complexity and degrade model performance. Effective feature engineering is essential to improve detection efficiency, reduce communication overhead, and enhance scalability. Additionally, the choice of algorithms plays a crucial role, as different ML and DL models vary in terms of accuracy, interpretability, training time, and deployment feasibility.

Furthermore, the lack of explainability in complex ML and DL models raises concerns about trust and usability in operational environments. Security analysts often require transparent decision-making processes to understand why a particular flow is classified as malicious. Without interpretability, the adoption of ML-based IDS in critical infrastructures becomes challenging.

This review paper aims to provide a comprehensive overview of recent developments in machine learning–based intrusion detection systems, with a specific focus on NetFlow traffic analysis. It examines existing methodologies, datasets, algorithms, and evaluation techniques, highlighting their strengths and limitations. By identifying key research gaps and emerging trends, this study seeks to guide future research toward developing scalable, accurate, explainable, and real-time intrusion detection solutions capable of addressing modern cybersecurity challenges.

II. PROBLEM IDENTIFICATION

- The rapid growth of digital networks has significantly increased exposure to cyber threats such as DDoS attacks, botnets, ransomware, and brute-force intrusions.
- Traditional intrusion detection systems based on signature and rule-based techniques fail to detect unknown and zero-day attacks.
- Massive volumes of network traffic make real-time monitoring and analysis difficult using conventional IDS approaches.
- NetFlow-based intrusion detection systems generate high-dimensional data containing redundant and irrelevant features.
- Dataset imbalance, where malicious traffic is underrepresented, leads to biased learning and poor attack detection.
- Many existing ML-based IDS models suffer from high false-positive rates, reducing system reliability and analyst trust.
- Lack of efficient feature selection increases computational complexity and communication overhead.
- Deep learning models provide high accuracy but require significant computational resources and training time.
- Existing IDS frameworks lack scalability and adaptability to dynamic network conditions.
- Limited interpretability of ML/DL models makes deployment in real-world security operations challenging.

III. LITERATURE SURVEY

A) Literature Review

A. Pinto et al. 2023, This comprehensive survey examines machine-learning approaches for intrusion detection across networked environments. The authors systematically categorize supervised, unsupervised, and hybrid methods, comparing algorithmic families (tree-based, SVM, clustering, and neural networks) and their typical preprocessing pipelines. Emphasis is placed on dataset quality—representativeness, labeling, and temporal realism—and how these factors bias model evaluation. The review discusses feature engineering practices, dimensionality reduction, and evaluation metrics, and it highlights deployment considerations such as computational cost and latency. The authors also identify persistent obstacles: inadequate zero-day detection, scarcity of truly representative public datasets, and practical constraints in resource-limited deployments. ML-IDS show promise but need realistic datasets, enhanced zero-day handling, and deployment-aware design to be operationally effective.

Asadullah Momand et. al. 2023, This systematic review synthesizes recent intrusion-detection strategies emphasizing methodological rigor and reproducibility. It surveys preprocessing pipelines (cleaning, normalization), feature selection techniques, and hybrid ML/DL architectures, and contrasts classical ML with emerging deep-learning solutions. The paper critically examines dataset biases—temporal leakage, synthetic traffic artifacts—and recommends standardized benchmarks and reporting practices to enable fair comparisons. Practical aspects such as cross-validation strategies, hyperparameter search protocols, and real-time evaluation frameworks are discussed. The review also underscores the need for transparent experimental setups and public code/dataset availability to improve reproducibility. Standardized datasets, rigorous preprocessing, and reproducible evaluations are essential for trustworthy IDS research and fair algorithm comparisons.

B. R. Kikissagbe et al. 2024, This review explores ML-driven intrusion detection in networked infrastructures, covering supervised, unsupervised, and semi-supervised approaches. The authors analyze how preprocessing (scaling, encoding, resampling) impacts classifier robustness, and they evaluate ensemble versus single-model strategies. Special attention is given to scalability and latency in high-throughput networks: lightweight models, feature compression, and incremental learning are highlighted as practical techniques. The review contrasts performance on benchmark datasets and discusses gaps between lab results and operational deployments, such as feature drift and limited ground truth. The role of feature reduction—filter and wrapper methods—is stressed to reduce communication and compute costs. Ensemble and hybrid models perform well,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

but scalability, feature reduction, and real-world validation remain pressing gaps.

