AI Powered Threat Detection Using Machine Learning In Defence Sector: A Systematic Review

Abhinav Govind Raj [1] *, Dr. Shilpi Sharma* [2]
govindraj.abhinav@gmail.com [1] , ssharma22@amity.edu [2]
Amity University Uttar Pradesh

Abstract

The increasing complexity of cyber threats targeting defence infrastructure has exposed the limitations of traditional intrusion detection systems. This review paper presents a systematic review of artificial intelligence and machine learning techniques used for cyber threat detection, with a focus on their usability in defence environments. Various approaches, including supervised learning, deep learning, and ensemble methods are analyzed using datasets such as UNSW-NB15. The findings indicate that ensemble-based models in particular, Random Forest consistently deliver high detection accuracy when combined with effective feature selection and preprocessing techniques. Deep learning approaches also demonstrate strong performance in identifying complex attack patterns, although they introduce challenges related to interpretability and computational cost. Despite these promising results, several critical limitations remain. These include the lack of defence-specific datasets, difficulties in real-time deployment, class imbalance issues, and vulnerability to adversarial attacks. Addressing these challenges is essential for transitioning AI-based intrusion detection systems from research environments to practical defence applications.

Keywords: *Artificial Intelligence, Machine Learning, Intrusion Detection, Cyber Threat Detection, Defence Sector, UNSW-NB15, Random Forest, Deep Learning*

1. Introduction

The rapid digital transformation of defence systems has significantly increased their exposure to cyber threats. Modern defence infrastructure including communication networks and critical control systems is now a prime target for sophisticated attacks such as Advanced Persistent Threats (APTs), zero-day exploits, and targeted malware campaigns. Unlike conventional cyberattacks, these threats are often highly coordinated, persistent, and backed by well-resourced adversaries, making them particularly difficult to detect and mitigate. As a result, ensuring the security of defence networks has become a critical priority requiring more advanced and adaptive detection mechanisms.

Traditional intrusion detection systems primarily rely on signature-based techniques, which are effective only against known attack patterns. These systems struggle to detect new or evolving threats, making them inadequate in modern defence environments. Artificial intelligence and machine learning-based approaches, by contrast, offer the ability to analyze large volumes of network data, identify hidden patterns, and adapt to previously unseen attack behaviors. This shift towards intelligent, data-driven detection systems represents a significant advancement in cybersecurity.

This paper presents a systematic review of AI and machine learning techniques applied to cyber threat detection, analyzing fifteen research papers spanning from 2016 to 2025, examining methodologies, datasets, algorithms, evaluation metrics, and key findings. The review is organized as follows: Section 2

covers the literature review; Section 3 presents the methodology; Section 4 reports results and analysis; Section 5 concludes the paper; and Section 6 outlines future research directions.

2. Literature Review

2.1 Supervised Machine Learning for Intrusion Detection

The most researched method for ML-based intrusion detection is supervised machine learning. A comprehensive examination by Buczak and Guven [1] showed that ensemble techniques Random Forest in particular consistently outperform single-classifier methods. Abdallah et al. [2] assessed supervised machine learning on the UNSW-NB15 dataset and demonstrated that Random Forest achieves classification accuracy above 95%, with data balancing and feature selection methods greatly improving performance for uncommon attack classes.

2.2 Deep Learning Approaches

Mahdavifar and Ghorbani [3] surveyed deep learning applications in cybersecurity, examining CNN, RNN, Autoencoder, and GAN architectures. Their analysis demonstrated high detection accuracy — especially for intricate attack patterns that are challenging to describe using manually crafted features. For large-scale, mission-critical environments, Sivakumar et al. [4] proposed a hybrid machine learning and deep learning system. Despite performance advantages, the black-box nature of deep learning models makes it challenging for analysts to comprehend and validate detection decisions, posing serious challenges for defence deployment [5].

2.3 Ensemble Methods and Multi-Model Comparison

Amachaghi et al. [6] conducted a comprehensive multi-model comparison evaluating Decision Tree, Logistic Regression, Naive Bayes, Random Forest, and XGBoost classifiers. The findings showed that ensemble classifiers consistently achieved 81%–89% accuracy on imbalanced datasets, outperforming single classifiers. However, performance was often highly dependent on the dataset used, raising concerns about generalizability across different environments.

