IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 October 2024, accepted 10 December 2024, date of publication 26 December 2024, date of current version 27 January 2025.

*Digital Object Identifier 10.1109/ACCESS.2024.3523409*

RESEARCH ARTICLE

# Adversarial Attack Detection in Smart Grids Using Deep Learning Architectures

**STEPHANIE NESS**
University of Vienna, Diplomatic Academy of Vienna, 1040 Vienna, Austria
e-mail: a01050675@unet.univie.ac.at

This work was supported through open-access funding provided by the University of Vienna.

**ABSTRACT** Smart grids themselves have emerged as vital structures of the up-to-date practical power systems or electricity networks that incorporate high technologies and information handling. Yet, they are more susceptible to an adversarial attack that can interfere with the critical functions like energy distribution and faults detection. This paper therefore proposes a new alternative to developing a DL and ML framework for identifying adversarial attacks on smart grids. After analyses of the performances of Logistic Regression, Perceptron, Gaussian Naive Bayes and Multi-Layer Perceptron, LSTM network has better results with an accuracy of 99.81%. The suggested framework strengthens smart grid immunity to cyber threats such as DoS attacks, back door injections, and adversarial perturbations while increasing energy distribution stability and security. For enhancing smart grid security, our results emphasize the importance of integration of ML and DL techniques and provide such an understanding of threat environment for future research and development on threat identification.

**INDEX TERMS** Adversarial attacks, smart grids, long short-term memory models, perceptron.

## I. INTRODUCTION

Smart grids have gone through a very dynamic development process that has enabled the development of enhanced and most efficient functions of power systems with incorporation of high technologies and data management systems. However, the specific interconnections, automation, coordinated control, and advanced data analytics, such as machine learning and deep learning, in smart grid operations bring new security risks [1]. There are problems called adversarial attacks, during which small alterations of the input data are made in order to deceive these models [2]. It can affect such important tasks as load forecasting, faults detection and energy distribution threatening the stability and security of the grid [3]. In this context, failure identification and counteraction of the above-stated adversarial threats contribute to attaining smart grid reliability [4]. It has been established that adversarial attacks can be damaging to the functionality of smart grids [5]. Said data control can result in wrong energy predictions that create a mismatch with the available supply. Fault detection systems could also be affected, and the possibility of timely identifying problems in the infrastructure could be hugely reduced, leading to more frequent outages [6]. In extreme conditions, adversarial attacks are fatal enough to stop the normal functioning of the grid or even bring about blackouts or equipment destruction, incurring severe losses to both the utility and the consumer. These vulnerabilities show why smart grid systems are required to protect against such adversarial threats.

One of the most important objectives of smart grid research is to develop methods for identifying adversarial attacks on smart grid systems. Smart grids mainly depend on real-time data and intelligent models for key areas such as energy control, load control, and fault control [7]. Should these adversarial attacks remain unnoticed, they can interfere with these operations and cause an imbalance in the required energy distribution, potentially causing blackouts and even extensive blackouts. Small and Medium-sized Enterprises (SMEs) like JD Power focus on customer satisfaction assessments to detect customer dissatisfaction generators and advanced application smart technology grids that enable the identification of potential problems that would otherwise cause system downtime and affect critical infrastructures. Furthermore, the identification and prevention of these attacks promote

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

users' confidence in smart grid technologies and enhance the capacity of operators to adopt automated mechanisms in managing smart grids without worrying about disruptions. In conclusion, the detection of adversarial attacks enhances smart grid-presented security that will enable these grids to protect the increasing energy demands from potential cyber adversaries. The contribution of this paper is as follow:

- In this paper a new approach based on deep learning for detecting adversarial attacks in smart grid systems is proposed. For this, the framework adopts several ML models combined with DL, which improve smart grid robustness against DoS, backdoor, and adversarial perturbation attacks to guarantee proper operation of the grid.
- The paper also offers a comparison of several other ML models including logistic regression, perceptron, Gaussian Naive Bayes, and multilayer perceptron, together with LSTM. This proves that LSTM has the maximum accuracy level of 99.81% hence providing a competitive tool to detect adversarial attack in smart grids compared to the traditional ML models.
- Altogether, this research presents novel ideas into the adversarial training and the robust data engineering making smart grids more equipped to identify and counter cyber attack. The proposed approach improves the security of smart grid because the system can adapt to the new cyber threats and continue to function securely.

This research is arranged into five sections as follows. The first introduces smart grids and describes how adversarial attacks impact their operation. The second section II revisits earlier work done on adversarial attacks in smart grids to find out the literature review of the work. The third section III is about the proposed approach of the study and draws out the envisioned method for addressing these types of attacks, with specific emphasis on the robustness and effectiveness of the detection process. The fourth section IV explains the experimental study performed to assess the effectiveness of the proposed approach, and the last section V demonstrates the enhancement in detection accuracy and the resilience of the system against adversarial attacks.

## II. LITERATURE REVIEW

The author in this paper [8] introduces a new approach to pro- tect deep learning models in SMs from adversarial attacks. In the current work, the researchers provide a method that introduces adversarial examples in order to train the model and thereby improve the resilience against such attacks. To do so, the model is trained on these 'glossed' examples, which aim to help the model improve its ability to detect between good and bad inputs. The results show that the proposed method dramatically improves the model's robustness against adversarial attacks and guarantees the smooth functioning of smart grids. The author in this paper [9] presents a new approach to IoT system design with a deep learning system to foster reliable demand-side management for smart grids. In the proposed approach, adversarial training methods are used to strengthen the deep learning models against possible attacks. Thus, the system becomes more immune to the different adversarial samples infused in the models, which helps maintain the security of the demand-side management business. The results show that the proposed framework effectively increases the dependability and reliability of IoT-based smart grid systems.

