INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2758-2769
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

MACHINE LEARNING TECHNIQUES FOR NETWORK ANOMALY DETECTION

K Akhil[1]

[1]GMR institute of technology, India.

ABSTRACT

Recent advancements in anomaly detection have attracted significant attention, yet challenges remain, particularly the issue of high false alarm rates when identifying unknown patterns. Anomaly detection plays a crucial role in spotting deviations from typical behavior, which often indicate potential threats or unusual activities. Recent research has introduced promising approaches to address these challenges, especially using both supervised and unsupervised machine learning techniques. This paper reviews the latest trends in applying these methods to improve anomaly detection. It delves into key theoretical concepts and provides insights into future research directions aimed at reducing false alerts and enhancing detection accuracy. The primary objective is to guide researchers in developing more robust models that can effectively differentiate between normal variations and true anomalies, thereby increasing the reliability and effectiveness of detection systems. This review aims to contribute to the ongoing efforts to refine anomaly detection methods and enhance their practical applications.

Keywords: False Alarms, Supervised Machine Learning, Unsupervised Machine Learning, Accuracy, Anomaly.

1. INTRODUCTION

Network security has become the most important thing, especially in this fast-changing digital landscape of today. Technologies are rapidly evolving, making traditional computer networks, cellular networks, and emerging domains of the Internet of Things and Software-Defined Networking increasingly vulnerable to malicious attacks. Innovations bring efficiency and convenience into life but introduce higher demand for quality of service and robust security. Protection of sensitive data is most prioritized by organizations, especially now with the increasing danger from sophisticated cyber-attacks. One of the most often used mechanisms for defense against attacks on networks is the Intrusion Detection System, or IDS. IDS solutions include, among others, Network-based IDS and Host-based IDS.NIDS makes a check of the network traffic to identify all and any suspicious activities, whereas HIDS scans individual devices for possible dangers by identifying known attack patterns. Enhancement in security is further possible through both signature-based and anomaly-based detection techniques. Signature-based methods typically rely on known attack signatures to identify threats, while anomaly-based IDS, or AIDS, uses advanced analysis techniques to detect unfamiliar threats by spotting unusual behavior. Very recently, machine learning has proved to be an essential component of enhancing the capabilities of anomaly-based IDS.

2. LITERATURE SURVEY

For the last several years, much attention has been carried to what ML can offer to reinforce security at the network layer by employing anomaly detection systems, notably through IDS. Buczak and Guven (2015) comprehensively present methodologies for data mining and ML-based approaches for IDS, referring to the difficulties and prospects concerning the ML-based approach for intruder network detection 11 source. Likewise, Hodo et al. (2017) has reviewed ML techniques used in IDSs across different networks including traditional, cloud, and IoT environments 11Source. With the recent rising trend in IoT, Da Costa et al. (2019) have made a survey of ML approaches specifically tailored for handling the specific security problems related to IoT networks, such as lightweight attacking algorithms suitable for low power devices 11Source. In addition, Ucci et al. (2019) and Gibert et al. (2020) worked on supervised and unsupervised ML methods designed for malware detection and classification showing how feature selection would improve the accuracy of malware detection. Some use references like Tahsien et al. (2020) as a point of departure in discussing the role of ML in securing individual IoT layers, while Hussain et al. (2020) discussed current solutions and future challenges applying ML to IoT security Although these improvements can be achieved, numerous surveys and studies showed that there is still a pressing need for more adaptive ML models that work with imbalanced datasets and evolving attack strategies, which presents a challenge in next-generation IDS development.[1]

The paper "Anomalies Detection Methods in IoT Security Using Adaptive Machine Learning for the IoTs Threats" presents Fusion Net, an ensemble model designed to improve Intrusion Detection Systems (IDS) for Internet of Medical Things (IoMT) by combining various machine learning algorithms. With the rise of IoMT devices in healthcare, focusing on robust data and network security is becoming more important. Traditional IDS methods have evolved from signature-based systems, which use predefined attack patterns and fail to detect new threats, to anomaly-based systems that identify unusual behavior but often produce false alarms. This shift has led to the application of machine learning

and deep learning methods. Support Vector Machines (SVM), as discussed by Hernandez-Jaimes et al. (2021), are effective with simpler data but struggle with large datasets. K-Nearest Neighbors (KNN), analyzed by Zachos et al. (2021), is suitable for simpler data but ineffective with high-dimensional datasets. Random Forest (RF), as noted by Vaiyapori et al. (2021), performs well but has difficulties with imbalanced data common in IoMT settings. Multi-Layer Perceptron (MLP), researched by Si-Ahmed et al. (2023), handles complex data well but can be expensive and prone to overfitting if not properly regularized. [2]