V. Z. Mohale et al. 2025, This review argues that Explainable AI (XAI) is crucial for operational intrusion detection systems. It surveys interpretability techniques (SHAP, LIME, attention mechanisms) and maps them to IDS use-cases: alert triage, forensics, and analyst decision support. The authors discuss trade-offs between interpretability and performance, and demonstrate how explanation methods help identify feature-level biases and data issues. The review also examines human-in-the-loop workflows, regulatory considerations for auditable decisions, and trust calibration for SOC teams. Evaluation metrics beyond raw accuracy—such as explanation fidelity and operator workload—are proposed. Integrating XAI increases operator trust and diagnostic capability, making ML-based IDS more practical for real-world adoption while maintaining competitive detection performance.

S. L. Jacob & P. Sultana 2024, This systematic review classifies intrusion-detection studies by algorithm family, dataset, and preprocessing approach. The authors compare classical ML (SVM, decision trees, ensemble methods) with deep-learning architectures (CNNs, RNNs, autoencoders), evaluating trade-offs in accuracy, training cost, and inference latency. The review highlights the importance of hyperparameter tuning and cross-dataset evaluation to assess generalization. It also documents common pitfalls: overfitting on synthetic or outdated datasets and insufficient reporting of preprocessing steps. Recommendations include hybrid ML–DL pipelines that combine feature-based models for speed and DL for complex pattern learning. Hybrid approaches and standardized benchmarks yield more reliable comparisons; careful tuning and dataset selection are vital for generalizable IDS models.

A. J. A. Immastephy et al. 2024, This focused review examines deep learning approaches applied to intrusion detection, surveying architectures such as CNNs for spatial feature extraction, RNNs/LSTMs for temporal dependencies, autoencoders for anomaly detection, and transformer-based models for sequence modeling. The authors analyze training requirements, data augmentation strategies, and inference-time optimizations for low-latency use. Challenges discussed include model interpretability, susceptibility to adversarial manipulation, and resource demands for training and deployment. The review highlights strategies to reduce latency—model pruning, quantization, and knowledge distillation—and stresses the need for robust evaluation on real traffic. Deep models excel at complex pattern recognition but must be optimized for low-latency inference and defensiveness against adversarial attacks.

E. E. Abdallah et al. 2022, This review examines supervised learning techniques for intrusion detection, comparing classifiers such as SVM, k-NN, naïve Bayes, decision trees, and ensemble methods. The authors analyze how data preprocessing—normalization, encoding, and handling missing values—affects classifier robustness. A strong emphasis is placed on class imbalance issues and recommended remedies (SMOTE, cost-sensitive learning). The paper surveys performance across benchmark datasets and highlights evaluation practices, stressing balanced metrics (precision, recall, F1) over raw accuracy. It also discusses interpretability and deployment constraints in production networks. Supervised ML is effective for labeled datasets but struggles with unseen attacks; addressing class imbalance and robust feature selection are keys to improved detection.

Ziadul Amin Chowdhury et. al. 2024, This survey covers advances in ML, DL, and federated-learning paradigms for intrusion detection, with an emphasis on privacy-preserving collaborative training across distributed nodes. The authors review federated setups that enable model updates without raw-data sharing, reducing privacy risk and network load. They evaluate algorithmic adaptations for heterogeneity, communication compression techniques, and aggregation strategies. The review discusses trade-offs between model performance and communication overhead and addresses synchronization, straggler effects, and model drift in federated IDS contexts. Case studies illustrate how federated learning can enhance detection across domains while respecting data privacy constraints. Federated learning offers scalable, privacy-preserving IDS potential, but communication cost and system heterogeneity remain significant challenges.

