IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 June 2024, accepted 28 June 2024, date of publication 2 July 2024, date of current version 10 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3422077

RESEARCH ARTICLE

A Traffic Analysis and Node Categorization-Aware Machine Learning-Integrated Framework for Cybersecurity Intrusion Detection and Prevention of WSNs in Smart Grids

TAMARA ZHUKABAYEVA[1,2], AISHA PERVEZ[3], YERIK MARDENOV[1,4], MOHAMED OTHMAN[5], (Senior Member, IEEE), NURDAULET KARABAYEV[1], AND ZULFIQAR AHMAD[6]

[1]International Science Complex "Astana," 020000 Astana, Kazakhstan
[2]Faculty of Information Technology, L. N. Gumilyov Eurasian National University, 010000 Astana, Kazakhstan
[3]Telecommunication Department, Hazara University, Mansehra 21300, Pakistan
[4]Higher School of Information Technologies and Engineering, Astana International University, 020000 Astana, Kazakhstan
[5]Department of Communication Technology and Networks, Universiti Putra Malaysia (UPM), Serdang 43400, Malaysia
[6]Department of Computer Science and Information Technology, Hazara University, Mansehra 21300, Pakistan

Corresponding authors: Yerik Mardenov (emardenov@gmail.com) and Zulfiqar Ahmad (zulfiqarahmad@hu.edu.pk)

This work was supported by the Science Committee of the Ministry of Education and Science of Kazakhstan under Grant AP19680345.

ABSTRACT Smart grids are transforming the generation, distribution, and consumption of power, marking a revolutionary step forward for contemporary energy systems. Communication in smart grid environments is majorly performed through Wireless Sensor Networks (WSNs). The WSNs enable real-time monitoring and management inside smart grids. However, the integration of digital technologies and automation in smart grids introduces cybersecurity challenges, including unauthorized access, data breaches, and denial of service attacks. To address these difficulties and maintain the reliability of smart grid infrastructure, this study proposes a comprehensive architecture for strengthening cybersecurity within WSNs operating in smart grid environments. By integrating traffic analysis, node categorization, and machine learning algorithms, the framework intends to effectively detect and prevent cyber threats. Extensive evaluation reveals that traffic analysis using the Random Forest model successfully predicts traffic load within WSNs, achieving a mean squared error (MSE) of 2.772350, a root mean squared error (RMSE) of 1.665038, a mean absolute error (MAE) of 1.099080, and a coefficient of determination (R^2) of 0.717982. In intrusion detection, the Random Forest model outperforms Decision Trees and Logistic Regression, with higher precision (0.99), recall (0.99), and F1 scores (0.98) across various attack types. Specifically, Random Forest achieves perfect recall (1.00) in identifying Flooding attacks, underscoring its capability to detect all instances of such intrusions. Leveraging the insights gathered from the WSNBFSF dataset, this study gives significant findings into proactive cybersecurity tactics, stressing the necessity of securing key infrastructure for the reliable and secure distribution of power to consumers.

INDEX TERMS Smart grids, WSNs, cybersecurity, intrusion detection and prevention, machine learning, traffic analysis.

The associate editor coordinating the review of this manuscript and approving it for publication was Wanqing Zhao.

I. INTRODUCTION

Smart grid is the latest technology breakthrough in power grids, where digital communication, sensors, and automation are harnessed to improve the efficiency of the operation of the whole system [1], [2]. Conventional grids, on the contrary, serve as unidirectional flow routes of electricity; smart grids create bidirectional flow among the different components, such as power plants, substations, meters, and consumer appliances [3], [4]. Through the fusion of real-time data analytics and management systems, smart grid technology improves the generation, distribution and consumption of energy and makes the systems more reliable, resilient and sustainable. It offers consumers the chance to make better decisions when it comes to energy consumption, it eases the integration of renewable electricity sources, and it supports the move to electric vehicles. Smart grid technology is an advanced form of technology that makes grid security faster by giving real-time detection and response capabilities to disruptions ranging from cyberattacks to natural disasters. It guarantees that the power supply to homes, businesses and sensitive infrastructures is on around the clock [3], [5].

WSNs constitute a pivotal component within smart grids. WSNs serve as the infrastructure backbone for real-time monitoring, control, and optimization [1], [4]. These networks are composed of multiple sensors that are connected and placed on various grid infrastructure items, like power plants, substations, distribution lines, and consumer premises. The sensors are used to record data not only from voltage levels, current flows, temperature, humidity and environmental conditions but also from other sources [1], [6]. Smart grid WSNs are designed to achieve the aim of remote monitoring and grid asset management. Through effective utilization sensors, utilities will be able to have comprehensive information about the performance and health of each grid component which will, in turn, enable them to easily identify a potential issue and get it resolved to enhance reliability and minimize the downtime. WSNs are a tool to improve grid efficiency by providing information regarding energy consumption patterns and participation in load balancing, demand response and energy efficiency programs [7]. WSNs can be credited for keeping the grid running in a steady and secure manner by closely monitoring the infrastructure looking for unusual behavior or security vulnerabilities. This will therefore allow the identification of cyberattacks or physical intrusions as well as the repair of broken equipment in a timely manner [8].

Cyber-attacks pose significant threats to the security and integrity of WSNs within smart grids, necessitating robust cybersecurity intrusion detection and prevention mechanisms [9]. These will include a strong cybersecurity measure, for example, the 24/7 monitoring of both network traffic and behavioral patterns of occurrence, to be able to pick out any form of suspicious activities or anomalies hinting at potential cyber-attacks. Being part of the smart grid operational infrastructure, WSNs are very vulnerable, and thus, they could expose the environment to such vulnerabilities as malicious attack or tampering like malware, denial of service, and unauthorized access. IDS includes sensor data and network analysis in the identification of unauthorized access, data manipulation, or abnormal patterns of communication [10], [11], [12], [13], [14]. This needs to be complemented by the use of proactive measures such as access control, encryption protocols, and authentication mechanisms in order to avoid the occurrence of cyber-attacks. Machine learning (ML) and Artificial Intelligence (AI) methods are a supplement to cybersecurity abilities that bring prediction of risks and automate responses toward the ever-changing threat. ML models based on historical data and known patterns of attacks help strengthen the cybersecurity of WSN within smart grids from potential cyber-attacks. ML models are adaptive and adjust to real-time feedback, with the advantage of having a proactive threat mitigation capability and dynamically changing security protocols [15]. The machine inductive learning not only helps in improving the WSN cybersecurity framework but also includes an enhanced ability for utility detection and preventing cyber threats [8], [16], [17], [18].

Traffic analysis plays a key role in fortifying the cybersecurity of WSNs within smart grids [19], [20]. It checks the pattern of network traffic over time to find if there are any irregularities, which show potential cyber threats. In this vein, utilities may monitor data packet flows searching for anomalies such as unexpected surges in traffic volume or aberrant transmissions of data that may point to a potential malicious activity. Traffic analysis can be a sentinel toward insider threat if it detects any unauthorized attempt for access or any other abnormal behavior within the boundary of the organization. In the context of a cybersecurity incident, traffic analysis significantly supports in incident reconstruction and devising incident response strategies that would prove to be effective [6], [19], [21]. The flowing analysis of the traffic influences the WSN resilience for further improvement and enhancement in the continuous and secured delivery of electricity to the consumers by being proactive in cybersecurity.

The proposed study aims to address cybersecurity challenges faced by WSNs within smart grids. Our goal is to implement and integrate such cutting-edge methods as traffic analysis, node classification and ML. To achieve that, the analysis of traffic and the identification of threats as well as the classification of nodes based on sensitivity is used in the study to increase the precision and efficiency of cybersecurity. ML provides such a framework with the power to adjust and grow to face the ever-changing threats. It gives way to forward-looking remediation measures. The aimed study also aims building up sustainable and customized cybersecurity technology that will specifically address the WSN challenges in smart grid settings. It is responsible for maintaining the security, authenticity, and faultlessness of the network for the sake of the community.

The proposed research study has the following main contributions:

- The study analyzes traffic patterns in WSNs of smart grid technology, categorizing nodes based on their sensitivity, importance, and workload.
- The study integrates ML technology to predict cybersecurity attacks within WSNs of smart grids, specifically

utilizing models including Random Forest and Logistic Regression.
- Based on traffic analysis and predictions made by ML, the study establishes an intrusion detection and prevention (IDP) system for WSNs in smart grids, demonstrating significant improvements in prediction accuracy and threat detection.

We organized the remaining part of the research article as follows: Section II covers the related work. Section III implements the system design and model in accordance with the proposed framework. Section IV presents a performance evaluation, and finally, Section V concludes the article with several future directions.

II. RELATED WORK

We review the related work in context with smart grid technologies, traffic analysis in WSNs, ML-integration of cybersecurity intrusion detection and prevention.

Smart grids (SG) and distributed energy generation through Internet of Things (IoT), as well as new technologies like Internet of Energy (IoE) and intelligent systems are garnering attention as a means of achieving low-carbon sustainable energy development [22]. The web facilitates the interoperability of intelligent energy systems, which boosts network efficiency and intelligent management while enabling automatic usage optimization. IoE is a fascinating subject that is closely related to SG, electrical mobility, IoT, communication systems, and energy efficiency. It helps to attain zero-carbon technology and green settings. In addition, the growing prevalence of processors used for mining virtual currency in residences and compact warehouses are some additional elements that need to be considered when analyzing electric energy usage and greenhouse gas emissions these days. Nevertheless, studies examining how the Internet might be used to assess energy misallocation and its potential impact on CO_2 emissions are frequently overlooked. The authors of this study [22] provide a thorough analysis of the development of SG in relation to the use of IoE systems and the important elements of IoE for decarbonization. Moreover, computational models that incorporate simulations are offered to assess the contribution of IoE to CO_2 emission reduction.

Load forecasting (LF) is becoming more popular as smart grids (SGs) become more successful [23]. The LF method provides assistance in terms of planning upgrades by SGs and decisions regarding the operation of power. This also contributes toward improving the ability to supply affordable and dependable power services. The accuracy of demand forecasting, especially with artificial intelligence (AI), is far better through the use of ML and deep learning (DL) techniques. In the selection of an appropriate and accurate LF approach to be used in SGs, it will require being chosen critically, examining a number of LF methods. The state-of-the-art forecasting methods are critically analyzed, and an evaluative comparison of time series-based, AI-based, classical, and clustering-based methods is carried out with reference to their performance and results in [23]. The work presented in [23] also identifies the optimum LF technique for a set of SG applications. The results of these methods concluded that the best among them for forecast performances were AI-based LF techniques using ML and Neural Network (NN) models. On average, the AI-based ML and NN models have higher values for RMS and Mean Absolute Percentage Error (MAPE) of the daily aggregated forecasts.

Building smart cities requires the timely collection and analysis of city traffic flow data, as urbanization is taking place at a breakneck speed. The study in [24] suggests a novel approach to using WSNs in collecting data on the flow of traffic, which records the position of the vehicle and its speed in addition to the flow of traffic. Based on this, the traffic flow data, a technique for analyzing data based on incremental noise addition, provided a chaotic identification criterion [24]. In this way, noise in the signal is introduced in the form of different intensities sequentially, followed by quantification of the signal complexity through the use of the delayed mutual information. The trend of complexity change of the mixed signal can enable a person to distinguish the properties of the signal. In order to prove this, numerical testing was carried out, and its result shows that periodic data, random data, and chaotic data have different complexity patterns with reference to the added noise increment. The implementation of this technology opens new avenues for the collection and analysis of traffic flow data [24].