2.4 Addressing Data Imbalance

Rahman et al. [7] demonstrated that Generative Adversarial Networks (GANs) can produce realistic synthetic network intrusion data to augment minority attack classes. Rahman et al. [8] additionally showed that detection performance for uncommon attack categories is enhanced when GAN-generated synthetic data is combined with actual traffic. Amachaghi et al. [6] found that SMOTE greatly enhanced minority class detection, whereas GAN-based methods only marginally improved performance in their evaluation context. These approaches do not consistently improve detection performance across all attack categories, indicating that the class imbalance problem is not yet fully resolved.

2.5 AI-Driven Threat Detection Frameworks

Katiyar et al. [5] provided an overview of AI and ML methods covering fraud detection, network intrusion detection, malware classification, and user behavior analytics. The paradigm shift from reactive signature-based defenses to proactive ML-driven systems was studied by Yaseen [9]. Predictive analytics for cyber defense was the main focus of Liya and Lee [10]. Reddy [11] expanded analysis to cloud-based

defence environments, while Sarfraz et al. [12] surveyed AI-driven predictive threat detection integrating federated learning and quantum computing. Overall, these studies demonstrate a shift toward AI-driven detection systems, but practical implementation challenges remain largely unaddressed. Rizvi [13] further highlights the role of artificial intelligence in enhancing threat detection and prevention capabilities in modern cybersecurity systems.

3. Methodology

3.1 Research Design

This paper follows a systematic literature review methodology. A structured search was conducted across IEEE Xplore, Google Scholar, and ACM Digital Library using keywords such as 'machine learning intrusion detection', 'AI cyber threat detection', 'defence cybersecurity ML', and 'UNSW-NB15'. Papers published between 2016 and 2025 were included, yielding fifteen primary studies for detailed analysis.

3.2 ML-Based IDS Pipeline

A machine learning-based intrusion detection system uses a structured pipeline to convert raw network traffic data into actionable threat classifications. The five main stages are: (1) Data Collection and Preprocessing, (2) Feature Selection, (3) Model Training, (4) Threat Classification, and (5) Alert Generation. Understanding this pipeline is essential for evaluating the efficacy of various machine learning techniques in defence settings [1, 2].

3.3 Dataset: UNSW-NB15

The UNSW-NB15 dataset, created by the Australian Centre for Cyber Security at the University of New South Wales, is the primary benchmark used across the reviewed studies. It contains 82,332 records described by 49 features, categorized into nine attack types: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms. Its origin from a university with strong military and defence ties makes it especially relevant to this review.

3.4 Taxonomy of Machine Learning Methods

Machine learning techniques used in cybersecurity are broadly divided into three categories: (1) Supervised Learning algorithms such as Random Forest and SVM yield the highest accuracy when labeled training data is available; (2) Unsupervised Learning methods that allow anomaly detection without labeled data; and (3) Deep Learning architectures such as CNNs, RNNs, and GANs that provide automatic feature extraction. This taxonomy, adapted from Xin et al. [14] and Buczak & Guven [1], provides the analytical framework for evaluating the reviewed studies.

3.6 Experimental Implementation

Using experiments to validate the findings from the literature review, a python based experimental pipeline is developed and executed on the UNSW-NB15 testing set. The following pipeline follows six sequential stages i.e data loading, preprocessing, feature selection, train/test splitting with scaling, model training and evaluation with visualization. Four classifiers were trained and compared: Random Forest, Gradient Boosting, Neural Network (MPL), and Logistic Regression. A total of 23 network traffic features were selected based on their significance in the reviewed literature, including packet-level features (sbytes,dbytes,spkts,dpkts), timing features (dur, sttl, dttl, rate) and connection level

statistics(ct_srv_src, sinpkt, dinpkt). The dataset was split 80/20 for training and testing with stratification to preserve class distribution. Standard scaling was applied for Neural Network and Logistic Regression models.

