scientific reports

Check for updates

OPEN

Advanced AI-driven techniques for fault and transient analysis in high-voltage power systems

Abdul Aziz[1,2,12], Muhammad Zain Yousaf[3,4,5,12], Feng Renhai[2]✉, Wajid Khan[2], Umar Siddique[6], Mehran Ahmad[6], Muhammad Abbas[2], Mohit Bajaj[7,8,9] & Ievgen Zaitsev[10,11]✉

Each substation is critically essential to the overall operation of the electrical power system. Potential dangers include thermal stress, noise, slip, trip, fall hazards, animal waste, and nonionizing radiation. These are the causes of joint failures of cables and overhead lines, failure of one or more phases of circuit breakers, and melting of fuses or conductors in one or more phases. These kinds of failures bring about a decline in the substation's level of dependability. On the consumer side, power cannot be received adequately because there are losses in the transmission line. To accomplish the objective of enhancing the voltage profile, the DG must be optimized. The substation's transient analysis utilizing a variety of factors, a study of faults and transients that occur in the substation and their effects using ETAP, and an optimization of a range of parameters using artificial intelligence techniques are all used for this goal. This paper offers the complete simulation of a 500kv substation. The simulation uses advanced software Electrical Transient Analyzer Program (ETAP) with detailed load flow analysis and short circuit study of the 500 kV substation system using ETAP software. From the ETAP-generated load flow details and the short circuit details, which are studied by varying loads or other parameters, these whole simulations are carried out multiple times using real-time data from the past eighteen months. A simulation data set contains data on both standard and different faulty conditions. In the 1st step, the normal and faulty conditions are classified. In the 2nd step, the reasons for fault occurrence include line-to-line, line-to-ground, and double line-to-ground using the Artificial Intelligence technique. In both steps, Catboost performs well, followed by Support Vector Machine and Logistic Regression. In the first step, Catboost classifies normal and faulty conditions with an accuracy of 98%, SVM is 96%, and Logistic regression is 93%. Again, in the 2nd step to identify different faulty conditions, the accuracies of Catboost SVM and Logistic Regression are 97%, 95%, and 92%, respectively.

Keywords Load flow, Short circuit analysis, Radial distribution system, Artificial intelligence, Catboost

Sustaining a continuous power supply for end users is vital, especially in contemporary cultures where industrial and household activities depend primarily on electricity. The basis upon which this constant power supply is based is the dependability and stability of high-voltage power systems. These systems have to regularly run within specified limits to avoid outages and other disturbances with possible significant societal and financial effects. Identifying and classifying faults in these systems have become more critical than ever, given the growing

[1]CECOS University of IT and Emerging Sciences Peshawar Kpk Pakistan, Peshawar, Pakistan. [2]School of Electrical and Information Engineering, Tianjin University, Tianjin, China. [3]Center for Renewable Energy and Microgrids, Huanjiang Laboratory, Zhejiang University, Zhuji 311816, Zhejiang, China. [4]School of Electrical and Information Engineering, Hubei University of Automotive Technology, Shiyan, China. [5]School of Aeronautics and Astronautics, Zhejiang University, Hangzhou 310027, China. [6]Computer Systems Engineering, UET Peshawar, Peshawar, KPK, Pakistan. [7]Department of Electrical Engineering, Graphic Era (Deemed to be University), Dehradun 248002, India. [8]Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan. [9]College of Engineering, University of Business and Technology, 21448 Jeddah, Saudi Arabia. [10]Department of Theoretical Electrical Engineering and Diagnostics of Electrical Equipment, Institute of Electrodynamics, National Academy of Sciences of Ukraine, Beresteyskiy, 56, Kyiv-57, Kyiv 03680, Ukraine. [11]Center for Information-Analytical and Technical Support of Nuclear Power Facilities Monitoring, National Academy of Sciences of Ukraine, Akademika Palladina Avenue, 34-A, Kyiv, Ukraine. [12]Abdul Aziz and Muhammad Zain Yousaf contribute equally to this paper. ✉email: fengrenhai@tju.edu.cn; zaitsev@i.ua

complexity of electrical grids resulting from integrating renewable energy sources and developing linked networks.

Because of developments in simulation tools, data analytics, and machine learning algorithms, conventional fault detection and classification approaches have significantly changed. More complex and precise fault detection systems, essential for preserving the integrity of high-voltage power systems, have been made possible by these technical advancements. This work emphasizes using modern machine learning techniques in concert with ETAP (Electrical Transient Analyser Program) for system design and simulation. The objective is to present a complete fault diagnosis and classification method in a 500 kV high-voltage power system. High-voltage power systems may develop several types of faults. With significant consequences related to this, these problems can substantially compromise the dependability and stability of the electricity grid. Among the various causes that could lead to them are malfunctioning equipment, harsh weather, or even human mistakes. These defects must be discovered quickly and precisely to minimize damage and ensure the power system runs uninterrupted. In recent years, increasing the accuracy and efficiency of fault detection systems with simulation-based analysis and machine learning models has gained increasing interest[1–3]. I constructed a 500 kV power system using ETAP, a potent instrument primarily used to operate and research power systems. Extensive load flow and short-circuit analysis made feasible by ETAP are essential for modeling complex electrical networks. These simulations generate the data required to train machine learning models under normal running conditions and failure scenarios. While the load flow analysis assures that the operational parameters of the system stay within tolerable limits under normal conditions[4], the short-circuit analysis simulates several failure scenarios to collect the needed fault data.

Because of its unique ability to manipulate large amounts of data and learn from patterns in the data, machine learning has emerged as a potential solution for diagnosing problems in energy systems. In this work, three different machine learning methods—Catboost, Support Vector Machine (SVM), and logistic regression—help us classify many types of problems in a 500 kV power system. The method chosen for this study is due to the effectiveness of recent studies in solving problems and the accuracy of power systems[5,6].

Logistic regression is especially useful for binary classification problems and is a statistical model that can identify trends. This relatively useful tool enables one to study the fundamental aspects affecting fault conditions in power systems in a simple and explainable way. In the framework of this work, the logistic regression method has been used to distinguish between normal and abnormal operating conditions in a 500 kV system. Its ability to provide such probabilities makes it a popular choice for early fault diagnosis, as it generates probabilities of different failure modes depending on the input data. This method leverages another machine learning technology, SVM. Widely used in fault detection and other related fields, SVM is well known for its flexibility and ability to analyze high-dimensional data. To effectively separate many classes, SVM searches the feature space for the best available hyperplane, increasing the margin between classes. SVM is very useful in power systems because it can accurately distinguish normal and fault states, even in cases where the boundaries between classes are not clearly defined.

The current study also uses CatBoost, a relatively recent addition to gradient-boosting techniques. CatBoost's remarkable success in various machine learning competitions and its effective handling of categorical data are well known. Since CatBoost can automatically handle categorical data with minimal training, it is a good candidate for solving power system problems. Furthermore, its overload resistance and ability to explain complex data relationships add to its power to accurately classify different types of faults in the 500 kV system[2].

This study provides the data to evaluate training and program strategies using the ETAP intervention. By combining multiple failure modes and normal operating conditions, the data can provide a complete picture of what will happen at the power plant. The training, testing, and validation techniques were also identified to prepare the data set. The preparation can be tested on data external to the model and trained under different conditions, thereby increasing the complexity and robustness of the model.

After training, precision and accuracy are standard metrics of machine learning model performance. With enhanced detection and classification capabilities, the new provides a comprehensive understanding of the energy model of the system. Examining the model's accuracy, precision, and recall can help assess its ability to routinely detect diseases without false positives or negatives, just like in software engineering. Determining the harmonic maximum between two numbers[7,8] generates the F1 score. This particular metric aggregates recall and precision. ETAP provides a consistent framework for modeling faults and normal situations in power plant management. Using these simulations, training machine learning models enables the identification and categorization of complex problems. Traditional methods were used before for statistical analysis. Still, machine learning techniques are awe-inspiring due to the many adverse effects of artificial intelligence and machine learning (ML) over these due to the ability to access large data generated by power plants. In addition, this approach allows us to use the data to discover relationships and trends that are not obvious using traditional analytical tools. With data-driven instruction, machine learning models can be continuously improved to improve accuracy and error detection.

This article presents the analysis study of a 500 kV Power plant. By utilizing the machine learning ML approach with ETAP, better results were obtained than with traditional methods.

(a) The algorithm used for the analysis classifies these faults more decently and accurately. Since it enhances the consistency and reliability of high-voltage power systems, this factor is quite essential. The combination of ETAP and ML classifies and finds the faults in this study and presents the complete method. It is very relevant to many power plants. This strategy may help lower the frequency and effects of power outages, improving the general power system resilience.
(b) Furthermore, this work may inspire the development of creative defect identification and categorization methods. The need for sophisticated defect detection techniques will only increase as power systems devel-

op and become more complex. This work presents a road map for developing sophisticated flaw detection systems, guiding further studies in this field. The difficulty of seeing faults mainly results from the necessity to compromise speed, accuracy, and simplicity of knowledge.

(c) Though they can occasionally be seen as opaque, machine learning models offer little insight into the fundamental mechanisms producing their predictions. Lack of interpretability in essential infrastructure like power systems might provide a significant obstacle to accepting fault-detecting systems grounded on machine learning. Using a simple and readily available approach, regression analysis, let us effectively solve this problem in our research. Using our method, we provide a structure that satisfies spatial and comprehension criteria and helps to find faults precisely.

(d) The need for interpretability in power system failure detection cannot be emphasized too often. Operators and engineers who want to make wise judgments on mistake elimination must be strong masters of the basic ideas guiding a model. This understanding is vital in critical events where choosing the wrong action might have significant consequences, including widespread power outages or damage to essential infrastructure. The primary factor is the speed of detection of faults and faulty parts, which are beyond comprehension. High-voltage power systems run in real time. Hence, any delay in fault detection can have significant effects. Machine learning models are often well-suited for real-time uses because they can rapidly handle vast volumes of data. Still, the model used could affect the detecting speed. In this work, we combine SVM with CatBoost to guarantee our system's accurate and quick fault detection. While CatBoost provides fast training and inference durations, SVM is well-known for its capacity to handle high-dimensional data effectively. For real-time defect detection, SVM is, therefore, the best option.