Network intrusion prevention for WSNs is an important area of research because of the rapid advancement of WSNs due to its popularity and the increased security issues brought about by their flexibility and ease of deployment [25]. One of the common types of the network attack is the denial of service (DoS) attack, and whose main mission is to bring down the target network. Such attacks would be devastating to WSNs devices because of the limited resources. The work proposed in [25] applies a detection technique for the anomaly of DoS traffic in WSNs through principal component analysis (PCA) and a deep convolution neural network (DCNN) since WSN is highly sensitive to security attacks and has a limitation in storage capacity at its nodes. The proposed model can be in a position to detect abnormal network traffic in WSNs with devices of limited storage capacity since it is lightweight compared to the deep learning structures that are normally used and has more efficiency in feature extraction capability. Validation of classification results and model effectiveness is being done by Receiver Operating Characteristic (ROC) curves, various classification metrics, and confusion matrices [25].

The latest advancements in system security and defense mechanisms have been made possible by the extensive application of ML-based intrusion detection system (IDS) techniques [26]. Shared networks and their attendant vulnerabilities have led to a notable rise in security concerns in smart grid computer settings. Even though the smart grid environment is experiencing significant security challenges because of its particular environmental vulnerabilities, ML-based IDS research in a smart grid is still relatively unexplored when compared to other network settings. The authors in [26] conducted a thorough survey on machine learning-based intrusion detection systems (ML-based IDS) in smart grids based on the following important factors: (1) the use of ML-based IDSs to address security vulnerabilities in transmission and distribution side power components of a smart power grid; (2) the creation of datasets and their application in the smart grid; (3) a variety of ML-based IDSs utilized in the smart grid environment by the surveyed papers; (4) metrics, complexity analysis, and evaluation testbeds of the IDSs applied in the smart grid; and (5) lessons learned, insights, and future research directions [26].

In a smart grid system, intrusion detection is important to provide a secure service and informing the system administrator of adversary threats via a high-priority alert message [27]. To precisely categorize different attacks on smart power grid systems, the study in [27] suggests an intelligent intrusion detection scheme. The suggested plan made use of feature selection based on binary grey wolf optimization. The non-linear, overlapping, and complex electrical grid features were extracted from the publicly accessible Mississippi State University and Oak Ridge National Laboratory (MSU-ORNL) dataset, and the ensemble classification approach was optimized for learning them. The suggested method's promising performance is shown by the experimental results for two class and three class problems utilizing a 10-fold cross-validation setup and a chosen feature subset. The robustness of the suggested strategy was justified by the noticeably better performance when compared to the current benchmark methodologies [27].

Literature review highlights important areas regarding technologies implemented within the smart grids and corresponding cybersecurity protocols. Starting with smart grid development, IoE exploration within the frameworks of smart grid development puts special emphasis on computational models and simulations for the evaluation of IoE's ability to reduce CO_2 emission [22]. Moreover, the role played by load forecasting has also been much appreciated in smart grid frameworks. This area specifically focuses more on the utilization of modern techniques empowered by AI, such as ML and neural networks, which could lead to more precise forecasting of demand. Research studies on the data for traffic flow also consider the methodology with a new approach based on WSNs. This model is based on the identification criteria for both noise and chaotic identification so as to find the complexity in traffic flow. Investigations would delve deep into the world of IDS customized for WSNs. PCA and DCNN are proposed methodologies that would serve, to say the least, in proficiently identifying anomalies [27]. The authors in [27] reviews the current trends and practices of the ML-based IDS applications used in a smart grid environment. It summarizes the data set preparation, diversity of ML-driven IDS architectures, and used evaluation metrics along with the desired future research directions. Authors finally propose an intelligent intrusion detection framework for smart power grids through feature selection techniques and ensemble classification methodologies to discern and classify varied forms of effective attacks.

III. SYSTEM DESIGN AND MODEL

This study proposes a traffic analysis and node categorization-aware ML-integrated framework for cybersecurity intrusion detection and prevention of WSNs in smart grids as shown in Figure 1. The proposed integrated smart grid environment framework has four basic components: smart grid environment, WSN in a smart grid, traffic analysis, and categorization of nodes. This work further includes a ML-integrated intrusion detection and prevention system for WSN in smart grids. The framework aims to contribute to security improvements in the Wireless Sensor Network (WSN) of a smart grid environment by integrating several integral parts. Basic infrastructure in the smart grid: It is to comprise utilities, substations, and distribution networks along with the power plants, all focusing toward the collection and analytics of data in real-time so that the best grid operation and reliability can be carried out. The data collectors include all those that help in aggregating the data by gathering information from sources such as smart meters and sensors to transmit them to the utility control centers for analysis. WSNs are such important components that are applied for monitoring and control in the sections of smart grid infrastructure, allowing remote sensing and data transmission, including that of voltage, current, temperature, and status of equipment. This implies facilitating automated metering, integration with distributed energy resources, and optimization of generation processes. They also facilitate monitoring in different substations and lines that go a long way in helping manage the grid effectively. An important component in the analysis, behavior of data transmission of WSNs, is classified based on their relevance as well as the node type and the traffic pattern within it. The approaches above explain the analysis of the volume, frequency, and kind of messages that take place over the network for any anomalies that may mean a potential cyber threat. Identification of sensitive nodes such as substations and control centers that are critical to the operation of the grid and therefore potential targets of attacks. Moreover, the distinction of high traffic nodes, e.g., collector nodes, from the normal nodes, e.g., sensor nodes, aids in better understanding the behavior of the network. This architecture includes the core intrusion detection and prevention system (IDPS) that employs machine-learning algorithms for the speedy discovery and real-time mitigation of any security breach. The IDPS will be in a position to detect any anomaly with a show sign of intrusion attempts by traffic dynamics analysis and giving a prediction of cyber threats, then start the necessary automatic responses or mitigation measures in order to save the smart grid infrastructure from losing its integrity. The system always learns and keeps adapting constantly to all changing threats with this continuous learning. This guarantees that it

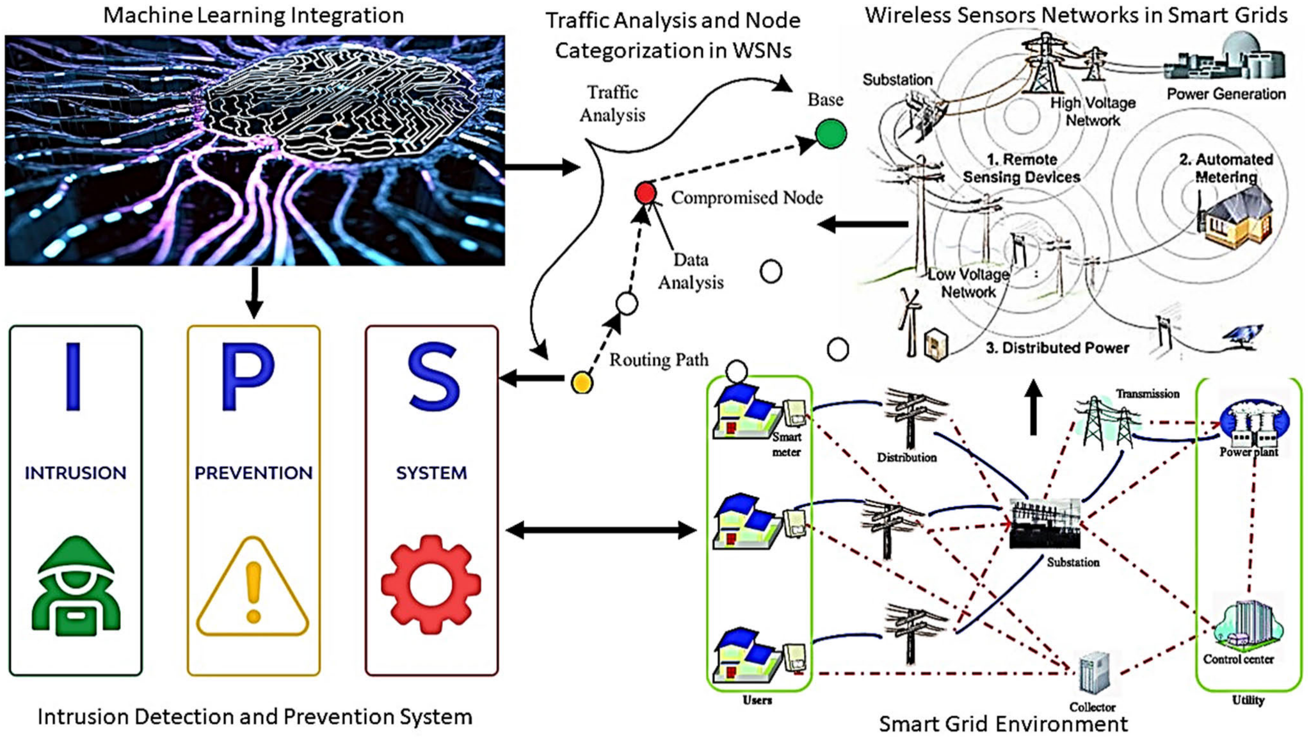


FIGURE 1. Traffic analysis and node categorization-aware machine learning-integrated framework for cybersecurity intrusion detection and prevention of WSNs in smart grids.

maintains effectiveness and it is easy to adapt to the other existing security infrastructures for defense-in-depth strategy.

A. SMART GRID ENVIRONMENTS

Smart Grid is an organized system aimed at coordinating diversified stakeholders and technologies that are pulling in the same direction to boost the effectiveness, reliability, and sustainability in electricity generation, distribution, and consumption. In fact, users are empowered within these smart grid frameworks by the fact that households, businesses, and industries—most of electricity use—are made up of users, thereby increasing the resources and insight-based numbers that would help empower these users in regulating their energy usage. Such empowerment might see them access real-time information on energy consumption, flexible pricing mechanisms, and programs that promote demand-response power usage-shifting by consumers on critical occasions of its necessity the most, such as during peak hours. The collector plays an essentially neural role in information consolidation from diverse sources in the smart grid network. It collects information from smart meters, sensors dispersed all over the grid infrastructure, and other monitoring devices. The collector further processes the data and channels it to the control center for analysis and making informed decisions about the utility.

The utility, which is most likely a power company or an electricity provider, develops, maintains, and operates the smart grid infrastructure from electricity generation down to its transmission and distribution. Smart Grid Environment allows utilities to deploy avant-garde technologies and data analytics in refining grid operations, improving reliability, and giving way to consumers' ever-changing needs. Smart meters are the latest in measurement tools installed at the consumer end, which can check the use of electricity around the clock. Unlike the traditional meters, smart meters have in-built communication abilities through which they can send consumption data back to the utility remotely. The capability, therefore, empowers the utilities to monitor usage trends, be it for management purposes or to implement demand-response initiatives, leading to accurate billing of customers for the energy actually consumed by the customer. The substation is a critical point in smart grid infrastructure because it works as an intermediate node between the voltage of the high-voltage transmission network and the distribution network. The substations manage voltage levels, flow of power, and stability control of the grid. In the smart grid field, the substations are installed with sensors and monitoring equipment to determine any deviations, improve the performance of the devices, and hence support the predictions and predictive maintenance steps.

The distribution network includes all the physical network that electricity follows from substations to the end users. This includes power lines, transformers, switches, and associated infrastructure. In smart grid, distribution assets will have equipped with sensors and communication which will be able to monitor the asset's status and detect faults, control and automate grid operations from remote locations. Conventional power plants will provide the central facility for the generation of electricity and, at the same time, accept the energies from all the sources including coal, renewables, hydroelectric power, or nuclear. Smarter power plants are seamlessly integrated into the grid using cutting-edge control systems and communication networks in the smart grid ecosystem. It will allow utilities to achieve the right balance between the dispatching schedules of the conventional power plants and electricity grid flexibility around the fluctuating demands of the grid, including the integration of renewable energy sources. The basis of the smart grid's grid communication technologies is the use of wireless communication technologies to enable data transmission and coordination. The devices provide a platform for communication to all elements within the grid: smart meters, sensors, collectors, substations, and utility control centers It paves the way for the monitoring, control and optimization of grid assets in the real time tracking and management, which is done through digital means that maximizes grid reliability, efficiency and responsiveness.