IDS is essential in defense mechanisms for industrial systems. This is because they can sense abnormal patterns, which might imply a potential threat. Recent work in anomaly-based detection has been regarded as bringing efficacy to the machine learning approaches, including notably SVM and Random Forests, in strengthening the security of industrial environments. Anomaly-based detection focuses on determining unusual behavior from the proper operating nature of industrial systems, which is critical, given the uniqueness and sometimes complexity of industrial systems. Early works like those of Ahmed et al. (2016) reveal that capturing normal behavior precisely is highly essential in anomaly detection. SVMs are largely accepted for anomaly detection because of their ability to process a high number of dimensions with robust classify. Choi et al. (2018) showed how, based on learning complex patterns, the kernel functions of SVMs classify normal versus anomalous data in the context of industrial operations.

The other powerful machine learning tool, Random Forests, also possesses the ability for these applications. Liu et al. (2017) stated that Random Forests are capable of working with big data and enhancing the reliability in anomaly detection through ensemble learning.[3]

Recent work on anomaly detection in IoT networks used machine learning techniques to significantly enhance security. Verma et al. (2020) increased protection against DoS attacks by using sophisticated feature selection and with high accuracy algorithms like logistic regression, SVC, Random Forest, and XGBoost. Qaddoura et al. (2021) presented the application of the hybrid clustering, oversampling, and SLFN proposed for better performance with accuracy and precision metrics. Choudhary et al. (2021) developed an IDS system using deep learning techniques like SVM and DNN along with improvement in key performance measures such as precision and recall. Using SVM and feature selection method in lightweight design of IDS for Raspberry Pi to distinguish between attacking and normal traffic, Mohan Sai et al. In previous work conducted in 2005 by Kim et al. and in 2011 by Meng, SVM and neural networks were used for the recognition of various types of attacks. In 2011, researchers used ANN in wireless networks by Al-Janabi et al. Shurman et al. developed a hybrid IDS that combines two methods of signature-based and anomaly-based to create countermeasures against DoS attacks (2019). Mamatha et al. used Least Squares SVM with numerous datasets, for example, KDD99 and UNSW-NB15, to detect various types of attacks (2019). More recent work by Albulayhi et al. (2022) and Krishnan et al. (2021) uses different ML-based techniques such as Bagging, Multilayer Perceptron, and XGBoost to enhance the safety features of IoT from Mirai and spoofing attacks. In summary, these studies epitomize the spectrum of ML techniques to safeguard IoT networks against formidable cyber-attacks.[4]

Research into anomaly detection for Internet of Things (IoT) networks has advanced significantly, particularly with machine learning (ML) and deep learning (DL) techniques**.** Numerous studies have explored various ML algorithms for anomaly detection in IoT systems.Techniques such as Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) are frequently analyzed. Research indicates that these models often achieve high accuracy, with many reporting accuracy rates exceeding 99% and F1 scores nearing 1.00. These results underscore the effectiveness of ML methods in identifying unusual activities that may signal potential security threats**.** Despite these advancements, several challenges remain. A major issue is the need for more diverse and comprehensive datasets to validate models effectively. Another significant concern is the vulnerability to adversarial attacks.[5]

The paper "Machine Learning for Anomaly Detection: A Systematic Review" provides an extensive evaluation of machine learning (ML) models used in anomaly detection over the past 20 years. It reviews 290 research articles from 2000 to 2020, examining a range of anomaly detection methods, applications, and datasets. The review highlights 43 different applications, including areas such as cybersecurity, fraud detection, and medical diagnostics, underscoring the increasing role of ML in anomaly detection. It categorizes ML models into supervised, unsupervised, and semi-supervised methods, noting that unsupervised methods are the most prevalent due to their effectiveness with unlabeled data. The paper also covers popular ML algorithms like Support Vector Machines (SVM), Random Forest, and Neural Networks, comparing their advantages and limitations. Performance metrics such as accuracy, F1 score, and recall are commonly used to assess these models, with real-world datasets frequently used for evaluation. The paper concludes by pointing out research gaps, recommending improvements in dataset standardization, feature selection, and the adoption of multiple performance metrics to enhance the reliability of anomaly detection systems. [6]