The author in this paper [10] develops a complete deep-learning model for anomaly detection and classification in the smart grid context. The proposed method finds and categorises different types of anomalies by combining autoencoder and classifier networks. In order to further illustrate the proficiency of the proposed model, the following advantages of the two architectures complemented each other to ensure that the model located and distinguished the known and unknown anomalies with high levels of accuracy and reliability. As the results show, the proposed approach is superior to existing methods and will be a useful tool for maintaining the security and dependability of smart grid operations going forward. The author in this paper [11] introduces robust data engineering with a deep learning approach for improving the security and reliability of the smart grids against PMU adversarial attacks. Using deep learning techniques, the proposed method is capable of analysing the PMU data while minimising the effect of the attacks. Thus, by presenting numerous experiments and evaluations, the paper proves the effectiveness of the claimed approach in enhancing smart grid systems' resilience and reliability against a range of adversarial threats.

The author in this paper [11] to improve the anti-PMU adversarial attack capability of smart grids, the paper puts forward an appropriate robust data engineering based on deep learning. The proposed method uses deep learning techniques and successfully detects and contains malicious attacks on the PMU data. The effectiveness of the proposed approach in enhancing the robustness and dependence of MAL smart grid systems on different types of adversarial threats is shown through various experiments and evaluations in the paper. This paper [12] puts forward a powerful learning-based strategy for identifying electricity theft adversarial evasion attacks in smart grids. The novel method presented here reduces the effect of such attacks, enabled by higher-level machine learning techniques designed to specifically solve the applications' goal of detecting concealed electricity theft. Using various experiments and a thorough evaluation of the scheme, the paper presents the benefits of the proposed approach in improving the security of smart grid systems from adversarial attacks.

This paper presents [13] a more efficient model, which is a deep learning-based cyber-attack detection in the IoT smart city environment that we can improve by introducing an adversarial training strategy. By fine-tuning the selected deep learning model with specifically designed adversarial samples, the approach attempts to enhance the generalization and recognition of new attacks. The studies verify the applicability and efficiency of the developed method in enhancing

the accuracy and stability of identification services in smart cities built on IoT technologies. This paper [14] presents a new adversarial machine learning technique to mitigate false data injection attacks on demand response programs in smart grids. The approach proposed here will try to enhance model robustness against adversarial examples by generating them and training the detection model with these. The results show that the proposed approach provides a way to mitigate the effects of false data injection attacks and improve the reliability of demand response systems of smart grid applications.

As analyzed in the cited papers, the presented approaches to detecting and preventing adversarial attacks in smart grids have some drawbacks. Recurrent problem is on the disclosure of adversarial training that while they make models robust against certain classes of adversaries, they are not very effective when it comes to robustness against other types of adversarial attacks or adversarial perturbations. In this case, the robustness of these models decreases when the models are applied in real environments, where data is less structured and uniform as is often the case with testing data. These involves some methods, especially those that use deep learning architectures as the base type, can be very demanding in terms of computation and thus making their real time implementation in large smart grid systems a challenge. Furthermore, it means that the security approaches employed within these frameworks may do not cover the other aspects of cyber threats except for adversarial examples, including stealth or insiders ones. Last but not the least, while the experiments have demonstrated high potential of applying these techniques, the actual application of these ideas could be burbed by practical difficulties associated with implementing new techniques into the highly developed smart grid environment without interrupting its work.

## III. PROPOSED APPROACH

This research proposes a new method for predicting adversarial attacks in smart grids using a DL approach. The datasets are divided into four different parts, and in the first part, different types of attacks are explained before combining all parts. Each segment is accompanied with three different types of attacks to mimic variety of attack situations. The model also implements both an ML and deep learning classifiers for the purpose of predictive demonstrations which in turn enables the evaluation of their defenses versus various attacks. This method takes advantage of the features of each part to design a versatile approach to predicting adversarial threats in smart grid scenarios for application.

Anything associated with essential systems such as smart grids deserves protection at all costs. Such approaches have prevailed in the past, though now they are incapable of identifying an increasing number of attacks efficiently. The current approach utilizes machine learning (ML) models for comparative analysis, as detailed in Algorithm 1 in addition, DLclassifier is described in Algorithm 2. This approach is proposed to improve the accuracy of the prediction in order to address the pertinent security issues confronting smart grids. It specifically encompasses the utilisation of techniques that involve infrastructures with more than one hidden layer as a show of the program's readiness to incorporate advancement in the neural network system. The main goal of this activity is to strengthen smart grid protection against known and potential cyber threats and guarantee both the adequacy of the smart grid against potentially dangerous actions and the stability of its systems against their probable consequences.

To minimize model overfitting and increase the forecast precision the data set has been divided systematically into four partitions namely. This division ensure that the learning is done equally and then there is the differences in terms of threats like; DoS backdoor injection and so on, and adversarial attacks. Due to this step the chances for model developing biasness towards specific attacks reduces to minimum, helping the model to reach higher prediction accuracy as shown in Figure 1.

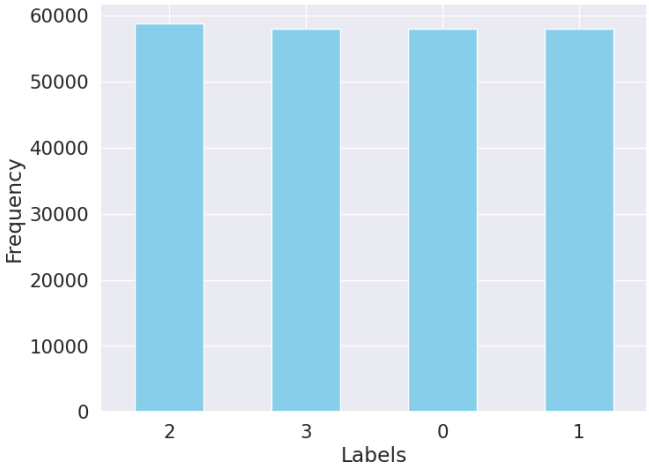


**FIGURE 1.** Overview of attacks on smart grids.

### A. DATA PREPROCESSING

Preparation that is done on the dataset in order to enhance its quality for the next steps to be followed next. In this stage, entries that have value missing recorded are omitted making the data more complete and reliable. In its wake, data exploration is made in order to discover the inherent structures which underlie the given data. Depending on the need for the next step in the analysis if any features are deemed not to be necessary for the data then they are removed hence enhancing the dataset.