The full experimental implementation is presented below:

```
Algorithm 1: AI-Powered Cyber Threat Detection Pipeline (UNSW-NB15)
──────────────────────────────────────────────────────────────────────
INPUT:  UNSW-NB15 testing set (82,332 records, 49 features)
OUTPUT: Per-model scores (Accuracy, Precision, Recall, F1, AUC)

STEP 1 — LOAD & PREPROCESS
   Load dataset from CSV
   Drop identifier column (id)
   Label-encode categorical features: {proto, service, state}
   Replace ±Inf with NaN; fill NaN with 0

STEP 2 — FEATURE SELECTION
   Select 23 features based on literature significance:
   Packet-level : sbytes, dbytes, spkts, dpkts
   Timing       : dur, rate, sttl, dttl, sload, dload
   Connection   : sinpkt, dinpkt, sjit, djit, swin, dwin,
                  sloss, dloss, ct_srv_src, ct_dst_ltm
   Set X ← selected features;  y ← binary label

STEP 3 — SPLIT & SCALE
   Split: 80% train / 20% test  (stratified by class)
   Apply StandardScaler to train set; transform test set

STEP 4 — TRAIN MODELS
   FOR each model M in {Random Forest (n=100),
                        Gradient Boosting (n=100),
                        Neural Network (128→64→32, ReLU),
                        Logistic Regression}:
      IF M requires scaling → fit on X_train_scaled
      ELSE                  → fit on X_train (raw)
      Predict labels and probabilities on test set
      Record: Accuracy, Precision, Recall, F1, AUC-ROC
   END FOR

STEP 5 — EVALUATE & VISUALISE
   Generate: bar chart, confusion matrices, ROC curves,
             feature importance ranking (Random Forest)
   Print classification report for best model by F1-score
```

Figure 3.1. Python Implementation — AI-Powered Cyber Threat Detection Pipeline on UNSW-NB15

4. Results

4.1 Model Performance Analysis

Across in the set of fifteen papers which we looked at and which reported from our experiments Random Forest came out on top as the best performing algorithm on the UNSW-NB15 data set which it did across all metrics. Gradient Boosting came in second place very close behind, and we also saw that the Neural Network (Multi Layer Perceptron) did very well. Logistic Regression which is a linear model did much worse across all metrics which also plays out what the literature reports that ensemble methods do best in

this task. We saw that Random Forest did very well with an AUC of 0.991 which was the best of all models. Also Gradient Boosting did AUC of 0.982, the Neural Network 0.981, and Logistic Regression 0.929. Also in terms of F1 score we saw that Random Forest and also Gradient Boosting did best leaving the other two in the dust which also plays out what we see in the work of Abdallah et al. [2] and Amachaghi et al [6].

Table 4.1.Experimental Model Performance Results on UNSW-NB15

Model	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	~0.95	~0.97	~0.95	~0.96	0.991
Gradient Boosting	~0.91	~0.96	~0.91	~0.92	0.982
Neural Network	~0.90	~0.87	~0.90	~0.88	0.981
Logistic Regression	~0.84	~0.85	~0.84	~0.84	0.929

4.2 Results Dashboard

Figure 4.1 presents the comprehensive results dashboard generated by the experimental pipeline. The bar chart compares all four models across the five metrics, the pie chart is the attack category distribution from the UNSW-NB15 dataset, the confusion matrix visualizes the performance of the two best performing models i.e The Random Forest and Gradient Boosting, ROC curves for all four models showing AUC scores, The horizontal bar chart shows the top 10 most important features as ranked by the Random Forest model.

