Machine Learning-Based Detectors of High Impedance Faults in Power Systems

Muhammad Mutiul Muhaimin[1,a], Md Maidul Islam[2,b], Abdulrahman Takiddin[1,c], Md Omar Faruque[1,d]
[1]Deptartment of Electrical and Computer Engineering, Florida State University, Tallahassee, FL, USA
[2]Electric Power Research Institute, Knoxville, TN, USA
[a] mmuhaimin@fsu.edu, [b] mislam@epri.com, [c] a.takiddin@fsu.edu, [d] mfaruque@eng.famu.fsu.edu

***Abstract*—Early research has found the detection of high impedance faults (HIFs) to be challenging due to low current ripple magnitudes. HIF faults create very slight deviations in sensor readings, often mistaken for non-fault switching events. Usually, HIF fault waveforms are seen to be very close to capacitor bank switching events. This paper compares the performance of several machine learning and deep learning-based approaches, namely, DT, RF, SVM, convolutional neural networks (CNN), gated recurrent units (GRU) and long short-term memory (LSTM) models for HIF detection. Five evaluation metrics: accuracy, precision, recall, F1 score, and area under curve are observed for comparative analysis. The SVM model shows accuracy scores of 84.000%, RF shows 99.634%, DT shows 100%, CNN shows 95.306%, and GRU and LSTM show 97.074% each. The shallow models prove to have their own strengths in terms of HIF detection. The deeper models are found to be more prone to overfitting compared to the shallower models.**

***Index Terms*—artificial intelligence, deep learning, high impedance fault, machine learning, power grid, power system disturbances.**

I. INTRODUCTION

At various locations in a power system (PS), the obtained output current (I) and voltage (V) may not be as expected. This happens when a line gets grounded abruptly through some impedance, and is termed fault. The occurence of faults in a PS may harm the infrastructure and machinery, causing service disruptions. Depending on fault impedance, faults can be classified into three types: bolted faults, low impedance faults (LIF) and high impedance faults (HIF).

What makes HIF detection challenging is, when HIFs (HIF_FAULT) occur, the VI readings show minor and random fluctuations from normalcy [1]–[3]. Such transients are analogous to capacitor bank switching events (CAP_SWITCH) and are undetectable by overcurrent relays or fuses [4]. 83% of all HIFs stay undetected [5]. HIFs may cause loss of power, damage to cables, initiate wildfires, and pose public safety hazards if the energized cables lie on a poorly conducting surface for a long time. Due to the advent of machine learning (ML) and deep learning (DL) algorithms, the issue regarding undetectability can be overcome. ML and DL algorithms can detect and learn fluctuations that are too insignificant to be detected either by vision or by setting a threshold.

The goal of the paper is to classify such "faulty" and "non-fault" data. The output of the paper will be a tool that can compliment protection relays and circuit breakers (CB) in the field of HIF detection. The tool will act as a secondary layer of protection along with the sensors and the relays. This will leverage the power of the existing hardware, while also providing a fail-safe for sensor noise and latency.

Existing fault detection methods can be divided into three sub-groups: hardware-based, model-based, and history-based methods [6].

A. Hardware-Based

Hardware-based solutions are fast and effective, but they depend upon sensors, making these expensive. There have been attempts to use harmonics information to detect HIF [7]. Authors of [8] proposed a proportional relay-based HIF detection technique. Authors of [9] proposed a second harmonic-based HIF indicator as an advanced functionality of smartmeters. However, these techniques are sensitive to what threshold the algorithm needs to consider for calling it HIF, meaning that it requires engineering experience. Again, the commercial relays available are designed for HIFs with arc, and not for those without it.

B. Model-Based

Such solutions stand out because they use mathematical or physical modeling for fault detection. This can relieve stress on finances. However, this kind of solution also necessitates accurate modeling and more computational capability. Authors in [10] analyzed HIF V-I characteristics using time-domain distortion for event detection. Authors of [11] proposed a mathematical morphology-based HIF detection strategy. Besides, authors of [12] and [13] attempted a finite impulse response (FIR) filter-based HIF detection methodology. The problem with analyzing the harmonics response is that HIF faults are non-stationary in nature.

C. History-Based

Traditionally, sensor waveforms were monitored manually or based on preset thresholds. With the advent of advanced sensors and micro-phasor measurement units (μPMUs), vast amounts of data are generated in real-time today, making manual observations impractical. Again, some event signatures are too subtle to detect manually. So, research involving sensor measurements is increasingly shifting from traditional to data-driven methods [5]. ML and DL are recognized for their ability to detect patterns with high accuracy from historical data without mathematical modeling. These algorithms can model complex systems but they demand large datasets and are computationally intensive [6].