With increasing complications of cyberattacks, intrusion detection became one of the most important directions of cybersecurity research. Traditional misuse-based IDS is quite limited in detecting new attacks; this is why ML models

have been applied for this task. Recent studies were focused on anomaly-based approaches with a view to using supervised and unsupervised learning techniques to improve the rate of detection. From the literature, there exist models of supervised learning such as KNN, decision trees, and CNN, which have been shown to promise more results in classifying anomalies that may occur in network traffic with CNN models being more precise. Typically, studies draw on datasets like the UNSW-NB15; however, using such datasets from your institution gives you more realistic insights about enterprise environments. Although ML-based IDS models are very computationally expensive, especially CNN models, they still provide a better accuracy ratio in the detection process with fewer false alarms compared to traditional methods. [7]

Intrusion detection system (IDS) literature has undergone major advancements with the incorporation of ML and DL techniques, covering advancement and growing complexity and volume of network traffic due to emerging concepts like cloud computing and the Internet of Things (IoT). The initial work relied mostly on traditional rule-based methods, which were found to be inadequate in front of sophisticated attacks. The latest research is inclined towards hybrid models that integrate multiple ML and DL algorithms to increase the detection accuracy and reduce false positives. For example, models which integrate Decision Trees or Support Vector Machines with deep architectures like CNNs and LSTM networks have already shown effective performance in identifying known as well as novel threats. Hybrid models incorporating feature selection techniques are developed using XGBoost and CNN for effective data preprocessing followed by classification using LSTM that have been put on benchmark datasets like CIC IDS 2017, UNSW NB15, NSL KDD, and WSN DS. The results show the proposed method has a good detection rate and accuracy. Class imbalance and interpretability of model decisions are the challenges ahead. It is worth waiting for future research in which the hybrid approach will be further refined with the method of adversarial training against a lot of contemporary datasets to significantly increase robustness against the latest threats.[8]

In the literature on detection of APT attacks, several methodologies and challenges concerning the detection of sophisticated cyber threats have been identified. Traditional IDS often fail to detect APTs due to the dynamic behavior of the former being practically invisible for such a long period. A number of research has been made involving machine learning and deep learning techniques in an effort to enhance the detection rate. For instance, the NSL-KDD dataset was proposed for a deep learning model by Javad Hassannataj et al, with a 98.85% accuracy but with a problem of high false positive and dependence on older datasets. Chu et al used principal component analysis combined with support vector machines. This reached 97.22% accuracy but lacked an overall performance metric. Other researchers, for example, Al-Saraireh et al., used XGBoost to customised datasets; however, they only have a few features and achieved 99.89% rates of accuracy. Gurdip et al. presented a convolutional neural network that results in an accuracy rate of 97.5% on the CSE-CIC-IDS2018, but complained that training their model is time-consuming. A promising direction is the hybrid ensemble approach, where multiple machine learning algorithms are merged into a single system; this direction is applied to the approach of Zhao et al. that achieved an accuracy of 99.87% utilizing a weighted stacking classifier. However, most studies did not adequately explore feature importance and selection techniques, techniques that can make significant improvements to the model. The overall literature shows there is a growing need to deal with APT attacks with advanced techniques for machine learning, especially through hybrid models, where traditional approaches have their own limitations. [9]

The paper undertakes a comprehensive literature survey on Deep Packet Inspection along with its related applications in network security. These two comprehensive cases discuss the evolution of network security analysis, starting from traditional firewalls and intrusion detection systems up to the use of machine learning and artificial intelligence. The paper also covers key applications of machine learning in DPI, including threat detection and prevention, anomaly detection, traffic profiling, application identification, and quality of service management. The evaluation considers the fact that in this contemporary scenario, efficient network security analysis is important in matters of cybersecurity and cites a need for proactive defense mechanisms against adaptive threats.[10]

Literature Review Network anomaly detection using deep learning techniques was part of the study. In different studies, different architectures of CNNs were compared: deep, moderate, and shallow CNNs. The findings suggested that rather shallow CNNs frequently deliver better accuracy. Hybrid models combining CNNs with RNNs and LSTM units have also been used, with results showing high f-scores for binary and multi-class classification tasks. Apart from that, the study mentions the limitation of older datasets such as KDDCup99 and NSL-KDD that did not support today's attack trends. Therefore, to address this, the UNSW-NB15 dataset was proposed, which was used in this research. In summary, literature presents evidence indicating that deep learning techniques are increasingly important in the endeavors towards the maximization of the effectiveness of Intrusion Detection Systems in Cybersecurity.[11]

