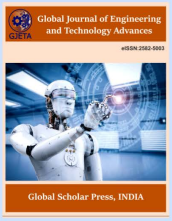
Global Journal of Engineering and Technology Advances
eISSN: 2582-5003
Cross Ref DOI: 10.30574/gjeta
Journal homepage: https://gjeta.com/



(RESEARCH ARTICLE)

Check for updates

Comprehensive Analysis of SCADA System Data for Intrusion Detection Using Machine Learning

Smart Idima *, Philip Nwaga and Patrick Evah

Department of Computer Science, School of Computer Science Stripes Hall 44, Western Illinois University.

Global Journal of Engineering and Technology Advances, 2025, 22(02), 064-089

Publication history: Received on 31 December 2024; revised on 09 February 2025; accepted on 12 February 2025

Article DOI: https://doi.org/10.30574/gjeta.2025.22.2.0027

Abstract

This report investigates the implementation of advanced machine learning models within Supervisory Control and Data Acquisition (SCADA) systems to enhance intrusion detection capabilities and system security. By utilizing models such as CatBoost and XGBRegressor, which excel in processing complex, non-linear data, the study demonstrates significant improvements in predicting and managing operational states in wind turbines. The incorporation of Explainable AI (XAI) techniques, particularly SHAP values, further provides transparency in model decisions, fostering trust among stakeholders. Recommendations are provided for effective model integration, deployment with XAI features, and necessary policy enhancements to ensure the secure, reliable, and ethical use of AI in critical infrastructure environments.

Keywords: Machine Learning; SCADA system; Intrusion Detection; Explainable AI (XAI); Critical Infrastructure; Cybersecurity; Policy Enhancements

1. Introduction

SCADA The SCADA system is an Industrial Control System or platform that controls processes across critical sectors in energy, water, transport, and others. These systems are not just technical applications; they ensure the operational integrity of services that nourish social order and economic stability in nations. SCADA involves near-real-time high-resolution data and control of processes concerning the smooth functioning of infrastructures that go into the gray areas of national security thrust in the aid of public welfare.

There is no denying the importance of SCADA in the infrastructure sector. It sets the stage for monitoring and control throughout the utility and service spectrum-from the sprawling electricity grid system through hydro treatment plants-optimizing operation efficiencies while assuring the reliability of services upon which the wider public depends. As discussed in the [1], real-time data and control are offered by SCADA in such a way that there is very little else remaining to keep services operating or obtain normality to essential services. In any situation when that does not happen, damaging ramifications, whether for society or for the economy, arise.

However, SCADA's growth over time has exposed them to greater vulnerability attributable to cyber threats. Modernization of these systems has brought with it added integration with corporate networks and the Internet, placing these types of systems in the limelight of a new set of cyber vulnerabilities; that interconnectedness, beneficial for operational efficiency and data management, opens a whole new array of avenues for cyber-attacks. Modern SCADA systems are a far cry from the older versions that did not communicate with anyone or anything. The majority of their operations are based on standard communication protocols and put more and more often within the wider framework of IT. With this goes a need for high-level security measures against internal and external threats.

* Corresponding author: Smart Idima

As highlighted in [2], the integration of internet connectivity and other networked operations has fundamentally altered the cybersecurity landscape for SCADA systems. Increased exposure thus increases the threat of conventional cyber threats on one hand and, on the other, introduces unheard-of complexities, for example, how might one secure a distributed network spread over a wide geographical area with multiple entry points for potential attacks? This mature evolution has thus brought forth a whole range of sophisticated cyber defense mechanisms, enforcing not the nature of isolated threats but addressing continuous and evolving threats that can impact the lifeblood of the economy and security of a nation.

The importance and relevance of installing and updating SCADA system cyber protection processes are, therefore, more than ever. Those processes must encapsulate an understanding of the technical domain of the systems and the threats present in the ever-dynamic cyberspace. An effective cybersecurity strategy should apply advanced security technology, develop serious upgrade and testing protocols, and be fully aware of threats as they arise to instill confidence in the integrity and survivability of any of the nation's critical infrastructures.

1.1. Problem Statement

SCADA systems are inextricably linked to the operation of modern infrastructures, such as power grids, water supply networks, and transportation systems, and thus are the key link to the existence of a great number of cyber threats that directly compromise their functioning. With the disruption of these systems, numerous serious economic and social implications arise, ranging from the loss of vital services to substantial levels of financial liability upon communities and entire nations. Adding to these risks are some very special characteristics of SCADA systems.

Firstly, Many SCADA systems were built upon technologies that are now regarded as legacy. These systems usually have outdated software, lacking any provisions to safeguard against modern threats. Indeed, many of these very old SCADA systems are inherently weak due to the absence of basic security provisions that have since become standard in rather more modern applications. The cost implications and the halt of operational activities during the life cycle of replacing such legacy systems with newer technological infrastructure discourage many duty holders from the management and upgrading of SCADA applications in their critical infrastructures. The fact that inherent vulnerabilities of these systems pose continuous threats to the security and reliability of critical operational processes has been reported in [4] [1].

Secondly, the operational requirements of SCADA systems, involving real-time functioning, considerably add on to the difficulty posed by the implementation of cybersecurity measures. These systems require virtually uninterrupted operations with the least downtime, capable of processing any data immediately so that critical services can function smoothly. Hence, whatever cybersecurity measures are developed must work without compromising the performance of the systems. This requirement frequently limits the implementation of stronger security measures that would include good levels of encryption and regular updates to the software, which would otherwise introduce temporary interruptions to system operations.

The interconnected nature of SCADA systems, which often integrate with various other networks and the Internet, increases their exposure to external threats. SCADA systems have increased connectivity, subsequently increasing the level of exposure to cyber threats, as the potential entry points multiply into several smaller cyber entry points and a larger attack surface. On the positive side, this connectivity increases operational and data-sharing efficiency. The caveat, however, is that these vulnerabilities in the network allow attackers to sneak into the very systems protected by it.

Additionally, the cyber-physical nature of SCADA systems means that cyberattacks can have direct physical consequences [3]. An attack could not only lead to data breaches or loss of operational control but also to physical damage to the infrastructure itself. This aspect dramatically raises the stakes, as failures can result in immediate safety hazards and environmental damage, compounding the social and economic impacts.

Considerations affecting the protection of SCADA systems got responses in a cocktail of forms, from strengthening and further securing the technologies to enforce tougher operational procedures and countermeasures. Above all, the proactive nature of cybersecurity should remain flexible, so that it can account for new attack vectors in light of the vulnerabilities arising from integration of legacy systems with contemporary technology. This strategy will keep the resilience of all critical infrastructures against an array of cyber threats, while also continuing to protect essential public services and people's livelihoods.

1.2. Research Objectives

The study aims to develop an advanced intrusion detection system based on Explainable Artificial Intelligence (XAI) to mitigate SCADA systems' weaknesses. The project intends to make existing systems more potent in terms of detection with the introduction of the XAI framework and further increase the transparency and trustworthiness of the decision-making process in these critical systems. The specific objectives of the research are as follows:

- Enhance Detection Capabilities: The project aims to employ advanced machine-learning algorithms that can perform real-time detection and response to cyber-attacks significantly. The goal is to identify real or potential threats rapidly, accurately, and without major disruption to the normal operation of SCADA systems. Therefore, the goal is the reduction of false positives and the improvement of real threat detection capacity with timely safeguards.
- Integrate XAI: Methods like SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) will be introduced to ensure that AI-driven decision-making will be comprehensible. Thereby human operators and regulatory authorities would have trust and ease of compliance when it comes to AI actions within SCADA systems. XAI consideration is crucial from the point of view of ensuring stakeholders can understand, trust, and perform risk management on automated systems, which are becoming increasingly important in critical infrastructure [2].
- Evaluate and Adapt Security Measures: The mission is to persistently measure and enhance the security community in establishing measures against feedback from real-world applications and simulated scenarios where attacks may or may not involve the agreed-upon threats. With the endeavor on, we must evaluate and modify the defense systems to meet the recent and upcoming threats. It is this constant evaluation and adaptation for the systems to be constantly effective and strong against ever-changing threats [4].

Along these objectives, the research prepares the ground for a major enhancement of existing cybersecurity measures in SCADA systems through modern-day capabilities of XAI. This will, in turn, improve the reliability and safety of the infrastructures controlled by those systems, thus securing them to fight back for the benefit of societal and economic stability.

2. Literature Review

2.1. Review of Existing Approaches in Intrusion Detection

Intrusion detection systems (IDS) are very much required for the detection and mitigation of threats in network-based environments, including critical infrastructure-like Supervisory Control and Data Acquisition (SCADA) systems. The evolution of intrusion detection has seen an unprecedented acceleration with the advent of machine learning and deep learning technologies, allowing for ingenious ways of detecting previously known and unknown threats.

2.1.1. Machine Learning in Intrusion Detection

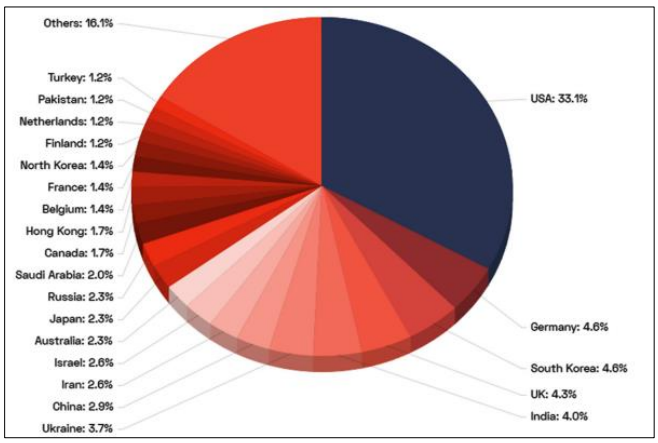


Figure 1 Most frequent targeted cyber warfare attacks between 2009 and 2019[5]

Machine learning algorithms are instrumental to develop intrusion detection capabilities. Traditional machine learning techniques like Support Vector Machines, Decision Trees, and Random Forests have been applied profusely based on

their general success with classification within SCADA systems. Such methods need feature engineering and are often accompanied by additional ensemble methods to increase the accuracy of detection and reduce false positives [5].

Figure 1 illustrates the distribution of cyber-attacks originating from various countries. The largest part, which is in dark blue, represents the United States, which constitutes 33.1% of total attacks and is thus the major source of cyber threats. Other considerable contributors include Germany and South Korea, which each account for 4.6%, very closely followed by the UK with 4.3% and India with 4.0%.

The graph additionally bands several more countries under the realm of "Others," contributing 16.1% to the total landscape of cyber-attacks. These countries include Ukraine at 3.7%, China at 2.9%, and Russia at 2.3%. This visualization captures the international context of cybersecurity threats and reinforces the fact that without international cooperation and firm cyber defenses in place in different countries, the mitigation of these threats will prove extremely difficult.