B. WSNs IN SMART GRIDS

WSNs are among the enabling technologies that improve the operation and competitiveness of smart grids, the modern systems which provide electricity in a smart way. These grids are based on the modern technology that balances and regulates the power generation, transmission and the consumption of electricity. The WSNs are of paramount importance when they enable the placement of remote sensing devices in the vicinity of the grid infrastructure. These sensors are so powerful that they can monitor multiple parameters in real-time; like the voltage, the current, the temperature, the humidity, and the status of the equipment. For instance, temperature sensors placed on the power lines can detect overheating, while vibration sensors can be installed on the transformers to detect mechanical faults. The data taken from these sensors are then wirelessly sent to the central control system where it is examined.

WSNs serve as a communication tool in automated metering which is the main application of smart grids. Smart meters (the ones which have been equipped with wireless communication capabilities) are installed at customer premises. After that, the electricity consumption data is sent in real-time. This does away with the device reading and allows the utilities to have a better and clearer monitoring of the energy consumed. Moreover, automated metering helps in implementing time-of-use pricing and demand-response programs, permitting consumers to be able to reduce their power consumption and adjust it to the signals informing them about pricing. Along with WSNs, the integration of the distributed power generation resources (solar panels, wind turbines, and microgrids) into the smart grid is one of the primary roles. Installing sensors in the DERs (distributed energy resources) is one of the key functions of these sensors. These sensors monitor in real-time parameters such as power, voltage, frequency, and other parameters, enabling communication between DERs and grid operators. This guarantees a better handling of distributed generation and grid stability; or in other words, it is about the fact that renewable energy sources can be smoothly connected to the grid while maintaining stability that is required for the grid to operate properly.

C. TRAFFIC ANALYSIS AND NODE CATEGORIZATION

In the context of smart grids, a deep dive into the complex analysis of data transmission dynamics in WSNs is the main objective of traffic analysis. The evaluation process involves analyzing the amount, frequency, and nature of communication flowing between nodes in the network. The extent of the traffic pattern changes depends on the different cases, like mesh topology, communication protocols, and conditions of the network operation. In the node categorization scheme several nodes are classified as sensitive due to the fact that they have critical role in grid work or vulnerabilities which can lead to great impacts on grid operations or security if they are compromised. These sensitive nodes encompass various components:

- Substation Nodes: While situated in substations, these nodes in fact play the most important roles in the grid of monitoring and coordinating the operations. A failure of any of these critical nodes could cause major grid fluctuations, which may lead to a cascading effect and eventually to a system-wide blackout.
- Control Center Nodes: These nodes are the most important data gathering point in utility control centers, the decision-making and analysis processes of which are the most effective. Through the agreement, the states may lose the grid security and power to control.
- Smart Meter Nodes: These nodes located at consumer premises are the main hardware components that provide entry points to collect and transmit consumption data. Unauthorized access to smart meters may occur, and the result could be privacy breach or tampering with consumption data.
- Communication Gateway Nodes: These nodes that spread across the grid allow easy communication between different components of the grid which may include collectors or concentrators. Broken nodes would be the reason for the communication and control mechanism failure that are vitally important for the grid.

In addition to that, there are certain nodes which are highly loaded due to their role as communication hubs or data aggregation points for all the nodes in the network notable examples encompass:

- Collector Nodes: These nodes, faced with the responsibility of amalgamating data from a variety of sources within the grid, including smart meters and sensors, are weighing down the grid with massive amounts of data traffic. Utilities can monitor these systems and pass the information to local control rooms for detailed research.
- Substation Nodes: The nodes will also be placed at the center of the grid infrastructure and have the role of

consolidating and forwarding data from sensors that are observing equipment health and grid parameters. Substations are nodes of the grid, always under monitoring, and therefore, they are exposed to a surge of traffic.
- Control Center Nodes: The role of the data collector nodes in data stream pipelines is to receive data from the collector nodes as well as other sources. They are the key data analysis tools that help the grid operators make informed decisions. Therefore, they mainly operate on the heavy traffic caused by data processing and control functions.

The WSNs nodes with low traffic, which is represented by sensor networks deployed across the grid infrastructure to capture parameters like voltage, current, temperature and humidity. These nodes include sensor nodes and actuator nodes. Sensor nodes act as an integral part of acquiring information related to the power grid and the environment. Therefore, their data transmission happens periodically, being set to a specific sampling interval or a given trigger event. Actuator nodes in the control grid are responsible for controlling the grid equipment and executing commands based on the received instructions. They may then only receive commands intermittently, while they also receive data from sensors that are far more frequently transmitted.

D. MACHINE LEARNING-INTEGRATED INTRUSION DETECTION AND PREVENTION SYSTEM

In the realm of cybersecurity, a ML-integrated intrusion detection and prevention system (IDPS) specifically designed for WSNs operating within smart grids presents an innovative strategy to fortify critical infrastructure against potential cyber threats. By harnessing the power of ML algorithms, which capitalize on traffic analysis and attack prediction functionalities, this IDPS is adept at swiftly identifying and neutralizing security breaches in real-time.

- Traffic Analysis: ML algorithms undertake a comprehensive analysis of the traffic dynamics within the WSN, establishing benchmarks for normal operational behavior. By scrutinizing key parameters such as packet rates, data volume, communication frequencies, and alterations in network topology, the system can pinpoint deviations indicative of suspicious activities or looming attacks. Notable examples include sudden surges in traffic or irregular communication patterns, which may foreshadow potential Distributed Denial of Service (DDoS) assaults or clandestine data exfiltration endeavors.
- Attack Prediction: Drawing insights from historical data, ML models have the capacity to forecast potential cyber threats by discerning underlying patterns and trends. These predictive models excel in identifying subtle precursors to specific attack modalities, encompassing reconnaissance efforts, malware propagation strategies, or command and control communications. Through the correlation of diverse network parameters and behaviors, the system can preemptively anticipate and counter emerging threats, forestalling their escalation into full-blown assaults.
- Anomaly Detection: ML algorithms exhibit prowess in anomaly detection, enabling the IDPS to flag deviations from standard network behavior indicative of intrusion attempts or malicious actions. These anomalies may manifest in diverse forms, such as anomalous data flows, unauthorized access endeavors, or unanticipated alterations in sensor readings. By continuously assimilating new information and adapting to evolving threat landscapes, the system remains poised to detect nascent attack vectors that conventional rule-based methodologies might overlook.
- Response and Mitigation: Upon the detection of suspicious activities or looming threats, the IDPS is primed to initiate automated responses or mitigation measures to quash the menace and fortify the integrity of the smart grid infrastructure. Depending on the gravity and nature of the threat, these countermeasures may encompass the isolation of compromised nodes, rerouting of traffic, blocking of malicious packets, or the activation of alerts to security personnel for further investigation and intervention.
- Adaptive Learning: The IDPS capitalizes on adaptive learning methodologies to fine-tune its detection capabilities over time. By continually assimilating fresh data inputs and feedback from security analysts, the system iteratively refines its models and algorithms to acclimate to emergent threats and shifting network dynamics. This adaptive learning paradigm ensures that the IDPS remains efficacious in identifying and mitigating both known and novel security risks in the dynamic milieu of smart grids.
- Integration with Existing Security Infrastructure: The ML integrated IDPS seamlessly integrates with extant security frameworks within smart grids, including firewalls, intrusion prevention systems, and Security Information and Event Management (SIEM) platforms. This interoperability fosters a holistic defense-in-depth strategy, bolstering the overall security posture of the smart grid ecosystem and engendering resilience against multifarious cyber threats.

E. ALGORITHM FOR TRAFFIC ANALYSIS AND NODE CATEGORIZATION-AWARE MACHINE LEARNING-INTEGRATED FRAMEWORK

The proposed procedural framework is depicted in Algorithm 1. The algorithm takes an input like data from WSNs and pre-trained models of ML. Output generated is a detailed analysis that covers traffic behaviors and node classifications, ML methods incorporated, and measures for detecting and preventing intrusion. First, it defines the boundaries for sending and receiving data, setting the ML algorithms ready to capture any unwanted intrusion. It will then probe further into how the data is being transmitted within the WSNs, finding more details into how much the

Algorithm 1 Traffic Analysis and Node Categorization-Aware Machine Learning-Integrated Framework

```
1.  Begin
2.  Input:    WSNs data, ML models,
3.  Output:   Traffic analysis, node categorization, ML-integration, IDP
4.  Procedure: Traffic analysis and node categorization-aware ML-integrated framework
5.  Initialization:
            • WSNs communication in smart grids
            • Parameters for data transmission and traffic analysis
            • ML algorithms for intrusion detection
6.  Traffic Analysis:
    for each node (i) in the WSNs
      Calculation of traffic volume $V_i$based on the number of packed exchanged:
                          $V_i = \sum_{j=1}^{n} P_{ij}$
      Calculation of the average packet rate $R_i$ for node i:
                          $R_i = \frac{V_i}{T}$
      Determination of the communication frequency $F_i$ by analyzing message exchange patterns
      Identification of anomalies in traffic patterns using ML algorithms
    end for
7.  Node categorization:
      Categorization of nodes based on their criticality and traffic patterns:
      Sensitive Nodes $S_i$:
      if node i is critical in smart grids WSNs or node i experiences maximal traffic,
                          $S_i = True$
      else
                          $S_i = False$
      end if else
      Normal Nodes $N_i$:
      if node i is not critical in smart grids WSNs and node i experiences minimal traffic,
                          $N_i = True$
      else
                          $N_i = False$
      end if else
8.  ML integration:
      Training of ML models
              Input features: traffic volume, packet rate, communication frequency, node category
              Output labels: normal behavior or anomaly
      Validation and optimization of the models using cross-validation techniques
      Integration of the trained models into the intrusion detection and prevention system
9.  Intrusion detection and prevention system:
      for each data packet received
              Feature extraction: traffic volume, packet rate, communication frequency, sender node category
              Prediction for the likelihood of intrusion using the ML models
        if the predicted probability exceeds a threshold
          Initiation of response actions
          Isolation of compromised nodes
          Redirection of traffic to mitigate the impact
          Notifying security personnel
        end if
      end for
10. end
```

data is being transmitted, how often, and how the devices are interacting. ML enables them to find any strange patterns in the data, which would enable them to catch potential security troubles. They then classify the devices into groups on the basis of how important they are and how those devices communicate. Therefore, it differentiates between the critical and the less critical devices to smart grid operation. Using the ML models, the algorithm has the capacity to forecast the probability of incidence of a security threat by analyzing the information collected. When it senses that a threat is likely to happen, it takes actions such as isolating the compromised devices or modifying the way data flows in order to limit the risk. These usually refer to a very detailed framework that uses highly analytic techniques and ML in trials to bring out a comprehensive approach for the proactive security of WSNs in making them very secure for smart grid environments.

IV. PERFORMANCE EVALUATION

We perform simulations and evaluate the performance of proposed framework with respect to the cybersecurity intrusion detection and classification.

A. EVALUATION METRICS

We evaluated the performance of the models implemented in the proposed framework using MSE, RMSE, MAE, and R^2 for traffic analysis, and accuracy, precision, sensitivity (recall), F1 score, specificity, and precision-recall curve for intrusion detection [28]. We calculated accuracy, precision, recall, and F1 score based on the following terms:

- True Positives (TP): The number of tuples that are really found to be intrusive at the end of the process.
- True Negatives (TN): The number of valid tuples that are found at the end of the detection process.
- False Positives (FP): The number of safe tuples that, at the conclusion of the detection process, are identified as intrusions.
- False Negatives (FN): The quantity of dangerous tuples that, at the conclusion of the detection process, are found normally.