### B. DATASET GENERATION

A dataset of smart grids that was compromised by malicious attacks in a systematic multi-class paradigm with classes to include DoS, benign, backdoor, and adversarial attacks was used in this research. The first dataset is of benign smart grids only and has been obtained from [15]. To measure the performance of the used and deployed models of LSTM, three different kinds of attacks were proposed: DoS, backdoor, and AT attacks. In [16], the authors found that the

inclusion of adversarial examples during model training also improves robustness. Preprocessing involves data cleaning which entails eradicating missing genes; data transformation, (categorical data encoding and normalization offeature values); and data reduction which entails selection of only benign samples. The training dataset is divided into four parts, out of which vulnerabilities are implemented to the parts to have an extra control on the model training. After all preprocessing is done, and the attacks are included, the individual data segments are joined for further analysis. The five techniques covered here seek to disrupt the availability, accuracy and efficiency of machine learning systems through adversarial attack, backdoor attack, and denial-of-service attack. Real-world invasions can place malicious triggers into the training phase, which can degrade or distort the specified model. On the other hand, DoS attacks can prevent the user access to the web application or online service because it saturates the machine learning systems with large amount of data and make it inoperable. Although these attacks have different objectives, all of them are potentially catastrophic to the dependability of machine learning systems.

**Algorithm 1** Train and Evaluate Machine learning Classifiers

1: Initialize an empty list `MLA` to store classifier names and accuracies
2: Define a list `classifiers` containing:
- LogisticRegression()
- GaussianNB()
- Perceptron()
- MLPClassifier()

3: Define a list `classifier_names` with corresponding names:
- "LogisticRegression"
- "GaussianNB"
- "Perceptron"
- "MLP Classifier"

4: **for** each `model, name in zip(classifiers, classifier_names)` **do**
5: Train the model using `model.fit(X_train, y_train)`
6: Make predictions using `pred = model.predict(X_test)`
7: Calculate accuracy: `accuracy = accuracy_score(pred, y_test) * 100.0`
8: Print `name` and accuracy
9: Append `(name, accuracy)` to `MLA`
10: Print error message for name
11: **end for**
12: Create a DataFrame `df` from `MLA` with columns `'model', 'accuracy'`
13: Print the DataFrame `df`

Developments in smart energy networks point to enhanced software, apart from creating operational risks. Backdoor

**Algorithm 2** Training and Evaluation of Tuned LSTM

1: Initialize LSTMModel with `num_classes`
2: Set loss function as `SparseCategoricalCrossentropy ()`
3: Set optimizer as `Adam ()`
4: Set `epochs` to 10
5: Set `batch_size` to 32
6: Compute `steps_per_epoch` as `len(X_train_scaled) // batch_size`
7: Initialize empty lists
8: **for** `epoch` in range(`epochs`) **do**
9: **for** `step` in range(`steps_per_epoch`) **do**
10: `start ← step * batch_size`
11: `end ← (step + 1) * batch_size` with `tf.GradientTape()` as `tape`
12: `predictions model(X_train_tensor[start :end])`
13: `loss ← loss_object(y_train_tensor[start :epredictions)` with
14: `gradients ← tape.gradient(loss, model.trainable_variables)`
15: `optimizer.apply_gradients(zip (gradients, model.trainable_variables))`
16: **end for**
17: `train_loss ← loss_object (y_train_tensor, model (X_train_tensor))`
18: `test_loss ← loss_object(y_test_tensor, model (X_test_tensor))`
19: Append `train_loss.numpy()` to `train_loss_history`
20: Append `test_loss.numpy()` to `test_loss_history`
21: `y_pred_train np.argmax (model (X_train_tensor).numpy(), axis =1)`
22: `y_pred_test ← np.argmax (model(X_test_tensor axis =1)`
`train_accuracy accuracy_score(y _train, y_pred_train)`
24: `test_accuracy`

```
    Paccuracy_score (y_test,
    y_pred_test)
25: Append train_accuracy
    to
    train_accuracy_history
26: Append test_accuracy
    to
    test_accuracy_history
27: Print f"Epoch epoch +
    1/epochs, Train Loss:
    train_loss:.4f, Train
    Accuracy:
    train_accuracy:.4f, Test
    Loss: test_loss:.4f, Test
    Accuracy: test_accuracy:.4f"
28: end for
```

injection attacks can also be a point of entry used by Cybercriminals as they get privileged access injecting code that allows them into the system [3]. This is rather dangerous as an attacker can cause a blackout, get the Profiler rights and manipulate electrical distribution or coerce the grid into power peaks to create havoc, or manipulate specific parts of this network to get information concerning energy consumption and grid management. They can lead to a domino effect which will lead to severe operational consequences. In denial of service (DoS) attacks, the attacker floods the system and networks with traffic to make the resources outside its functionality unavailable to authorized users for some time or forever [17].

In connection with smart grid networks, these attacks can flood the infrastructure with excess messages, which hinders the critical messages of this network, including those related to data acquisition and sending commands [18]. The supply of electricity is a continuous process all over the network, and any breakdown may affect consumers in the society. In extended DoS attacks, electricity blackouts are resulted from halting of the automated control system and billing [19]. Moreover, adversarial attacks are usually characterized by small perturbations that are indistinguishable from the original input; a small change in the inputs given to machine learning models will create wrong outputs. Temptations to change or manipulate data received by the smart grid sensors or signals that control actions of these algorithms may be given by attackers. If these adversarial perturbation attacks are carried out their consequences could be disastrous to the security and stability of smart grid operations. Through subtle interference over sensor data or signals from the communication layer, an attacker can deceive and mislead the decision made hence incorrect energy distribution and lack of proper failure rectification in the smart grid system. Real threats such as cyber-terrorists and cyber-espionage wishing to compromise the smart grid systems can only be thwarted by mature/anomaly detectors and constantly updated models.