Deep learning, a subset of ML, has brought in capabilities of learning and inference from massive amounts of data, without explicit programming. Techniques like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been utilized in intrusion detection with superior performance in recognizing complex patterns or anomalies in network traffic [6].

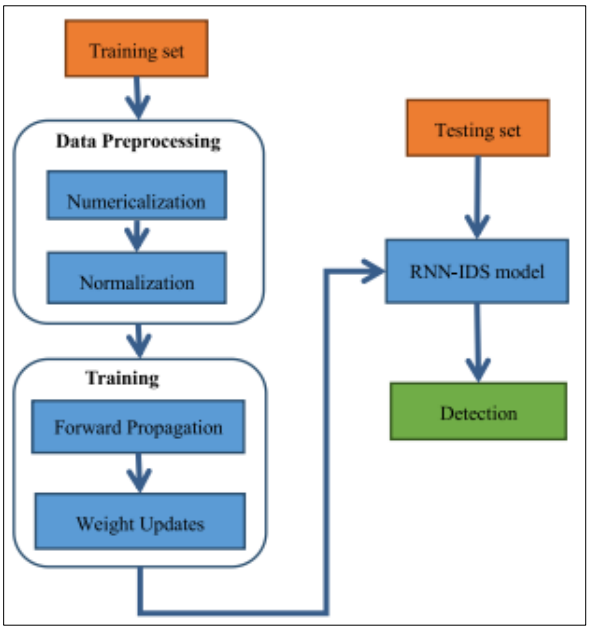


Figure 2 Block diagram of proposed RNN-IDS [6]

Figure 2 illustrates a flowchart depicting the process of using a Recurrent Neural Network (RNN) for an Intrusion Detection System (IDS), termed RNN-IDS. This is the blueprint for the crucial steps of the machine-learning pipeline to detect cyber threats utilizing deep learning techniques.

It all starts with the Training Set, which is data that has been collected and has multiple features required to train the RNN model. This data then goes through Data Preprocessing, which includes actions like Numericalization transforming text or categorical data to numerical values and Normalization—scaling of numerical values to a specific range to make data processing consistent.

With that, the other phase that comes in immediately after is the Training phase, where the RNN model is trained through steps such as Forward Propagation, whereby input data is passed through the neural network to obtain predictions, which are compared with the desired actual output and errors calculated thereafter. The learning process now runs through Weight Updates: this means subsequent iterative corrections on weight in the network to minimize error over time and thus improve the accuracy of prediction.

After sufficiently training the model, it is now validated against an independent Testing Set. It is extremely important to assess the generalization and effectiveness of the RNN-IDS in this stage. Testing aside, it also becomes very significant;

should the model pass testing, the RNN-IDS is now put to work for Detection by recognizing the presence of any new incoming data as potential intrusions, assisting with the mechanisms of defense in cybersecurity. This flowchart is generic and applicable to supervised learning models in machine learning whereby the importance of a systematic approach in carving out an effective predictive model is highlighted.

2.1.2. Role of Deep Learning in SCADA Security

Deep learning has made its mark into intrusion detection, especially in Supervisory Control and Data Acquisition applications that keep critical infrastructures like the power grid and water distribution networks operational. A salient feature of DL, which sets it apart from other paradigms, is its functioning on great amounts of unstructured data, thus uncovering minute and complex patterns that traditional machine learning may not pick. Such features find nestled applications in SCADA domains, where the integrity and availability of operational data are crucial, and even an insignificant- looking anomaly may indicate a certain level of security breach or cyber-attacks.

The RNNs represent a particular class of deep learning models that would be extremely advantageous for applications in SCADA systems since this deal primarily with sequential data. RNNs would have a considerably higher performance level in time-series data analysis, which is basically the nature of data found in any industrial control system. In such data sequences, atypical behavior would certainly indicate the occurrence of either a serious cyber threat or system failure. Due to their temporal functioning, RNNs can detect anomalies in sequential data that find a continuous flow in SCADA systems, thereby making them a comprehensive mechanism for real-time anomaly detection [7].

Furthermore, attacks continue evolving, hence the deep learning models can be trained for detection of these new kinds of attacks which are more adaptable to emerging threats than other traditional models that often necessitate manual retraining or redesigning. Hence, the adaptability of DL models is an important aspect in a constantly changing environment such as the one in which SCADA systems find themselves due to cybersecurity threats. A predictive model based on historical and real-time data would thus improve its accuracy in SCADA system security against known and unknown attack vectors.

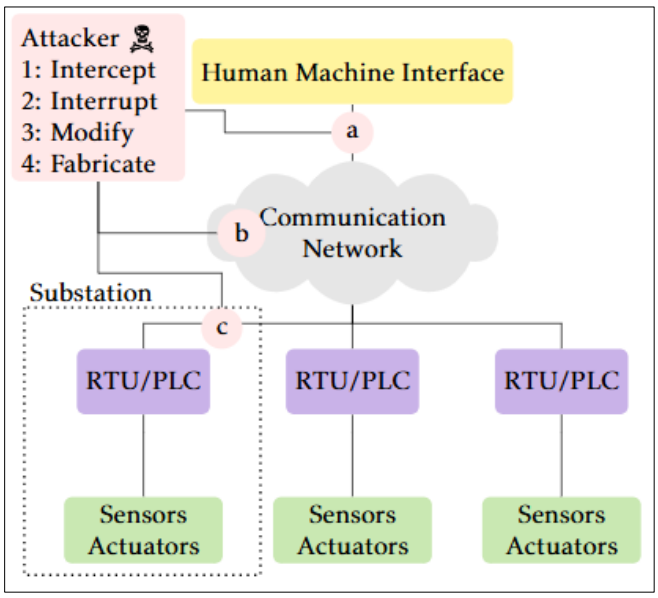


Figure 3 Attack model demonstrating four network attacks, denoted as (1–4), against a simplified SCADA architecture with three attack targets (a–c) [5]

Figure 3 provides a schematic overview of some of the possible threats to cybersecurity in a SCADA system (Supervisory Control and Data Acquisition), stress is also laid, on whatever line the attacker tries to compromise the system. Intercept or interrupt or change or make false attacks may be carried out by the attacker against different components of the system. For instance, the Human Machine Interface (HMI) can be intercepted (a), which means that through interception, unauthorized access to system controls is made along with some sensitive data. The communication network could then be interrupted (b), stopping the data communication between the HMI and the Remote Terminal Units or Programmable Logic Controllers (RTUs/PLCs). Finally, the RTUs/PLCs are subject to modifications and fabrications (c) with respect to their management of sensors and actuators in substations causing operating failures or

giving false readings from the physical processes. This illustration underlines the need to secure each layer of a SCADA network against assorted cyber threats.

These advancements offer the construction of more autonomous intelligent systems of threat detection and response upon which critical infrastructure sectors and IM systems have come to rely for major improvements in their security posture. This development represents mainly an advancement in the application of DL to SCADA and signals an important shift toward more sophisticated and data-driven approaches governing industrial cybersecurity measures, as corroborated by the opinions expressed in the referenced article.

2.1.3. Integration Challenges and Solutions

The integration of ML and DL into SCADA systems comes with its challenges; the prime reason for this being the variability and complexity due to the type of network traffic, as well as the real-time processing requirements.

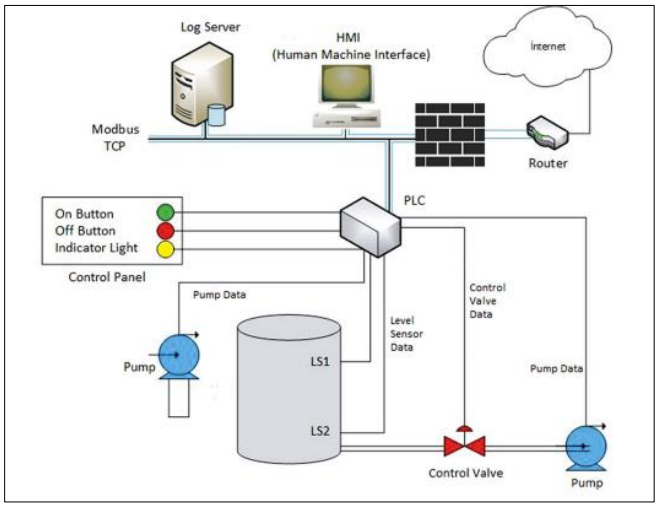


Figure 4 Representation of a SCADA system [9]

Figure 4 is a schematic representation of a typical SCADA (Supervisory Control and Data Acquisition) system layout used for monitoring and controlling industrial and infrastructure processes. This diagram illustrates the components that come into play in pump control-integration and communication between the various elements:

- Human Machine Interface (HMI): This is the user interface for operator interaction. HMI displays data, logs, and controls for human operators to monitor the process and make adjustments as necessary. HMI connects to a log server for storing and retrieving operational data.
- Programmable Logic Controller (PLC): At the core of this system is the PLC, which is responsible for executing control instructions based on the program written by the user. It receives data from various sensors, processes this data, and sends control commands to the machinery.
- Control Panel: This includes the On and Off buttons along with indicator lights which show the operational status (e.g., running, stopped) of the pumps. These manual controls allow operators to start or stop the pump and observe its status directly from the control panel.
- Pump and Control Valve: The diagram shows a pump which is used to move fluid from one place to another. The control valve, which is operated based on commands from the PLC, regulates the flow of the fluid being pumped. This is crucial for maintaining desired levels of output and safety.
- Level Sensors (LS1 and LS2): These sensors monitor the level of the fluid in the tank. Their data is crucial for ensuring that the fluid does not overflow or deplete beyond safe operational levels. The sensors send data back to the PLC to help automate the control of the pump and valve based on predefined criteria.
- Pump Data and Control Data: These are data points and commands transmitted between devices. Pump data likely includes operational parameters such as speed and output pressure, while control data would be commands from the PLC to adjust the pump and valve as necessary.
- Modbus TCP: This is a common communication protocol used in industrial environments, highlighted here as the method of communication between the HMI, PLC, and possibly other components. It allows for the reliable exchange of command-and-control data over TCP/IP networks.

- Router and Internet: These components suggest that the system is capable of remote access or is part of a larger networked system. This connectivity can be used for remote monitoring and control, or for integrating the SCADA system with other enterprise systems for analytics and advanced management.

Hybrid models combining ML and DL have been proposed to leverage the strengths of both approaches, ensuring robust detection capabilities while managing the computational overhead. Furthermore, federated learning approaches applied in SCADA systems are tackled regarding privacy and latencies involved in processing data for almost real-time IDS applications in critical settings [9].