(e) Examining several fault kinds in the power system is vital for this work. High-voltage power systems have several faults, each with unique characteristics and effects. Whereas line-to-line faults involve a short circuit between two phases, line-to-ground faults arise when one phase contacts the ground. More complex double line-to-ground faults are typified by two stages involving the ground. Every one of these defects presents various challenges for classification and detection; thus, models that can precisely distinguish between them are needed. In the current article, we concentrate on three primary forms of faults: double line-to-ground (LLG), line-to-ground (LG), and line-to-line (LL). Common in high-voltage power systems, these faults could affect system dependability and stability. Our models may give operators useful information by precisely categorizing these defects, helping them take the necessary action to reduce the effect of the problems. This capacity is crucial in the framework of renewable energy integration, where the fluctuation of generating sources might provide extra difficulties for defect detection.

Integrating renewable energy sources into power systems is becoming increasingly important as the demand to reduce greenhouse gas emissions and support home power generation rises. Still, the erratic and unknown character of renewable energy sources such as sun and wind create additional challenges to the long-term survival of energy systems. In this case, maintaining dependable control becomes much more critical as any faults might perhaps cause the bridge to collapse totally. This study presents an ideal approach for tackling these difficulties as it provides a dependable and accurate instrument for spotting and classifying problems with renewable energy sources. This study finds a strong basis for using ETAP for system design and simulation. Widely acknowledged as a superior tool for power system analysis, ETAP's variety of features makes it ideal for simulating complex electrical systems. Here, we use ETAP's ability to precisely replicate a 500 kV power system in steady-state and off-state environments. These simulations guaranteed that our machine learning models were ready to manage the motion of actual energy systems, generating enough data for their training and testing. Research accuracy and precision are much enhanced by using ETAP. Fault analysis depends on a strong knowledge of the operational traits of a power system under normal conditions. Shows understanding of the parameters required for the system to run within certain times: voltage levels, energy flows, and material losses. Using fault tolerance analysis helps one to examine the system's behavior under steady conditions, much like in an electrical engineer's case.

It produces a wide range of failure modes and provides the required information for fault detection training of machine learning models. Apart from its outstanding analytical capacity, ETAP provides a spectrum of instruments to evaluate analytical findings. These inspection instruments are pretty helpful in pointing out possible areas of concern and providing a thorough knowledge of power system properties in many different contexts. Using the means of our study, we improve the accuracy of our machine learning models and provide a thorough understanding of several faults using ETAP prediction techniques. In hybrid power systems, fault diagnosis becomes a very effective technique using fully integrated machine learning models and ETAP simulations. ETAP provides a quick and exact power system performance analysis. Using this information, machine learning techniques improve the precision in fault state prediction and trouble spot identification. This research provides a comprehensive analysis of the key elements causing the lack of application of maintenance and recovery strategies, therefore improving the security and safety of the power system. The knowledge acquired by the research has more general relevance for power system analysis and capacity distribution choices. Given the present tendencies in renewable energy sources and the spread of linked grids, power systems are changing. Consequently, the need for fault tolerance methods will only grow. The method described in this paper provides a complete and exact means for spotting and fixing problems in hybrid power systems, hence advancing next-generation needs-finding systems. In a 500 kV high voltage power system, which is under consideration as an input for the research, this work offers a complete method for problem identification and analysis. It uses ETAP capabilities for system analysis and prediction and machine learning approaches. This study provides a quick and easy method to investigate and repair integrated power grid faults. It considers the first either normal or faulty, and then It generally finds three types of faults: double line-to-ground (LLG), line-ground (LG), and line-to-line (LL) because of the data provided by these faults. Three machine learning techniques, Boosting, Support

Vector Machine (SVM), and Catboost, are used to classify the faults. The research results can help improve the reliability and stability of power systems, thus impacting their preparation and energy management.

Literature review

Design and analysis of electrical power systems extensively rely on simulation technologies such as ETAP. Detailed modeling of complicated networks made possible by ETAP helps load flow analysis, fault simulations, and short-circuit study. These features are essential for creating complete datasets for defect detection. Machine learning models may be trained. The efficiency of simulating several failure scenarios in high-voltage systems using ETAP was not enough because only analysis was required to determine whether it was a fault. The segmentation of the faults, like unsymmetrical faults, should be considered; the method used meanwhile only focuses on the faults or the response time of the machine learning model[9]. to the fault[10].

In the meantime, much research has either concentrated on traditional fault detection methods or has employed few datasets, thereby not entirely using ETAP's simulation features. This work closes the discrepancy by training machine learning models on ETAP-generated data. This guarantees that the models experience a broad spectrum of failure situations, enhancing their prediction accuracy and resilience.

Recent advances in machine learning techniques to power system fault analysis have contributed much. The paper is titled "Enhancing HVDC Transmission Line Fault Detection Using Disjoint Bagging and Bayesian Optimization with Artificial Neural Networks," where it discusses the detection of faults in high voltage direct current (HVDC) transmission lines by applying AI techniques, thereby exploring very little Bayesian optimization on real noisy datasets.

Incorporating AI into the ETAP allows this work to make its model more potent in dealing with fault analysis for noisy real-world datasets compared to using Bayesian techniques in isolation. High-voltage power networks must be used if effective power distribution across large areas is to occur. Though widely used, conventional object identification methods such as convolutional networks and traffic flow models struggle to fit current applications. The integration demands more sophisticated fault detection methods of renewable energy sources and grid network expansion[8].

The article "Machine Learning Applications in Power System Fault Diagnosis Research Advancements and Perspectives " also offers an overview of ML applications within the context of fault diagnosis on power systems, focusing on the need for comprehensive datasets corresponding with real-world operational conditions. To bridge this gap, we apply ETAP to more accurately generate exhaustive datasets mirroring these conditions. Liu et al. underlined, for instance, the need to integrate data-driven approaches with conventional techniques[4,11]. Notwithstanding these developments, research on integrating powerful machine learning models with simulation tools such as ETAP (Electrical Transient Analyser Program) with fault detection in high-voltage systems is significantly lacking.

While machine learning can recognize complicated patterns and manage vast amounts of data, its application in fault detection has acquired pace. Among the most often applied machine learning techniques in this field are logistic regression, SVM, and Catboost. Logistic regression is a basic but effective method for binary classification problems, that is, between normal and faulty conditions. Particularly helpful for categorizing various kinds of errors, SVM—known for its efficiency in high-dimensional space, uses categorical data and training on big datasets; CatBoost, a gradient-boosting method, has demonstrated exceptional performance[12,13].

Both SVM and Logistic Regression were used to aim for fault classification in power distribution networks; nonetheless, SVM showed better accuracy than Logistic Regression[2]. The benefits of CatBoost are that, especially in cases involving large datasets that are imbalanced, it has better accuracy in defect identification than current methods[14,15]. This study aims to close a gap in the literature utilizing thorough experiments combining ETAP-generated data with these machine-learning algorithms. Despite these positive findings, the literature lacks any similar research. Effective fault identification depends on the careful use of machine learning algorithms. Logistic regression, support vector machines, and boosts are suitable for various activities depending on their benefits and drawbacks. Comparative research can provide perceptive details about their performance in several contexts. A comparative study included Logistic Regression, SVM, and CatBoost for fault identification in a 220 kV transmission line. The research found that whereas logistic regression offered simplicity and interpretability, SVM was better at managing multi-class classification problems. Catboost, however, produced the most remarkable accuracy—especially for large datasets[16–18].

Though this comparison is enlightening, no research has been done contrasting these techniques with data generated by simulation tools such as ETAP. This work bridges this gap and provides a more complete knowledge of each model's fault detection capabilities by evaluating the efficacy of Logistic Regression, SVM, and Catboost using ETAP data.

Fault-type classification is crucial for identifying issues with power systems and guarantees quick corrective action. Common varieties of faults include line-to-ground (LG), double line-to-ground (LLG), and line-to-line (LL). To maintain system dependability and avert additional damage, one must precisely find these defects.

Studies on defect classification have examined several machine-learning techniques. Because SVM can manage non-linear relationships in data, it has proved quite helpful in differentiating between complex fault types. Logistic regression is useful in systems that must make judgments fast, even if they are simpler, as it provides a straightforward way to classify mistakes. Catboost has great classification accuracy for different fault types[19]. It works remarkably well in conditions involving large, imbalanced datasets using its gradient boosting method. The research still lacks, in the meantime, on the integration of these methods with ETAP simulations for fault-type classification. This work closes this gap by training and testing the algorithms using ETAP-generated data, enhancing their accuracy in classifying many failure types in a 500 kV system. In finding faults, there are still many challenges, even with significant development. One of the primary challenges is the need for large-scale datasets that accurately show the broad spectrum of fault situations that could arise in practical power

systems. Although ETAP can replicate a broad spectrum of events, more research is needed to ascertain how successfully these simulations may be used to train machine-learning models[8,20].

Another challenge is enhancing the performance of machine learning models to enable their management of the complexity of modern power systems. Many of the present versions either lack design concerning the particular needs of high-voltage systems or are too simple. Future studies should aim to produce computationally efficient models with accuracy. The study leverages ETAP simulations to provide a complete dataset with a broad spectrum of fault scenarios, overcoming these difficulties. After that, Logistic Regression, SVM, and Catboost models are trained using the dataset, offering a strong and accurate fault-detecting mechanism for high-voltage power systems.

Techniques for locating faults in distribution networks include traveling wave, intelligent, impedance, and time domain approaches. All these methods have disadvantages, however. Traveling wave-based techniques may encounter issues with high sampling frequencies, intricate structures, and database requirements[2]. , whereas intelligent techniques may encounter issues with intricate structures and the need for a sizable and precise database[21–23]. To extract characteristics for different fault kinds, Refs[24–29]. Employ a wavelet transform in conjunction with an artificial neural network. The goal is to identify the precise position of the fault and the problematic part. This method's primary disadvantage is that it depends on the network structure being studied; if there is even a slight change in the structure, fresh training data must be created to train the artificial neural network.