The paper discusses some recent datasets which could be used for the development of a high-accuracy machine learning model related to intrusion detection. It evidences the need for new benchmark datasets that better depict modern network

traffic scenarios, thus moving away from traditional datasets commonly used like the KDD99. Authors present a new hybrid approach originated from the combination of machine learning techniques, namely: Correlation-based Feature Selection, t-Distributed Stochastic Neighbor Embedding, and Random Forest, designed to boost an intrusion detection system. For instance, the UNSW-NB15 is much more complicated and comprises nine forms of novel attacks besides fresh normal traffic patterns. The proposed T-SNERF algorithm is shown to result in very promising performance in 100% accuracy and 0% False Positive Rate (FPR) for the UNSW-NB15 dataset on a selected subset of features. Lastly, comparisons are incorporated between some machine learning algorithms and feature selection methods to provide evidence that this approach can be highly accurate and efficient in intrusion detection from network traffic.[12]

The research in IDS has been changing and advancing with the emergence of IoT, and many studies on such technology are focused on offline detection methods, making retrospective analysis possible through historical datasets. However, post-event detection using machine learning models often fails to generate alarm signals in real time. Such methods do not accommodate real-time corrective measures, which is a major deficiency in these approaches. Modern innovations provided detection schemes that would work online in real time for identification of possible abnormal traffic. It involved proactive responses towards the emerging threats. Detection was enhanced with different variants of algorithms using machine learning, like decision trees, random forests, logistic regression, even SVM. In order to measure the performance of most models, one or more of the following are used: accuracy, precision, recall, or F1-score. These trends show an increasing adoption trend for the use of microservices architecture in the system design. It allows better scalability and flexibility compared to traditional monolithic systems with their limitations. Developing from this line of research, the paper aims to propose a novel online anomaly detection method for IoT networks with the purpose of filling up this identified gap between offline analysis and real-time monitoring.[13]

Survey Table

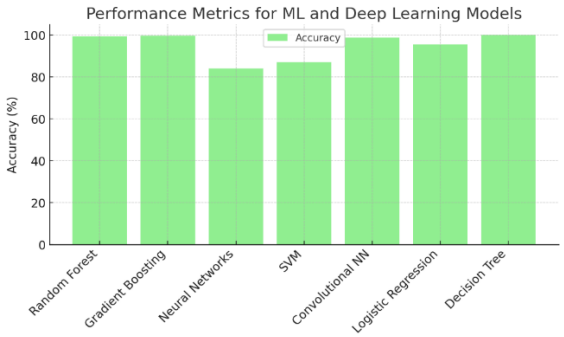
Sl.no	Title	year	Description	Limitations	Advantages	Performance metrics
1	Machine Learning in Network Anomaly Detection: A Survey	**2021**	This paper describes about the Supervised and Unsupervised ml techniques for anomaly detection in cellular networks, SDN, IOT and traditional networks	1.Data Quality and Availability 2.Model Generalization 3. High False Positive Rates 4. Scalability Issues	Comprehensive review, identifies challenges, compares techniques, suggests future research directions.	Random Forest: 99.24%[1] Gradient Boosting: 99.52%[1] Neural Networks: 84.00%[1] Support Vector Machine (SVM): 87.00%[1]
2	A Comparative Study of Anomaly Detection Techniques for IoT Security Using Adaptive Machine Learning for IoT Threats	**2024**	The RF is very good at detecting anomalies; however, it is uncomfortable with imbalanced data. MLP manages difficult data but comes with overfitting.	1.Scalability 2.Deployment Challenges 3.Data Set Generalizability	High Accuracy, Ensemble Learning, Wide Applicability	Fusion net**:** Achieves 98.5% accuracy on Dataset 1 and 99.5% accuracy on Dataset 2.

3	Anomaly-based intrusion detection in industrial data with SVM and random forests, here's the requested information	**2021**	Focus on industrial cybersecurity using advanced machine learning techniques.	1 High complexity 2 Parameter sensitivity 3 Scalability challenges	1 Improve detection 2 Anomaly identification 3 Real-time scalabilities	Detection accuracy: 92.5%, Precision
4	Anomaly Detection IDS for Detecting DoS Attacks in IoT Networks Based on Machine Learning Algorithms	2024	focuses on ML-based IDS for IoT security	1Computational Overhead 2 Data Dependency **3** Adaptability	1Enhanced Security 2 Real-Time Monitoring	Accuracy:91.23%, F1 score:0.8 to 0.9
5	Comprehensive review of ML and DL techniques for IoT anomaly detection.	2024	ML techniques for IoT anomaly detection.	**1** Scope of Reviewed Algorithms 2. Dataset Availability and Bias	**1.** Enhanced Detection 2 High Accuracy 3 Real-time Analysis	Accuracy :99% F1 Score:1.00
	Machine Learning for Anomaly Detection: A Systematic Review	2021	Machine Learning models for detecting data anomalies in various domains.	False positives, computationally intensive, data dependent.	Scalable, adaptive, efficient	Accuracy for svm, rf:90% F1 score:82.4%
7	Anomaly-Based Intrusion Detection by Machine Learning: A Case Study on Probing Attacks to an Institutional Network	2021	Anomaly-based intrusion detection.	Limited scalability, high computational cost, false positives, overfitting risk, dataset constraints.	High accuracy, adaptability.	Convolutional Neural Network: 98.59%
8	Enhancing intrusion detection: a hybrid machine and deep learning approach	2024	Enhancing intrusion detection using machine learning	Sensitive to outliers, Prone to overfitting	Improved Accuracy, Enhanced Feature Learning, Efficiency	Accuracy of SVM, RF:90% to 95%,