2.1.4. Future Directions

The cybersecurity developments regarding SCADA systems are witnessing fast changes toward creating more autonomous and adaptive security mechanisms. It has been extensively argued in the contemporary literature, as shown by [10]. that there is still a larger trend toward giving a fair share of advanced AI capabilities to security systems. The AI-based mechanism does not work just as a response to mitigating threats to security but also plays an active role in predicting and preventing them. This active intervention allows the system to assertively try to mitigate possible attacks before it is ever launched against the infrastructure. Therefore, an acceptable level of security for the infrastructure is assured.

Future research is likely to concentrate on various directions for security improvement. Most importantly, these will be improvements in AI efficiency optimizing algorithms to reduce processing times and computational resources as a priority for real-time threat detection and response. Another important area will be reducing AI requirements for big data sets, as many current paradigms for AI, chiefly those based in deep learning, require very large data sets for training. This requirement limits the implementation of AI in environments where data acquisition poses challenges or where privacy issues constrain the data available. Transfer learning, few-shot learning, and synthetic data generation might provide routes for solving some of these limitations.

Generalization ability across different network environments also constitutes an area of critical interest for future research. SCADA systems share common characteristics, but depending on the application and the industry, these systems could be configured to operate in very different ways and to face a variety of threats. Thus, the major undertaking will be to develop Artificial Intelligence systems that are able to adapt and perform excellently under these varying conditions, without the necessity of cumbersome retraining or customization. This may call for the evolution of some higher-level models with context-awareness and adaptation mechanisms to respond to real worldly operational circumstances; this would contrast with increasing-moderately configurable modular systems farmed into a minimum amount of input toward configuration.

Furthermore, ethical and security considerations surrounding the AI models themselves are of increasing concern. With the transformation of these systems toward increasing autonomy, guaranteeing the transparent and secure operation of these systems would be paramount. Researchers will most likely work toward increasing the interpretability of AI decisions and toward securing AI-based security systems against specific adversarial attacks meant to confuse or mislead them.

In general, the future direction of research into AI-based SCADA security systems is predicted to be the rise of systems that are smarter, faster, and more adaptable. This advance is expected to make significant contributions to the field within cyberspace and render critical infrastructures much more secure and resilient against an increasingly sophisticated and dynamic landscape of cyber threats.

An analysis of the current methods and advances demonstrates that intrusion detection for SCADA systems is a dynamic process with machine learning and deep learning becoming core to redefining security frameworks in the future.

2.2. Importance of XAI in Cybersecurity

Explainable AI (XAI) is becoming more and more important in cybersecurity to increase the transparency of AI systems, which is necessary to understand, trust, and properly govern AI recommendations in environments with security sensitivity. From a cybersecurity point of view, therefore, integrating XAI emphasizes the decision-making process of the AI systems so that human operators can comprehend and thus trust AI output values. As discussed in [11], XAI offers a way to dissect the black-box nature of conventional AI models, providing a clear understanding of how and why certain decisions are made, particularly in detecting and responding to cyber threats. This is critical for the identification and mitigation of imminent threats in a reliable and accurate manner.

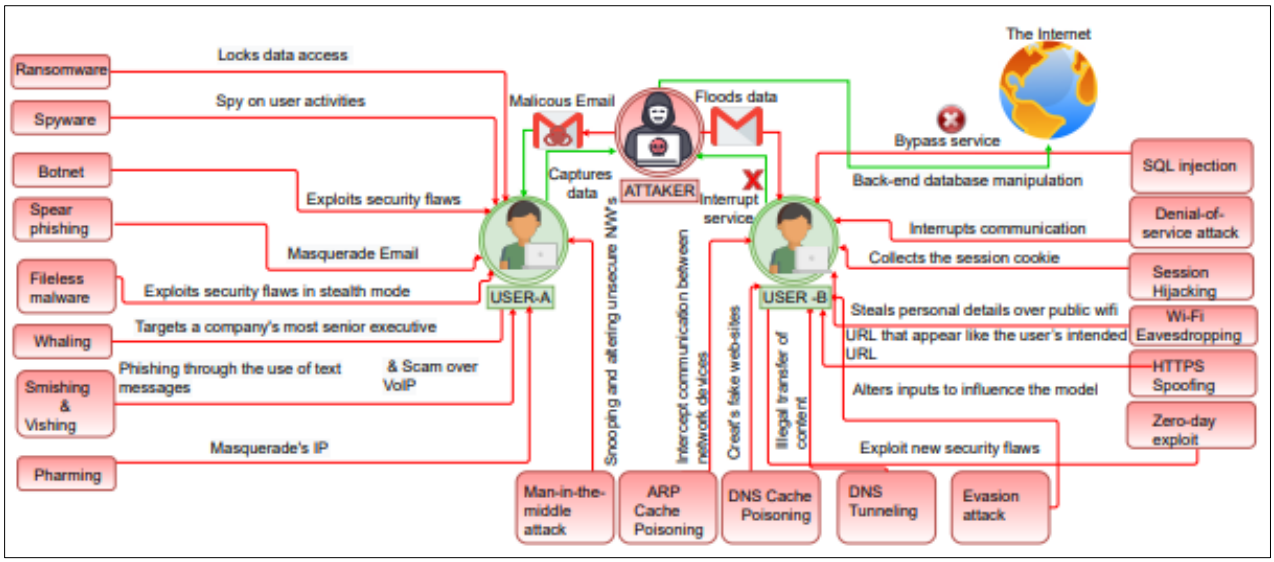


Figure 5 Types of cyber attacks [11]

Figure 5 shows various cyberattacks and their methodologies in the framework of SCADA systems, indicating how attackers leverage the vulnerabilities of ransomware to SQL injection. Methods employed within this regard include intercepting data to redirect users to malicious sites-a very dynamic threat landscape indeed.

The application of Explainable AI (XAI) for capacity building in cybersecurity entails transparent rendering of the decision-making processes undertaken by AI-enabled security systems. By identifying and explaining potential threats, XAI allows prompt and precise modification of defense strategies while assisting in regulatory compliance. By clarifying the internal workings of AI, the efficacy of AI in cyberattack detection and mitigation is enhanced, facilitating forensic investigations of incidents and building threat awareness among users. The integration of XAI capabilities into cybersecurity serves to enhance the detection mechanism, while equally maintaining a solid platform for proactive and reactive defense.

2.3. Emerging Trends of XAI in Cybersecurity

Figure 6 The research directions and opportunities for xAI [12]

XAI is being increasingly recognized for countering elaborate dilemmas in cybersecurity by justifying accountability on AI-driven systems. In [12], it is noted that XAI not only aids in enhancing detection accuracy but also ensures the

operators can validate and trust the automated detections made by AI systems. The paper also describes a systematic framework that would be used to explore future research directions and show opportunities for XAI in the field of cybersecurity; this constitutes an emerging trend in developing AI systems that are both competent and comprehensible and friendly. These trends imply that the most resilient cyber defense mechanism envelops XAI as an honest actor working for the provision of detailed insight into the behavior and decisions of AI models, thereby leading to a better understanding and better security practices.

Figure 6 shows different research directions and prospects for Explainable AI (XAI) and emphasizes the main domains for development to promote transparency, efficacy, and the ethical perspectives of AI applications. These domains include the development of algorithms that can be considered fair or unbiased with respect to various social indicators, data transparency, privacy concerns, and AI empowerment in cybersecurity. It also stresses the user-centered design, multidisciplinary research, and the definition of custom evaluation metrics for the XAI domain under different fields. The infographic suggests, thus, a holistic way forward for XAI, where technical barriers and the aligning widespread societal implications shall be considered.

3. Data Description

3.1. Dataset Overview

The dataset utilized in this study was collected via a SCADA system installed in a wind turbine in Turkey, thus demonstrating its ability to monitor and safeguard critical energy infrastructure. The dataset consists of thirteen different measurements taken with a ten-minute resolution and serves as an important domain of analysis for the performance and security of energy production processes. The ten-minute intervals capture a variety of measurements that are necessary for operational monitoring and cyber threat detection.

3.1.1. The dataset, in brief, is characterized by the following features.

- **Date/Time:** Each record corresponds to the time, being considered in ten-minute steps, and this gives the dataset a sequential time reference that is crucial for time-series analyses. Timeliness here allows transient anomalies that might indicate cyber-attacks or system faults to be detected.
- **LV ActivePower (kW):** This feature tells us how much power was actually produced by the wind turbine. It serves to assess whether the turbine is operating properly, as well as to provide insight on possible system compromise or inefficiency where that production level is unusual.
- **Wind Speed (m/s):** Measured at the hub height of the turbine, this parameter directly influences the turbine's power generation capability. The fluctuations in wind speed data could well assist in differentiating between normal environmental variations and incidents of anomalies that could point to data integrity issues within the SCADA.
- **Theoretical_Power_Curve (KWh):** This curve, according to the wind turbine manufacturer, predicts the amount of power output the turbine should deliver at preceding wind speeds; abnormalities or dropping in comparisons between actual output and theoretical values may point to cyber interference.
- **Wind Direction (°):** This refers to the direction from which the wind is blowing at the hub height of the turbine, thereby influencing the orientation and efficiency of the particular turbine. Any sudden unexpected change in wind direction that goes uncorroborated by any external data source can be an indication of possible unauthorized tampering with the system data.

This dataset's comprehensive nature ensures that researchers have access to both the operational parameters necessary for routine performance analysis and the detailed data required to identify and investigate cybersecurity threats. Hence monitoring the deviations from the expected patterns in terms of power output, wind speed, and turbine orientation, wherein anomalies are not visible through traditionally known monitoring parameters, will enhance cybersecurity against possible threats in the energy sector.

3.2. Feature Engineering

The process of feature engineering for the dataset has included some main steps to effectively prepare data for machine learning model training:

3.2.1. Date Time Conversion

The 'Date/Time' column is converted into a datetime format for easier handling and lessening the workload for time-series data indexing.

3.2.2. Resampling

Resampling analysis of the data was on different frequencies (hourly, daily, weekly, and monthly) to uncover certain patterns on different timescales. The resampling stage helps the identification of long-term trends and effects of time on system behavior.

3.2.3. Loss Calculation

A 'Loss' column is created by subtracting the actual power from the theoretical power to quantify discrepancies that could indicate inefficiencies or anomalies.