A literature review reveals a considerable discrepancy in integrating simulation tools such as ETAP with machine learning techniques for fault identification. Most current research has either employed restricted datasets or concentrated on conventional fault detection methods, neither of which adequately reflects the complexity of contemporary power networks. Furthermore, although much research on machine learning-based defect detection has been done, not much has looked at using ETAP-generated data. This article fills these voids by suggesting a fresh approach combining powerful machine learning methods with ETAP simulations. Using ETAP-generated data guarantees that the machine learning models are trained on an extensive dataset reflecting the spectrum of fault circumstances that could arise in a 500 kV system. After that, the performance of logistic regression, SVM, and Catboost is evaluated to understand their fit for fault detection in high-voltage power systems. Three times are the enhancements made by this work:

- It shows how well merging ETAP simulations with machine learning techniques detects faults.
- It thoroughly compares Logistic Regression, SVM, and CatBoost, stressing their advantages and drawbacks.
- It offers a road map for further studies in this field, pointing out essential obstacles and possible fixes for raising the fault detection system accuracy.

Particularly with machine learning approaches, the field of fault identification in power systems has seen remarkable development. Still, research incorporating these approaches with simulation tools like ETAP is greatly needed. This effort intends to meet this demand by using data given by ETAP to train and assess advanced machine-learning models. The findings of this work have significant consequences for the building and operation of high-voltage power systems, primarily in terms of their reliability and capacity to resist breakdowns. Table 1 below shows some of the comparisons between the previous methodologies and the current I have done a lot of work to carry out the best results due to the need for fault detection in the power system. It's a single case we have used it for 500kv as a study we can use it on any system.

Materials and methods

The material methodology of this system shows the load flow and transient analysis of the 500kv Bus bar system of Tarbela with the application of ETAP and AI techniques. First, the real-time data is collected from the Tarbela power station. The system is designed using ETAP taking the input of the data collected from Tarbela, the load flow analysis of the 500kv system is carried out after the load flow of the multiple faults generated that was being occurred from 2020 to 2022, these faults generated by varying loads on the bus tripping generators, from ETAP we got two data set one was the data in which the system was steadily in the expected condition (Normal data) the other was the data from all the faults (Faulty data), using both a data set is generated. The overall data was split into train, test, and validation; here, 70% of the data was used for training, and the remaining 30% was used for testing and validation, 15% each. This data is trained with different AI models like logistic regression (LR),

S#	Contribution	Research gap	Comparison with this work
1	Fault detection in HVDC transmission lines using AI techniques [4].	Limited exploration of Bayesian optimization in real-world noisy datasets.	Integration of AI with ETAP provides a more robust framework for fault analysis in noisy, real-world datasets compared to Bayesian approaches alone.
2	Overview of ML applications in fault diagnosis within power systems [8].	Need for comprehensive datasets that reflect real-world operational conditions.	Used ETAP to generate extensive datasets that mirror operational conditions more accurately, addressing this gap.
3	Survey on ML methods for fault detection in power transmission lines [5].	Gaps in real-time deployment of ML models and their scalability in operational settings.	Demonstrates a practical implementation of AI with ETAP for real-time data, and scalable fault detection solutions.
4	Exploration of ensemble models in detecting losses in smart grids [6].	Challenges in adapting models to different grid architectures and varying data quality.	Our methodology potentially adapts more dynamically to various grid architectures due to the comprehensive simulations from ETAP.
5	Use of CatBoost in detecting electricity theft within power systems [7].	Insufficient models for handling non-technical losses in diverse geographic and system contexts.	Extends the application of AI and different AI algorithms in power systems beyond theft detection to comprehensive fault analysis, showing versatility in different contexts.

Table 1. Comparisons table.

support vector machine (SVM), and Catboost. These three models were used to predict the faults. Therefore, the classification was first made into two classifications (Normal, Faulty) and then into three classes (Line to ground, Double Line to Ground, and Line to Line). So, on the previous data, the system is trained whenever the faulty condition comes over to tell whether it is healthy or faulty or which fault it is. The proposed methodology for this research is illustrated in Fig. 1. It provides a systematic approach to analyzing the 500 kV substation using ETAP and artificial intelligence techniques. The flow starts from data collection, ETAP simulation, and data preprocessing to model training and evaluation.

Subject area (500kv)

The 500-kilovolt system of the Tarbela hydroelectric power plant, located close to Islamabad on the Indus River, is now operating. It has a sizable reservoir that covers a surface area of 250 square kilometers, making it the second-largest reservoir in terms of volume in the world. This power plant's construction was finished in 1976, and currently, it has a 3,478 MW installed capacity. The assembly ring, excitation machinery, auxiliary equipment, and secondary equipment of the Tarbela HPP are now being renovated. Units #5 and #6's static excitation systems underwent renovations by ANDRITZ HYDRO in 2014. Static excitation systems will now be placed on rotating exciter units #1 through #4, while brand-new static excitation systems will replace the current ones on units #7 and #8. The agreements for units #1–#4 and #7–#8 are evidence of ANDRITZ Hydro's standing in the hydropower sector of Pakistan. All six units will be delivered gradually beginning in 2016. It also intends to increase the power-producing capacity by adding units 9–14. A 500kv and 220kv systems, totaling 14 units, comprise the system's two components: No of buses = 2 main buses.

The data used for this analysis was acquired from the monitoring unit of the Tarbela hydropower project, which is under the management of Pakistan's Water and Power Development Authority (WAPDA). The study concentrated on information gathered between 2020 and 2022, focusing on variables including voltage, current, voltage angle, current angle, and power factor. These variables were extracted from historical daily records. This data is used to design the model through ETAP.

Design through ETAP

The dataset was produced using data provided by the Tarbela power plant. The whole design of the present substation was developed in ETAP, as shown in Fig. 2. To guarantee the effective and secure functioning of a 500 kV substation, specific essential actions must be taken during the design phase utilizing ETAP (Electrical Transient Analyzer Program). The design process begins with a single-line diagram, including load flow and short circuit analysis.

The foundation is a single-line diagram that shows the general architecture of the substation as well as its key elements, including transformers, circuit breakers, switches, and transmission lines. After placing the schematic, a load flow study is carried out to determine how electricity moves throughout the substation under steady-state circumstances. This study aids in identifying voltage profiles, possible bottlenecks, and the loading levels of different pieces of equipment. This study will give us the normal data that should be trained for AI models.

Load flow and short circuit analysis

The design was made in ETAP load flow analysis was carried out to study the behavior of the substation. Load flow is performed at normal conditions when the substation is working at a normal state. Multiple faults were

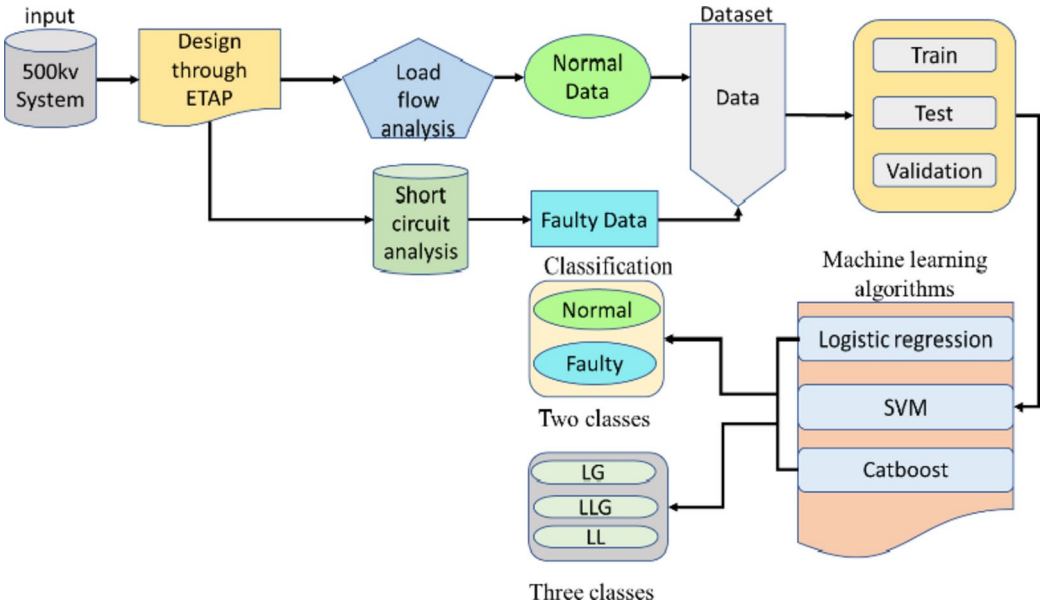

Fig. 1. The proposed methodology for the whole research.