### C. CLASSIFICATION MODELS

The principal notion around which a few advanced analytics tools, referred to as Machine learning (ML) algorithms, were designed is the ability to analyze data and search for patterns and correlations in it. These algorithms consist of one or the other of many methods, of which reinforcement learning, supervised learning and unsupervised learning are parts. The introduction of ML has impacted a variety of industries transforming complicated automation and improving choice-making in various fields which includes financial predicting, cars without drivers, picture recognition, and interpreting human language.

#### 1) LOGISTIC REGRESSION

Logistic Regression (LR) is a technique of modeling the probability that an outcome belongs to one of two groups as part of binary classification issues. It adds up the input variables by a weight, where the sum is then put through a logistic function stated in Equation 1 below which gives an output that is a probability, between 0 and 1. It is important to note the contribution level of each variable and this makes LR quite a very insightful model of prediction evident with the detailed results above showing distinct impacts for each variable. Despite that, it eases the predictive analysis and decision making and it is broadly applied across numerous disciplines, including medicine and health care, economics and business, and social sciences, where it may find the multiple logistic regression.

$$f^2(x) = a\left(\omega_0^{(2)} + \sum_{i=1}^{U_1} \omega_i^{(2)} f_1^{(1)}(x)\right)^{1/3} \tag{1}$$

Logistics regression is not the same as linear regression in that it does not estimate the direct relationship between the independent and dependent variables or the intercept where the slope plans the y axis but estimates the log of odds ration. If the data has non-linear relationships this can damage the results as logistic regression doesn't necessarily capture such complexity. Another complication is that if the features are highly related, to each other, a condition called multicollinearity. This can lead to high variability of the coefficient estimates, this will lower the reliability of the model's prediction.

#### 2) PERCEPTRON

Based on the Perceptron model, we first introduce the basic idea of machine learning as a well-established field of study for artificial intelligence. It has only the layer of output neurons, which uses input features to produce one-bit output either 1 or 0. This model works on the basis of passing the input through a weighted sum and applying a transformation through the activation function normally the step function in decision making whether the weighted input sum should be greater than a certain predefined threshold. This in

mathematics can be represented as in Equation 2.

$$y = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} \omega_i x_i + b > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The Perceptron ypothesis space is a linear array of perceptron that learn directly from the training data by modifying its weight in order to minimize the classification errors of the training data set. This is done using an iterative update rule of the weights that are used in computing the output; the rule updates the weights based on some discrepancy between actual output and the predicted output. While the Perceptron can identify linear versions in the data, it fails to do so with non-linear versions, making it less useful in more complicated problems. However, the Perceptron model is simplistic in nature, but it plays an important role in learning the basic concepts of the neural networks and, therefore, helps in introducing architectures of the Multi-Layer Perceptrons (MLPs) and deep learning. Thus, it fits the bill as a huge model in the large field of Machine Learning because of its capacity to explain the dynamics of supervised learning and classification.

#### 3) GAUSSIAN NAÏVE BAYES

Naïve Bayes (NB) is a classification algorithm for classification issues in machine learning. When calculating the probabilities, it assumes that all features of the data are mutually exclusive, which is less complex. For the case of continuous data, another form the Gaussian variant assumes is that distribution of each feature is normally distributed, with the curve of the distribution resembling the shape of a bell. As described with rather simple assumptions, this algorithm shows good performance in real life situations especially given the large number of variables where the independence assumption holds. GNB works best in scenarios as spam filtering and language model where the independence characteristics is reasonable. Owing to its performance and scalability, it is commonly used in classifying data in multiple industries of automotive importance while disregarding interaction between features. The example of the conditional probability is given by Equation 3.

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k).p(\mathbf{x} \mid C_k)}{p_{\mathbf{x}}} \tag{3}$$

Gaussian Naïve Bayes which supposes that the features can possess a specialized variance and that the traits have no correlations, and are normally distributed. However, if the data distribution does not approximate this assumption, the operation of the model can be negatively affected. However, one of the most common violations of assumptions in many samples is that of equal variances in two distribution. The extent to which this assumption is violated can be alleviated by the use of the transformation techniques.

#### 4) MLP CLASSIFIER

MLP Classifier is widely used for classification because of its unique ability to detect a high level of complexity with the given data. Being a network of relevance with multiple interrelating nodes, it operates in like manner to an artificial neural network as shown in Equation 4. MLP uses a process known as supervised learning and comprises an input layer, one or more layers than can be referred to as hidden layers, and an output layer. This architecture becomes useful when we need complex and flexible models to capture linearity and non-linearity in data.

$$f^{(2)}(x) = a\left(\omega_0^{(2)} + \sum_{i=1}^{U_1} \omega_i^{(2)} f_i(x)\right)^{1/3} \tag{4}$$

Most commonly MLPs is used in tasks for classification as it facilitates complex decision boundaries in Image recognition, and speech analysis. Nevertheless, great care should be taken when fitting the model so that it does not give an overly complex picture. Teaching MLPs may time-consuming and leave high computational demands on the algorithms especially with large datasets and complex architectures of the MLP. Besides, When the network is complex compared to the data available for the training set, MLPs tend to overfit and deliver a poor accuracy on the unseen data. As with any other machine learning technique, MLPs have a great number of hyperparameters that substantially limit the character of the resultant network and its learning capabilities hence their proper tuning is possible through previous experience and trial/error experimentation.