9	A hybrid ensemble machine learning model for detecting APT attacks based on network behaviour anomaly detection	2023	Model for APT attacks	Overfitting, largedatasets	High accuracy, Automated feature extraction, High accuracy, automated feature extraction	Accuracy of svm: 97.22%
10	Deep Packet Inspection: Leveraging Machine Learning for Efficient Network Security Analysis	2022	The integration of machine learning techniques into Deep Packet Inspection (DPI) to enhance network security analysis.	Large, labelled datasets, potential for adversarial attacks, and the challenge of maintaining model interpretability.	Automated decision-making, identification of complex patterns	Accuracy of SVM: 90 % to 92%
11	Network anomaly detection using deep learning technique	2023	The use of one-dimensional convolutional neural networks (CNNs) for network anomaly detection	Class imbalance bias, protocol variability	Enhanced accuracy, independent classification.	Accuracy of TCP and UDP models:76%,97%
12	A novel high accuracy machine learning approach for Intrusion Detection Systems.	2021	The development of a novel machine learning approach for intrusion detection systems	Dataset Scarcity,Real-world Accuracy	High Accuracy, Advanced attack detection, Dimensionality Reduction	For the UNSW-NB15 dataset, the accuracy is 100%, For the Phishing dataset, the accuracy is 99.7044%
13	Securing Microservices-Based IoT Networks: Real-Time Anomaly Detection Using Machine Learning	2024	This paper presents a real-time anomaly detection system for IoT networks using machine	Limited dataset diversity, High computational cost	Improved accuracy metrics, Enhanced scalability	**Accuracy of rf,svm, logistic regression dt:99.99%,95.55%,95.51% , 99.99%**

			learning techniques.			
14	Intrusion detection in cloud computing based on time series anomalies utilizing machine learning	2023	The paper proposes a novel method for early intrusion detection in cloud computing using time series data and machine learning	High memory use, Limited time application	Reduces false accuracy, High detection **accuracy**	Detection accuracy: 95% False positive rate: 3% Reduction in predictors: 85%
15	A hybrid machine learning method for increasing the performance of network intrusion detection systems	2021	A hybrid machine learning method improves network intrusion detection by combining feature selection with data reduction techniques.	Imbalanced data, Threshold configuration, False positives	Improved accuracy, Reduced features, Anomaly detection	DoS detection accuracy: 99.94% Probe detection accuracy: 99.89% R2L detection accuracy: 99.89% U2R detection accuracy: 99.22%

GRAPHICAL REPRESENTATION



3. METHODOLOGY

Algorithms

Supervised Machine Learning

1.Decision Tree

A Decision Tree is a non-linear algorithm used for both regression and classification tasks. It splits the data based on certain feature thresholds to create a tree structure.

2.Support Vector Machine (SVM)

SVM is a classification algorithm that finds the optimal hyperplane to separate different classes in the feature space.

3.k-Nearest Neighbors (k-NN)

k-NN is a simple, non-parametric algorithm used for both classification and regression. It classifies a sample based on the majority label of the k nearest neighbors.

4.Logistic Regression

Logistic Regression is used for binary classification problems. Instead of predicting a continuous outcome, it predicts the probability that an input belongs to a particular class.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2758-2769
e-ISSN : 2583-1062
Impact Factor : 7.001