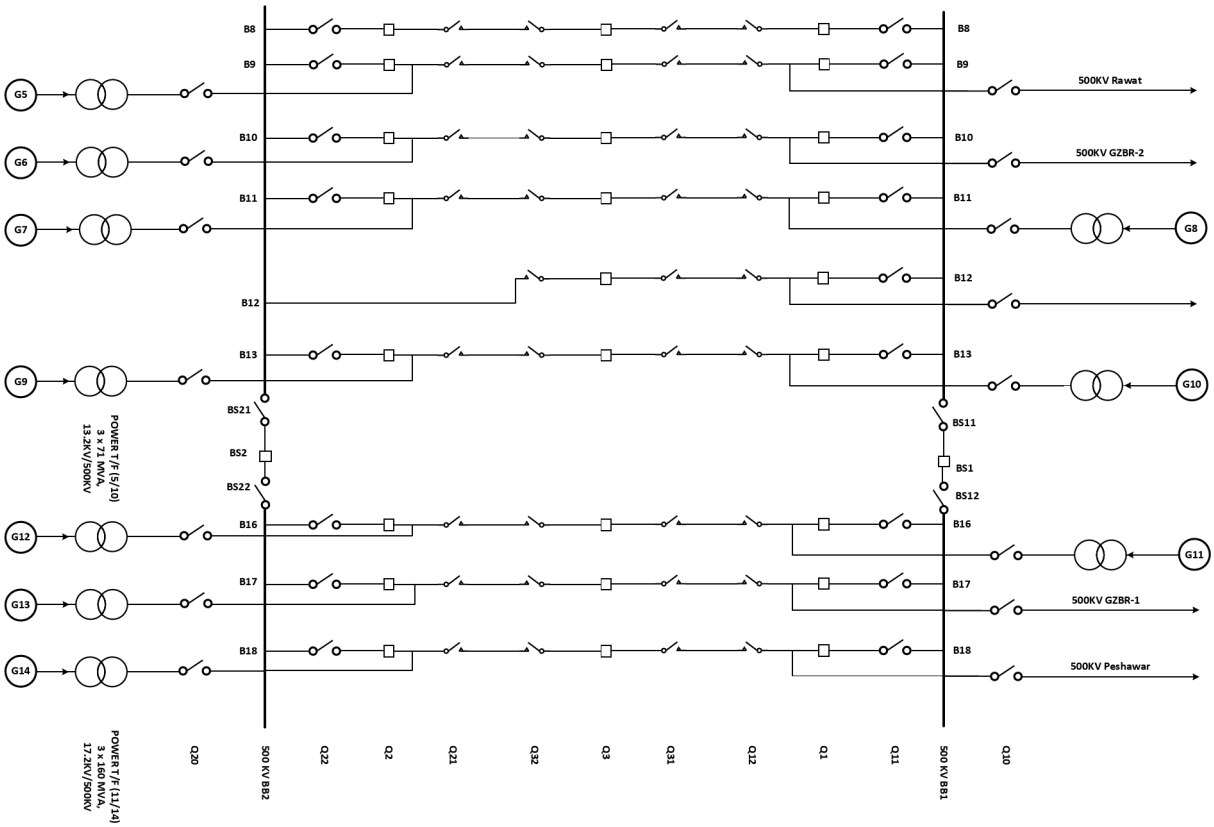

Fig. 2. Design of 500kv substation using ETAP.

generated at the substation taking different parameters into account like varying loads and tripping the generation units. These studies help us to generate a data set to study normal and abnormal behavior (LG, LLG, LL).

The short circuit anlysis (faults) can be mathematically found by the following equations using sequence impedance like positive sequence, impedance, negative impedance, and zero impedance.

The positive sequence impedance consists of three phasors having the same magnitude.

$$Z_1 = r_a + (X_a + X_d)_{\mathrm{i}} \tag{1}$$

$$R = \frac{\delta\, L}{A} \tag{2}$$

$$X_a(f, GMR) = 1.032\left(\frac{f}{60}\right) log\left(\frac{1}{GMR}\right)\frac{\Omega}{miles} \tag{3}$$

$$X_d(f, D_{eq}) = 1.032 log\left(\frac{D_{eq}}{1ft}\right)\frac{\Omega}{miles} \tag{4}$$

$$D_e(f, \rho\,) = 2160\sqrt{\frac{\rho}{f}}ft.Hz \tag{5}$$

When it comes to negative sequence impedance is an identical phase sequence as the originals, three phasors of equal magnitude with a phasor angle of 120 degrees.

$$Z_2 = r_a + (X_a + X_d)_{\mathrm{i}} \tag{6}$$

$$R = \frac{\delta\, L}{A} \tag{7}$$

$$X_a(f, GMR) = 1.032\left(\frac{f}{60}\right) log\left(\frac{1}{GMR}\right)\frac{\Omega}{miles} \tag{8}$$

$$X_d(f, D_{eq}) = 1.032 log\left(\frac{D_{eq}}{1ft}\right)\frac{\Omega}{miles} \tag{9}$$

$$D_e(f,\rho) = 2160\sqrt{\frac{\rho}{f}}ft.Hz \tag{10}$$

The zero-sequence impedance can be calculated by using the mathematical expression

$$Z_0(r_c, f, D_e, GMR) = (3.r_c + \frac{0.00477}{Hz}.f + 1j\frac{0.01397}{Hz}.log(\frac{D_e}{GMR})).\frac{\Omega}{Mile} \tag{11}$$

$$D_e(f,\rho) = 2160\sqrt{\frac{\rho}{f}}ft.Hz \tag{12}$$

To calculate the faults these faults can be calculated by the following mathematical expressions. For the calculation of line-to-ground faults, we use the general equation which is given as follows.

$$\overrightarrow{If} = \frac{3\overrightarrow{E_R}}{\overrightarrow{Z_1} + \overrightarrow{Z_2} + \overrightarrow{Z_0}} \tag{13}$$

Same as for the double line to ground faults the fault calculation can be done by the following mathematical methods

$$\begin{bmatrix} Ia(0) \\ Ia(1) \\ Ia(2) \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & a & a^2 \\ 1 & a^2 & a \end{bmatrix}\begin{bmatrix} 0 \\ Ib \\ Ic \end{bmatrix}, \tag{14}$$

$$If = I\overrightarrow{b} + \overrightarrow{I}c \tag{15}$$

$$Vb = Vc = If{*}Zf \tag{16}$$

$$V\vec{0} = V\vec{1} = V\vec{2} = \frac{V\vec{a}}{3} \tag{17}$$

$$I\vec{a} = I(1)\vec{} + I(2)\vec{} + I(0)\vec{} = 0 \tag{18}$$

$$I\overrightarrow{f} = 3I\overrightarrow{a}(0)\Big) \tag{19}$$

$$I\vec{f} = \frac{-3Z\vec{0}}{Z\vec{2} + Z\vec{0}} * \frac{E\vec{r}}{Z\vec{1} + \frac{Z\vec{2}Z\vec{3}}{Z\vec{2}+Z\vec{3}}} \tag{20}$$

$$I\overrightarrow{f} = \frac{-3Z\overrightarrow{2}*E\overrightarrow{r}}{Z\overrightarrow{0}*Z\overrightarrow{2} + Z\overrightarrow{0}*Z\overrightarrow{2} + Z\overrightarrow{1}*Z\overrightarrow{2}} \tag{21}$$

Sometimes due to stormy conditions, threes occur three phase faults the current equation for three-phase faults will be given as

$$I\overrightarrow{f} = \frac{-\mathrm{i}\sqrt{3}*\overrightarrow{E}_R}{Z\overrightarrow{1}*Z\overrightarrow{2}} \tag{22}$$

Data set generation

The steady-state behavior of electrical networks is studied and analyzed using a basic technique called load flow analysis in power system engineering. It assists in calculating the magnitude of the voltage, the phase angles, the active power (actual power), and the reactive power (reactive load) at various nodes (buses) within the power system. From this analysis, we get normal data. Then at multiple points, the faults are generated which gives the faulty data. We combine both the normal and abnormal data to make the data set to be trained to AI models.

Algorithms selection for experiments

In this investigation, three different approaches to machine learning were utilized to build models and make predictions regarding the occurrence of problems in the substation. The performance and expertise of algorithms utilizing artificial intelligence can be improved through the utilization of training data. Logistic regression, support vector machine (SVM), and Catboost are the three algorithms that are utilized in this process.

Logistics regression

Regression analysis examines relationships between a dependent variable and independent variables. It involves model specification, data collection, estimation, validation, and interpretation. Regression helps identify links between variables and make predictions. Proper technique and analysis are vital to draw accurate conclusions. The method's effectiveness is validated using a 500kv bus test system[29]. A voltage drop caused by inadequate generation and other disruptions may be simulated using the technology. The original approach was used to

obtain the situations illustrated here, though. The loads on substations are many. Substation problems are caused by altering the loads.

Support vector machine (SVM)

For the last few layers of the CNN model, the SVM model was used instead. High-accuracy predictions are made by the SVM at a higher training pace, using features that are taken from the Xception, InceptionV3, InceptionResNetV2, NASNetLarge, and DenseNet201 models. Since n is the number of features in the dataset, the SVM represents all of the dataset's values in n-dimensional space[30]. Within a dataset, the method assigns absolute coordinates to each feature value. Afterward, it seeks to construct several hyperplanes, or lines of separation, between the values of related classes[31]. The best hyperplane for increasing the difference between classes is then chosen by the model. The space between each class's nearest samples, or support vectors, and the hyperplane is known as the margin. Nonlinear and linear SVM algorithms are the two different varieties. An SVM algorithm that is linear or nonlinear may be selected. When the dataset can be partitioned along linear dimensions, linear support vector machines are used. But if the dataset cannot be split linearly, a nonlinear support vector machine is used. Usually utilized are flat, rectangular, poly (poly), and round-robin (RBF) basis fibers. A range will be chosen depending on the dispute itself as well as the present facts. Support Vector Machines (SVMs) reflect the input data in a high-dimensional space whereas RBF and poly kernels are applied in this work to divide the data. Using this knowledge, SVMs may then locate the hyperplane optimizing the many classes in the training set. Applying the learning algorithms developed from machine learning theory to the SVM learning algorithm lets it learn from training data and extend to fresh untested data. The idea of latency—the link between adjacent data points in every class and the latent class—is among the fundamental ideas of SVM. SVMs seek hyperplanes that maximize this domain to improve the classification performance.

Catboost

In 2018, XGBoost (Extreme Gradient Boosting) and Catboost (CB) emerged as powerful frameworks for Gradient Boosting Decision Tree (GBDT), significantly enhancing the technique's capabilities. While both tools augment GBDT, Catboost specifically addresses gradient bias and prediction shifts. It handles categorical and numerical data, performing regression, ranking, binary classification, and multiclass classification tasks. Catboost uniquely incorporates categorical variables through order-boosting, maintaining an unbiased gradient-boosting strategy while efficiently managing categorical data. Employing one-hot encoding for features with a limited number of classes, Catboost optimizes preprocessing before tree separation, proving successful across various classifications. The Catboost algorithm introduces the Minimum Variable Samples (MVS) training system, leveraging weighted sampling regularization to create boosting models from random forest data. It combines decision tree and random forest settings, optimizing hyperparameters to determine threshold parameters for the RF model, thereby improving boosting settings by utilizing more data points and storing hyperparameters. Catboost's versatility and optimization make it a valuable tool for predictive modeling.

Evaluation tools

Confusion matrics

The accuracy, precision, sensitivity, specificity, and AUC of the suggested models were calculated using Eqs. (23)–(27), correspondingly, based on the confusion matrices from each system: Metrics in the confusion matrix that indicate the quantity of properly detected histological data of the substation are True Positive (TP) and True Negative (TN). The amount of mislabeled faults is represented by false positives (FP) and false negatives (FN).