#### 5) LONG SHORT-TERM MEMORY (LSTM) NETWORK

An LSTM is a derivate of the recurrent neural network that was proven to be particularly successful in recognizing data patterns in sequences. LSTMs are really built with specific gate control mechanisms to make it easier for the function to learn long-distance dependencies and enhance the circumstances in speaking in language and in time series data series. Because the links carry out weight updates through the algorithms and backpropagation method, the LSTMs minimize on the errors during training. These networks are heavily used in recent applications like natural language processing, speech-recognition, time-series forecasting; owing to the complex relational patterns within large data. The hedge word occurrence which can be predicted by an LSTM is presented in Equation 5.

$$h_t = \mathbf{LSTM}(x_t, h_{t-1}) \tag{5}$$

The LSTM model converts the training and testing data into TensorFlow tensors and uses the Keras API in the TensorFlow to build up this network architecture. LSTM layers have been adopted as the core model, and they are able to learn temporal features and relationships inherent in the data. To optimize, the Adam optimizer was used and furthermore, the loss function is the sparse categorical cross-entropy. Moreover, cross-validation is used when 20 percent of the training data remain unseen and the model is trained for 50 epochs with a batch of 62. Finally the model's performance on a newly obtained test dataset is used to give out the test loss

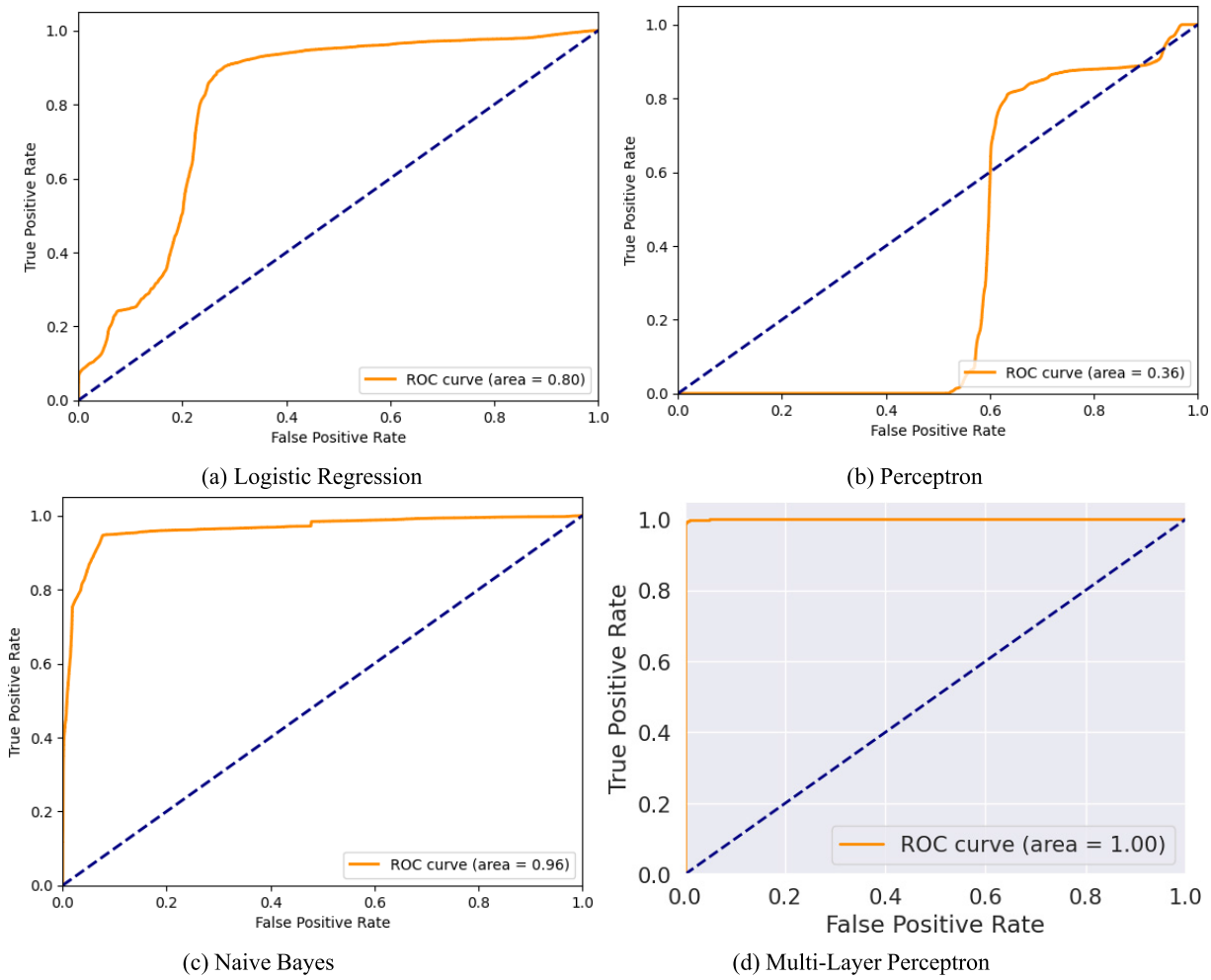


**FIGURE 2.** ROC for machine learning classifiers.

and accuracy. Specifically, LSTM networks face difficulties in terms of accurate temporal dependency modeling, and management of dimensions of datasets and overfitting and less interpretability because of the complexity that is inherent in them. However there are several issues that come into contest when used on this dataset, for instance the costs of training and optimization of deep architectures like LSTMs, and depth of models required for effective processing of data.