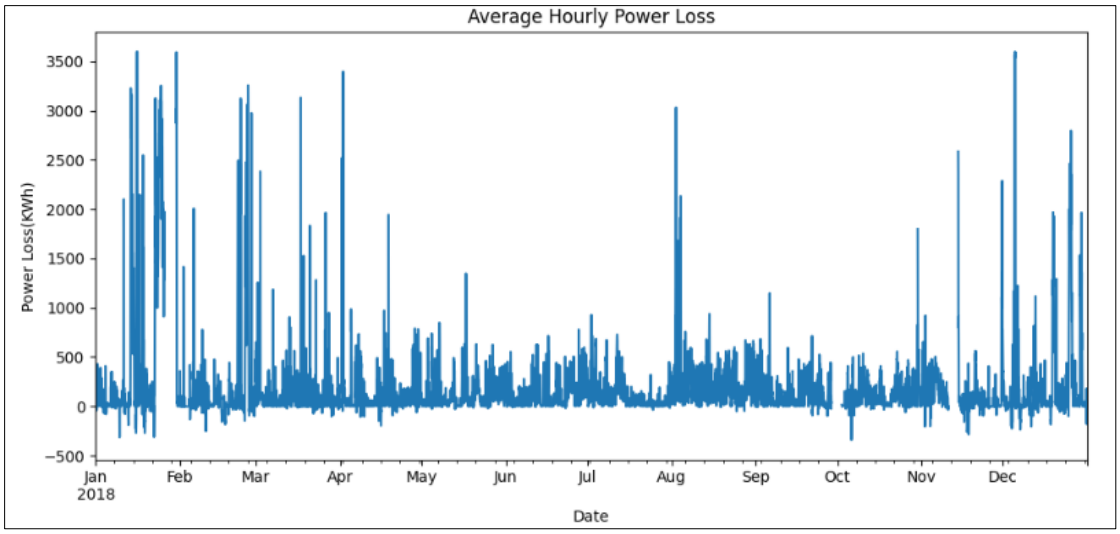


Figure 7 Average Hourly Power Loss

Figure 7 presents the wind turbine kWh (Kilowatt-hour) power lost over the year 2018. Power loss levels are exhibited to be fluctuating and peaking sporadically over the year. During the months of January, March, October, and December, substantial spikes in power loss have been observed. The chart suggests that at times during these months, the power loss crossed 2500 kWh, suggesting possible problems or irregularities with the turbine operation. Variability and occurrence of these peaks might be important for troubleshooting problems related to equipment failure, environmental effects on turbine performance, or perhaps even cyber-attacks affecting the power output.

3.2.4. Categorical Direction Conversion

Initially at directional angles, wind direction is simplified into categorical units, i.e., N, NE, E, SE, etc., hence facilitating the analysis of wind patterns relevant to power output.

3.2.5. Outlier Treatment

Outliers are treated by determining cutoff values based on the interquartile range, downweighting their influence without compromising the robustness of the model.

3.2.6. Normalization

Normalization of variables is applied so that the model is not biased toward variables with very large magnitudes.

3.2.7. Interpolation

The interpolation process is carried out to fill missing values so that the integrity of data is maintained, thus ensuring that the machine-learning models will not be trained on data with missing values.

All of these feature engineering steps are tailored to make the dataset more useful for detecting power-output-related anomalies due to cyber threats, thus providing a solid basis for the predictive model that will help identify potential security breaches in real time.

4. Methodology

4.1. Data Preprocessing

4.1.1. Handling Missing Values

The dataset contains missing entries for different reasons, such as sensor malfunctions and data transmission errors. To handle this, missing values are interpolated using linear interpolation methods that assume a linear transition between the points immediately before and after the missing data. This method best suits time-series data, in which change between consecutive data points can be often assumed to be gradual.

4.1.2. Outlier Detection

Outliers pose a significant risk of corrupting the results of the data analysis and model training. Outliers are attempted to be detected using various statistical methods like the Interquartile Range (IQR) method. In the IQR method, if a data point is below 1.5 times the IQR from the first quartile or above 1.5 times the IQR of the third quartile, that data point is taken as an outlier. Outliers are either rectified or discarded according to the varying situations so that they would not pollute the general analysis.

4.1.3. Normalization

Data normalization is transforming the feature data by scaling it so that it fits into a predefined scale, such as ranging between 0 and 1, or -1 to 1. This is because it is very important for those few machine learning models, which are highly affected by the scale of their input data, that is neural networks. Through normalization, the contribution of all the features used for the final prediction can be considered with somewhat equal importance, while effectively fast-tracking the convergence of the training process.

4.2. Exploratory Data Analysis (EDA)

4.2.1. Statistical Summaries

The initial step in EDA is to generate statistical summaries for each feature consisting of mean, median, mode, standard deviation, and range. The statistical summaries of the dataset reveal very important trends and changes in operational characteristics of a wind turbine over two periods. For Power and Theoretical Power, the increase in both mean and variance from the first period to the second period indicates more efficient or stable operating conditions for the turbine. Wind Speed also has an increasing average speed, which has much less variance, further indicating more uniform wind conditions. Losses between theoretical and actual power output have decreased, both in terms of average and variance, which indicates greater system accuracy and efficiency. Changes in the x_com and y_com characteristics, which probably have to do with wind direction, with a notable change in mean values and a decrease in variance for x_com suggest a change in prevailing wind directions or a change in turbine orientation. This metric is crucial for understanding turbine performance, diagnosing operational issues, and improving predictive maintenance approaches.

4.2.2. Visualization

Different forms of visualization are made use of to better comprehend the data:

Histograms

Used to view the distributions of various features, helping to identify skewness and the presence of outliers.

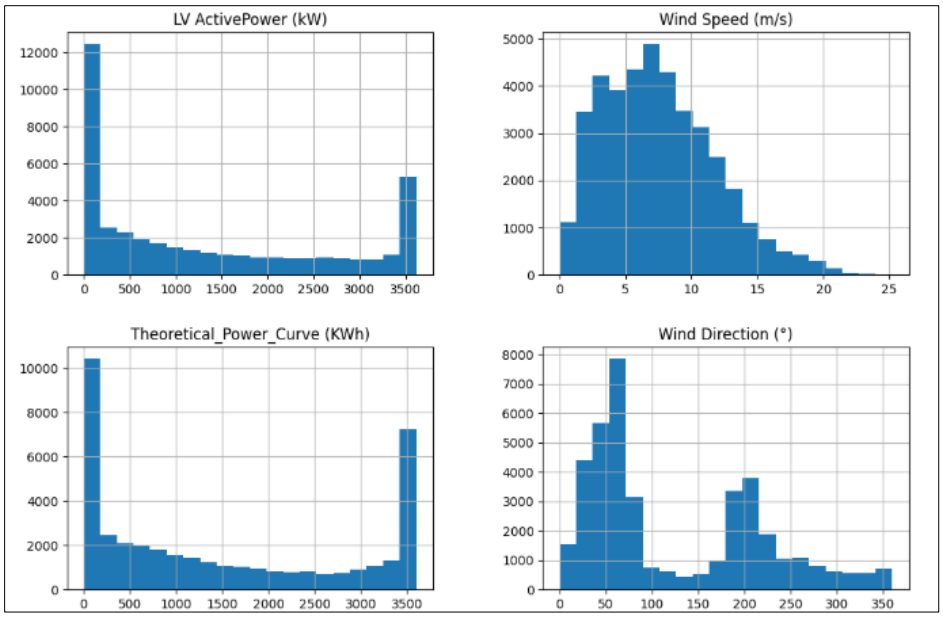


Figure 8 Histograms of numerical columns

The histograms in Figure 8 represent the distributions of four important variables in the dataset for a wind turbine: LV Active Power, Wind Speed, Theoretical Power Curve, and Wind Direction. Here is a summary of the distributions from the histograms:

- **LV ActivePower (kW):** The histogram shows a multimodal distribution having two main peaks, one near the minimum range and the other at the higher range. This indicates that the turbine generally operates at these two power output ranges, presumably in two different operational modes or under two different wind speed conditions.
- **Wind Speed (m/s):** Wind speeds were found to be unimodal, generally low with high prevalence, and somewhat skewed to the right. This means that low winds would occur more frequently; however, there is a longer tail toward higher wind speeds. The mode is around 5-6 m/s, substantiating this as the range of winds most frequently experienced by the turbine.
- **Theoretical Power Curve (KWh):** The Theoretical Power Curve histogram again displays a bimodal distribution, similar to that of LV ActivePower. This represents the theoretical model of power output against wind speed, where the two peaks represent the turbine manufacture's designated most productive wind generation speed.
- **Wind Direction (°):** This wind direction histogram shows a more complex multimodal distribution with multiple peaks that correspond to the preferred wind directions. Some wind directions around the turbine seem to be more typical owing to certain geographical factors and local wind patterns affecting the site.

These histograms are critical for the characterization of the operational and environmental interactions of the wind turbine and lay a foundation for further performance optimization and anomaly detection analysis.

4.2.3. Scatter Plots

Considerable efforts have been made to explore the relationships between pairs of variables, particularly for wind speed and its power output.

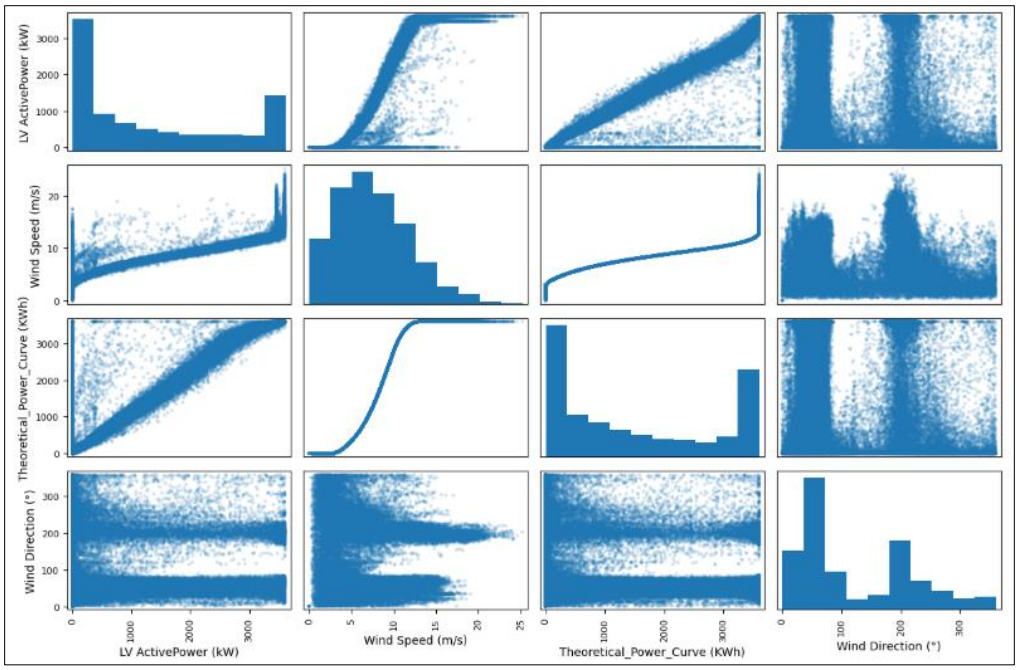


Figure 9 Scatter Plot

Figure 9 presents several scatter plots and histograms from the SCADA data of a wind turbine to examine the relationships and distributions of important variables, namely LV ActivePower, Wind Speed, Theoretical Power Curve, and Wind Direction. The histograms reflect the distribution characteristics of each of the variables: LV ActivePower appears bimodal, suggesting two distinct operating levels; Wind Speed appears to be positively skewed, implying low-speed conditions with occasional very high speeds; Theoretical Power appears to rise very steeply and then flatten, which matches the expected wind turbine power curve; finally, Wind Direction displays a variety of frequencies, suggesting the importance of some prevailing wind directions.