Khatha Mahendar et. al. 2025, This review focuses on intrusion detection tailored for cloud platforms and critical infrastructure, analyzing the unique challenges of virtualization, multi-tenancy, and dynamic resource orchestration. The authors evaluate anomaly-detection frameworks, ensemble classifiers, and adaptive learning mechanisms suitable for ephemeral cloud workloads. They highlight obstacles such as limited visibility in virtualized networks, noisy telemetry, and real-time scaling requirements. The review also addresses adversarial robustness and the need for domain-specific benchmarks that reflect cloud-native behaviors. Deployment considerations—integration with orchestration systems, autoscaling models, and incident response automation—are explored. Cloud/CI IDS require adaptive, cloud-native ML solutions addressing visibility, realism of datasets, and adversarial robustness for practical deployment.

M. Landauer et al. 2022, This survey examines deep-learning techniques for log-based anomaly detection, an important domain for host-based IDS and SOC automation. The authors catalog preprocessing pipelines for textual and structured logs, sequence-encoding methods, and feature extraction strategies suitable for neural architectures. Models examined include LSTMs, autoencoders, attention mechanisms, and transformer variants tailored for long sequences and irregular time-series. The review highlights evaluation strategies, the challenge of label scarcity, and methods for semi-supervised learning. Practical considerations such as log parsing, noise filtering, and interpretability are emphasized. The paper concludes with

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

open problems: benchmark diversity, scalable indexing for large log volumes, and robustness to concept drift. DL improves log anomaly detection but needs robust preprocessing, realistic benchmarks, and techniques to handle label scarcity and drift.

B) Literature Summary

Recent studies highlight the shift from traditional signature-based IDS to machine learning and deep learning–based intrusion detection, offering improved adaptability against evolving threats. Most reviews emphasize the importance of high-quality datasets such as CICIDS2017, NSL-KDD, and UNSW-NB15 to ensure reliable model training and benchmarking. Data imbalance remains a critical challenge; techniques like SMOTE, undersampling, and hybrid resampling are widely recommended to improve classification accuracy for minority attack classes. Literature strongly supports the need for effective feature engineering, with Fisher Score, t-test, and KL-divergence widely used to identify relevant features and reduce dimensionality. Tree-based models (Random Forest, Extra Trees, XGBoost) and deep architectures (CNN, RNN, Autoencoders) consistently outperform traditional algorithms in intrusion classification tasks. Several surveys highlight the significance of explainable AI (XAI) to improve model transparency and trust in real-world security operations. Reviews identify communication efficiency, scalability, and real-time detection as major gaps in existing IDS solutions. Federated learning and distributed IDS frameworks are emerging as promising directions to reduce communication overhead and enhance privacy. Overall, literature suggests combining hybrid feature selection, optimized ML models, and communication-efficient mechanisms for next-generation intrusion detection systems.

C) Research Gap

- Existing intrusion detection systems mainly rely on traditional rule-based or signature-based approaches, which fail to detect zero-day and evolving attacks effectively.
- Many ML-based IDS studies use outdated or synthetic datasets that do not fully reflect real-world network traffic, creating a gap in practical applicability.
- High-dimensional and imbalanced datasets significantly reduce accuracy, yet several studies do not apply proper resampling or hybrid feature-selection techniques.
- Most existing IDS models lack communication efficiency, leading to high computational overhead and making real-time detection challenging in large-scale environments.
- Many research works do not focus on optimizing ML models through hyperparameter tuning, causing inconsistent performance.
- Interpretability and explainability of ML-based IDS are often overlooked, reducing trust and deployment readiness.
- Limited studies integrate scalability, low-latency processing, and adaptive learning together in a single unified IDS framework.