## IV. EXPERIMENTAL ANALYSIS AND RESULTS

This section focuses on different types of anomaly detection approaches in electrical networks, which play a crucial role in preserving the latter's protection. Seven specific techniques, including LR, GNB, perceptron, MLP and DNN, are discussed to recognise attacks in smart grid networks. The paper provides a comprehensive exploration of these methods with special reference to their efficiency, pointing out the downsides and benefits of each. The conclusions drawn from this paper will be useful in enhancing security against adversarial manipulations of power systems. The means utilised for assessing the performance of ML models are useful in defining how accurate the model is in relation to the data. Some parameters are accuracy, precision, recall, and F1-score. The following parameters demonstrate the model's capacity to identify abnormality. They also indicate how effectively the model is able to transfer the lessons learned to the new cases, which is an important factor for successful anomaly detection in smart energy networks. The performance metrics, therefore, make it easier for researchers to evaluate the efficiency of a model quantitatively. It will also assist decision-makers in determining the right measures for enhancing network security. Lastly, this framework also guarantees that declared anomaly detection models are compliant with the security measures of smart grids.

Training accuracy relates to how effectively the model gains the training set it gets from the input. The validation accuracy is defined as a ratio of right predictions by the created model for the given validation data set that has not been used when building the model. In training, an empirical risk minimization function quantifies the difference between expectation Outputs and actual Outputs. It gives a figure that

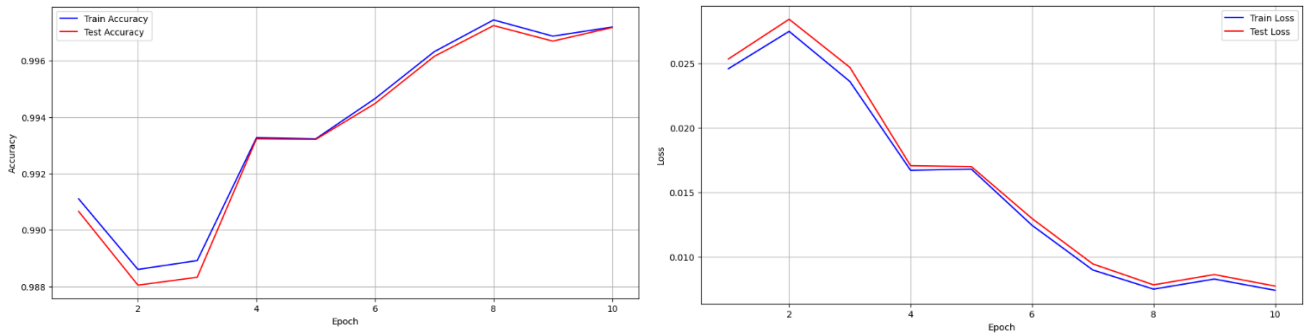


**FIGURE 3.** Training accuracy and loss curves.

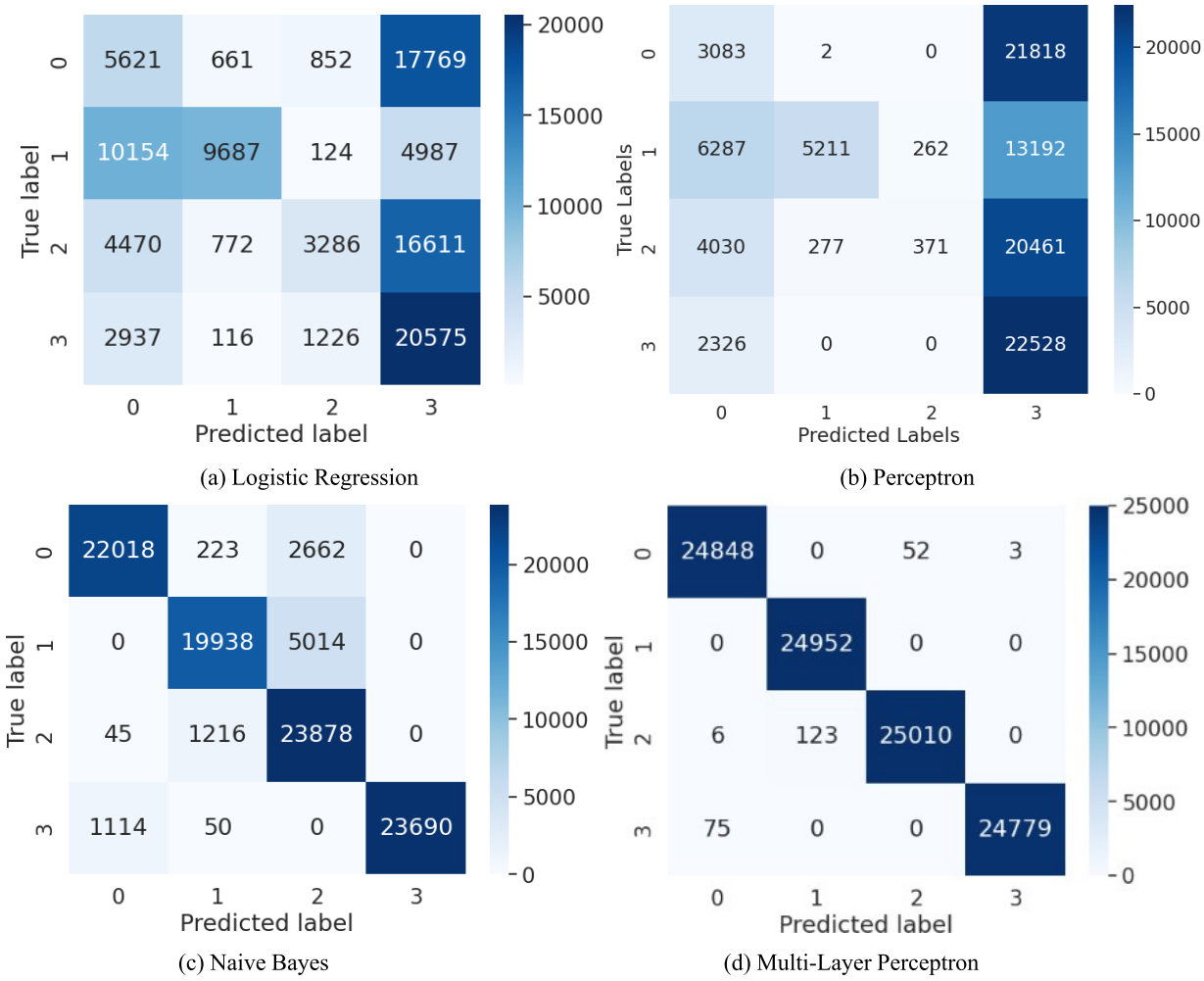


**FIGURE 4.** Confusion matrix for machine learning classifiers.

signifies the difference between the forecasts and the goals. Validation loss is applied to define how accurate is your model in predicting results for data it has not witnessed in training. During or after the training phase, the model finds out patterns from a given data set and produces an estimate. The ratio of true positive results relative to the false positive results differs depending on the various cut-off points adopted in classification. This relationship is illustrated graphically in a ROC curve as depicted by Figure 2.