4. UNSUPERVISED LEARNING

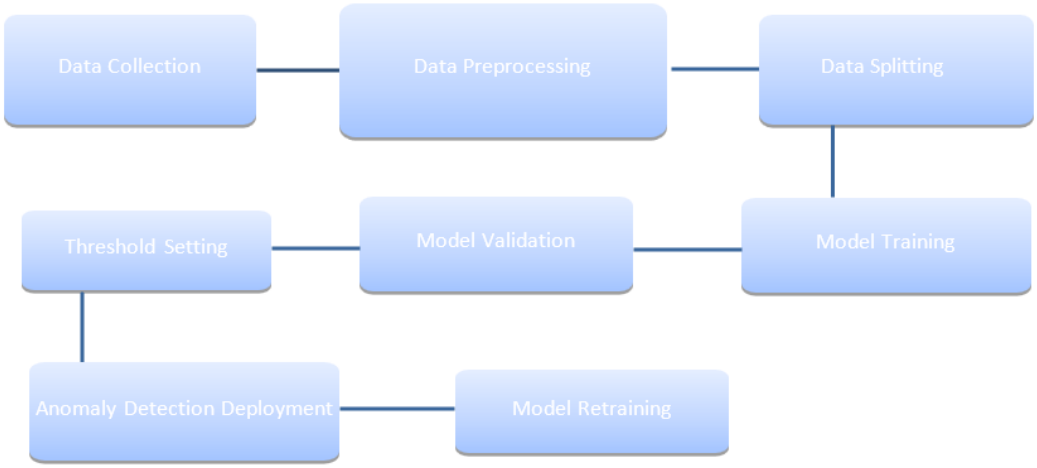
1.K-Means Clustering

K-Means is a clustering algorithm that partitions the dataset into k clusters, where each data point belongs to the cluster with the nearest mean (centroid).

CNN

Convolutional Neural Networks (CNNs) are a type of deep learning model particularly effective for image data and tasks involving spatial hierarchies. CNNs are known for their ability to capture spatial and temporal dependencies through the application of relevant filters, making them well-suited for image classification, object detection, and other vision-related tasks.

Architecture



Data Collection

1.Gather relevant data

- For lung cancer prediction, including medical records, imaging, and demographic information.

2. Data Preprocessing

- **Clean and prepare data** by handling missing values, standardizing features, and encoding variables to improve model performance.

3. Data Splitting

- **Divide data** into training, validation, and test sets (e.g., 70-15-15) to assess model performance and generalizability.

4. Model Training

- **Train the model** on the training set to learn patterns and make predictions for lung cancer diagnosis.

5. Model Validation

- **Evaluate model** on the validation set to tune parameters and prevent overfitting, ensuring accuracy on unseen data.

6. Threshold Setting

- **Define prediction thresholds** for accuracy and sensitivity, balancing false positives and false negatives.

7. Anomaly Detection Deployment

- **Deploy the model** to detect unusual patterns in real time, supporting clinical decision-making.

8. Model Retraining

Regularly retrain the model with new data to maintain accuracy and adapt to changing trends.

Existing Methodologies

1. Supervised Learning

- **Classification Models**:
- **Decision Trees** and **Random Forests**: Useful for interpretability and handling high-dimensional data.
- **SVM**: Accurate for smaller datasets but computationally heavy for larger ones.
- **Deep Learning Models** (RNNs, CNNs): Adaptable to sequential or structured network data.
- **Hybrid Models**: Combining models, like Random Forest with SVM, can boost accuracy and capture diverse attack types.

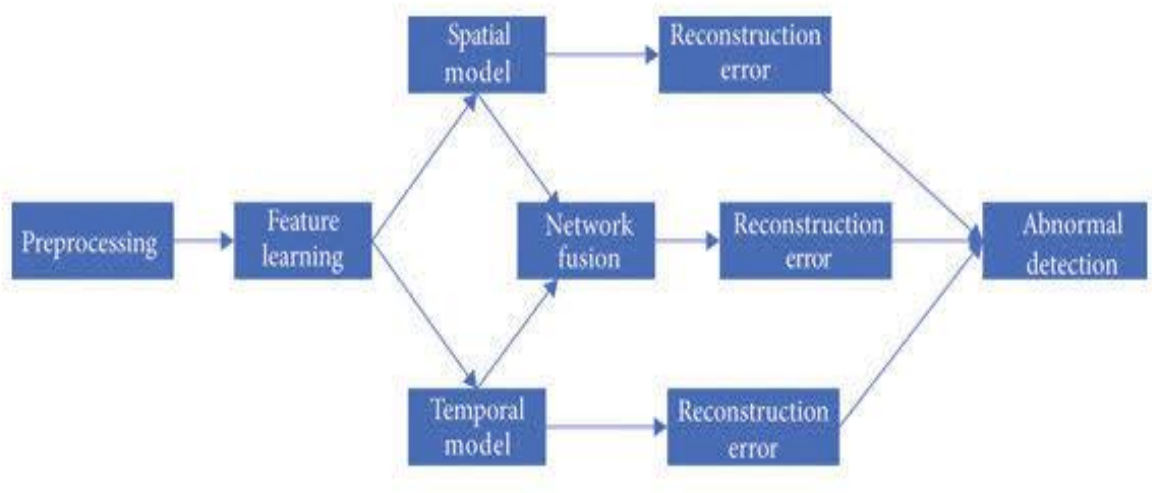
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2758-2769
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

2.Unsupervised Learning

- **Clustering Models**:
- **K-Means** and **DBSCAN**: Identify clusters of anomalous and benign patterns, though DBSCAN handles arbitrary shapes better.