Most importantly, from the scatter plots, LV ActivePower has been positively correlated with wind speed; with wind speed higher, power output will also be higher; Theoretical Power against wind speed gives more or less the turbine output expected as functions of wind speed; and the plots involving wind direction strongly indicate lesser direct impact on power output, but much of its dispersion can represent the effect on turbine efficiency from the different environmental parameters. Visualization is incredibly important in analyzing turbine performance, maintenance decisions, and operational anomalies, therefore leading into wind turbine dynamics at detail.

4.3. Correlation Analysis

Correlation coefficients are calculated to measure the strength and direction of the relationship between pairs of variables. This analysis helps in selecting features that have a significant impact on the target variable, which, in this case, is the power output of the turbine.

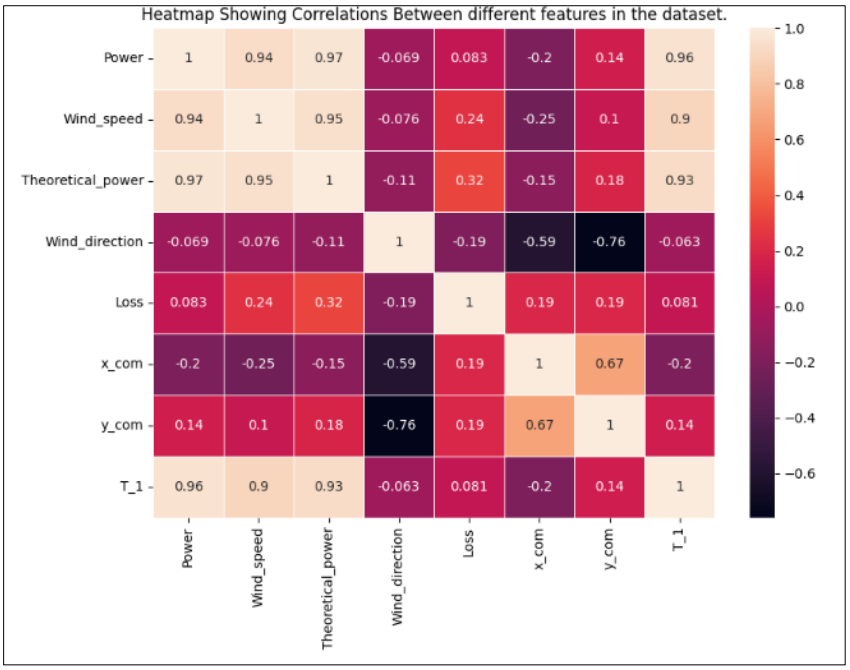


Figure 10 Correlation Plot

Figure 10 presents a heatmap showing the correlation coefficients between different features in a dataset collected from a SCADA system managing a wind turbine. Correlation coefficients range from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and values close to 0 indicate no linear relationship.

4.3.1. Key observations from the heatmap include:

- **Power, Wind Speed, and Theoretical Power**: These features exhibit very high positive correlations with each other (ranging from 0.94 to 0.97), suggesting that as wind speed increases, both the actual power and the theoretical power output of the turbine increase correspondingly.
- **Wind Direction**: Shows generally low correlations with power and wind speed (around -0.069 to -0.076), indicating that the direction of the wind has little linear influence on these variables.
- **Loss**: This feature, likely representing the difference between actual and theoretical power outputs, shows moderate positive correlations with wind speed (0.24) and theoretical power (0.32), suggesting that losses may increase as these values rise.
- **X_com and Y_com**: These could represent components of wind direction or other derived features, showing varied correlations with other features. Notably, they have negative correlations with wind direction (-0.59 and -0.76 respectively), which might suggest a calculation or derivation from the wind direction.
- **T_1**: Shows very high correlations with power and wind speed, indicating it is likely another power-related feature, possibly a time-lagged version of power or similar measurement.

Overall, the heatmap is a very informative tool to gather the relationships between the variables from the dataset and so assist in feature selection and the modeling strategy to predict turbine performance and detect anomalies.

4.4. Machine Learning Models

4.4.1. Model Selection

The analysis of wind turbine performance and anomaly detection is done using a host of machine learning models that have their strengths and weaknesses being tailored to address various concerns of the dataset at hand:

Linear Models

- Linear Regression: Serves as a foundational model in statistical and machine learning applications. It assumes a linear relationship between the input variables (features) and one output variable (target). It is clear, interpretable, and simple, hence the very reason able to use it as a benchmark for assessing the performance of

various complex models. Linear regression fits a straight line (hyperplane in higher dimensions) to the data points that best describe this relationship, whereby the contribution of each feature to the output is made clear.

Ensemble Models

- Random Forest Regressor: This model constructs a number of decision trees during training; each individual tree gives a particular prediction, and the average of all predictions becomes the final prediction. It reduces overfitting risks common in a single decision tree and is very effective for large datasets, providing a higher level of accuracy while maintaining the ability to model complex interactions within the data.
- Extra Trees Regressor: Stands for Extremely Randomized Trees, an ensemble learning method that randomizes cuts in the data more aggressively than the Random Forest. By using the whole dataset to grow the trees and making splits at random rather than the best split among a subset, it usually results in faster training times and can model very complex decision boundaries.
- Gradient Boosting Regressor: Operates by sequentially adding predictors to an ensemble, each one correcting its predecessor. This model is highly flexible and no distribution assumption on the data is required. Each new tree helps to correct errors made by previously trained trees. Gradient Boosting has proven effective in a variety of practical applications, offering state-of-the-art performance on many problems.
- AdaBoost Regressor: Short for Adaptive Boosting, it begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but adjusts the weights of instances that were incorrectly predicted to increase their influence on the next predictor. This sequential attention to the instances that are harder to predict yields improved model accuracy.
- Decision Tree Regressor: Uses a tree-like model of decisions and their possible consequences. It is simple to understand and interpret and can handle both numerical and categorical data. As a standalone model, it can be very sensitive to small changes in the data, leading to high variance, which makes it a good candidate for ensembling techniques.

Advanced Boosting Models

- XGBRegressor and XGBRFRegressor: These models are extensions of gradient boosting designed for speed and performance. XGBRegressor uses a more regularized model formalization to control over-fitting, which provides better performance. XGBRFRegressor combines the random forest algorithm with gradient boosting techniques, introducing randomness into the base model structure and combining this with gradient boosting's powerful optimization.
- CatBoostRegressor: Tailored for handling categorical variables very efficiently, CatBoost avoids the extensive data preprocessing typically required by other algorithms to convert categories into numbers. It also implements symmetric trees that optimize prediction time and model complexity, making it highly effective and fast with a focus on reducing overfitting.

Support Vector Machine

- SVR (Support Vector Regression): Implements the SVM approach for regression problems, where the goal is to fit as many instances as possible between the decision boundaries while limiting margin violations. The SVR shows effectiveness when working in high-dimensional zones along with instances where dimension numbers surpass sample numbers which becomes vital for complex industrial forecasts such as wind speed and power output predictions.

Multiple diverse predictive models were used to study wind turbine dynamical behavior by analyzing linear as well as sophisticated nonlinear connections in the data. Under the precise experimental conditions established by setting random_state=42, the research identifies the most impactful predictors of turbine operation. This predictive method simultaneously delivers better accuracy results and reveals performance-determining factors that maintain critical value for operational enhancements along with preventive maintenance decisions.

4.4.2. Feature Engineering

The modeling process becomes more effective through feature engineering since it introduces valuable data and makes existing patterns more prominent. The procedures of feature engineering generate substantial performance gains for wind turbine power output models by helping them identify subtle variable behaviors and relationships.

4.4.3. Feature Creation

Temporal Features: It is important to incorporate features that accurately represent time as they influence different aspects of power generation. The hour of the day feature, for example, accounts for daily variations in the wind speed as well as the power output. Moreover, the day of the week captures cycles of other periodic variables such as maintenance activities or operational changes. Furthermore, the month feature captures seasonal changes in wind production and power generation that are critical for strategic long-term planning and efficient operations. These temporal features aid in understanding and forecasting changes in power output across different time intervals, improving the model's accuracy and detail, and allowing for more precise scaling examinations.

4.4.4. Interaction Terms

Wind Speed and Direction Interaction: Both the speed and the direction at which the wind is blowing need to be considered when the turbines are being operated. Adding wind speed as a feature and wind direction as another may not provide a comprehensive understanding of their effectiveness in power generation. We can have these features interact with each other by combining wind speed and direction so that the model receives specific information regarding these different propellant aspects. For instance, the set physical parameters of the turbines may require some wind speeds to be more effective while rotating in particular directions. Interaction terms are very useful in capturing such nonlinear dependencies, which may have been missed if these features were considered individually.

The engineered features can now be added to the dataset for training the machine learning models, where the temporal dynamics and interactive effects are well represented in the realm of predictive modeling. This not only enhances the prediction power of the model but also gives a greater understanding of the factors responsible for the power generation dynamics of wind turbines.

4.5. Model Evaluation

4.5.1. Performance Metrics

Performance metrics quantify the effectiveness of machine learning models. These metrics offer objective measures to evaluate and compare different models or to assess improvements in the same model over time.

RMSE (Root Mean Square Error)

RMSE is one of the most commonly used measures in regression. This measures the square root of the average of squared differences between prediction and observations. It gives a sense of the error magnitude in the prediction of the model.

Thus, each error is squared and RMSE gives higher weight to large errors. Therefore, it becomes useful when large errors are especially undesirable. Lower RMSE indicates that a model fits the dataset better. RMSE has the same units as the variable targeted and thus is easy to interpret and communicates meaningfully within the context of the scale of the target.

R^2 Score (Coefficient of Determination)

It is a statistical measure that represents the proportion of variance for a dependent variable that's explained by an independent variable or variables in a regression model.

An R^2 score of 1 indicates that the regression predictions perfectly fit the data. Values closer to 0 suggest that the model fails to accurately capture the essential trends of the target variable. This value is a worse fit than simply predicting the mean of the target value. This metric is particularly helpful for comparing the performance of regression models. High R^2 implies a high explanatory power of the model over the data variability.

4.5.2. Validation Techniques

The validation methods are key to creating a reliable and robust machine-learning model by measuring its performance on an independent set of data.

Cross-Validation

Cross-validation consists of partitioning the data set into several subsets and subsequently training the model on each combination of k−1 subsets while retaining the last subset for testing. Applying this method several times will allow for tests on numerous alternate combinations of subsets.