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \times 100\% \tag{23}$$

$$Specificity = \frac{TN}{TN + FP} \times 100\% \tag{24}$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \tag{25}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{26}$$

$$AUC = \frac{Sensitivity}{Specificity} \tag{27}$$

In substation fault analysis, conditions are classified as normal, LG, LLG, and LL faults. Detecting these faults is critical, and precision in categorization is achieved with a confusion matrix.TP (True Positive): Faulty conditions correctly identified as faults.TN (True Negative) Normal conditions that are accurately identified as normal (False Positive) Normal conditions mistakenly identified as faults' (False Negative) Fault conditions wrongly categorized as normal. A confusion matrix organizes the actual against expected circumstances into a tabular form, capturing the classification system's performance. While inaccurate identifications are discovered in the off-diagonal cells, proper categorizations show along the main diagonal of the matrix. This arrangement offers an understanding of the model's accuracy and performance in identifying genuine conditions, be they defective or normal. Therefore, a basic instrument for assessing the general dependability and diagnostic accuracy of fault-detecting systems in substations is the confusion matrix.

Receiver operating characteristic (ROC)

The (ROC) receiver operating characteristic is the visual tool for observing the binary classification system's sensitivity. The ROC gives us the relationship between true and false prediction rates (TPR&FPR) as the classifier changes. So far, machine learning is the most common application of this data analysis tool. Comparing the number of false positive samples and correctly identified negative samples to the total number of genuine positive samples helped determine the FPR and TPR. Comparatively evaluating several criteria and evaluating the classifier's sensitivity and specificity measuring capacity makes excellent use of the ROC curve.

Area under the curve (AUC)

Binary classification system performance may be evaluated with the graphical tool AUC. We plot the actual positive and false positive rates to show the interaction between the two feet. Whereas the false positive rate shows the proportion of incorrectly identified negative observations, the actual positive rate shows the percentage of correctly classified positive observations. A better AUC value denotes classifier performance of excellence. The AUC measures the classifier's ability to distinguish between negative and positive classifications. It is between 0 and 1, where 1 denotes a perfect classifier, and 0.5 represents a random classifier. An AUC of 0.7 or more is considered a decent classifier in most situations.

Results

Methods of optimization for the ETAP-analyzed and ETAP-designed 500 kV system are now in place—techniques like system bifurcation, system reconducting, and ideal capacitor placement fall under this category. The study of the system seeks to determine the most effective way to operate it by identifying the magnitude as well as the phase angle of the voltage at each bus, the actual and reactive electrical power flowing in each line, the losses in a particular line, and any over or under-load conditions. The system assessed the faults at various buses using several loss optimization techniques, including logistics regression, SVM, and cat boost.

Outcomes through logistic regression

The advantage of logistic regression is that it is straightforward to use. Additionally, it has a tremendous probabilistic interpretation that allows us to gauge how confident the model is for both situations 'outcomes. The results obtained using this algorithm show an accuracy of 93.% for two classes, as depicted in the confusion matrix and ROC (Fig. 3(a)), and an accuracy of 92% for three classes (Fig. 3(b)), which is an improvement over previous work.

Outcomes using the SVM algorithm

In most cases, support vector machine models, often known as SVMs, are thought to be more effective than logistic regression, mainly when working with non-linearly separable data. SVMs, on the other hand, come with the requirement of picking a kernel function and the parameters connected with it, which makes them more challenging to grasp and tweak. In addition, support vector machines (SVMs) can also be computationally intensive, especially when trained on massive datasets. Therefore, in this scenario, we used support vector machines (SVMs) for both the condition and the comparison tasks, and the results showed that they performed marginally better than logistic regression. The accuracy of SVM for two classes is 96% (Fig. 4(a)), and for three courses, it is 95% (Fig. 4(b)), which is an improvement over previous work.

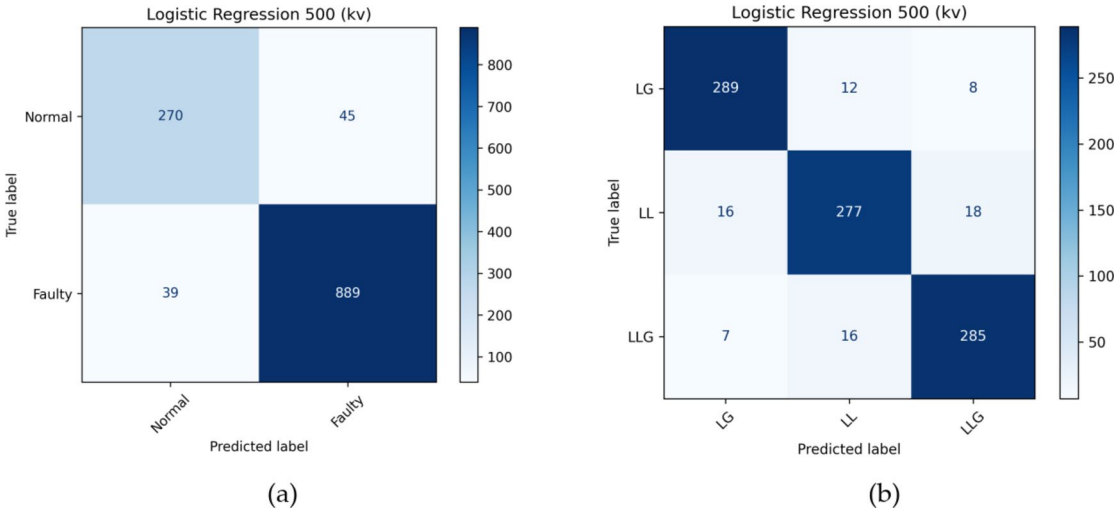

Fig. 3. Logistics regression results of both the classes, (**a**) Show the two class classifications and (**b**) show the three class classifications.

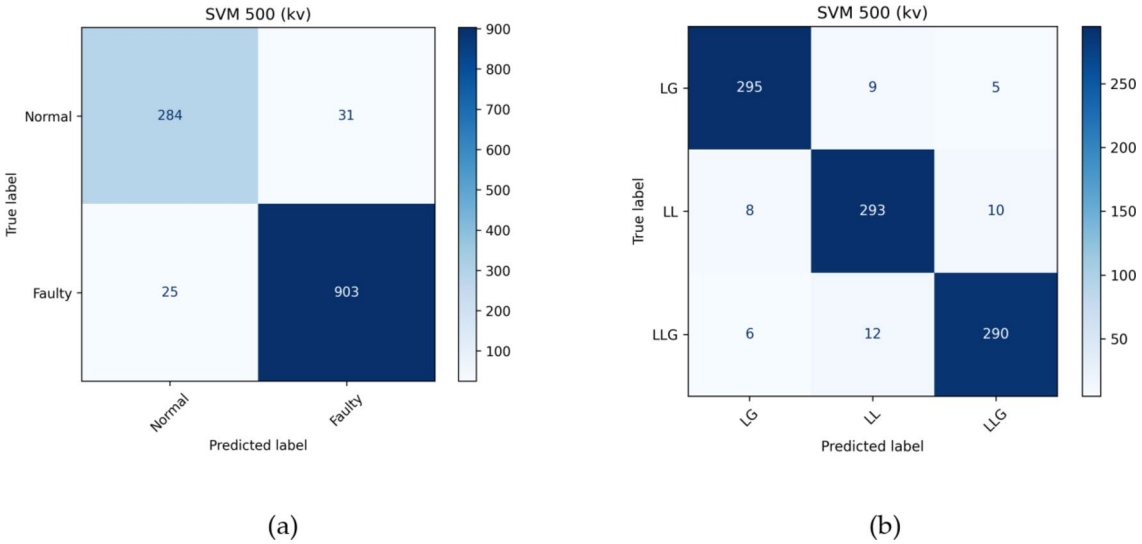

Fig. 4. Support Vector Machines result in the confusion matrix of both the classes, (**a**) Show the two class classifications and (**b**) Show the three class classifications.

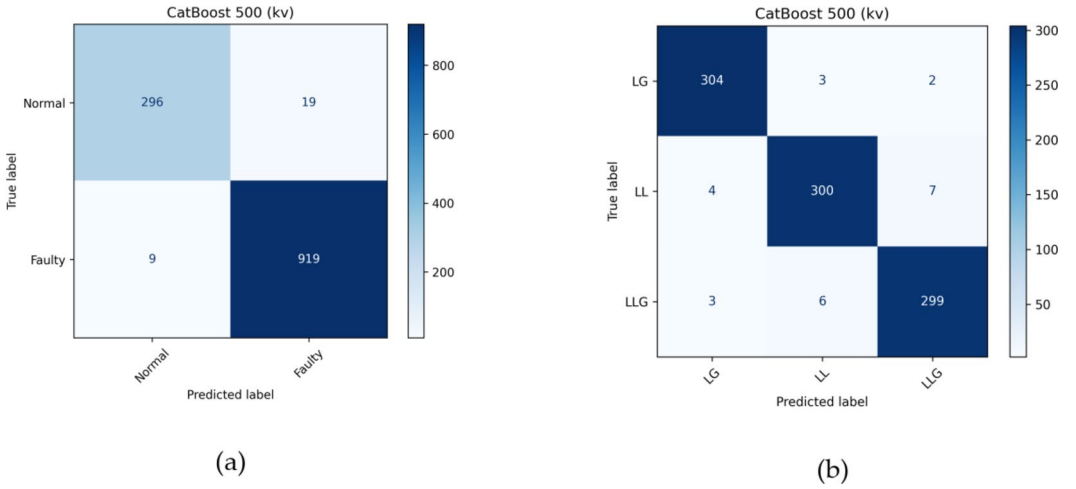

Fig. 5. CAT-Boost results in the confusion matrix of both the classes, (**a**) Shows the two class classifications and (**b**) Shows the three class classifications.

Outcomes over Catboost

With capabilities like parallel processing and handling missing information, Catboost is known for being quick and effective. Additionally, it features a variety of hyperparameters that may be adjusted to enhance model functionality. Using Catboost on datasets with categorical variables may be more straightforward because it can natively handle categorical features. As a result, when we applied it to both scenarios, the system's accuracy outperformed both SVM and Logistic Regression. These are the outcomes of the classification: for two classes, an accuracy of 98% was achieved (Fig. 5(a)), and for three classes, an accuracy of 97% was achieved (Fig. 5(b)), which is an improvement over previous work. Among all the algorithms, the best results were obtained using CatBoost.