When assessing the effectiveness of classification models, accuracy is a commonly used evaluation parameter. It evaluates the overall accuracy of the model predictions by figuring out the proportion of correctly predicted cases among all the instances in the dataset [29]. It is calculated with the help of equation 1.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision is a way to measure how well a classification model works. It finds how good the model is at making positive predictions by counting the number of true positives out of all positive predictions, or true positives plus fake positives [29]. It is calculated with the help of equation 2.

$$P = \frac{TP}{TP + FP} \tag{2}$$

Sensitivity is a way to measure how well a classification model works. This number is also known as the recall or true positive rate. The sensitivity of the model measures how well it can find every single positive case in the dataset [29]. It is calculated with the help of equation 3.

$$R = \frac{TP}{TP + FN} \tag{3}$$

The F1 score demonstrates how well classification models perform when balancing precision and recall. It is particularly useful when there is an imbalance between the number of true positives and false negatives in the dataset [29]. The F1 score is calculated using equation 4.

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{4}$$

A precision-recall curve has also been used during evaluation. It is a graphical representation of the trade-off between precision and recall for different classification thresholds. The precision-recall curve is created by varying the classification threshold of the model and determining the precision and recall at each threshold. A higher area under the precision-recall curve (AUC-PR) indicates better performance for the model.

B. DATASET

In order to evaluate the working of proposed framework, WSNBFSFDataset [30], a publicly available dataset on a Kaggle website has been used. The WSNBFSF dataset is a database that contains all the significant data that is helpful in the study of breaches in the WSNs with emphasis on the black holes, flooding, and selective forwarding attacks which are a subset of the DDoS attacks and are the most common type of attacks among network security researchers. This dataset is deliberately designed for research and analysis purposes to facilitate academicians and scholars in the study of the attack that exist within wireless sensor network (WSN). The dataset is comprised of 16 different features which are believed to be the attributes or the qualities that can be used to describe the model and its performance. This dataset has a significant number of rows, 312,106, which is a huge volume of data. Upon undergoing requisite preprocessing measures, such as data cleansing and formatting, the dataset is meticulously organized to delineate four distinct categories of network traffic: Blackhole attack traffic, Flooding attack traffic, Selective Forwarding attack traffic, and Normal traffic, which are used as a base measurement of the usual network activities. Black holes pose a significant threat to WSNs and influence network performance. They contribute to network congestion through packet drops, impeding proper communication and negatively impacting the functionality of WSNs. Their presence near WSN nodes strains energy resources as nodes expend more energy attempting to retransmit lost packets or seek alternative routes, thereby shortening battery lifespan and affecting network longevity. Beyond operational disruptions, black holes represent a major security threat by serving as gateways to other attacks such as selective forwarding or sinkhole attacks, aiming to disrupt communication, steal data, or gain unauthorized access to critical information for smart grid operations. Flooding attacks inundate WSNs with unwanted data packets, which is very risky. They clog the network by slowing down the flow of data and communication and exhausting resources such as bandwidth and processing power, thus affecting network performance. These attacks consume the node's energy, reducing the operational life and sustainability of the network. Thus, making it impossible to use networks, flooding attacks lead to denial-of-service situations and disrupt important operations. Selective forwarding attacks in WSNs are aimed at disrupting the communication between nodes, and as a result, data loss, disruption of communication, and compromise of security are experienced. Attackers selectively drop or forward packets, which leads to disruption and possibly depleting the energy sources as the nodes try to recover. This manipulation can result in poor network performance and data corruption, which is very dangerous to the WSN in terms of reliability and efficiency. Figures 2, 3 and 4 shows packet size, rest energy and source IP port distributions of the dataset respectively. Figure 5 shows traffic distributions across nodes. Figure 6 shows the top 20 nodes by total number of packets.

We conducted data preprocessing measures to ensure the dataset's quality and suitability for analysis. These techniques include data cleansing and formatting. We also employed feature engineering to extract relevant information and enhance the dataset's predictive capabilities. The dataset was carefully organized to delineate four distinct categories of network traffic: Blackhole attack traffic, Flooding attack traffic, Selective Forwarding attack traffic, and Normal traffic. This organization facilitated precise checks and measurements of performance for different detection and mitigation techniques within WSNs. We utilized machine learning packages and libraries to facilitate the preprocessing tasks and ensure the dataset's readiness for analysis.

C. EXPERIMENTAL DESIGN

The experiments were performed by implementing four ML models i.e., Random Forest [31], Decision Tree [32], Gradient Boosting [33], [34], [35], and Linear Regression [36], [37], [38] for traffic analysis. We used the performance parameters of MSE, RMSE, MAE and R^2 for traffic analysis. We also performed the experiments by implementing two ML models i.e., Decision Tree [32] and Random Forest [31] for intrusion detection and prediction. We used the performance parameters of accuracy, precision, recall and F1 score for intrusion detection and prediction. The dataset has been divided into two parts: the training set and the test set. The training set comprised 80% of the total records in the dataset. It was used to train the proposed models. On the other hand, the test set comprised 20% of the total number of records. It was used to test and validate the proposed model. Cross-validation was performed through the "cross_val_score" function from scikit-learn. All experiments are implemented in Python on a GPU based environment with 2.11 GHz CPU and 16 GB of RAM. Predefined ML packages and libraries including Pandas, Numpy, Seaborn, Sklearn, LabelEncoder, OneHoTencoding and Matplotlib have been implemented.

D. RESULTS AND DISCUSSION

1) TRAFFIC ANALYSIS

We performed traffic analysis by evaluating MSE, RMSE, MAE and R^2 values for four ML models. The models are designed to forecast traffic load based on the following factors:

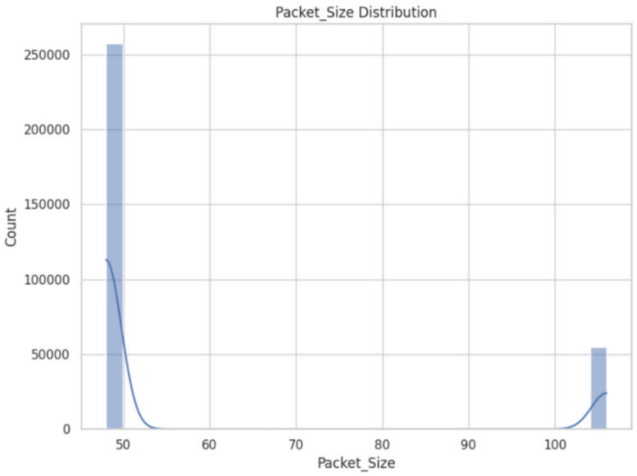


FIGURE 2. Packet size distribution.

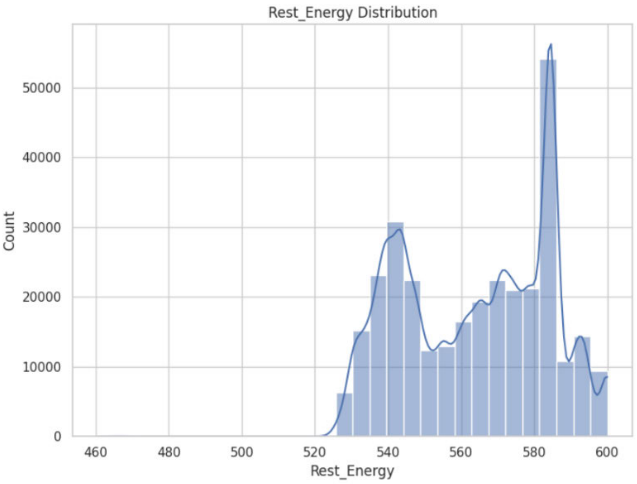


FIGURE 3. Rest energy distribution.

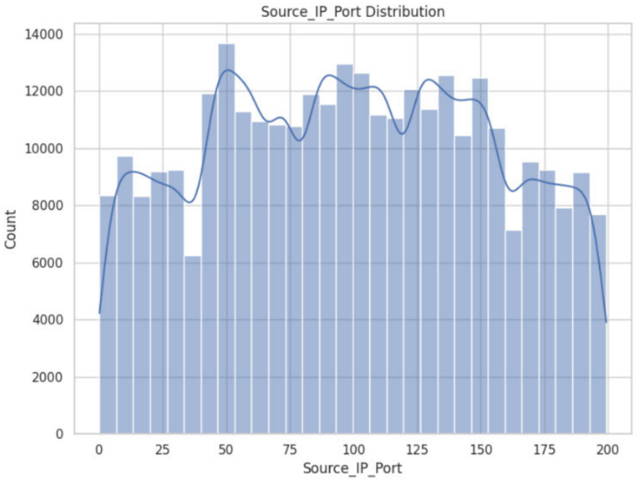


FIGURE 4. Source IP port distribution.

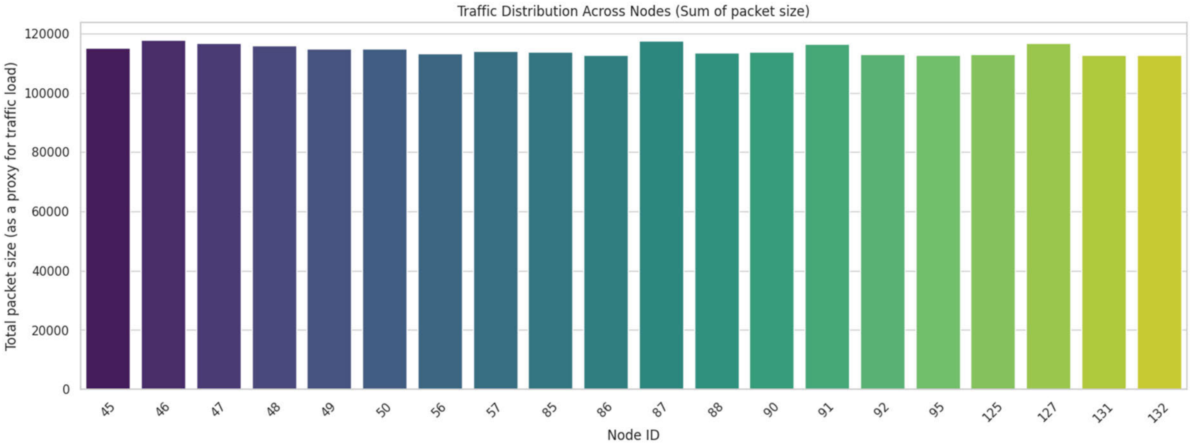


FIGURE 5. Traffic distributions across nodes.