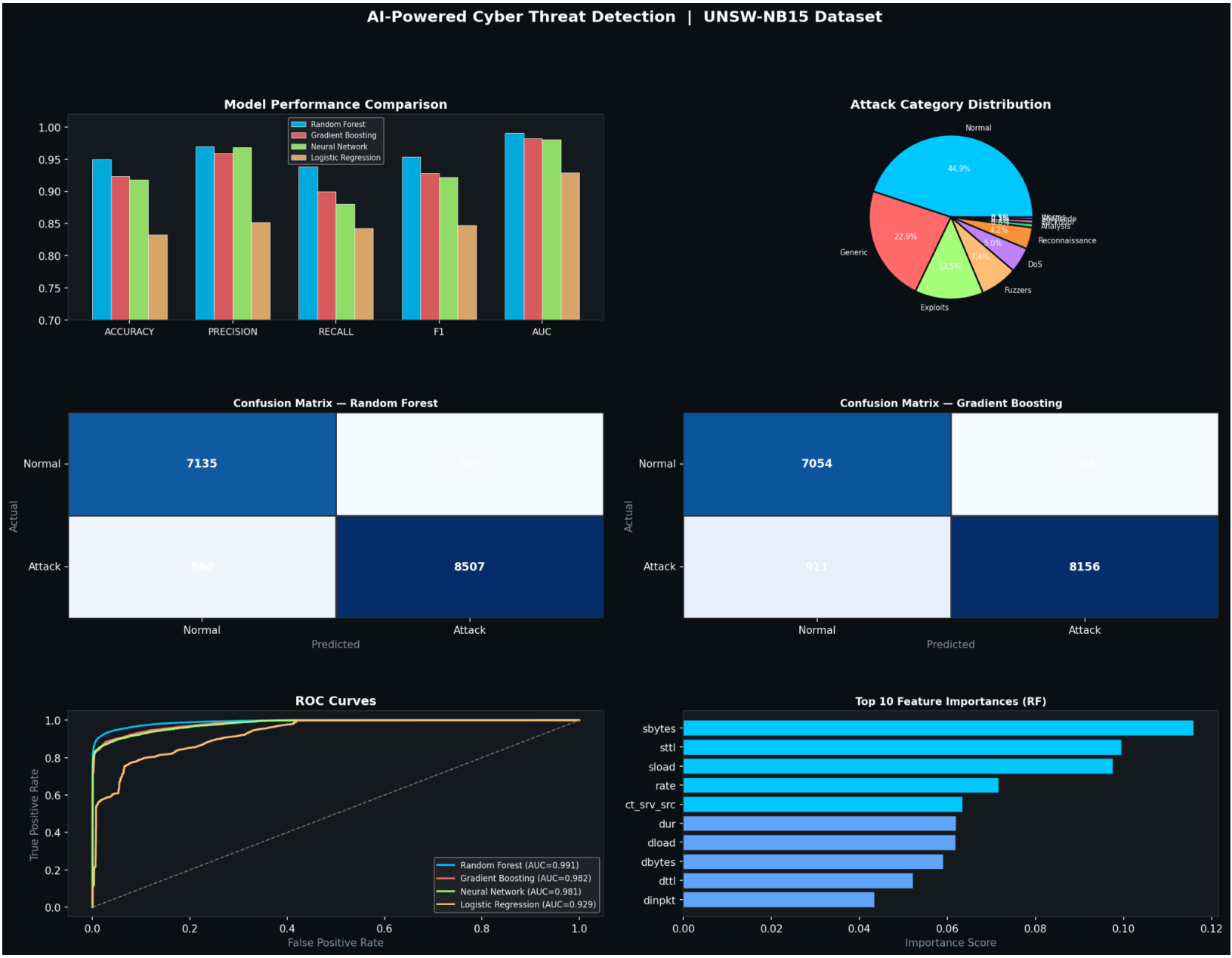

Figure 4.1: AI-Powered Cyber Threat Detection Dashboard — UNSW-NB15 Dataset (Random Forest AUC: 0.991)

4.2 Feature Importance Analysis

Random Forest analysis reports that sbytes (source bytes) , sttl (source time to live), and sload (source load) are the top three which is used to perform the task of differentiating between attacks and normal traffic. These network level byte and timing features also report in agreement with what was put forth by Abdallah et al. [2] who reported that traffic volume and connection duration features are key. Also we see that packet level and flow level statistical features do better than protocol level categorical features for the UNSW-NB15 dataset.

4.3 Confusion Matrix Analysis

In the case of Random Forest and also in the case of Gradient Boosting we see that which models did very well at classifying almost all of the normal and attack traffic. Random Forest put out 7,135 normal and 8,507 attack instances ,with a few misclassifications. Gradient Boosting reported 7,054 normal and 8,156 attack which had more false negatives which is to say that it did not catch as many attack instances as Random Forest. Also these results play out as what has been reported in the past literature.

5. Conclusion

This paper reviewed existing research on AI and machine learning techniques for cyber threat detection, with a focus on their relevance to the defence sector. The analysis confirms that machine learning-based intrusion detection systems, particularly those using ensemble methods such as Random Forest, outperform traditional signature-based approaches in detecting both known and unknown threats.