Receiver operating characteristic (ROC) and AUC

Figure 6 shows a ROC curve to evaluate and compare the performance of three machine learning models—Logistic Regression, Support Vector Machine (SVM), and CatBoost—on a binary classification task for the "LG" class at 500 kV. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for different classification thresholds, where the x-axis represents the FPR (incorrectly classified negatives) and the y-axis represents the TPR (correctly classified positives). Each model's performance is summarized by its AUC (Area Under the Curve), which indicates its ability to distinguish between the classes. Logistic Regression (AUC = 0.95), SVM (AUC = 0.97), and CatBoost (AUC = 0.98) all outperform random guessing, which is represented by the diagonal black line (AUC = 0.5). CatBoost achieves the best performance, as its curve is

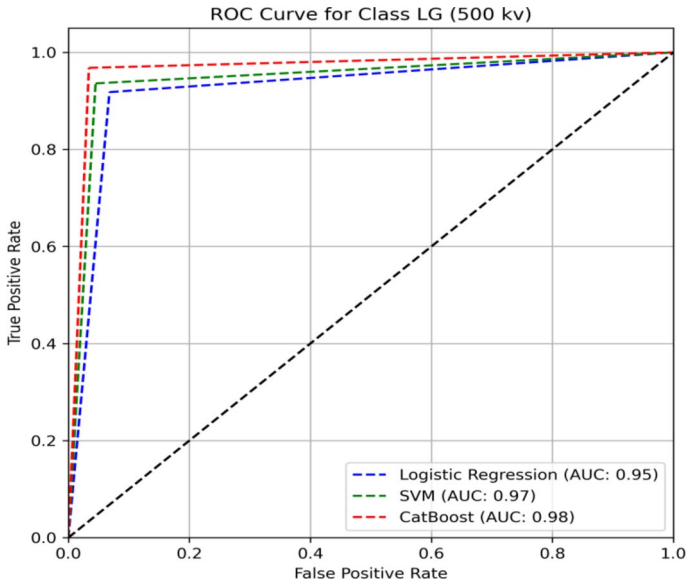

Fig. 6. ROC curves for LG at 500kv.

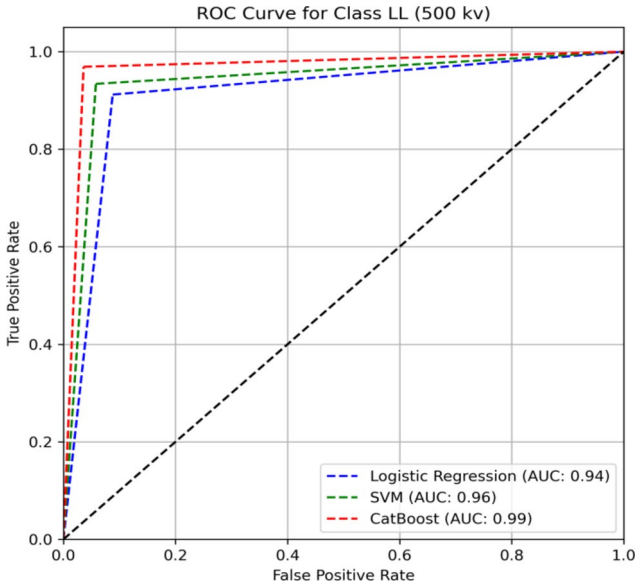

Fig. 7. ROC curves for LL at 500kv.

closest to the top-left corner, indicating it has the highest TPR for a given FPR, followed by SVM and Logistic Regression.

Figure 7 shows the ROC curve for the "LL" class at 500 kV, where three machine learning models—Logistic Regression, SVM, and CatBoost are compared. The plot shows the TPR on the y-axis against False Positive Rate FPR on the x-axis for varying thresholds. The closer the curve is to the top-left corner, the better the model's performance. Logistic Regression, represented by the blue dashed line, has an AUC of 0.94; SVM, shown by the green dashed line, has a slightly better AUC of 0.96; while CatBoost, represented by the red dashed line, achieves the highest AUC of 0.99. The black diagonal line indicates random guessing (AUC = 0.5), and all models perform better than random guessing, with CatBoost outperforming the other two models in terms of the highest AUC and closest approach to the top-left corner.

Figure 8 shows the ROC curve for the "LLG" class at 500 kV, where three machine learning models—Logistic Regression, SVM, and CatBoost- are compared regarding classification performance. The ROC curve plots the TPR on the y-axis against the FPR on the x-axis for varying classification thresholds. The closer the curve is to the top-left corner, the better the model's performance. The area under the curve (AUC) is used to quantify model performance, with higher AUC values indicating a better ability to distinguish between the two classes. In this case, Logistic Regression has an AUC of 0.93, SVM has an AUC of 0.95, and CatBoost has the highest

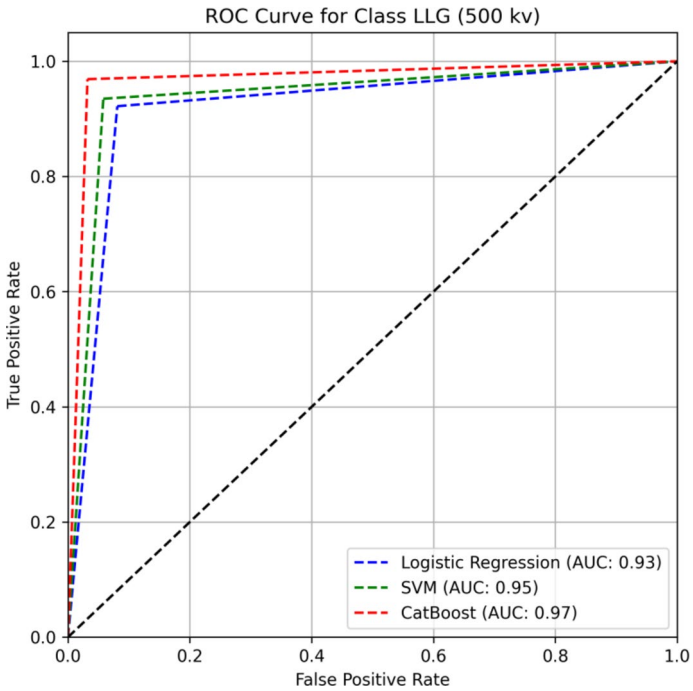

Fig. 8. ROC curves for LLG at 500kv.

AUC of 0.97. All three models perform better than random guessing, represented by the diagonal black line (AUC = 0.5). Still, CatBoost outperforms the other models in terms of AUC and its ROC curve's proximity to the top-left corner, followed by SVM and Logistic Regression.

To determine the validity of the algorithm's reaction to different voltage level values, the introduced models were trained and tested based on the simulated data on systems at 500 kV, 440 kV, and 220 kV using the CatBoost, Support Vector Machine (SVM), and Logistic Regression ML algorithms. To determine their performance flexibility, the models were further evaluated for 2-class- and 3-class classification problem settings. At 500 kV, when approaching the bulk data patterns, CatBoost worked as intended, providing 98% accuracy for the 2-class problem and 97% for the 3-class problem. SVM persisted with superior though slightly lower accuracy of 96% for the 2-class and 95% for the 3-class problem. Functional Logistic Regression was significantly less accurate, with 93% and 92% for 2-class and 3-class problems, respectively. The identification accuracies were slightly lower for models at 440 kV; still, CatBoost was the most stable across the given voltage systems, hypothesized to be 97% accurate for classification tasks. The test accuracy for the 2 class and 3 class was 95% and 94%, respectively, indicating that SVM was again very precise though slightly less stable than CatBoost. In Logistic Regression, however, the accuracy decreased to 90% for the 2-class problem and 91% for the 3-class problem. This indicates the approach's usefulness only on systems with little variability, such as the medium-voltage system presented in this paper. The accuracy for 2-class and 3-class problems at the lowest voltage level of 220 kV further declined to 95% and 94% of CatBoost, but CatBoost outperformed other models again. SVM provided 92% and 91%, while logistic regression reduced to 90% for two-class problems and 88% for three-class problems. Therefore, it can be concluded that CatBoost is the most accurate and deployment-ready model for a broad-range system voltage system. SVM is another interesting solution, but it is somewhat less efficient, while the performance of Logistic Regression prevents it from being used in projects that require high flexibility and accuracy. Table 2 and the ROC of the best algorithms are shown in Fig. 8 to validate the algorithm's response.

Figure 9 shows the ROC curve for the "LG" class, comparing the performance of CatBoost across three different voltage levels: 220 kV, 440 kV, and 500 kV. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings, with the ideal model having a curve closest to the top-left corner of the plot. CatBoost at 500 kV achieves the highest AUC of 0.98, followed by CatBoost at 440 kV (AUC = 0.97) and CatBoost at 220 kV (AUC = 0.95). All three models outperform random guessing, represented by the diagonal black line (AUC = 0.5), with the 500 kV model demonstrating the best overall performance in distinguishing between the classes.

Figure 10 shows the ROC curve for the "LL" class, comparing the performance of the CatBoost model at three different voltage levels: 220 kV, 440 kV, and 500 kV. CatBoost at 500 kV demonstrates the best performance, achieving an AUC of 0.99, followed by CatBoost at 440 kV with an AUC of 0.96 and CatBoost at 220 kV with an AUC of 0.94. All models outperform random guessing, represented by the diagonal black line (AUC = 0.5), with the 500 kV model showing superior discriminative power in distinguishing between the classes.

Figure 11 shows the ROC curve for the "LLG" class, evaluating the performance of the CatBoost model across three voltage levels: 220 kV, 440 kV, and 500 kV. The ROC curve plots the TPR on the y-axis against the FPR on the x-axis for different decision thresholds, with a curve closer to the top-left corner indicating better

System	Model	Accuracy for (2-class)	Accuracy for (3-class)
500 kv	CatBoost	98%	97%
	SVM	96%	95%
	Logistic regression	93%	92%
440 kv	CatBoost	97%	97%
	SVM	95%	94%
	Logistic regression	90%	91%
220 kv	CatBoost	95%	94%
	SVM	92%	91%
	Logistic regression	90%	88%

Table 2. Comparison of algorithms' response to different voltage levels.