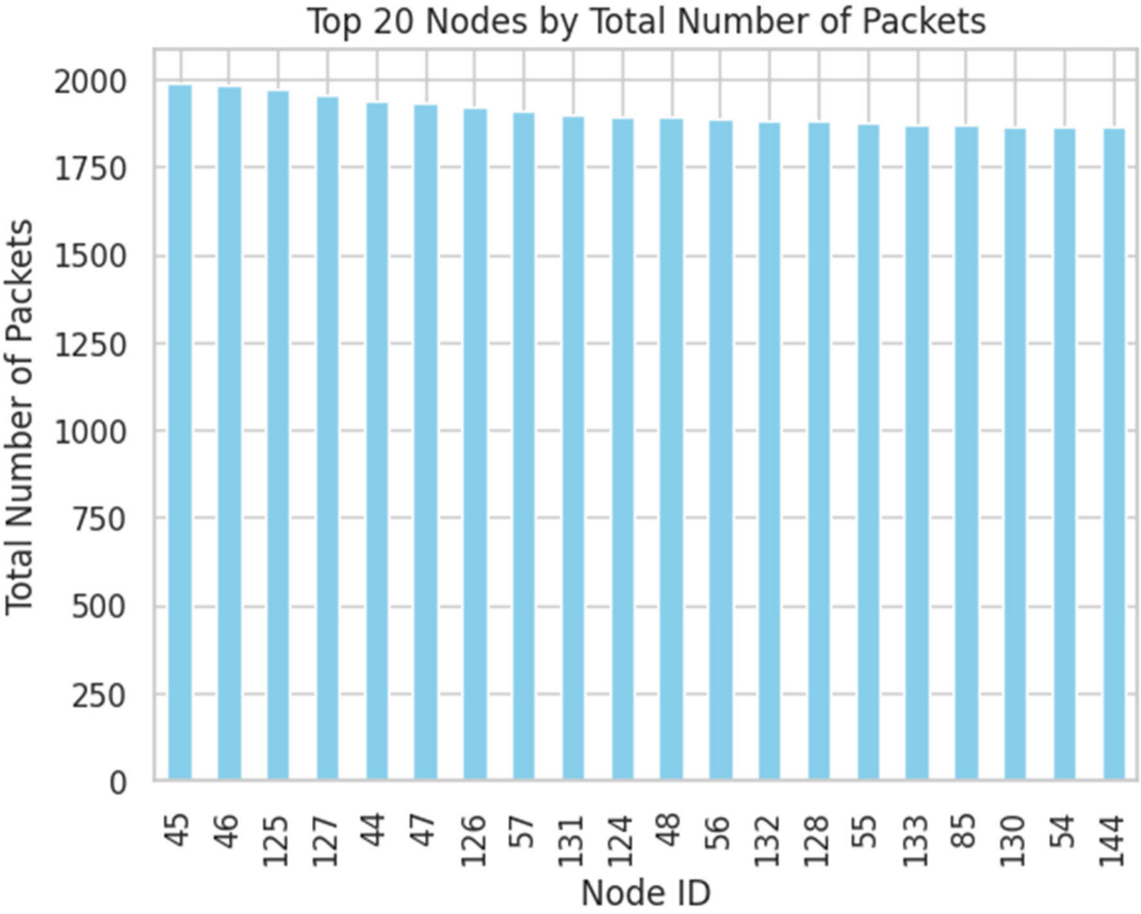


FIGURE 6. Top 20 nodes by total number of packets.

- **Packet Size:** This refers to the size of the packets being transmitted, directly influencing the overall traffic load. The highest packet size observed in the dataset is 117766 bytes, and it belongs to node ID 46 as shown in Figure 5. In smart grids, packet size determines the traffic flow by affecting congestion, energy consumption, and data transmission. Larger packet sizes, for instance, the observed 117766 bytes, may cause more congestion in the network since they need more bandwidth to transfer. This congestion can lead to a slow transfer of data and, generally, poor network performance. Also, larger packets may consume more energy during transmission; hence, the energy resources in the wireless sensor network (WSN) nodes will be depleted faster. Hence, it is vital to comprehend and control packet size to avoid congesting the traffic flow and to guarantee the proper functioning of smart grid systems.

- **Rest Energy:** This represents the residual energy levels of nodes, which can serve as an indirect indicator of their activity levels and, consequently, the traffic load they generate.
- **Source and Destination IP Ports:** These identifiers for the source and destination of network traffic, while not directly related to size, provide insights into traffic patterns and aid in modeling network behavior.

The models were trained using historical data to discern the relationship between the features and the average packet size per node. The developed code predicts traffic on nodes by estimating the average packet size for each node (Packet Size Average), which serves as a proxy for the volume of traffic handled by individual nodes. This approach is deemed reasonable as packet size is recognized as a significant factor influencing network load and traffic congestion. Data aggregation is conducted at the node level to summarize traffic conditions, thus establishing a clear target for prediction. Table 1 shows traffic analysis with MSE, RMSE, MAE and R^2. Random Forest emerged as the most effective model, demonstrating the lowest values across all evaluation metrics. It achieved a MSE of 2.772350, an RMSE of 1.665038, an MAE of 1.099080, and R^2 of 0.717982. On the other hand, Linear Regression yielded the least satisfactory results, showing the highest values across all metrics. Although Decision Tree displayed competitive outcomes in terms of RMSE and MAE, its performance lagged Random Forest concerning MSE and R^2. While Gradient Boosting showed better performance than Linear Regression, it still fell short compared to Random Forest and Decision Tree models across all evaluation criteria. These findings highlight the effectiveness of ensemble learning methods like Random Forest in achieving superior predictive accuracy compared to individual models. They underscore the limitations of linear regression in capturing the complexities inherent in the dataset.

TABLE 1. Traffic analysis with MSE, RMSE, MAE and R^2.

Parameters Models	MSE	RMSE	MAE	R^2
Random Forest	2.772350	1.665038	1.099080	0.717982
Decision Tree	3.659311	1.912932	0.972159	0.627755
Gradient Boosting	6.043927	2.458440	2.077709	0.385179
Linear Regression	9.555260	3.091158	2.661118	0.027988

Figure 7 shows the top 20 nodes with highest predicted traffic, where x-axis represents node ID, while y-axis represents predicted average traffic load. Node 102 stands out as having the highest predicted traffic, with a value of 62.285296. Nodes 109, 104, and 110 closely follow, each with predicted traffic values exceeding 62. This pattern suggests that these nodes are likely to bear the brunt of the network's traffic load according to the predictive model's analysis. Moving along the list, nodes 101, 90, and 105 also demonstrate notable predicted traffic values, indicating substantial anticipated activity within these nodes of the network. As we descend further down the list, the predicted traffic values gradually diminish. However, nodes such as 103, 111, and 98 still exhibit relatively high traffic predictions, albeit slightly lower than the top-ranked nodes. Nodes like 89, 97, and 91, among others, maintain this descending trend, showcasing varying levels of predicted traffic across different nodes in the network. These findings offer valuable insights into the expected traffic distribution across network nodes, which can inform decisions regarding network management and resource allocation to optimize overall network performance.

2) INTRUSION DETECTION AND PREDICTION

The Logistic Regression and Random Forest models are used to detect intrusions with Python libraries including Pandas and Sklearn. Table 2 shows the values of accuracy, precision, recall, and F1 score for Decision Trees (DT) and Random Forests (RF) in discerning intrusions within WSNs employed in smart grids. Through an evaluation encompassing various attack types including Blackhole, Flooding, Selective Forwarding, and Normal behavior, both DT and RF models exhibit diverse levels of precision, recall, F1-score, and accuracy. Noteworthy is the consistent superiority of Random Forest over Decision Trees, as evidenced by its higher precision, recall, and F1-scores across most attack categories. Particularly striking is Random Forest's exceptional precision and recall in identifying Flooding attacks, with a perfect recall score underscoring its capability to detect all instances of such intrusions. Both models demonstrate exceptional performance in accurately classifying Normal behavior, thus indicating a high level of accuracy. While Decision Trees exhibit marginally lower accuracy compared to Random Forest overall, they still present commendable performance, especially in identifying instances of Normal behavior. These outcomes highlight the efficacy of ML algorithms, particularly Random Forest, in bolstering intrusion detection within WSNs embedded in smart grid frameworks.

Figures 8 and 9 show the confusion matrices for DT and RF respectively. Figures 10 and 11 show the precision-recall curves for DT and RF models respectively and figures 12 and 13 show the receiver operating characteristic for DT and RF models respectively. The confusion matrices provide a comprehensive overview of the model's effectiveness in categorizing different classes. In the context of the Decision Tree's matrix, the diagonal elements reflect the

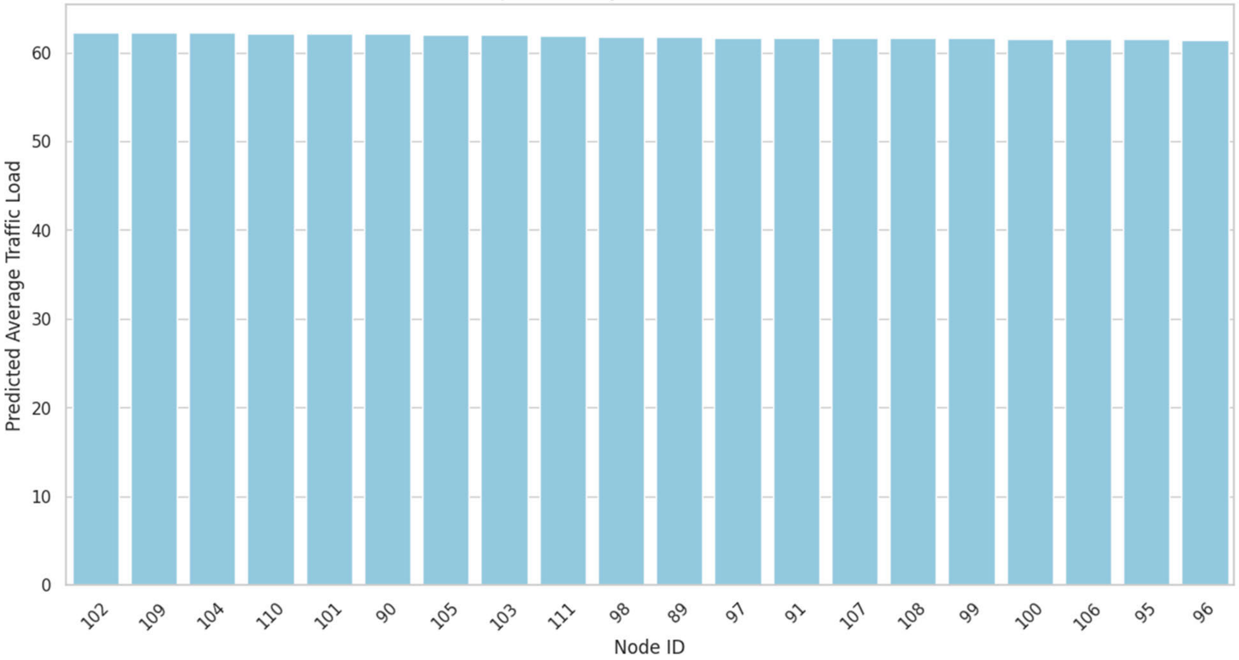


FIGURE 7. Top 20 nodes with highest predicted traffic.

TABLE 2. Accuracy, precision, recall and F1 score for Decision Tree (DT) and Random Forest (RF).

Class	Precision		Recall		F1-score		Accuracy	
	DT	RF	DT	RF	DT	RF	DT	RF
Blackhole (0)	0.65	1.00	0.66	0.77	0.65	0.87	0.95	0.99
Flooding (1)	0.89	0.98	0.90	1.00	0.89	0.99		
Selective Forwarding (2)	0.70	1.00	0.69	0.77	0.70	0.87		
Normal (3)	0.97	0.99	0.97	1.00	0.97	0.99		
Macro average	0.80	0.99	0.80	0.88	0.80	0.93		
Weighted average	0.95	0.99	0.95	0.99	0.95	0.98		

number of instances accurately classified within each class. For instance, it correctly identifies 1561 instances belonging to the "Blackhole" class, 5262 instances classified under "Flooding," 1032 instances categorized as "Forwarding," and 51146 instances assigned to the "Normal" class. Conversely, off-diagonal elements denote instances where misclassifications occur. Notably, it mislabels 115 instances of the "Blackhole" class as "Flooding," 63 instances of "Flooding" as "Forwarding," and so forth. Similarly, in the Random Forest's matrix, diagonal elements signify correct classifications, while off-diagonal elements indicate misclassifications.

The Random Forest model achieves higher counts of accurate classifications across all classes compared to the Decision Tree. For instance, it correctly identifies 1810 instances of the "Blackhole" class, 5853 instances of "Flooding". It demonstrates a tendency to make fewer misclassifications, particularly notable in the "Forwarding" class, where the Decision Tree exhibited more errors. This observation suggests that the Random Forest model possesses superior discriminative capacity and generalization ability relative to the Decision Tree, evident through its higher accuracy and reduced frequency of misclassifications.