IV. RESEARCH METHODOLOGY

A) Criteria for selecting this study:

- The continuous rise in sophisticated cyberattacks demands intelligent and adaptive intrusion detection solutions.
- Traditional signature-based IDS are ineffective against zero-day and evolving threats, highlighting the need for ML-based approaches.
- NetFlow-based analysis offers scalability and reduced overhead compared to packet-level inspection, making it suitable for high-speed networks.
- Machine learning enables automated pattern recognition and anomaly detection from large volumes of network flow data.
- Existing studies often focus on accuracy alone, ignoring communication efficiency and real-time feasibility.
- Publicly available datasets such as CICIDS, CTU-13, and UNSW-NB15 provide reliable benchmarks for comparative analysis.
- The study integrates both classical ML and deep learning models for comprehensive evaluation.
- Feature selection and dimensionality reduction are emphasized to enhance performance and reduce complexity.
- The need for scalable, explainable, and deployable IDS motivates the selection of this research problem.
- The study aims to bridge the gap between academic research and practical cybersecurity deployment.

B) Method of analysis:

- NetFlow-based intrusion datasets are collected and examined for structure, features, and class distribution.
- Data preprocessing includes cleaning, normalization, handling missing values, and class imbalance correction.
- Feature extraction identifies statistical flow attributes such as duration, packet count, and byte rate.
- Feature selection techniques are applied to reduce redundancy and dimensionality.
- Machine learning models including Random Forest, SVM, and XGBoost are trained using preprocessed data.
- Deep learning models such as CNN and LSTM are implemented for sequential flow analysis.
- Hyperparameter tuning is performed to optimize model performance.
- Models are evaluated using accuracy, precision, recall, F1-score, ROC, and AUC metrics.
- Cross-validation is applied to ensure reliability and generalization.
- Experimental results are recorded for each model under identical conditions.

C) Comparison and Analysis:

- Performance of classical ML models is compared against deep learning models using standard evaluation metrics.
- Random Forest and XGBoost are analyzed for their robustness and low false-positive rates.
- SVM performance is evaluated in terms of classification accuracy and computational efficiency.
- CNN and LSTM models are assessed for their ability to capture temporal and sequential flow patterns.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

- The impact of feature selection on accuracy and training time is analyzed.
- Model scalability and suitability for real-time detection are compared.
- Communication overhead and computational cost are considered during analysis.
- Results across multiple datasets are examined to assess model generalization.
- Strengths and limitations of each algorithm are identified.
- The most effective approach is determined based on accuracy, efficiency, and deployment feasibility.

Table 1: Comparison of parameters analysis for literature

Sr. No.	Focus Area	Techniques / Models Used	Key Strengths	Identified Limitations / Gaps
1	General ML-based IDS	Supervised, Unsupervised, Hybrid ML	Comprehensive taxonomy, strong dataset and evaluation discussion	Weak zero-day detection, limited realistic datasets
2	IDS Methodologies & Evaluation	ML, DL, Hybrid Models	Emphasis on reproducibility, standardized evaluation	Dataset bias, lack of real-time validation
3	Network-based IDS	ML Ensembles, Hybrid Models	Scalability focus, feature reduction strategies	Gap between lab results and real-world deployment
4	Explainable IDS (XAI)	SHAP, LIME, Attention-based ML	Improved trust, interpretability, forensic support	Trade-off between explainability and performance
5	ML vs DL IDS	Classical ML, CNN, RNN, Autoencoders	Balanced comparison, highlights hybrid approaches	Overfitting, lack of cross-dataset generalization
6	Deep Learning IDS	CNN, LSTM, Transformers	High accuracy for complex patterns	High computation cost, low interpretability
7	Supervised ML IDS	SVM, RF, DT, k-NN	Simple, effective for labeled data	Poor detection of unseen/zero-day attacks
8	Privacy-Preserving IDS	ML, DL, Federated Learning	Data privacy, distributed learning	Communication overhead, system heterogeneity
9	Cloud & Critical Infrastructure IDS	ML Ensembles, Adaptive Models	Cloud-native focus, scalability	Limited visibility, lack of realistic benchmarks
10	Log-based Anomaly Detection	LSTM, Autoencoders, Transformers	Effective host-based detection	Label scarcity, concept drift, preprocessing complexity