Two important things are used to measure how well ML and DL models do tasks like detection: accuracy and loss. Accuracy is how many of the predictions were correct. Loss departure reveals the entire blunders that occurred. This means there must be a balance between achieving the

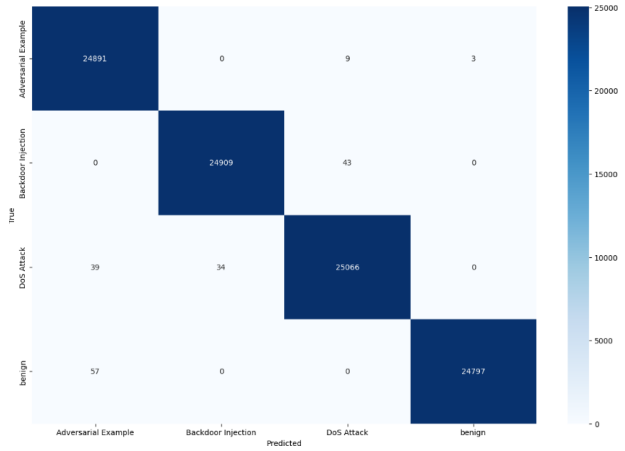


**FIGURE 5.** Confusion matrix of the LSTM model.

highest level of accuracy and incurring the least loss to create detection systems that are right and consume the least power. As mentioned in the above figures, the model was enhanced to minimise the loss, as seen in figure 3. The same structure here underlines that neural networks always learn and adjust their rules for the better.

The present investigation employs the ML and DL models to forecast potential attacks on smart grids. The most effective of them seems to be the LSTM, which constantly achieves the highest training accuracy and minimal training loss, as depicted in Figure 4. It demonstrates how effectively the LSTM can parse intricate and subtle features of data to detect threats to smart grid systems. Challenges of adversarial attacks exist, but the structure and effectiveness of the LSTM provide optimistic opportunities for enhancing the security of smart grid structures.

Evaluation of a model is central to assessing it with regard to how the latter copes with data. Accuracy, which is a basic measure of correct classification, percentages the number of correct results out of all the results to offer a broad perspective of the performance made. This information assists the model during model loss training as the expected values differ from the actual values, increasing the accuracy of the models being built. The ROC curve of the model defines the receiver operating characteristic, graphically illustrates the thresholds for classifying genuine and false cases, and gives a detailed and critical view of the prediction of the confused matrix model. It mentions specific details of False positive, False negative and True positive. This data is useful when wanting to compare one or several classes or groups of samples in relation to the model. These analytical metrics give a detailed performance of the model. Finally, it contributes positively to the calibration and the enhancement of the model for actual implementation.

The confusion matrix shows that the ML models failed to classify a limited number of instances, as shown in figure 5.

On the other hand, LSTMs present better performances and higher accuracy even though the ML models do not have any flaws in LSTM classification. Predictions are valid,

**TABLE 1.** Model evaluation results (%) for adversarial attack detection.

| Models | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| Logistic Regression | 51 | 39 | 37 | 39.23 |
| Perceptron | 51 | 31 | 24 | 31.24 |
| Gaussian Naive Bayes | 91 | 90 | 90 | 89.66 |
| Multilayer Perceptron | 100 | 100 | 100 | 99.38 |
| LSTM Model | 100 | 100 | 100 | 99.81 |

espacially when smart networks This gap is important when it comes to targeted adversarial attacks on LSTM to retrieve reliability on complex data models and pinpoint risks in the smart grid domain.

The smart grid application necessitates a large dataset capable of accommodating a multitude of features. This study employed a large dataset specifically tailored for smart grids. The efficiency of the dataset for identifying adversarial attacks on smart networks was tested using various classifiers, including MLP, perceptron, LR, and GNB models, which are familiar in the realm of machine learning. A deep-learning method was also integrated to advance the analysis. The proposed DL model not only achieved higher efficiency and accuracy than the competitive ML models but also eradicated the need for a more complex neural network, thereby enhancing the smart grid architecture against abuse attacks.

By using different ML models, as well as an LSTM model, adversarial attacks on smart grids were detected. Out of all the models, LSTM Which yielded an accuracy of 99.81%. This accuracy was higher than that of standard algorithms for ML such as MLP (99.38%) close to LSTM model, NB(89.66%), LR (39.23%), and Perceptron (31.24%), as shown in Table 1. Such a significant performance difference shows the extent to which DL methods can detect and mitigate threats from several attacks on smart grid systems including adversarial, DoS, and backdoor. From these discoveries primary importance is concisely in transports new comprehensive methodologies to advance smart grid system security and fixity against alternating cybersecurity threats.