The approach presented in this study is multi-faceted, and it integrates traffic analysis, node categorization, and ML based IDPS to achieve more robust cyber defense in WSNs of the smart grid. The performance of this framework was evaluated through an analysis of two main aspects: traffic routes and intrusion detection. The study analyzed traffic patterns for four ML models to evaluate whether they were effective in predicting traffic load on the WSN nodes. They were based on parameters such as the size of a packet, remaining energy

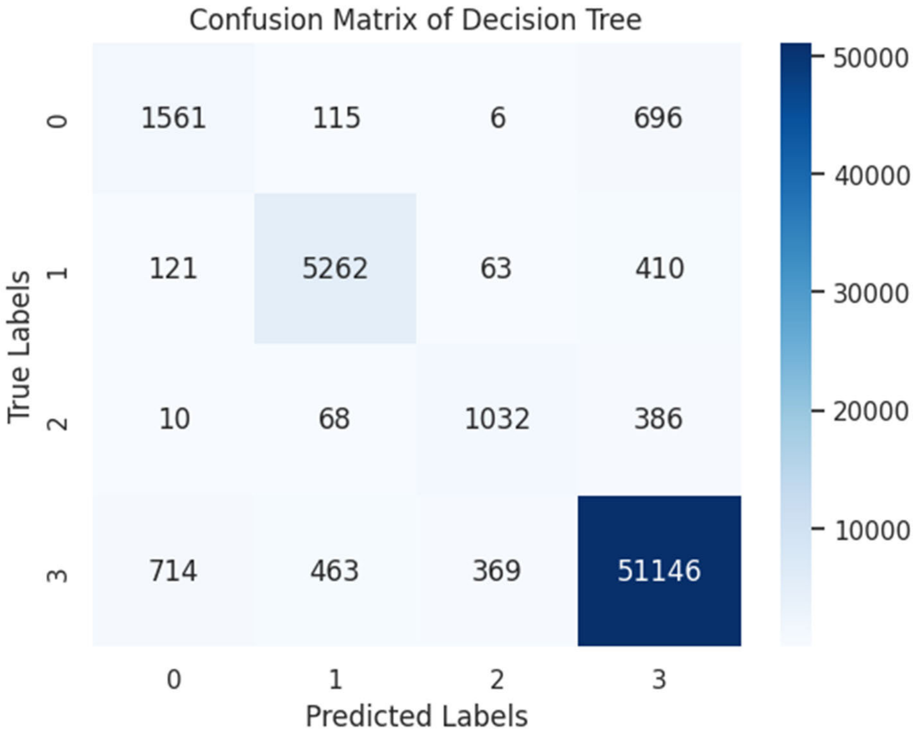


FIGURE 8. Confusion matrix of decision tree.

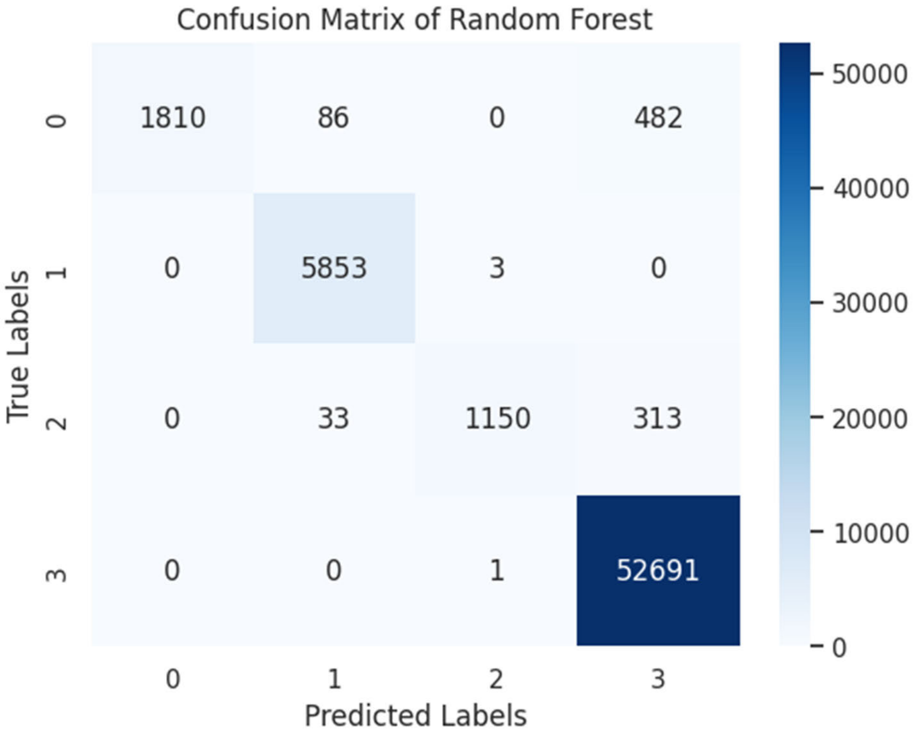


FIGURE 9. Confusion matrix of random forest.

levels, and source/destination IP ports. Results reveal that Random Forest model has higher predictive accuracy than the others. MSE, RMSE, MAE and R^2 are the indicators of this. In contrast, Linear Regression gave us fewer desirable results,

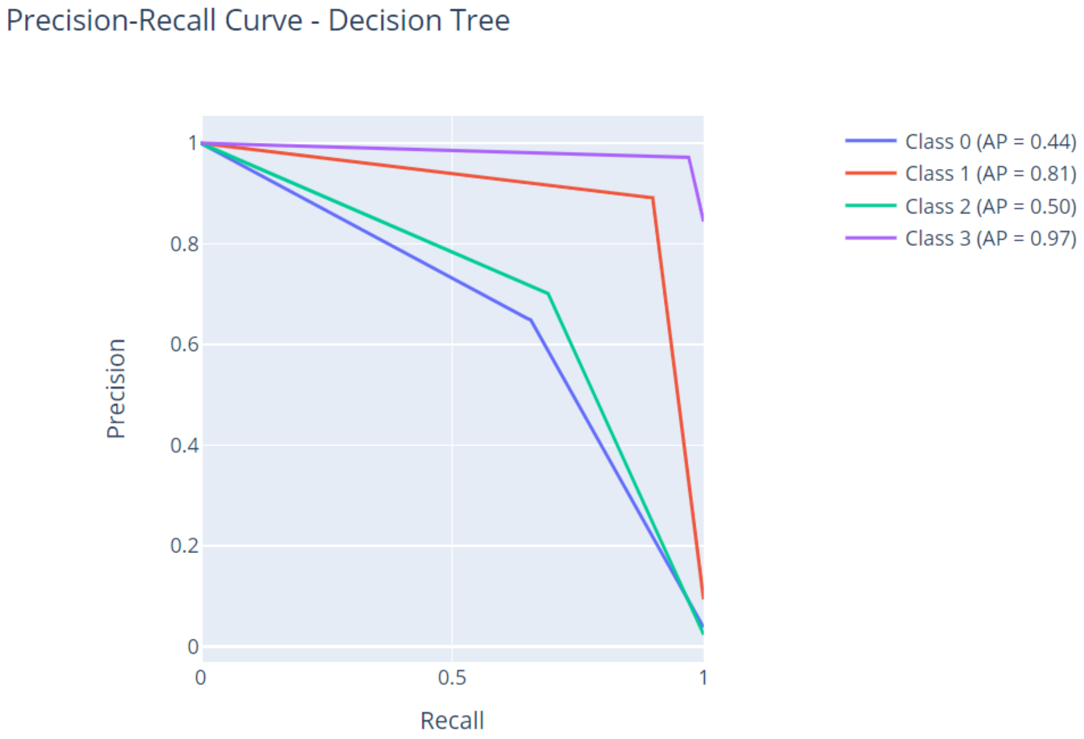


FIGURE 10. Precision-recall curve of decision tree.

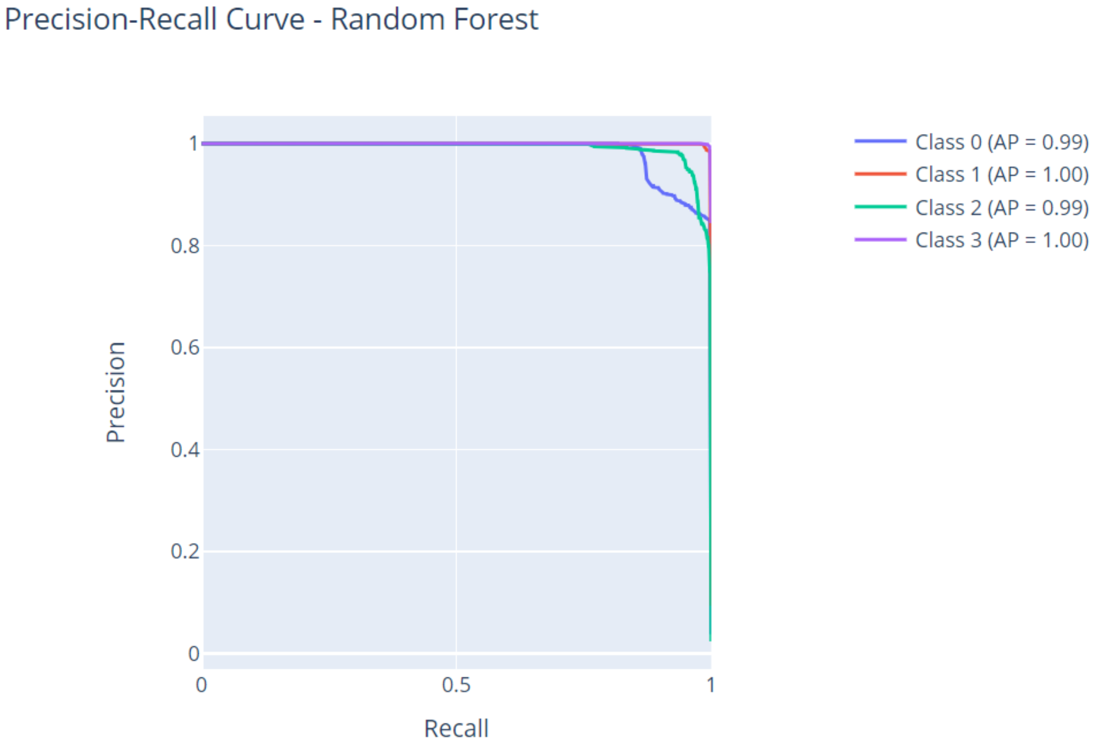


FIGURE 11. Precision-recall curve of random forest.

thus it is evident that it is not capable of capturing the dataset's complexities. The analysis found out some nodes, particularly Node 102, had the highest predicted traffic volume, therefore, it would be helpful for network management and