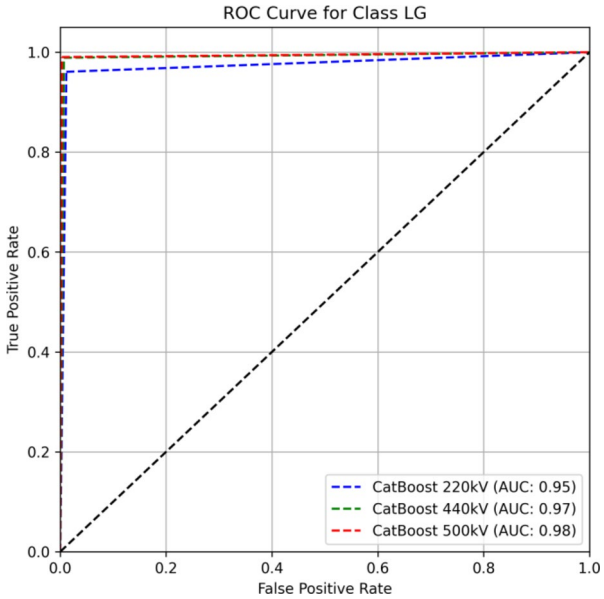

Fig. 9. ROC curves for LG at different voltage levels.

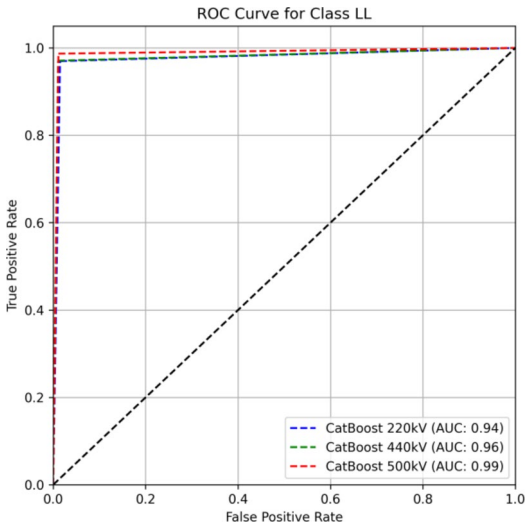

Fig. 10. ROC curves for LL at different voltage levels.

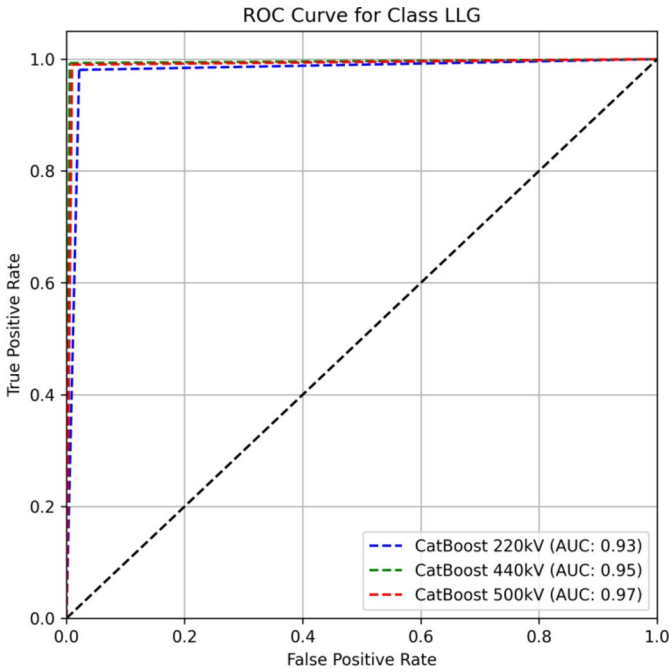

Fig. 11. ROC curves for LLG at different voltage levels.

performance. Each voltage level is represented by a distinct dashed line: CatBoost 220 kV (blue), CatBoost 440 kV (green), and CatBoost 500 kV (red). Each model's Area Under the Curve (AUC) indicates its classification ability, with higher AUC values reflecting better performance. CatBoost at 500 kV achieves the highest AUC of 0.97, followed by CatBoost at 440 kV with an AUC of 0.95 and CatBoost at 220 kV with an AUC of 0.93. All three models outperform random guessing, represented by the diagonal black line (AUC = 0.5), with the 500 kV model demonstrating the best overall discriminative power in distinguishing between the two classes, though the models at 440 kV and 220 kV also perform competitively.

Discussion

After reviewing the literature, I found that many grounding fault line selection techniques are accessible and have enhanced the accuracy of grounding fault line selection after thoroughly analyzing a large amount of literature. These techniques still need to be improved and have some limitations. The conventional short-circuit current approach is extensively used and can be used for various grounding fault types. It can promptly identify the location and kind of grounding fault. Furthermore, this method's computation is easy and doesn't require complicated system modeling or extra hardware[32]. However, since this approach is sensitive to the characteristics of the system model and cannot find the defect precisely, it may make mistakes when there are significant faults. When dealing with complicated systems or several faults simultaneously, the time-domain inversion approach works well for precisely locating, classifying, and locating the grounding fault. This technique is very durable and resistant to the system. However, this approach needs a lot of processing power and storage capacity, which could require extra computer assistance. With no further calculations or intricate modeling needed, the impedance ratio method is a reasonably straightforward technique that needs to compare two recorded values. It may be used with other selection techniques and is relevant to both (SP) single-phase and many-phase (multi-phase) problems. On the other hand, this approach places a heavy burden on the electrical grid as it needs precise and consistent grounding resistance values[33,34]. It is also unsuitable for pinpoint fault detection; it can only identify the fault area. Ground defects may be accurately detected and identified using the model matching technique based on the system model. Available computer simulation tools may be examined and simulated with high adaptability to many concurrent problems.

Nevertheless, this approach has many shortcomings, including the demand for accurate parameter estimates, system modeling, and significant processing and storage requirements. It is also incapable of pinpointing the exact site of the issue. However, precise and thorough data collection is difficult, time-consuming, and may result in inaccurate defect detection. Effective fault detection requires the development of algorithms that can manage a variety of problem circumstances and provide dependable findings. To achieve real-time requirements, effective algorithm design and quick data processing are also necessary for prompt fault responses. In contrast, insufficient algorithm performance results in imprecise issue detection, and insufficient data collecting causes false alarms and jeopardizes safety. Furthermore, system stability and fault isolation are hampered by the absence of real-time functionality. Artificial intelligence (AI) in grounding fault line selection may be pretty advantageous in today's world. Power systems engineering domain knowledge and AI algorithmic competence may be used in this partnership to provide more valuable and efficient AI methods for ground fault line selection. Large-scale real-time data processing is a strength of AI algorithms, which leads to quicker and more precise grounding

fault identification and diagnosis[35,36]. Furthermore, using many data sources and characteristics simultaneously, Artificial intelligence (AI) techniques may analyze faults holistically and adopt a more reliable and preventative approach to fault analysis. Furthermore, AI-based techniques provide a flexible approach to fault management since they may be easily adjusted to various power system designs and grounding problems. Ultimately, AI-based techniques may promote innovation and development in fault management and related domains by advancing AI technology and its applications in power systems.

Different grounding fault line selection approaches often have distinct advantages and disadvantages; thus, choosing one that best fits the particular circumstances is essential. A thorough checking and validation process would be carried out to guarantee the perfectness and dependability of the selected approach, and parameters such as the needs of the (PG) power grid, precision, and expenses should all be considered. Furthermore, the accuracy or dependability of the selection process may be improved by combining many fault line selection methodologies.

Because of the intricate architecture of the distribution network(DN), making a mathematically challenging model that accurately captures its features is difficult. However, machine learning algorithms such as logistic regression, SVM, and Catboost can effectively handle the two nonlinear problems. Because of their fault tolerance and strong self-adaptation, there is hope for their application in the slight current grounding system fault.

Using the data of the past three years being collected from WAPDA Pakistan, the information is compared and explored to determine which phase is faulty by creating a database of fault characteristics, combining the fundamental principles of genetic algorithms, and depending on the algorithm's healthy function. The data is classified into two classes: Normal data (Healthy condition) and faulty data (system at fault). Furthermore, the data is further divided into three courses if there are unsymmetrical faults (Line to ground, double line to ground, line to line). From there, calculations are performed to find the overall optimal solution. Reliability in fault line selection is achieved by AI algorithms (logistic regression, SVM, and Catboost), which process fault information by providing relative fault line selection as the output and employing fault characteristic values as input. Concurrently, AI algorithms (logistic regression, SVM, and Catboost) may fuse a range of fault criteria to pick the low-current grounding solution.

Even though ground fault line selection techniques have advanced significantly over time, there are still some obstacles due to technological constraints. Furthermore, many line selection techniques exhibit a more significant external impact, reducing dependability and detection sensitivity. Also, there is insufficient pertinent data. Power management and scholarly research will benefit from this study. AI-based grounding fault line selection technology can enhance power management by reducing the effect of power outages on consumers, increasing system safety and stability, and cutting down on maintenance time and expenses. Additionally, by monitoring and anticipating grounding failures, power management departments may increase the dependability and effectiveness of power systems. Additionally, using AI (logistic regression, SVM, and Catboost) in the power system will make the system quick decision-making; the previous techniques used only concerned fault analysis and estimation through software like PSCAD or Simulink[37]. On the performance of the models, it is possible to assess the accuracy and the ROC-AUC score. In the classification of control samples and targeted types of malignant cells, CatBoost is the most accurate, with 98% accuracy; in the differentiation of Classes LL and LG, with an accuracy of 97%, and in the ROC-AUC, with 0.99 for Class LG and 0.98 for Classes LL and LLG. This is shown to render it best placed to sort classes with the fewest errors between the two sets. Compared to CatBoost, SVM is slightly worse: 0.96 for 2-class and 0.95 for 3-class classification, and high AUC values (0.97, 0.95, and 0.96 for the classes correspondingly). However, it should be noted that despite a good result, the SVM algorithm cannot compete with the stable and high-quality performance presented by CatBoost. The worst performance on Logistic Regression is the lowest accuracy amongst the three- 93% for 2-class and 92% for 3-class as their ROC-AUC score is 0.95, 0.92, and 0.94, respectively. It also performs less than the SVM and CatBoost models; it shows how this model works on intricate connections within the provided dataset. Further analysis on other voltage levels, namely 220 kV and 440 kV datasets, are discussed in the results section to substantiate these algorithms' constructive and stable performance. This additional test should help ensure these models continue working under different circumstances. To achieve this benchmarking and determine how the models perform when exposed to these various voltage-level data, the study assumes different voltage-level values that demonstrate the models' behavior when adapted to a different setting. This analysis is also essential for modeling generalization and ensuring the developed models are optimized for different operational environments.