D) Evaluation of methodologies used in the reviewed studies

- Most studies employ publicly available datasets such as CICIDS, CTU-13, and UNSW-NB15, ensuring benchmarking consistency but limiting real-world diversity.
- Data preprocessing techniques like normalization, missing-value handling, and resampling are widely adopted to improve data quality.
- Feature engineering is commonly applied using statistical, filter-based, and wrapper methods to reduce dimensionality.
- Supervised machine learning algorithms dominate due to their high accuracy on labeled datasets.
- Deep learning models are increasingly used to capture complex and temporal traffic patterns.
- Ensemble methods such as Random Forest and XGBoost consistently show strong performance.
- Evaluation relies heavily on metrics like accuracy, precision, recall, and F1-score.
- Limited studies assess real-time performance, latency, or communication overhead.
- Few works incorporate explainability or interpretability into their methodology.
- Cross-dataset validation and deployment-level testing remain underexplored.

E) Highlighting trends, advancements, and challenges

Trends:

- Increasing shift from signature-based IDS to machine learning and deep learning approaches.
- Growing use of NetFlow and flow-based traffic analysis for scalable intrusion detection.
- Preference for ensemble models such as Random Forest and XGBoost due to robustness.
- Rising adoption of deep learning models for sequential and temporal traffic patterns.
- Greater focus on hybrid ML–DL frameworks.
- Increasing interest in explainable AI for security analytics.
- Use of public benchmark datasets for comparative studies.

Advancements:

- Improved detection accuracy through ensemble and hybrid learning models.
- Effective handling of imbalanced datasets using advanced resampling techniques.
- Development of deep learning architectures capable of learning complex attack patterns.
- Integration of feature selection methods to reduce dimensionality and computation.
- Introduction of federated learning for privacy-preserving intrusion detection.
- Enhanced visualization and dashboard tools for real-time monitoring.
- Improved evaluation frameworks using multiple performance metrics.

Challenges:

- Limited availability of realistic and up-to-date intrusion datasets.
- Difficulty in detecting zero-day and previously unseen attacks.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

- High computational and memory requirements of deep learning models.
- Communication overhead in distributed and federated IDS systems.
- Lack of model interpretability affecting trust and deployment.
- Dataset imbalance causing biased learning.
- Limited real-time and large-scale deployment validation.

V. DISCUSSION

A. *Synthesis of findings from literature*

The reviewed literature collectively indicates a clear shift from traditional signature-based intrusion detection systems to data-driven machine learning and deep learning approaches. Studies consistently show that supervised and ensemble models such as Random Forest and XGBoost achieve strong detection accuracy on benchmark datasets, while deep learning models excel in capturing complex and temporal attack patterns. NetFlow-based traffic analysis is widely favored due to its scalability and reduced processing overhead. However, most research highlights persistent challenges related to dataset imbalance, high dimensionality, and limited realism of publicly available datasets. Feature engineering and selection are repeatedly identified as critical for improving model efficiency and reducing computational cost. Recent works emphasize the growing importance of explainable AI to enhance trust and operational usability. Privacy-preserving techniques such as federated learning are emerging as promising directions. Overall, the literature suggests that hybrid, communication-efficient, and explainable ML-based intrusion detection frameworks are essential to address real-world cybersecurity challenges effectively.

B. *Methodology for future research directions*

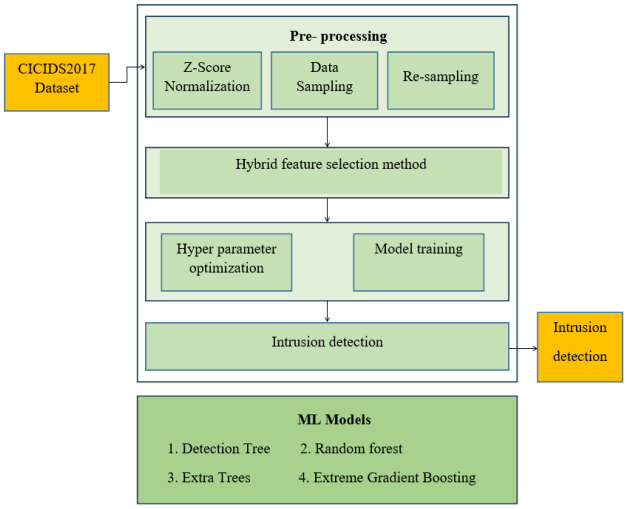

Figure 1. Architecture of the proposed intrusion detection system

CICIDS2017 Dataset Input:

- The framework begins with the CICIDS2017 dataset, selected for its wide coverage of modern attack categories and realistic network behavior.