Cross-validation reduces the chances of overfitting by ensuring that model performance is evaluated systematically on all available data. For this reason, it gives a much more complete view of how well the model performs over various subsets of data, thereby leading to a generalizable model. The more consistent a model's performance upon these walks, the more it will probably be stable and reliable.

Performance Plotting

Performance plots, such as residual plots and actual versus predicted plots, provide a visual means of assessing how closely model predictions correlate with true observed data.

- Residual Plots: These plots display residuals-the differences between observed and predicted values-on the vertical axis against predicted values on the horizontal axis to help identify patterns in the residuals.
- Prediction vs. Actual Comparison Plots: These plots show a scatter plot of predicted versus actual values and provide a good visual representation for judging predictions, where ideally if perfect predictions were made, the points form a diagonal line.
- Significance: These visual methods become extremely powerful tools to point out discrepancies, biases, or variances that are hardly seen through quantitative metrics alone; thus, are information for diagnosing problems of heteroscedasticity, outliers, or model bias.

By quantitative metric assessments with strong validation methodologies, a modeling practitioner ensures modeling performance statistically as well as its credibility and reliability with respect to real-world us.

5. Results

5.1. Model Performance

The evaluation of machine learning models for wind turbine power output prediction through R^2 Score and RMSE measurements showed substantial differences between their performance capabilities.

A model fit assessment occurs mainly through the R^2 score which demonstrates how well the independent variables explain dependent variable changes. The model-based Lithium Proton Exchange Price Assessment process showed the CatBoostRegressor achieved the best R^2 score at 98.256% while the XGBRegressor followed closely with 98.145%. These models effectively identify the majority of variance from the dataset through their exceptional proficiency which makes them top choices for predictive work requiring complex interactions combined with nonlinear relationships. The LinearRegression along with AdaBoostRegressor demonstrate major reductions in their ability to interpret dataset variability because they achieved R^2 scores of 90.607% and 86.687% respectively. Their performance deteriorated because the wind turbine data requires handling nonlinear interactions among various factors better than these models can achieve.

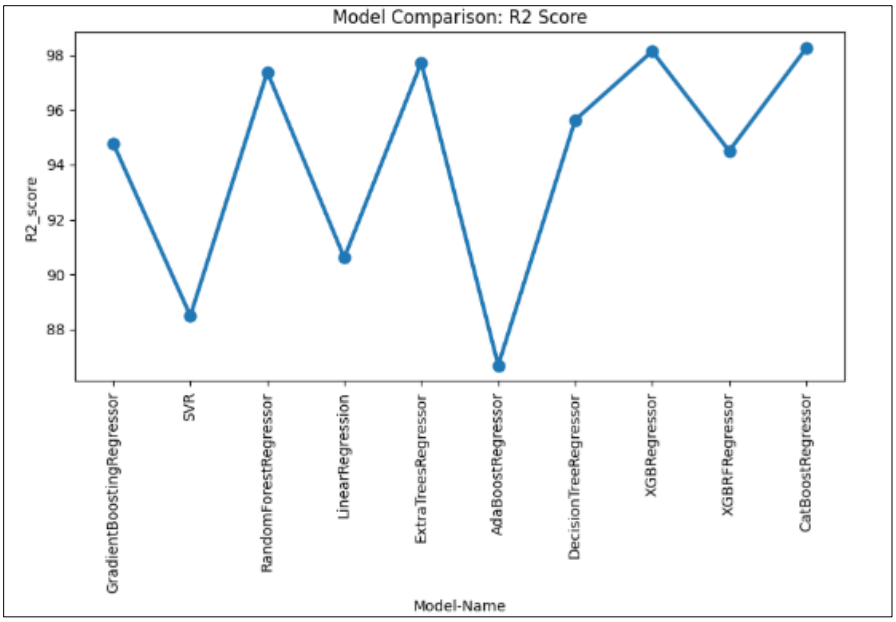


Figure 11 Model Comparison: R2 Score

Overall, the RMSE metric gauges the typical size of prediction error magnitude while extending the R^2 score to show absolute prediction discrepancies between models. Precision levels are evident from prediction accuracy because the CatBoostRegressor (172.501) and XGBRegressor (177.888) outperformed all other models with the lowest RMSE values while leveraging their capability to accurately predict actual values. The regression models LinearRegression and AdaBoostRegressor exhibited lower precision because they produced RMSE values of 400.341 and 476.620 which indicate a significant difference between predicted and actual values. The RMSE values reveal the practical limitations of these models since they fail to adequately represent the complex patterns found in the dataset.

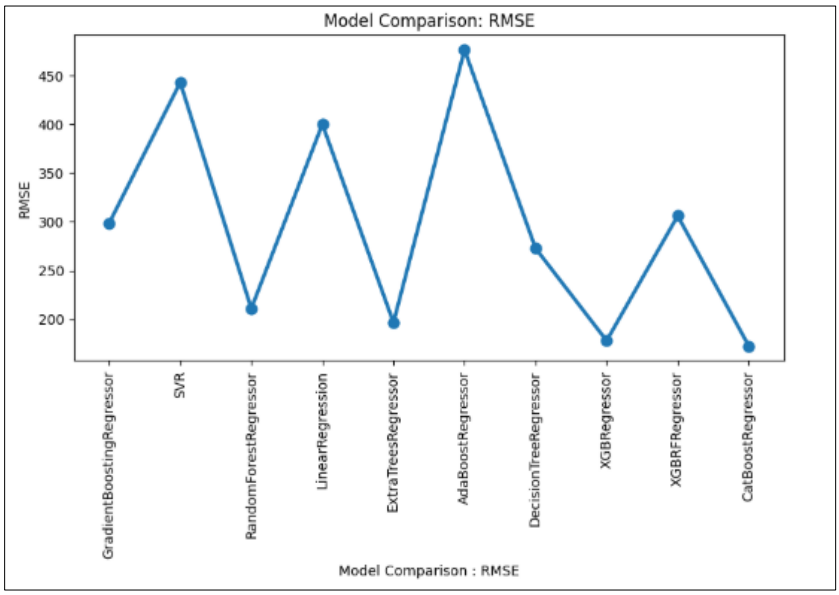


Figure 12 Model Comparison: RMSE

The comparison between R^2 values and RMSE shows us why we need to select the appropriate model through combining fit quality assessment with measurement of predictive accuracy. R^2 evaluates the strength of model data fit against original patterns yet RMSE delivers concrete assessment of prediction errors on average thus providing better practical performance indicators. The predictive models CatBoost and XGBoost demonstrate excellent capabilities for performing complex tasks because they excel at both data adaptation and creating reliable predictions. Models with

simpler or limited capabilities to handle nonlinearity and interaction require upgraded features or additional engineering to develop their ability for comparable tasks.

5.2. Feature Importance

The analysis of feature importance remains vital to understanding the predictive power of each variable in modeling thus helping researchers discover variables that create the strongest impacts on the modeled output. The data collection and feature engineering program gains direction through this method by targeting vital elements that maximize predictive models and performance speed. Understanding feature importance enables better model refinement through guided approaches for both additional data acquisition and feature enhancement activities that result in performance enhancement.

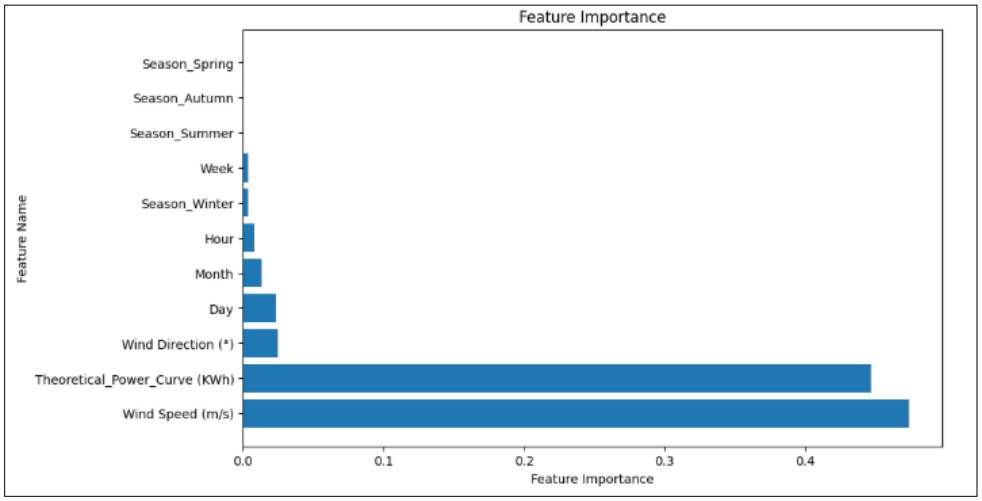


Figure 13 Feature Importance

5.2.1. Key Influential Features

- The Wind Speed measurement at m/s represents the utmost important factor because it defines power output from wind turbines. Wind speed plays a key role as a vital model predictor since it directly affects the generated power output levels.
- The theoretical power curve demonstrates KWh worth of value as the feature which follows immediate ranking after wind speed. The significant importance of this measure shows that real power output closely follows theoretical predictions for power hence helping identify any irregularities affecting turbine performance.
- The angle of wind flow remains significant even though its impact is lower than theoretical power production and wind speed. Turbine power output depends on its wind alignment while utilizing the wind stream because it affects efficiency levels in harnessing wind speed.

5.2.2. Seasonal and Temporal Features

- Temporal Features (Day, Month, Hour): These features include Day, Month, and Hour which measure time differences in wind behavior and output performance. The period during which measurements are taken affects wind speed through thermal changes while the monthly period reveals seasonal patterns in wind patterns.
- Seasonal Features (Season Winter, Season Summer, etc.): The model demonstrates that seasonal variations appear less important than direct measurements of wind speed and theoretical power in forecasting power output. These features enhance the models' capability to research wider time-related operational effects that impact turbine performance.

The feature importance plot shows that wind speed and theoretical power exert high influence on power output predictions of wind turbines. The incorporation of these findings produces better planning methods in operational management and anomaly detection which helps maintain turbine performance consistency throughout diverse conditions. The significance of temporal and seasonal features remains slight, but they offer information that effectively aids both forecasting accuracy and the identification of power output anomalies.

5.3. Shap Analysis

The machine learning explanation method known as SHAP (SHapley Additive exPlanations) values delivers exceptional capabilities to analyze feature contributions in prediction outcomes. The concept of SHAP values originates from game theory to analyze model predictions through a complete comparison of features with and without each other. The method provides both fair attribution and complete insights about output effects through feature analysis thus improving model transparency alongside interpretability.

Each feature contributes to the prediction through SHAP values by quantifying its effect against baseline predictions that use average data values. The SHAP value shows a positive value when the feature enhances the predicted outcome but displays a negative value when it reduces the prediction. SHAP generates an extensive overview of feature impact on model predictions through the collection of dataset-wide values.