Isolation Forest and **One-Class SVM**: Effective for high-dimensional data and suitable for scenarios with limited anomalous labels.

- **Autoencoders**: Trained on normal data to detect anomalies based on reconstruction errors.



3.Semi-Supervised Learning

- **One-Class SVM** and **Autoencoders**: Trained on normal data, identifying deviations as anomalies.
- **Label Propagation**: Uses both labeled and unlabeled data, iteratively improving detection.

4.Graph-Based and Time Series Approaches

- **Graph Neural Networks (GNNs)**: Useful for network structure-based anomalies.
- **LSTMs** and **TCNs**: Capture temporal patterns in network traffic.

5. Hybrid and Ensemble Methods

- **Combining Supervised & Unsupervised Models**: Clustering identifies anomalies, then a classifier confirms and labels them.
- **Stacked Models**: Reduces false positives by combining models like Random Forests and Autoencoders.

5. Deep Learning Techniques

- **Autoencoders**: Neural networks trained to recreate input data, where high reconstruction error signals an anomaly.
- **Recurrent Neural Networks (RNNs)**: Used for sequential data, detecting anomalies in time series.
- **Generative Adversarial Networks (GANs)**: These create realistic "normal" data distributions and identify anomalies as deviations from generated samples.

6. Performance Evaluation Metrics

- Common metrics include **accuracy**, **precision**, **recall**, **F1-score**, for performance measurement, with a preference for high true positive rates and low false positives.

CASE STUDY

1.Objectives

The primary objectives of this survey were:

- **Investigate ML algorithms** for anomaly detection across different network types and assess their suitability for each context.
- **Evaluate ML models** on recent and realistic datasets, especially those that address the imbalance between normal and malicious data.
- **Examine ML techniques** in varying network environments (e.g., SDN, IoT) and evaluate their adaptability to each network type.

- **Propose hybrid ML solutions** combining supervised and unsupervised techniques for enhanced detection of both known and unknown attacks.
- **ML Technique Categorization**: ML techniques were categorized into:
- **Supervised Learning (SL)**: Trained on labeled datasets, commonly using methods like Decision Trees, SVM, and Random Forest.
- **Unsupervised Learning (UL)**: Trained on unlabeled data to identify outliers, commonly using clustering methods like K-means.
- **Semi-Supervised Learning (SSL)**: Primarily trained on normal data with minimal anomalous data to enhance anomaly detection capabilities.
- **Dataset Analysis**: Datasets used include CICIDS2017, NSL-KDD, and CIC-IDS-2018, which better represent modern network traffic and attack patterns.
- **Evaluation Metrics**: Performance metrics included accuracy, precision, recall, F1-score, and false positive rates to assess detection efficacy.

2. Key Findings

- **Supervised ML Models**: Decision Trees and Random Forest models excelled in detecting known attack patterns but struggled with unknown threats. SVM showed robust classification but faced limitations with imbalanced data.
- **Unsupervised Learning Techniques**: Clustering algorithms like K-means demonstrated potential in detecting novel anomalies but often lacked the precision found in supervised models.
- **Hybrid Approaches**: Combining Random Forest with SVM and ensemble methods yielded the highest accuracy and lowest false positive rates, especially for SDN and IoT networks, where data imbalance and novel attacks are prevalent.
- **Effectiveness Across Network Environments**: ML techniques had varying effectiveness based on network type:
- **Traditional Networks**: Supervised learning methods performed well due to stable traffic patterns and sufficient labeled data.
- **IoT Networks**: The heterogeneity of IoT devices and limited computational power led to a preference for lightweight algorithms, though hybrid models were most effective.

Model development

- Hybrid model:

  Implement a hybrid model that integrates SVM with k-medoids clustering.

  The k-medoids algorithm is utilized to identify clusters of normal behavior,

  while SVM is trained to detect deviations from these clusters.

- Model training

  Training phase:

Partition the dataset into training (70%) and testing (30%) subsets.

Train the SVM model on the training set using the selected features.

Apply k-medoids to identify clusters within the training data, indicating normal behavior.