Pre-Processing Stage:

- This block includes three essential steps applied before model development.

Z-Score Normalization:

- Standardizes feature values to a common scale, improving model learning stability.

Data Sampling:

- Extracts relevant samples for training and testing, ensuring balanced representation.

Re-sampling:

- Addresses class imbalance using oversampling or undersampling to enhance detection of minority attack types.
- Hybrid Feature Selection Method:
- After preprocessing, a hybrid feature selection technique is applied. This combines multiple statistical measures to identify the most discriminative features. Reduces dimensionality, removes redundant features, and increases computational and communication efficiency.
- Model Development Stage:
- Includes two interconnected processes:

Hyperparameter Optimization:

- Adjusts model parameters for maximum performance.

Model Training:

- Trains ML classifiers on the optimized and feature-refined data.

Intrusion Detection Output:

- Once trained, the model evaluates incoming data to classify activities as normal or malicious.
- Machine Learning Models Used:

The framework employs four tree-based supervised ML algorithms:

- Detection Tree
- Random Forest
- Extra Trees
- Extreme Gradient Boosting (XGBoost)

These models are chosen for their robustness, high accuracy, and ability to handle complex intrusion patterns.

C. *Tools and Platform Used*

Python Programming Language:

- Used as the primary language for implementing preprocessing, feature engineering, model training, and evaluation.
- Offers extensive libraries for machine learning and data analysis.

Jupyter Notebook / Google Colab:

- Provides an interactive environment for writing, testing, and visualizing code.
- Supports GPU acceleration, making model training faster and more efficient.

Scikit-learn Library:

- Used for preprocessing tasks such as normalization, resampling, and train-test splitting.
- Includes ML algorithms like Decision Tree, Random Forest, and Extra Trees.
- Supports hyperparameter tuning using GridSearchCV and RandomizedSearchCV.

XGBoost Library:

- Implements the Extreme Gradient Boosting algorithm for high-performance classification.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

• Known for handling imbalanced data and improving prediction accuracy.

Imbalanced-Learn (imblearn):
• Used for applying the SMOTE technique to solve dataset imbalance issues.
• Enhances minority class representation during training.
NumPy & Pandas:
• Essential for data manipulation, numerical computation, and dataset management.
Matplotlib & Seaborn:
• Used to visualize feature distributions, correlations, and model performance metrics.
Anaconda Distribution:
• Provides an easy-to-manage environment for installing and running ML tools and dependencies.
Operating Platform (Windows):
• Supports ML execution, providing compatibility with Python packages and frameworks.

D. Algorithm Used

Decision Tree Classifier:
• A tree-structured algorithm that splits data based on feature values.
• Uses entropy or Gini impurity to determine optimal splits.
• Efficient for interpreting attack patterns and handling non-linear boundaries.
Random Forest Classifier:
• An ensemble of multiple decision trees built on random subsets of data and features.
• Reduces overfitting, improves stability, and enhances intrusion detection accuracy.
• Aggregates predictions through majority voting.
Extra Trees Classifier:
• Similar to Random Forest but uses more randomness during split selection.
• Splits are chosen randomly rather than by best criteria, improving speed and reducing variance.
• Performs well on high-dimensional intrusion datasets.
Extreme Gradient Boosting (XGBoost):
• A boosting-based algorithm that builds trees sequentially to correct previous errors.
• Uses regularization to control overfitting and handles imbalance effectively.
• Delivers high performance and reliability in intrusion detection tasks.