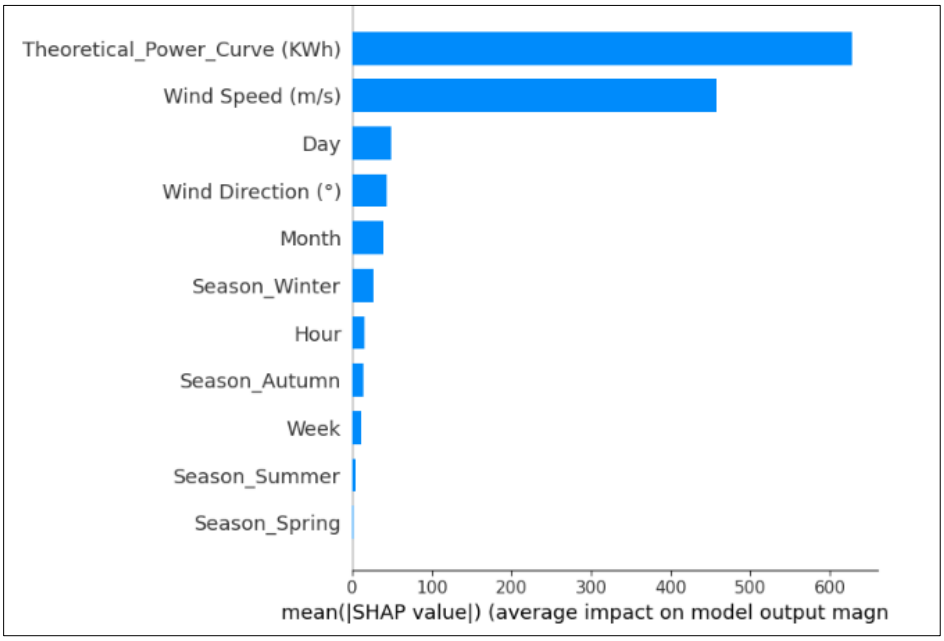


Figure 14 Shapley plot

The SHAP plot shows how the model output changes because of each feature averaged across the entire dataset. The provided plot shows us this information:

- Theoretical Power Curve (KWh) emerges as the feature with the highest SHAP value which establishes its fundamental contribution to prediction results. Higher theoretical power levels produce increased predicted power outputs making it an essential determining factor in prediction accuracy.
- Wind Speed demonstrates a substantial SHAP value which validates its fundamental influence on power generation processes. The power output directly follows the rise of wind speed which follows natural physical rules.
- These temporal variables of Day and Month together with Hour demonstrate a medium level of influence on prediction results through their SHAP values. These parameters indicate that electrical production levels vary according to daytime and month duration and individual day conditions reflecting both operational settings and environmental conditions.
- The Season_Winter and Season_Autumn seasonal indicators have minor effects on model predictions as their respective SHAP values indicate minimal influence compared to wind speed characteristics.

SHAP value analysis reveals the quantitative relationship between features and model predictions allowing users to efficiently understand advanced model responses. The analysis becomes necessary for stakeholders to understand what influences predictive models, particularly in vital applications such as power generation since model decision understanding is essential. The data visualization from SHAP value plots provides critical information that helps refine models to improve both accuracy and reliability throughout predictive analytics systems.

5.4. Cross Validation

Cross-validation stands as a reliable statistical model assessment technique that splits data into partitions then applies models one after another to successively test each subset. The technique reveals how consistent and dependable machine learning models perform between different data sections since this capability enables verification of model independence from data splitting methodology.

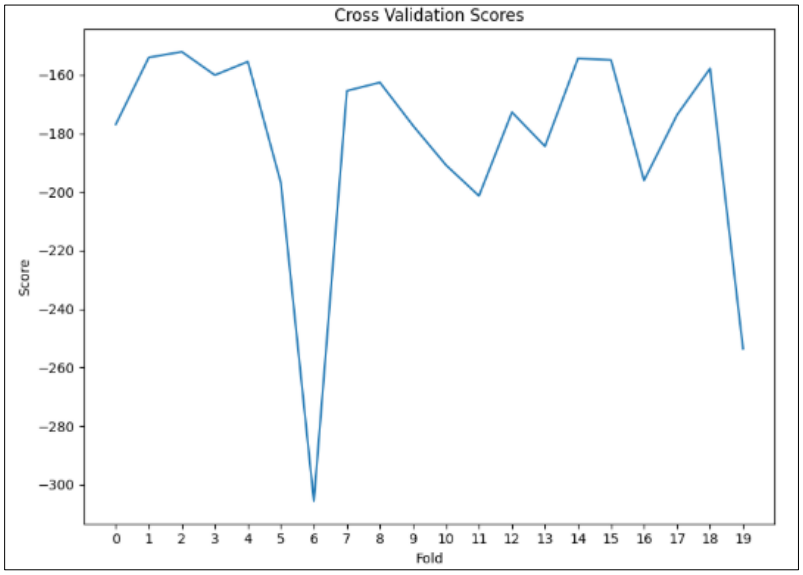


Figure 15 Cross Validation Scores

The performance results for the model span all 20 different folds as presented in Figure 15. The model scores distribute from -160 to -280 points during assessment except for one mid-fold measurement point reaching -280. This variation indicates some variability in model performance depending on the specific subset of data being tested. The substantial dip suggests a scenario where the model might not have performed well, possibly due to outliers, more complex data patterns in that fold, or insufficient model complexity to capture the underlying trends.

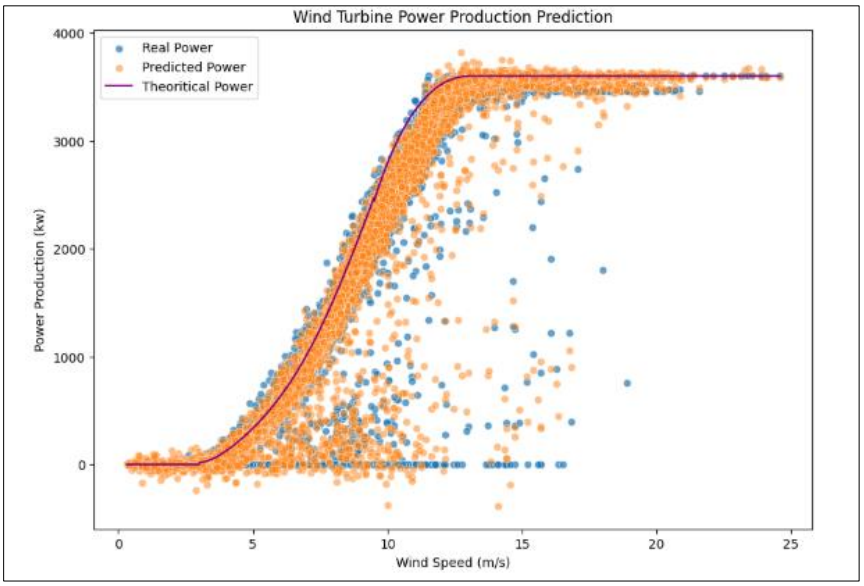


Figure 16 Wind Turbine Power Production Prediction

The relatively consistent scores across other folds, however, suggest that the model generally provides stable predictions. Most of the variability is contained within a 120-point range, which, while notable, does not indicate extreme variance. Such results emphasize the importance of using robust models and possibly enhancing them with

regularization techniques or more sophisticated modeling approaches to reduce sensitivity to variations in the training data.

Figure 16 provides a visual representation of the model's predictions against actual data points across varying wind speeds. The orange dots represent predicted power values, while the blue dots indicate real power measurements. The purple line shows the theoretical power curve provided by the manufacturer.

This plot illustrates that the model predictions closely align with the theoretical expectations at lower wind speeds but start to show greater dispersion as wind speed increases. This spread suggests the model's difficulty in capturing the effects of higher wind speeds on power production, where factors other than wind speed may influence the output, or where the turbine might be operating in a non-optimal regime due to mechanical limitations or safety protocols.

- At an LV ActivePower of 8057 kW, the model predicts 3348.91 kW compared to the theoretical power of 3055.65 kW. This discrepancy indicates overestimation at higher power outputs, where the model may not accurately predict power efficiency losses or turbine cut-out mechanisms.
- For LV ActivePower at 8059 kW, predictions and theoretical power values (2647.51 kW and 2746.12 kW respectively) are relatively closer, suggesting better model accuracy in mid-range operational scenarios.
- A notable case is when the LV ActivePower is 0 kW, where the prediction is at 23.82 kW against a theoretical 779.12 kW, likely reflecting scenarios such as turbine downtimes or maintenance periods not being effectively captured by the model.
- When analyzing a more moderate power output at 14774 kW, the predicted power (1334.57 kW) slightly underestimates the theoretical power (1276.83 kW), suggesting the model might underestimate the efficiency at lower wind speeds.
- At an LV ActivePower of 49374 kW, with predictions and theoretical values closely matched (533.00 kW vs 557.50 kW), the model shows an effective grasp of lower power generation scenarios.

The cross-validation analysis, together with the visualization of model predictions, highlights key aspects of the model's performance and areas for improvement. While the model is generally stable and aligns well with theoretical expectations under typical conditions, the variability in cross-validation scores and the dispersion of predictions at higher wind speeds suggest that enhancing the model's robustness and its ability to handle outliers and high variability scenarios could further improve its reliability and accuracy. Model improvement should include extra features or enhanced modeling techniques alongside model tuning which will help capture wind turbine power production better.

6. Discussion

The evaluation of machine learning models for wind turbine power production predictions produces essential results about SCADA system protection. The industrial automation systems known as SCADA need precise data predictions to operate efficiently and maintain their operational security. The precise forecasting of power output through environmental and operational variables serves as a critical element for SCADA security because it supports maintenance planning and enables anomaly detection to preserve power supply stability.

6.1. Effectiveness of Different Models and Feature Sets

Multiple machine learning models performed differently in evaluation tests according to results which demonstrated CatBoost and XGBRegressor models provided maximum accuracy rates. These models process both non-linear patterns and complicated patterns between features which align directly with typical SCADA system characteristics. These models show excellent performance while dealing with the specific features of wind speed and theoretical power curve which prove their potential to model intricate industrial processes.

6.1.1. Advanced Models: CatBoost and XGBRegressor

The CatBoost along with XGBRegressor algorithms outshine other models because of their remarkable analytic performance which proves effective at handling SCADA systems' normal nonlinear patterns in addition to complex feature relationships. Critical infrastructure systems require precise operation and the capability to detect complex patterns between operational and environmental factors. Advanced features within these predictive models deliver their main strength alongside complex algorithmic capabilities.