Conclusions

In conclusion, the research's main objective was optimizing the 500kv substation using ETAP (Electrical Transient analyzer program) and Artificial intelligence, which is covered in this paper. The usage of modern technologies like ETAP and AI could be beneficial for distribution and generation systems. For the betterment of the consumers and their requirements, the system must be efficient and reliable. For that reason, these tools can provide speed, efficiency, reliability, accountability, accuracy, and the detection of faults in the proper interval of time for the rapid investigation of the transformer and other equipment and the faulty condition within the substation would enable with the combination of ETAP and Artificial Intelligence Technology. The system can process large amounts of data and identify any disturbance in real time by applying artificial intelligence algorithms. These techniques can identify all the problems before time by using the previous data and act provocatively to make them less severe, reduce downtime, and protect the unessential power supplies. AI and ETAP can be a preventative method for the maintenance and preservation of unexpected equipment damage, as well as unplanned power supply drops due to network problems. As the potential faults can be detected in advance due to continuous monitoring of the 500kv substation and the data pattern analysis, it enables the pre-maintenance acts and the reduction of the capital cost associated with the replacement of the equipment. This proactive approach increases the power distribution's overall efficiency and reliability. So, power distribution

highly benefits the optimization of the substation from the combination of AI and ETAP technologies. This improves transient and fault detection, drastically improving the system's performance and customer satisfaction. These recommendations help with a cost-effective maintenance process by preventing equipment damage and reducing unscheduled outages. AI and ETAP will likely become more critical as the technology develops in the context of improving electricity distribution systems. In short, by applying these technologies, the monitoring and downtime are reduced, and the system will be efficiently working and easy to tackle.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 25 November 2024; Accepted: 10 February 2025
Published online: 15 February 2025

References

1. Markovic, N. et al. Hybrid fault detection in power systems. In. *2019 IEEE International Electric Machines & Drives Conference (IEMDC)* (IEEE, 2019).
2. Vaish, R. et al. Machine learning applications in power system fault diagnosis: Research advancements and perspectives. *Eng. Appl. Artif. Intell.* **106**, 104504 (2021).
3. Yousaf, M. Z. et al. Enhancing HVDC transmission line fault detection using disjoint bagging and bayesian optimization with artificial neural networks and scientometric insights. *Sci. Rep.* **14** (1), 23610 (2024).
4. Liu, Z. et al. Key target and defect detection of high-voltage power transmission lines with deep learning. *Int. J. Electr. Power Energy Syst.* **142**, 108277 (2022).
5. Feng, Y. et al. Robust support vector machines for classification with nonconvex and smooth losses. *Neural Comput.* **28** (6), 1217–1247 (2016).
6. Hussain, S. et al. A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection. *Energy Rep.* **7**, 4425–4436 (2021).
7. Sirojan, T. et al. Sustainable deep learning at grid edge for real-time high impedance fault detection. *IEEE Trans. Sustain. Comput.* **7** (2), 346–357 (2018).
8. Yousaf, M. Z. et al. A novel dc fault protection scheme based on intelligent network for meshed dc grids. *Int. J. Electr. Power Energy Syst.* **154**, 109423 (2023).
9. Gilanifar, M. et al. Multi-task logistic low-ranked dirty model for fault detection in the power distribution system. *IEEE Trans. Smart Grid*. **11** (1), 786–796 (2019).
10. Feng, R. et al. Nonintrusive load disaggregation for residential users based on alternating optimization and downsampling. *IEEE Trans. Instrum. Meas.* **70**, 1–12 (2021).
11. Feng, R. et al. Uniform physics informed neural network framework for microgrid and its application in voltage stability analysis. *IEEE Access* (2025).
12. Javaid, N. et al. Employing a machine learning boosting classifiers based stacking ensemble model for detecting non technical losses in smart grids. *IEEE Access* **10**, 121886–121899 (2022).
13. Saqib, S. M. et al. Deep learning-based electricity theft prediction in non-smart grid environments. *Heliyon* **10**(15) (2024).
14. Park, D. et al. LiReD: a light-weight real-time fault detection system for edge computing using LSTM recurrent neural networks. *Sensors* **18** (7), 2110 (2018).
15. Feng, R. et al. Saturated load forecasting based on improved logistic regression and affinity propagation. *Electr. Power Syst. Res.* **237**, 110953 (2024).
16. Shakiba, F. M. et al. Application of machine learning methods in fault detection and classification of power transmission lines: a survey. *Artif. Intell. Rev.* **56** (7), 5799–5836 (2023).
17. Yousaf, M. Z. et al. Bayesian-optimized LSTM-DWT approach for reliable fault detection in MMC-based HVDC systems. *Sci. Rep.* **14** (1), 17968 (2024).
18. Feng, R. & Cui, J. AINL: network topology identification of multiple communication modes via active intercepting and node locating. *IEEE Access* **11**, 121399–121409 (2023).
19. Shang, H. et al. Research on a Transformer Vibration Fault diagnosis Method based on Time-Shift Multiscale Increment Entropy and CatBoost. *Entropy* **26** (9), 721 (2024).
20. Khan, W. et al. Rotor angle stability of a microgrid generator through polynomial approximation based on RFID data collection and deep learning. *Sci. Rep.* **14** (1), 28342 (2024).
21. Li, Y. et al. A fault pattern and convolutional neural network based single-phase earth fault identification method for distribution network. In *2019 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia)* (IEEE, 2019).
22. Chen, K., Huang, C. & He, J. Fault detection, classification and location for transmission lines and distribution systems: a review on the methods. *High. Voltage*. **1** (1), 25–33 (2016).
23. Czumbil, L. et al. Analysis of load flow and short-circuit issues in a retrofitted 110/20 kV Romanian substation. In *201714th International Conference on Engineering of Modern Electric Systems (EMES)* (IEEE, 2017).
24. Chan, W. L. et al. A distributed on-line HV transmission condition monitoring information system. *IEEE Trans. Power Deliv*. **12** (2), 707–713 (1997).
25. Jacob, R. A., Senemmar, S. & Zhang, J. Fault diagnostics in shipboard power systems using graph neural networks. In *2021 IEEE 13th International Symposium on Diagnostics for Electrical Machines, Power Electronics and Drives (SDEMPED)* (IEEE, 2021).
26. Li, Z. et al. Transient Stability Analysis of Electrical Power Systems using Polynomial Approximation based Galerkin Method. In *2023 5th International Conference on Power and Energy Technology (ICPET)* (IEEE, 2023).
27. Hendawi, E. A high performance grid connected PV system based on HERIC transformerless inverter. *Indonesian J. Electr. Eng. Comput. Sci.* **20** (2), 602–612 (2020).
28. Jiang, H., Jia, M. & Lin, L. Adaptive ant colony algorithm based global optimization control of voltage/reactive power in the substation. In *2008 Fourth International Conference on Natural Computation* (IEEE, 2008).
29. Awalin, L. J. et al. Fault distance identification using impedance and matching approaches on distribution network. *Indonesian J. Electr. Eng. Comput. Sci.* **8** (3), 770–778 (2017).
30. Barker, P. P. & De Mello, R. W. Determining the impact of distributed generation on power systems. I. Radial distribution systems. In *2000 Power Engineering Society Summer Meeting (Cat. No. 00CH37134)* (IEEE, 2000).
31. Boza, P. & Evgeniou, T. Artificial intelligence to support the integration of variable renewable energy sources to the power system. *Appl. Energy*. **290**, 116754 (2021).
32. Huang, C. et al. Analysis of short-circuit current characteristics and its distribution of artificial grounding faults on DC transmission lines. *IEEE Trans. Power Delivery*. **33** (1), 520–528 (2017).

33. Shilong, L. et al. Fault line selection of single phase grounding fault in small-current ground system based on reactive current. In *2019 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia)* (IEEE, 2019).
34. Su, H. et al. An incremental learning framework for human-like redundancy optimization of anthropomorphic manipulators. *IEEE Trans. Industr. Inf.* **18** (3), 1864–1872 (2020).
35. Zhao, Y. et al. Artificial intelligence-based fault detection and diagnosis methods for building energy systems: advantages, challenges and the future. *Renew. Sustain. Energy Rev.* **109**, 85–101 (2019).
36. Mohammed, M. K. et al. Optimization and fault diagnosis of 132 kV substation low-voltage system using electrical transient analyzer program. *Int. J. Electr. Comput. Eng. (IJECE)* **13** (3), 2375–2383 (2023).
37. Lei, Y. et al. A systematic novel implementation on photovoltaic power, wind energy, and solar energy models. In *Second International Conference on Electronic Information Engineering and Computer Communication (EIECC 2022)* (SPIE, 2023).

Author contributions

All listed authors have substantially contributed to the manuscript and have approved the final submitted version, the description of each author's specific work and contributions are as follows: A.Z: Conception and design of study, Design of experiments, Analysis and interpretation of data, Writing - original draft. F.R: Algorithm optimization, Performance evaluation, Writing - review & editing. W.K: Construction of simulation network, Conducting experiments, Analysis and interpretation of data, Writing - review & editing. U.S: Theoretical analysis, Algorithm complexity analysis, Writing - review & editing. M.A: Literature review, Code debugging, Data processing, Writing - review & editing. M.Z.Y: Conception of network model and system architecture, Writing - review & editing. M.A: Performance evaluation, Writing - review & editing. M.B: Project administration, Supervision, Resources, Writing - Review & Editing. I.Z: Project administration, Supervision, Resources, Writing - Review & Editing.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.R. or I.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025