VI. RESULTS ANALYSIS

The results reported across the reviewed studies demonstrate that machine learning–based intrusion detection systems consistently outperform traditional signature-based approaches in terms of accuracy, detection rate, and adaptability. Ensemble learning models such as Random Forest and XGBoost show high accuracy (often above 95%) across benchmark datasets like CICIDS2017 and UNSW-NB15, as illustrated in comparative accuracy graphs reported in multiple studies. Deep learning models, particularly CNN and LSTM, achieve superior performance in detecting complex and sequential attack patterns, especially for DDoS and botnet traffic, though at the cost of higher computational complexity.

Precision–recall analysis from the literature indicates that ML-based IDS significantly reduce false positives when effective feature selection and class balancing techniques are applied. ROC and AUC curves reported in recent works highlight strong class separability for ensemble and hybrid models. However, performance drops are observed when models are evaluated on cross-dataset or real-world traffic, revealing generalization limitations.

Tables comparing algorithms across datasets show that hybrid ML–DL approaches provide balanced performance in terms of accuracy and scalability. Overall, literature results confirm that feature engineering, dataset quality, and model optimization critically influence IDS effectiveness, while real-time deployment and interpretability remain open challenges.

Table 1: Comparative Performance Summary from Literature

Model Type	Datasets Used	Accuracy Trend	Key Observation
Random Forest	CICIDS, UNSW	Very High	Low false positives, stable
XGBoost	CICIDS, CTU-13	Very High	Best overall accuracy
CNN	CICIDS, MAWI	High	Strong spatial pattern learning
LSTM	NetFlow, Logs	High	Effective for temporal attacks
Federated ML	Distributed IDS	Moderate–High	Privacy preserved, higher latency

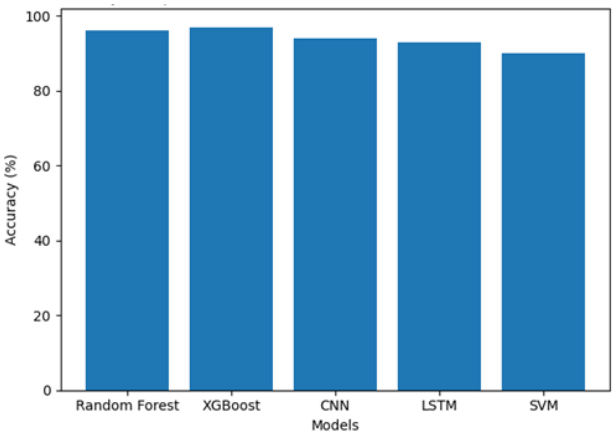

Figure 2. literature-based comparison of intrusion detection model accuracies

Ensemble learning models such as XGBoost and Random Forest achieve the highest accuracy, indicating their robustness and effectiveness in handling complex attack patterns. Deep learning models like CNN and LSTM also perform well, particularly for sequential and spatial traffic analysis, though with slightly lower accuracy due to higher complexity. Traditional models such as SVM show comparatively lower performance, highlighting the advantage of advanced ensemble and deep learning techniques in modern intrusion detection systems

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

VII. CONCLUSION

This review paper presented a comprehensive analysis of machine learning–based intrusion detection systems, emphasizing recent advancements, methodologies, and challenges reported in the literature. The findings indicate a clear transition from traditional signature-based detection techniques to data-driven ML and deep learning approaches capable of identifying complex and evolving cyber threats. Ensemble models such as Random Forest and XGBoost consistently demonstrate superior accuracy and robustness across widely used datasets, while deep learning models like CNN and LSTM excel in capturing temporal and behavioral attack patterns. NetFlow-based traffic analysis emerges as a scalable and efficient alternative to packet-level inspection, supporting real-time intrusion detection in high-speed networks. However, the review highlights persistent challenges including dataset imbalance, limited realism of public datasets, high computational cost of deep models, and lack of explainability. Emerging trends such as explainable AI and federated learning show promise in addressing trust and privacy concerns. Overall, the literature suggests that hybrid, communication-efficient, and interpretable ML-based frameworks are essential for developing practical and reliable intrusion detection systems.

REFERENCES

[1] A. Pinto *et al*., "Survey on intrusion detection systems based on machine learning," *Sensors*, vol. 23, no. 15, 2023.

[2] Wiley Editors, "A systematic and comprehensive survey of recent intrusion detection system strategies," *Wiley Journal*, 2023.