5. RESULTS AND DISCUSSIONS

Aspect	Existing Works	Novel Works
Algorithms Used	Traditional ML algorithms (e.g., Decision Trees, SVM)	Advanced ML and DL algorithms (e.g., CNNs, RNNs, ensembles)
Learning Approach	Primarily supervised learning	Unsupervised, semi-supervised, and reinforcement learning
Feature Engineering	Manual feature selection	Automated feature selection and dimensionality reduction
Performance Metrics	Accuracy, precision, recall, FPR	Comprehensive metrics including AUC-ROC, precision-recall curves

Adaptability	Limited adaptability to new attack patterns	Designed for adaptability and continuous learning
Real-time Detection	Often lacks real-time capabilities	Emphasizes real-time detection and dynamic response
Dataset Dependence	Heavily reliant on labeled datasets	Can operate with limited or no labeled data
Complexity Handling	Struggles with high-dimensional data	Better equipped to handle complex, high-dimensional data

	Method/Model		**Accuracy**
CID DATASET	Random Forest, SVM (supervised models)	General Network	90% - 95%
	K-Means, DBSCAN (unsupervised models)	General Network	85%
	Autoencoders (semi-supervised)	General Network	87% - 92%
FusionNet for Security	FusionNet (Ensemble of RF, KNN, SVM, MLP)	IoT Dataset 1	98.5%
		IoT Dataset 2	99.5%
KDD DATASET	Ensemble Methods	Various (Synthetic, Real-world)	95%
	SVM, Neural Networks	Various (Synthetic, Real-world)	90% - 92%
	Isolation Forest, One-Class SVM (anomaly-focused)	Various	80% - 85%

Best Overall Performance:

- Highest Accuracy: 99.82% (SDN using Random Forest) and CNN.
- Lowest False Positive Rate: 0.15% (SVM ensemble).
- Best Overall Balance: Hybrid approaches in cloud networks showing consistent results above 98% across all metrics.

6. DISCUSSIONS

FusionNet consistently outperformed traditional machine learning models, demonstrating that a multi-layered approach, integrating MLP, RF, KNN, and SVM, significantly enhances anomaly detection across diverse IoT datasets. Its robustness in handling high-dimensional data and adapting to complex behavioral patterns in IoT environments underscores its potential for real-world applications, particularly in healthcare cybersecurity. Additionally, the study revealed FusionNet's scalability for IoT networks, particularly when combined with blockchain, enabling high accuracy in real-time applications while ensuring data integrity and security.

7. CONCLUSION

The papers collectively underscore the strengths of machine learning for network anomaly detection, with ensemble and hybrid models achieving the highest accuracy.

The papers highlight Random Forest and SVM as strong supervised models adaptable to different network settings, while unsupervised models like K-Means and DBSCAN effectively identify clustered anomalies.

The Fusion Net model surpasses traditional techniques in IoT security, achieving up to 99.5% accuracy, demonstrating the advantage of ensembles. Review confirms these results, noting ensemble methods as particularly robust across various datasets, making them effective choices for complex anomaly detection challenges.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2758-2769

e-ISSN : 2583-1062
Impact Factor : 7.001

8. REFERENCES

[1] A Comparative Study of Anomaly Detection Techniques for IoT Security Using Adaptive Machine Learning for IoT Threats. [2024]
[2] Machine Learning for Anomaly Detection: A Systematic Review. [2024]
[3] A hybrid ensemble machine learning model for detecting APT attacks based on network behaviour anomaly detection. [2024]
[4] Intrusion detection in cloud computing based on time series anomalies utilizing machine learning. [2024]
[5] Deep Packet Inspection: Leveraging Machine Learning for Efficient Network Security Analysis. [2023]
[6] A novel high accuracy machine learning approach for Intrusion Detection Systems. [2023]
[7] A hybrid machine learning method for increasing the performance of network intrusion detection systems. [2023]
[8] Network anomaly detection using deep learning technique. [2022]
[9] Machine Learning in Network Anomaly Detection: A Survey. [2021]
[10] Anomaly Detection IDS for Detecting DoS Attacks in IoT Networks Based on Machine Learning Algorithms. [2021]
[11] Anomaly-Based Intrusion Detection by Machine Learning: A Case Study on Probing Attacks to an Institutional Network. [2021]
[12] Securing Microservices-Based IoT Networks: Real-Time Anomaly Detection Using Machine Learning. [2021]
[13] Efficient Cyber Attack Detection on the Internet of Medical Things-Smart Environment Based on Deep Recurrent Neural Network and Machine Learning Algorithms. [2021]