The UNSW-NB15 dataset remains one of the most suitable benchmarks for evaluating modern intrusion detection systems due to its representation of contemporary attack scenarios and its relevance to defence-related applications. However, the study also highlights several important limitations that restrict real-world applicability: the absence of defence-specific datasets, challenges in handling imbalanced data, lack of model interpretability, and constraints in real-time deployment.

6. Future Work

Future research must prioritize practical deployment challenges rather than focusing solely on improving model accuracy metrics. The following directions are recommended:

Defence-Representative Synthetic Datasets: The creation of defence-specific synthetic datasets using advanced GAN architectures is the most immediate priority. Federated learning techniques can also enable model training across distributed data sources without centralizing sensitive information, a viable approach for developing defence-relevant models while adhering to security protocols [12].

Explainable AI (XAI) Integration: Future IDS systems should provide human-comprehensible explanations of detection decisions via SHAP-based feature attribution or similar post-hoc explainability methods. This is essential to satisfy the transparency and accountability requirements of defence environments [5].

Adversarially Robust Training: Developing and evaluating adversarially robust training methods that maintain high detection rates even under sophisticated evasion attempts is a prerequisite for deploying ML-based IDS in defence settings [9].

Advanced Multi-Class Classification: Research should focus on fine-grained multi-class classification particularly for rare attack types using techniques such as hierarchical classification, cost-sensitive learning, and ensemble stacking to support operational response decisions [2].

Quantum and Federated Learning Integration: Emerging approaches such as quantum-enhanced ML and privacy-preserving federated learning hold promise for overcoming current scalability and data-sharing barriers in defence cybersecurity [12].

References

[1] Buczak, A. L., and Guven, E.: A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176, (2015).

[2] Abdallah, E. E., Eleisah, W., and Otoom, A. F.: Intrusion Detection Systems using Supervised Machine Learning Techniques: A survey. Procedia Computer Science, 201, 205–212, (2022).

[3] Mahdavifar, S., and Ghorbani, A. A.: Deep Learning for Cybersecurity: A Survey. arXiv preprint arXiv:1906.05799, (2019).

[4] Liya, B., and Lee, A.: AI-Powered Threat Detection: Enhancing Cyber Defense Through Predictive Analytics, (2025).

[5] Katiyar, N., et al.: AI and Cyber-Security: Enhancing threat detection and response with machine learning. Educational Administration: Theory and Practice, 30(4), 6273–6282, (2024).

[6] Amachaghi, E. N., Abdulkareem, S. A., Foh, C. H., Mi, D., and Shojafar, M.: Improving intrusion detection in O-RAN with synthetic data generation: A GAN and SMOTE approach. Telematics and Informatics Reports, 20, 100269, (2025).

[7] Rahman, S., et al.: SYN-GAN: A robust intrusion detection system using GAN-based synthetic data for IoT security. Internet of Things, 26, 101212, (2024).

[8] Rahman, M. A., Francia, G. A., and Shahriar, H.: Leveraging GANs for synthetic data generation to improve intrusion detection systems. Journal of Future Artificial Intelligence and Technologies, 1(4), 429–439, (2025).

[9] Yaseen, A.: AI-driven threat detection and response: A paradigm shift in cybersecurity. International Journal of Information and Cybersecurity, 7(12), 25–43, (2023).

[10] Sivakumar, J., et al.: AI-driven cyber threat detection: enhancing security through intelligent engineering systems. Journal of Information Systems Engineering and Management, 10(19), 790–798, (2025).

[11] Reddy, A. R. P.: The role of artificial intelligence in proactive cyber threat detection in cloud environments. NeuroQuantology, 19(12), 764–773, (2021).

[12] Sarfraz, M., Sumra, I. A., Khalid, B., and Fatima, E.: AI-driven predictive threat detection and cyber risk mitigation: A survey. Journal of Computing & Biomedical Informatics, 8(2), (2025).

[13] Rizvi, M.: Enhancing cybersecurity: The power of artificial intelligence in threat detection and prevention. International Journal of Advanced Engineering Research and Science, 10(5), 055–060, (2023).

[14] Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H., and Wang, C.: Machine learning and deep learning methods for cybersecurity. IEEE Access, 6, 35365–35381, (2018).

[15] Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A., and Xu, M.: A survey on machine learning techniques for cyber security in the last decade. IEEE Access, 8, 222310–222354, (2020).