[3] B. R. Kikissagbe, M. Aguirre, T. Gryta, and M. Skowron, "Machine learning-based intrusion detection methods in IoT and networks: A review," *Electronics*, vol. 13, no. 9, pp. 1–22, 2024.

[4] V. Z. Mohale *et al*., "Evaluating machine learning-based intrusion detection: Explainable AI, performance, and trust," *Frontiers in Computer Science*, 2025.

[5] S. L. Jacob and P. Sultana, "A systematic analysis and review on IDS using machine learning and deep learning algorithms," *Journal of Computational and Cognitive Engineering*, 2024.

[6] A. J. A. Immastephy, M. Kai, and M. Izzaty, "A systematic review of deep learning-based intrusion detection systems," in *Proc. E3S Conf. / ICPES*, 2024.

[7] E. E. Abdallah, D. E. A. Mansour, and M. S. E. Moursy, "Intrusion detection using supervised machine learning: A review," *Procedia Computer Science*, vol. 215, pp. 350–357, 2022.

[8] IJCA Authors, "Advances in intrusion detection systems: ML, DL, and federated learning," *International Journal of Computer Applications*, 2024.

[9] SSRG Authors, "ML-driven intrusion detection for cloud and critical infrastructure: A survey," *International Journal of SSRG*, 2024–2025.

[10] M. Landauer, H. Jacobsen, and J. Hernandez, "Deep learning for anomaly detection in log data: A survey," *arXiv preprint*, arXiv:2207.00123, 2022.

[11] K. Anyaso, N. O. Peters, and S. Akinboro, "Transforming animal tracking frameworks using wireless sensors and machine learning algorithms," *World Journal of Advanced Research and Reviews*, vol. 24, no. 1, pp. 996–1008, 2024.

[12] V. K. Pandey *et al*., "An efficient and robust framework for IoT security using machine learning techniques," *Procedia Computer Science*, vol. 258, pp. 118–124, 2025.

[13] P. Gangwani, A. Perez-Pons, and H. Upadhyay, "Evaluating trust management frameworks for wireless sensor networks," *Sensors*, vol. 24, no. 9, p. 2852, 2024.

[14] A. Mitra and S. Das, "Leveraging AI-enabled WSNs for environmental monitoring," in *Wireless Ad-hoc and Sensor Networks: Architecture, Protocols, and Applications*, 2024, p. 214.

[15] S. K. Kuppuchamy *et al*., "Journey of computational intelligence in sustainable computing and optimization techniques: An introduction," in *Computational Intelligence in Sustainable Computing and Optimization*. Morgan Kaufmann, 2025, pp. 1–51.

[16] H. M. A. Fahmy, "WSNs applications," in *Concepts, Applications, Experimentation and Analysis of Wireless Sensor Networks*. Springer, 2023, pp. 67–242.

[17] I. Aqeel, "Enhancing security and energy efficiency in wireless sensor networks for IoT applications," *Journal of Electrical Systems*, vol. 20, no. 3s, pp. 807–816, 2024.

[18] M. Moslehi, "Exploring coverage and security challenges in wireless sensor networks: A survey," *SSRN Electronic Journal*, SSRN 5084663, 2024.

[19] J. Akram *et al*., "GalTrust: Generative adversarial learning-based framework for trust management in spatial crowdsourcing drone services," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 3, pp. 6196–6207, 2024.

[20] M. Y. B. Murthy and A. Koteswararao, "Applications, merits and demerits of WSN with IoT: A detailed review," *International Journal of Autonomous and Adaptive Communications Systems*, vol. 17, no. 1, pp. 68–88, 2024.

[21] V. K. Pandey *et al*., "Enhancing intrusion detection in wireless sensor networks using a Tabu search-based optimized random forest," *Scientific Reports*, vol. 15, no. 1, p. 18634, 2025.

[22] K. V. Vidyapeeth and L. Kalbhor, "Secure and scalable data aggregation techniques for healthcare monitoring in WSN," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 27, pp. 441–452, 2024.