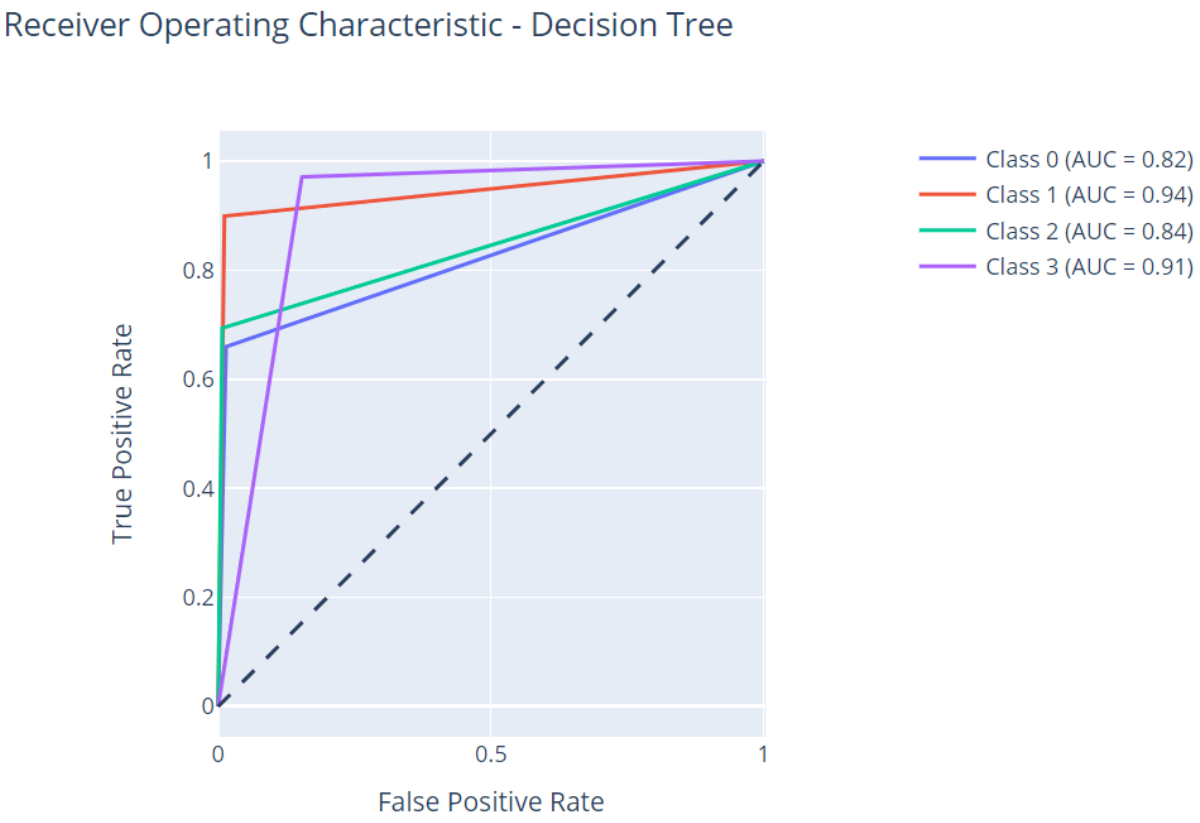


FIGURE 12. **Receiver operating characteristic – decision tree.**

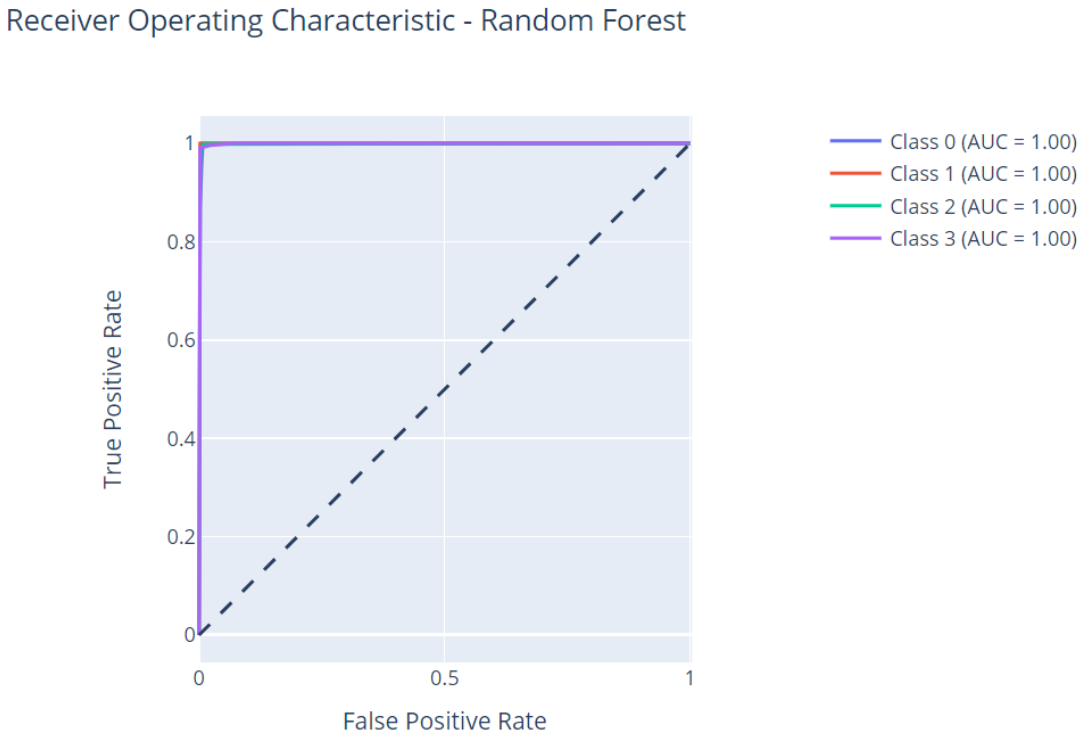


FIGURE 13. **Receiver operating characteristic – random forest.**

resource allocation. Intrusion detection study was conducted by using Logistic Regression and Random Forest models to recognize intrusions within WSNs used in smart grid systems. As Table 2 shows, Random Forest was unrivalled in its performance, having higher recall scores than Decision Trees in all types of attacks, most notably in the case of Flooding. Confusion matrices display the good performance of the Random Forest algorithms, with less misclassifications and higher accuracy as compared to Decision Trees. Apart from that, the accuracy will be determined by making use of precision-recall curves and ROC curves, which will also confirm that Random Forest is effective in intrusion detection. The outcomes of the research put emphasis on the utility of the suggested framework in increasing the cybersecurity levels of smart grid WSNs. Integration of traffic analysis, node identification, and ML based IDPS provides early detection and prevention of intrusions that effectively secure the integrity of the critical infrastructure. Moreover, the study highlights the significance of ensemble learning methods like Random Forest in achieving superior predictive accuracy and intrusion detection capabilities. These insights are invaluable for informing the development and implementation of robust cybersecurity solutions tailored to smart grid environments, thereby mitigating potential cyber threats, and ensuring the reliability and resilience of essential infrastructure.

The potential challenges observed in the WSN dataset, aside from traffic issues, include its implementation in real-world smart grid environments. These challenges encompass cost overhead, hardware limitations, data management, and the necessity for real-time analysis through high-performance computational resources. A Cloud to Edge computing environment has the ability to address these challenges by enabling real-time analyses with minimal latency and cost. Careful consideration of resource utilization is essential to handle large-scale data and varied types of sensor inputs while maintaining adequate performance.

V. CONCLUSION

The study offers a comprehensive framework aimed at improving cybersecurity within WSNs operating in smart grid environments. It aims to achieve detection and prevention mechanisms for cyber threats through the amalgamation of node classification, traffic analysis, and ML techniques. The research focused on detecting and preventing intrusions in WSNs using logistic regression and random forest models. Both models had shown impressive results in the right identification of different attack types like Blackhole, Flooding as well as Selective Forwarding attacks and normal behavior too. Random Forest as a classifier especially for the given metrics has proved to be very efficient and has shown its capability of enhancing intrusion detection systems. The proposed framework was implemented with the WSNBFSF dataset. The experimental design incorporated a ML model with a traffic analysis function and an intrusion detection capability. Different performance measures were used in the evaluation such as MSE, Root RMSE, MAE, R^2, accuracy, precision, recall and, finally, F1 score. The proposed framework is expected to be an important contribution to cybersecurity measures in the smart grid environment by utilizing modern techniques like traffic analysis, node categorization, and algorithmic learning. The findings underscore the importance of proactive cybersecurity strategies in safeguarding sensitive infrastructure and ensuring the reliable and secure delivery of electricity to consumers. The implications of this study are significant, as it provides a foundation for enhancing cybersecurity in smart grids through advanced ML techniques. By implementing this framework, smart grids can better detect and prevent cyber threats, thus ensuring the stability and reliability of electricity delivery. However, several limitations need to be addressed for practical implementation. Potential challenges include cost overhead, hardware limitations, data management, and the need for real-time analysis through high-performance computational resources. Edge devices typically have limited computational power compared to centralized servers, necessitating careful resource utilization to handle large-scale data and varied sensor inputs while maintaining performance.

In future, we will focus on refining ML models by integrating time series analysis and advanced communication protocols. We aim to explore hybrid algorithms to improve risk assessment and anomaly identification, especially against more sophisticated attack types. We also plan to implement fog computing-based fuzzy logic systems to optimize 5G communication technology performance in smart grids, further expanding the framework's applicability and effectiveness.

COMPETING INTERESTS

The authors declare no competing interests.

DATA AVAILABILITY

The dataset supporting the findings of this study is publicly available on a Kaggle website with title, WSNBFSFDataset [30].

REFERENCES

[1] M. Alonso, H. Amaris, D. Alcala, and D. M. Florez R., "Smart sensors for smart grid reliability," *Sensors*, vol. 20, no. 8, p. 2187, Apr. 2020, doi: 10.3390/s20082187.

[2] B. S. Torres, L. E. Borges da Silva, C. P. Salomon, and C. H. V. de Moraes, "Integrating smart grid devices into the traditional protection of distribution networks," *Energies*, vol. 15, no. 7, p. 2518, Mar. 2022, doi: 10.3390/en15072518.

[3] S. R. Biswal, T. Roy Choudhury, B. Panda, B. Nayak, and G. C. Mahato, "Smart meter: Impact and usefulness on smart grids," in *Proc. IEEE 2nd Int. Conf. Appl. Electromagn., Signal Process., Commun. (AESPC)*, Nov. 2021, pp. 1–6, doi: 10.1109/AESPC52704.2021.9708492.

[4] F. E. Abrahamsen, Y. Ai, and M. Cheffena, "Communication technologies for smart grid: A comprehensive survey," *Sensors*, vol. 21, no. 23, p. 8087, Dec. 2021, doi: 10.3390/s21238087.

[5] F. Bonavolontà, V. Caragallo, A. Fatica, A. Liccardo, A. Masone, and C. Sterle, "Optimization of IEDs position in MV smart grids through integer linear programming," *Energies*, vol. 14, no. 11, p. 3346, Jun. 2021, doi: 10.3390/en14113346.

[6] M. Ali, A. I. Jehangiri, O. I. Alramli, Z. Ahmad, R. M. Ghoniem, M. A. Ala'anzy, and R. Saleem, "Performance and scalability analysis of SDN-based large-scale Wi-Fi networks," *Appl. Sci.*, vol. 13, no. 7, p. 4170, Mar. 2023, doi: 10.3390/app13074170.

[7] M. Rashed, I. Gondal, J. Kamruzzaman, and S. Islam, "State estimation within IED based smart grid using Kalman estimates," *Electronics*, vol. 10, no. 15, p. 1783, Jul. 2021, doi: 10.3390/electronics10151783.

[8] M. Plauth, L. Feinbube, and A. Polze. (2017). *A Performance Evaluation of Lightweight Approaches To Virtualization*. [Online]. Available: http://www.thinkmind.org/

[9] A.-A. Bouramdane, ''Cyberattacks in smart grids: Challenges and solving the multi-criteria decision-making for cybersecurity options, including ones that incorporate artificial intelligence, using an analytical hierarchy process,'' *J. Cybersecurity Privacy*, vol. 3, no. 4, pp. 662–705, Sep. 2023, doi: 10.3390/jcp3040031.

[10] A. Corallo, M. Lazoi, and M. Lezzi, ''Cybersecurity in the context of Industry 4.0: A structured classification of critical assets and business impacts,'' *Comput. Ind.*, vol. 114, Jan. 2020, Art. no. 103165, doi: 10.1016/j.compind.2019.103165.

[11] I. de la Peña Zarzuelo, ''Cybersecurity in ports and maritime industry: Reasons for raising awareness on this issue,'' *Transp. Policy*, vol. 100, pp. 1–4, Jan. 2021, doi: 10.1016/j.tranpol.2020.10.001.

[12] L. Drazovich, L. Brew, and S. Wetzel, ''Advancing the state of maritime cybersecurity guidelines to improve the resilience of the maritime transportation system,'' in *Proc. IEEE Int. Conf. Cyber Secur. Resilience (CSR)*, Jul. 2021, pp. 503–509, doi: 10.1109/CSR51186.2021.9527922.