## V. CONCLUSION AND FUTURE SCOPE

The research explores a dataset of smart grids under adversarial attacks that include multi-category classifications, such as DoS, benign, backdoor, and adversarial perturbation attacks. Having set aside the experiment with the smart grid and populating the dataset exclusively with benign instances of the system, the efficacy of the implemented models was investigated. Different kinds of attacks were incorporated in order to assess the weaknesses that exist in the grid system. While using both DL and ML methods simultaneously, the enhancement of the models' capability to analyze smart grid security attacks was achieved. The evaluation included basic models including the logistic regression (LR), Perceptron, naive Bayes (NB), and multilayer perceptron (MLP) with accuracy varying from 31% to 99%. Compared to the benchmark, the Long Short Term Memory Model achieved the

highest accuracy of 99.81%, indicating the applicability of the proposed approach for timely detection of possible attacks and the improvement in smart grid safeguard measures. This enhanced accuracy underscores the importance of protecting smart grid structures against new forms of cyber threats. It must be noted that different degrees of attack prediction accuracy and various detection methodologies can still be optimized. Therefore, this work paves the way in outlining future adversarial risks to smart grids and benefit the safeguarding of significant energy infrastructures.

## REFERENCES


[1] M. Massaoudi, H. Abu-Rub, S. S. Refaat, I. Chihi, and F. S. Oueslati, "Deep learning in smart grid technology: A review of recent advancements and future prospects," *IEEE Access*, vol. 9, pp. 54558–54578, 2021.

[2] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Trans. Intell. Technol.*, vol. 6, no. 1, pp. 25–45, Mar. 2021.

[3] S. M. A. A. Abir, A. Anwar, J. Choi, and A. S. M. Kayes, "IoT-enabled smart energy grid: Applications and challenges," *IEEE Access*, vol. 9, pp. 50961–50981, 2021.

[4] L. Phiri, "A framework for cyber security risk modeling and mitigation in smart grid communication and control systems," Ph.D. dissertation, Dept. Eng., Univ. Zambia, Lusaka, Zambia, 2023.

[5] J. Tian, B. Wang, J. Li, and Z. Wang, "Adversarial attacks and defense for CNN based power quality recognition in smart grid," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 2, pp. 807–819, Mar. 2022.

[6] A. E. L. Rivas and T. Abrão, "Faults in smart grid systems: Monitoring, detection and classification," *Electr. Power Syst. Res.*, vol. 189, Dec. 2020, Art. no. 106602.

[7] D. K. Panda and S. Das, "Smart grid architecture model for control, optimization and data analytics of future power networks with more renewable energy," *J. Cleaner Prod.*, vol. 301, Jun. 2021, Art. no. 126877.

[8] J. Hao and Y. Tao, "Adversarial attacks on deep learning models in smart grids," *Energy Rep.*, vol. 8, pp. 123–129, May 2022.

[9] M. Elsisi, C.-L. Su, and M. N. Ali, "Design of reliable IoT systems with deep learning to support resilient demand side management in smart grids against adversarial attacks," *IEEE Trans. Ind. Appl.*, vol. 60, no. 2, pp. 2095–2106, Mar. 2024.

[10] I. Siniosoglou, P. Radoglou-Grammatikis, G. Efstathopoulos, P. Fouliras, and P. Sarigiannidis, "A unified deep learning anomaly detection and classification approach for smart grid environments," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 2, pp. 1137–1151, Jun. 2021.

[11] T. Berghout, M. Benbouzid, and Y. Amirat, "Towards resilient and secure smart grids against PMU adversarial attacks: A deep learning-based robust data engineering approach," *Electronics*, vol. 12, no. 12, p. 2554, Jun. 2023.

[12] A. Takiddin, M. Ismail, and E. Serpedin, "Robust data-driven detection of electricity theft adversarial evasion attacks in smart grids," *IEEE Trans. Smart Grid*, vol. 14, no. 1, pp. 663–676, Jan. 2023.

[13] M. M. Rashid, J. Kamruzzaman, M. M. Hassan, T. Imam, S. Wibowo, S. Gordon, and G. Fortino, "Adversarial training for deep learning-based cyberattack detection in IoT-based smart city applications," *Comput. Secur.*, vol. 120, Sep. 2022, Art. no. 102783.

[14] Z. Guihai and B. Sikdar, "Adversarial machine learning against false data injection attack detection for smart grid demand response," in *Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids (SmartGridComm)*, Oct. 2021, pp. 352–357.

[15] S. Zidi, A. Mihoub, S. M. Qaisar, M. Krichen, and Q. A. Al-Haija, "Theft detection dataset for benchmarking and machine learning based classification in a smart grid environment," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 1, pp. 13–25, Jan. 2023.

[16] Q. Tian, K. Kuang, K. Jiang, F. Wu, and Y. Wang, "Analysis and applications of class-wise robustness in adversarial training," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 1561–1570.

[17] A. Huseinovic, S. Mrdovic, K. Bicakci, and S. Uludag, "A survey of denial-of-service attacks and solutions in the smart grid," *IEEE Access*, vol. 8, pp. 177447–177470, 2020.

[18] M. Z. Gunduz and R. Das, "Cyber-security on smart grid: Threats and potential solutions," *Comput. Netw.*, vol. 169, Mar. 2020, Art. no. 107094.

[19] D. L. Burgos, "DoS threat to smart grids: Review, analysis, and challenges," NTNU, Trondheim, Norway, Tech. Rep. 11250/3154983, 2024.


**STEPHANIE NESS** received the B.Sc. degree from London School of Economics and Political Science, U.K., the M.S. degree from Harvard University, Cambridge, MA, USA, and the L.L.M. degree from The University of Law, London, U.K. She is an experienced Senior Research Engineering Manager, leading teams in developing cloud security solutions and advanced analytics for energy systems. As a Senior Consultant at Maschinenhirn GmbH, she advises on AI integration and strategic cloud security initiatives for leading Austrian companies. She has extensive project management and technical leadership experience. She is actively involved in advancing energy systems through innovative technologies. With over 25 peer-reviewed publications, her research interests include cloud security, data analytics, artificial intelligence in power systems, and the integration of quantum computing into smart grids.

Open Access funding provided by 'University of Vienna' within the CRUI CARE Agreement