- CatBoost has the best features among predictive models because it processes complex data with categorized inputs while bypassing intensive preprocessing stages. This model's performance stands out because it

achieves high accuracy along with reduced overfitting, which makes it ideal for SCADA systems operating under extensive condition variations.
- The XGBRegressor stands out for its effectiveness through its gradient boosting frameworks because it provides both configurability and effective non-linear modeling abilities. Regularization is built into the model framework to stop overfitting which enables its effective use in SCADA systems with dynamic operational areas

The two forecasting methods use ensemble learning principles to merge predictions obtained from distinct algorithms for enhancing precision and reliability. The tested models display strong performance results in the study thus demonstrating their capability to deliver accurate predictions across diverse operational settings that SCADA systems need for stability and security maintenance.

6.1.2. Simpler Models: Linear Regression and AdaBoostRegressor

The performance of Linear Regression and AdaBoostRegressor models fell behind the more complex models implemented in this study. Linear and Boosted regression methods perform poorly in handling SCADA system complexity since they fail to detect nonlinear patterns in such data.:

- The simple nature of Linear Regression yields useful baseline estimations yet fails to capture complex feature relationships effectively. The linear model becomes less useful due to its flexibility restrictions in SCADA systems where precise predictions directly affect operational safety along with efficiency.
- The AdaBoostRegressor model demonstrates enhanced performance over linear regression through its boosting methodology for weak learners but fails to deliver sufficient accuracy in complex real-time applications that demand sophisticated relationship modeling.

6.1.3. Importance of Comprehensive Feature Sets

Wind speed and theoretical power measurement with time-of-day variables form an essential part of comprehensive feature sets which the research demonstrates as crucial for analysis success. Model performance benefits substantially from these specific features which detect critical operational dynamics and time-related patterns that occur in SCADA systems. The assessment considers these variables because they help the predictive models understand SCADA system operational dynamics which often face strong impacts from environmental conditions combined with time-based factors.

The successful results from this study prove that choosing correct models with specialized features best serves SCADA system requirements. The complex analytical abilities of CatBoost and XGBRegressor make them superior choices for handling non-linear processes because they excel with intricate data types. Compared to other models, complex predictive assignments should receive enhanced attention because less complex models either need additional development features or remain limited to basic predictions. The proper distinction between models and features is essential for creating efficient and reliable predictive models operating securely in SCADA systems, thus requiring a strategic method to optimize performance within critical infrastructure environments.

6.2. SHAP Value Analysis and XAI Methodologies

The application of SHAP values enables users to comprehend the exact impact features have on model prediction outcomes. Explainable AI (XAI) delivers key importance to SCADA systems because stakeholders need visibility to confirm automated decisions made through AI-driven remove breaking points. The SHAP analysis demonstrated that wind speed along with theoretical power values play the most dominating role in wind turbine operations according to operational priorities. The identification of these influencing factors enables model improvement together with the creation of clear descriptions for SCADA operators to better understand system security protocols.

6.3. Practical Implications for Real-World SCADA Systems

The implemented research findings deliver considerable value to genuine SCADA system applications in operational environments. High precision power output predictions along with monitoring benefits system resilience by enabling efficient responses toward potential disruptions and performance optimization. Explainable Artificial Intelligence such as SHAP allows organizations to create decision-making trails in their machine learning systems thereby meeting critical infrastructure security requirements and regulatory standards.

Predictive modeling functions enable both preventative maintenance and anomaly detection leading to less equipment downtime and prevention of failures which might cause security breaches. The implementation of models providing

accurate predictions together with interpretability capabilities helps SCADA operators produce decisions based on data-driven principles and human understanding needed to build trust in automated systems.

The research demonstrates that SCADA system security requires both advanced machine learning solutions with strong performance and a complete selection of features. Implementing XAI methodologies improves prediction interpretability and trustworthiness because such methods are essential for deploying AI in security-sensitive SCADA systems. The research demonstrates advanced machine learning methods can boost critical infrastructure operational security while improving efficiency thus building more dependable SCADA system frameworks.

7. Conclusion

The research evaluated multiple machine learning models to assess their ability in SCADA system power output prediction for critical wind turbine operational maintenance purposes. The evaluation showed how CatBoost together with XGBRegressor surpassed basic models because their capability to process intricate non-linear data and multiple operational variable connections produces superior results. The models succeed best in SCADA systems since they deliver precise and dependable predictions needed to preserve system stability and operational efficiency. Explainable AI (XAI) methodologies with SHAP value analysis improves model decision interpretation while simultaneously promoting transparent automation of systems operating in critical infrastructure.

Research and development work follows a distinct direction for the coming years. Future research must target the inclusion of real-time data elements including environmental events and system operational reports because they will improve predictions under SCADA system dynamic operating conditions. Analyzing AI integration with present cybersecurity approaches would establish dual security layers that strengthen both predictive functions and SCADA frameworks. The security and operational efficiency of global critical infrastructure depends heavily on advancing machine learning models while developing extensive and high-performance feature sets during continual research and development.

7.1. Recommendations

7.1.1. Implementing Machine Learning Models in SCADA Systems for Intrusion Detection

The integration of machine learning models into SCADA systems provides major benefits to enhance intrusion detection capabilities. The sophisticated technologies utilize complex pattern analysis to detect security threats by identifying abnormal data behavior which strengthens protective system measures. Warehouse managers must choose appropriate models which integrate smoothly with existing system architectures alongside methodology for maintaining accurate predictions by means of adaptive learning systems.

- Machine learning models CatBoost and XGBRegressor should be selected first for deployment in SCADA systems. Advanced machine learning models excel at processing complex SCADA system interactions because of their ability to detect operational scenarios effectively. The accuracy of detection models improves when developing training data from a wide variety of operational states and all potential system fault conditions.
- Existing SCADA architectures will benefit from integration of machine learning models because developers should create interfaces that work with legacy systems while upholding current cybersecurity standards. The integration process should feature tools enabling real-time data assessment while simultaneously using model-generated information in both operational control measures and threat handling methods.
- The organization should establish standardized processes for applying fresh threat signatures and new data into their model training systems. Regular updating and training of machine learning models will ensure their operational effectiveness through SCADA system progression along with emerging vulnerabilities.

7.1.2. Deploying Machine Learning Models with XAI Features in SCADA Environments

The XAI features that explain the decision-making processes in machine learning models need deployment across SCADA systems to maintain system transparency and accountability. The accessibility and understandability of machine learning processes through XAI increases the trustworthy nature of the decisions operators can make using it. The deployment must emphasize XAI solution integration which reveals AI decision-making mechanisms to stakeholders, so they establish greater trust and embrace AI technology.

- XAI techniques with SHAP value analysis should become a part of your operation to provide transparent visibility into machine learning decision-making processes. The active display of system operation facilitates

trust between operators and oversight agencies since all parties can evaluate automated decision-making methods.

- XAI interfaces require development into accessible interfaces that present explanations in plain terms for operators to use model-based decisions. Interface systems should deliver predication models alongside comprehensive explanations about feature significance for security-related decisions so operators can develop optimal system protection strategies.
- The operators of SCADA systems receive training sessions together with workshops about new AI tools through which they learn to understand and use XAI methodology insights. The educational curriculum will produce a needed link between artificial intelligence innovation and real-world operational needs.

7.1.3. Policy Changes and Enhancements

The Machine learning integration in SCADA systems calls for both policy changes and enhancements to safeguard critical infrastructure. Robust security policies need development to monitor responsible AI technology utilization because they will protect SCADA systems from modern cyber threats. The policies need to promote innovation together with full compliance of implementations to industry standards and protection of privacy and data integrity.

- Standardized AI Security Protocols need development through advocacy actions alongside industry participation for establishing framework standards in critical infrastructure including SCADA systems. The protocols must establish specifications which cover data processing rules and training procedures together with operational outcome goals and security precautions to achieve safe outcomes from AI deployments.
- Implementation of AI in critical systems needs a complete regulatory framework so you must establish partnerships with regulatory bodies. Such a framework must clarify guidelines to defend the security and maintain the integrity of SCADA systems by addressing data privacy and model accountability alongside ethical AI standards.
- The development of advanced AI applications for SCADA systems should receive support through industry partnerships that include both academia and government involvement. The partnership between stakeholder organizations produces better protective strategies against cyber dangers and advanced analytical models together with standardized operational methods which reinforce critical infrastructure safety while boosting operational effectiveness.

Advanced machine learning models and XAI features serve as key recommendations to strengthen SCADA system security and operational efficiency alongside reliability. These measures help infrastructure systems meet current challenges and cyber threats effectively.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] P. B. A. B. P. E. K. J. H. S. K. S. Yulia Cherdantseva, "A review of cyber security risk assessment methods for SCADA systems," computers & security, vol. 56, pp. 1-27, 2016.

[2] S. BELLAMKONDA, "Cybersecurity in Critical Infrastructure: Protecting the Foundations of Modern Society," International Journal of Communication Networks and Information Security (IJCNIS), vol. 12, no. 2, pp. 273-280, 2020.

[3] R. D. Y. O. P. I. G. R. Tsochev, "Key problems of the critical information infrastructure through SCADA systems research," Proceedings of SPIIRAS, vol. 6, no. 18, p. 1333–1356, 2019.

[4] A. G. B. M. Ž. T. A. M. Borja Garcia de Soto, "Construction Cybersecurity and Critical Infrastructure Protection: Significance, Overlaps, and Proposed Action Plan," 2020.

[5] G. K. ·. G. Kumar, "Machine learning and deep learning methods for intrusion detection systems: recent developments and challenges," Soft Computing, vol. 25, p. 9731–9763, 2021.

[6] Y. Z. J. F. A. X. H. CHUANLONG YIN, "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks," Digital Object Identifier, vol. 5, 2017.

[7] H. L. *. a. B. Lang, "Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey," Applied Science, vol. 9, no. 20, 2019.

[8] F. A. R. S. T. E. Rocio Lopez Perez, "Forget the Myth of the Air Gap: Machine Learning for Reliable Intrusion Detection in SCADA Systems," EAI Endorsed Transactions on Security and Safety, vol. 6, no. 9, 2019.

[9] Z. T. G. A. C. K. Tolgahan Öztürk, "Machine learning based intrusion detection for SCADA systems in healthcare," Network Modeling Analysis in Health Informatics and Bioinformatics, vol. 11, no. 47, 2022.

[10] K. O. A. M. A.-M. S. R. a. K. O. A. A. Oyeniyi Akeem Alimi, "A Review of Research Works on Supervised Learning Algorithms for SCADA Intrusion Detection and Classification," Sustainability, vol. 13, no. 17, 2021.

[11] R. H. J. S. B. S. P. R. P. K. R. M. G. Y. J. G. H. M. A. T. R. G. GAUTAM SRIVASTAVA, "XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions," ACM Computer Surv, vol. 1, 2022.

[12] A. P. R. K. M. C. Marek Pawlicki, "Advanced insights through systematic analysis: Mapping future research directions and opportunities for xAI in deep learning and artificial intelligence used in cybersecurity," Neurocomputing, vol. 590, 2024.