[13] A. Corallo, M. Lazoi, M. Lezzi, and P. Pontrandolfo, ''Cybersecurity challenges for manufacturing systems 4.0: Assessment of the business impact level,'' *IEEE Trans. Eng. Manag.*, vol. 70, no. 11, pp. 3745–3765, Nov. 2023, doi: 10.1109/TEM.2021.3084687.

[14] S. Kumar and R. R. Mallipeddi, ''Impact of cybersecurity on operations and supply chain management: Emerging trends and future research directions,'' *Prod. Oper. Manage.*, vol. 31, no. 12, pp. 4488–4500, Dec. 2022, doi: 10.1111/poms.13859.

[15] A. Naeem, M. S. Farooq, A. Khelifi, and A. Abid, ''Malignant melanoma classification using deep learning: Datasets, performance measurements, challenges and opportunities,'' *IEEE Access*, vol. 8, pp. 110575–110597, 2020, doi: 10.1109/ACCESS.2020.3001507.

[16] M. Nankya, R. Chataut, and R. Akl, ''Securing industrial control systems: Components, cyber threats, and machine learning-driven defense strategies,'' *Sensors*, vol. 23, no. 21, p. 8840, Oct. 2023, doi: 10.3390/s23218840.

[17] K. Shaukat, S. Luo, V. Varadharajan, I. Hameed, S. Chen, D. Liu, and J. Li, ''Performance comparison and current challenges of using machine learning techniques in cybersecurity,'' *Energies*, vol. 13, no. 10, p. 2509, May 2020, doi: 10.3390/en13102509.

[18] A. Butnaru, A. Mylonas, and N. Pitropakis, ''Towards lightweight URL-based phishing detection,'' *Future Internet*, vol. 13, no. 6, p. 154, Jun. 2021, doi: 10.3390/fi13060154.

[19] T. N. Gia, A.-M. Rahmani, T. Westerlund, P. Liljeberg, and H. Tenhunen, ''Fault tolerant and scalable IoT-based architecture for health monitoring,'' in *Proc. IEEE Sensors Appl. Symp. (SAS)*, Apr. 2015, pp. 1–6, doi: 10.1109/SAS.2015.7133626.

[20] D. Rahbari and M. Nickray, ''Low-latency and energy-efficient scheduling in fog-based IoT applications,'' *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 27, no. 2, Mar. 2019, Art. no. 52, doi: 10.3906/elk-1810-47.

[21] A. Rastogi and A. Bais, ''Comparative analysis of software defined networking (SDN) controllers—In terms of traffic handling capabilities,'' in *Proc. 19th Int. Multi-Topic Conf. (INMIC)*, Dec. 2016, pp. 1–6, doi: 10.1109/INMIC.2016.7840116.

[22] M. Ghiasi, Z. Wang, M. Mehrandezh, S. Jalilian, and N. Ghadimi, ''Evolution of smart grids towards the Internet of energy: Concept and essential components for deep decarbonisation,'' *IET Smart Grid*, vol. 6, no. 1, pp. 86–102, Feb. 2023, doi: 10.1049/stg2.12095.

[23] H. Habbak, M. Mahmoud, K. Metwally, M. M. Fouda, and M. I. Ibrahem, ''Load forecasting techniques and their applications in smart grids,'' *Energies*, vol. 16, no. 3, p. 1480, Feb. 2023, doi: 10.3390/en16031480.

[24] H. Wang, M. Ouyang, Q. Meng, and Q. Kong, ''A traffic data collection and analysis method based on wireless sensor network,'' *EURASIP J. Wireless Commun. Netw.*, vol. 2020, no. 1, p. 2, Dec. 2020, doi: 10.1186/s13638-019-1628-5.

[25] C. Yao, Y. Yang, K. Yin, and J. Yang, ''Traffic anomaly detection in wireless sensor networks based on principal component analysis and deep convolution neural network,'' *IEEE Access*, vol. 10, pp. 103136–103149, 2022, doi: 10.1109/ACCESS.2022.3210189.

[26] N. Sahani, R. Zhu, J.-H. Cho, and C.-C. Liu, ''Machine learning-based intrusion detection for smart grid computing: A survey,'' *ACM Trans. Cyber-Phys. Syst.*, vol. 7, no. 2, pp. 1–31, Apr. 2023, doi: 10.1145/3578366.

[27] M. Panthi and T. Kanti Das, ''Intelligent intrusion detection scheme for smart power-grid using optimized ensemble learning on selected features,'' *Int. J. Crit. Infrastructure Protection*, vol. 39, Dec. 2022, Art. no. 100567, doi: 10.1016/j.ijcip.2022.100567.

[28] W. M. S. Yafooz, Z. B. A. Bakar, S. K. A. Fahad, and A. M. Mithon, ''Business intelligence through big data analytics, data mining and machine learning,'' in *Data Management, Analytics and Innovation* (Advances in Intelligent Systems and Computing), vol. 1016, N. Sharma, A. Chakrabarti, and V. Balas, Eds., Singapore: Springer, 2020, doi: 10.1007/978-981-13-9364-8_17.

[29] A. Kumari, R. K. Patel, U. C. Sukharamwala, S. Tanwar, M. S. Raboaca, A. Saad, and A. Tolba, ''AI-empowered attack detection and prevention scheme for smart grid system,'' *Mathematics*, vol. 10, no. 16, p. 2852, Aug. 2022, doi: 10.3390/math10162852.

[30] Kaggle. *WSNBFSFDataset*. Accessed: Feb. 10, 2024. [Online]. Available: https://www.kaggle.com/datasets/celilokur/wsnbfsfdataset

[31] N. Farnaaz and M. A. Jabbar, ''Random forest modeling for network intrusion detection system,'' *Proc. Comput. Sci.*, vol. 89, pp. 213–217, Jan. 2016, doi: 10.1016/j.procs.2016.06.047.

[32] H. Dabiri, V. Farhangi, M. J. Moradi, M. Zadehmohamad, and M. Karakouzian, ''Applications of decision tree and random forest as tree-based machine learning techniques for analyzing the ultimate strain of spliced and non-spliced reinforcement bars,'' *Appl. Sci.*, vol. 12, no. 10, p. 4851, May 2022, doi: 10.3390/app12104851.

[33] A. V. Konstantinov and L. V. Utkin, ''Interpretable machine learning with an ensemble of gradient boosting machines,'' *Knowl.-Based Syst.*, vol. 222, Jun. 2021, Art. no. 106993, doi: 10.1016/j.knosys.2021.106993.

[34] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, ''A comparative analysis of gradient boosting algorithms,'' *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

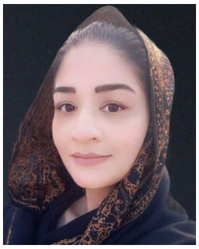
[35] A. I. Ahmed Osman, A. N. Ahmed, M. F. Chow, Y. F. Huang, and A. El-Shafie, ''Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia,'' *Ain Shams Eng. J.*, vol. 12, no. 2, pp. 1545–1556, Jun. 2021, doi: 10.1016/j.asej.2020.11.011.

[36] K. Shah, H. Patel, D. Sanghvi, and M. Shah, ''A comparative analysis of logistic regression, random forest and KNN models for the text classification,'' *Augmented Human Res.*, vol. 5, no. 1, p. 12, Dec. 2020, doi: 10.1007/s41133-020-00032-0.

[37] N. Mozaffaree Pour and T. Oja, ''Prediction power of logistic regression (LR) and multi-layer perceptron (MLP) models in exploring driving forces of urban expansion to be sustainable in Estonia,'' *Sustainability*, vol. 14, no. 1, p. 160, Dec. 2021, doi: 10.3390/su14010160.

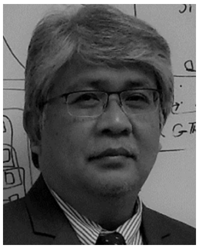
[38] T. G. Nick and K. M. Campbell, ''Logistic regression,'' in *Topics in Biostatistics* (Methods in Molecular Biology), vol. 404, W. T. Ambrosius, Ed., Humana Press, doi: 10.1007/978-1-59745-530-5_14.

TAMARA ZHUKABAYEVA received the Ph.D. degree from Satbayev University, Kazakhstan. She is currently an Associate Professor of informatics, computer engineering, and management with L. N. Gumilyov Eurasian National University, Astana, Kazakhstan. She has published more than 70 scientific and educational-methodical works: in Kazakhstan and countries far and near abroad, including a foreign edition from the Clarivate Analytics Database, Scopus. She is the author and co-author of educational publications and scientific monographs, has an innovative patent and copyright certificates for intellectual property rights. She is an Associate Member of the Universal Association of Computer and Electronics Engineers, has membership in scientific societies in The Society of Digital Information and Wireless Communications (SDIWC), and the Universal Association of Computer and Electronics Engineers.



AISHA PERVEZ received the Bachelor of Science (B.S.) degree in electronics engineering from COMSATS University Islamabad, Abbottabad Campus, in 2012, and the Master of Science (M.S.) degree in telecommunication and networks, in 2022. She joined Hazara University, Mansehra, Pakistan, in 2012, and has gained extensive teaching experience at the Computer Science and Information Technology Department and the Telecommunication Department at the undergraduate level. Currently, she is a Faculty Member with the Telecommunication Department, Hazara University. Her research interests include information security, cybersecurity, cloud computing, wireless sensor networks (WSNs), traffic analysis, and the Internet of Things (IoT).

YERIK MARDENOV is currently pursuing the Ph.D. degree majoring in computer science. He works as a Senior Lecturer with Astana International University and a Researcher with International Science Complex "Astana." He specializes in such advanced areas as information security, attacks, and wireless sensor networks.



MOHAMED OTHMAN (Senior Member, IEEE) received the Ph.D. degree (Hons.) from the National University of Malaysia. He is currently a Professor of computer science with the Department of Communication Technology and Networks, Universiti Putra Malaysia (UPM). Prior to that, he was the Deputy Director of the Information Development and Communication Centre, where he was in charge of the UMPNet Network Campus, uSport Wireless Communication Project, and the UPM Data Center. He was also a Visiting Professor with South Kazakhstan State University, Shymkent, and the L. N. Gumilyov Eurasian National University, Astana, Kazakhstan. He is also an Associate Researcher and a Coordinator of high-speed machines with the Laboratory of Computational Science and Informatics, Institute of Mathematical Science, UPM. He has also filed six Malaysian, one Japanese, one South Korean, and three U.S. patents. He has published more than 300 international journals and 330 proceeding articles. His main research interests include computer networks, parallel and distributed computing, high-speed interconnection networks, network design and management (network security, wireless, and traffic monitoring), consensus in IoT, and mathematical models in scientific computing. He is a Life Member of Malaysian National Computer Confederation and Malaysian Mathematical Society. He was awarded the Best Ph.D. Thesis in 2000 by Sime Darby Malaysia and the Malaysian Mathematical Science Society. In 2017, he received an Honorary Professorship from SILKWAY International University (formerly known as South Kazakhstan Pedagogical University), Shymkent, Kazakhstan.

NURDAULET KARABAYEV is currently pursuing the master's degree with Astana IT University. He concurrently serves as a Junior Researcher with International Science Complex "Astana." His research interests include various domains, including cybersecurity, wireless sensor networks (WSNs), artificial intelligence (AI), and the Internet of Things (IoT).

ZULFIQAR AHMAD received the M.Sc. degree (Hons.) in computer science from Hazara University, Mansehra, Pakistan, in 2012, the M.S. degree in CS from COMSATS University Islamabad, Abbottabad, Pakistan, in 2016, and the Ph.D. degree in computer science from the Department of Computer Science and Information Technology, Hazara University, in 2022. He is currently serving as a lecturer with the Department of CS & IT, Hazara University. He is the author of several publications in the fields of fog computing, cloud computing, high-performance computing, and scientific workflow execution and management. His research interests include scientific workflow management in cloud computing, the Internet of Things, fog computing, edge computing, cybersecurity, and wireless sensor networks (WSNs).