Authors of [14] used a decision tree (DT) model trained on statistical and harmonic features extracted from current waveforms. Besides, authors of [15] trained SVMs with variational mode decomposition (VMD) and singular value decomposition (SVD) signatures of HIF events to isolate HIF events from normal conditions. The use of signature extraction algorithms for fault detection using ML is not a novel concept. Authors of [16] used intrinsic mode functions (IMF) from empirical mode decomposition (EMD) to train artificial neural networks (ANN) for the purpose of HIF detection. Authors of [2] added Hilbert Huang transform (HHT) with long short-term memory (LSTM), optimized with vector-based ensemble strategies for HIF detection. Authors of [17] used a bidirectional gated recurrent units (GRU) algorithm for HIF detection and achieve good accuracy. While these techniques provided better detection strategies against traditional HIF detection methods, they didn't provide a comparison against the data-centric approaches. Authors of [18] attained over 99% accuracy in LIF and HIF detection on transmission lines using LSTMs. But, authors of [17] use the 13-bus IEEE distribution network, and authors of [18] used the 14-bus IEEE transmission network. The authors did not show the performance of their models upon a much larger network like the IEEE 123-bus distribution network. Authors of [19] used a fully-connected CNN model for HIF detection. The benefits of these models are that, they incorporate multi-phase faults and also achieve very high accuracy. However, almost 95% of all HIF instances are single phase to ground faults [5], [20]. Authors of [3] integrated Maximal Overlap Discrete Wavelet Transform (MODWT) and deep bidirectional LSTM (DBLSTM) for HIF detection in a mesh network and achieved 99.1% accuracy. All datasets considered in the previous models, apart from the one in [18], only had HIF events and normal operation instances. However, the challenge in detecting HIF events is, the signatures look similar to normal switching events, like CAP_SWITCH, and the previous works did not test the model performances when the 0 class was CAP_SWITCH. Authors of [21] used CWT features to train a transfer learning based convolutional neural networks (CNN) algorithm for HIF detection. Authors of [1] had acknowledged the additional complexity that is introduced to the mechanism due to digital signal processing (DSP) tools, and therefore, disregarded it hybridizing transformers with CNN. This method requires a lot of training data.

Authors of [22] and [23] took into consideration that photos may be distorted very little, and this can make it a challenge for the machine to distinguish between real and fake. This problem is closely related to the problem in this study. The authors utilize neural networks and transformers to tackle this challenge. The authors of [24] and [25] recognized the challenges in obtaining and labeling data for PS events. As a result, both studies adopted clustering solutions, assigning labels based on cluster properties. Authors of [24] proposed a convolutional autoencoder-based k-means clustering algorithm and those of [25] proposed a bi-directional anomaly generative adversarial network (Bi-AnoGAN) for PS event clustering. But unsupervised learning models cannot be evaluated against known benchmarks as they aren't validated against metrics like accuracy (ACC), precision (DR), recall (RE), F1 score and area under curve (AUC).

The contributions of this paper are as follows:

- We develop the IEEE 123-node benchmark model with added photovoltaics (MOD123) for fault analysis.
- We employ ML and DL techniques with hyperparameter optimization based on three-fold cross validation for HIF detection.
- We conduct a comparative analysis among proposed ML and DL architectures against known benchmarks to validate the superiority of shallow models for HIF detection, with DT outperforming SVM by 19.048%, random forest (RF) by 0.004%, CNN by 4.925% and GRU and LSTM by 3.014%.

II. DATA PREPARATION

The dataset has been developed by obtaining sensor readings from a modified IEEE 123 node test feeder with added photovoltaics [26] on MATLAB Simulink. Sensor locations are found using an optimum sensor placement algorithm [27]. Then HIF_FAULTs and CAP_SWITCHes are triggered on certain buses and time-series data is recorded at each of the sensors. The antiparallel diode model has been used for HIF simulation.

Each sensor locations have three-phase (3ϕ) V-I sensors. A total of 3,352 readings are collected for HIF_FAULT, and 3,547 readings for CAP_SWITCH with a 50 μs time-step (τ) and saved as *.mat files. Each instance of the dataset also contains a "bus" variable that stores locus of event. This is unnecessary as the goal is to detect whether certain transients are caused due to HIF_FAULT or CAP_SWITCH, and not to identify the location. Also, the work only considers single-phase HIFs because, 95% of all HIF instances is single phase to ground fault [5], [20].

The *.mat files are then processed into *.csv files after dropping the "bus" variables and the voltage readings. This is because, the HIF event signatures are more prominent in the current readings [14]. A complete outline of the processed dataset can be seen from Fig. 1. Each reading in the dataset is 0.7 s long, with the switching event being initiated at 0.2 s and cleared at 0.4 s. Two samples of three-phase current readings at bus 57 during a CAP_SWITCH event and a HIF_FAULT event are shown in Fig. 2. A train-validation-test split of 70%-10%-20% is done.

III. PROPOSED HIF DETECTION

The methodology adopted for the proposed ML and DL models can be obtained from Fig. 3.

At first, the dataset is preprocessed for the respective ML and DL algorithms. For all DL-specific tasks, the TensorFlow

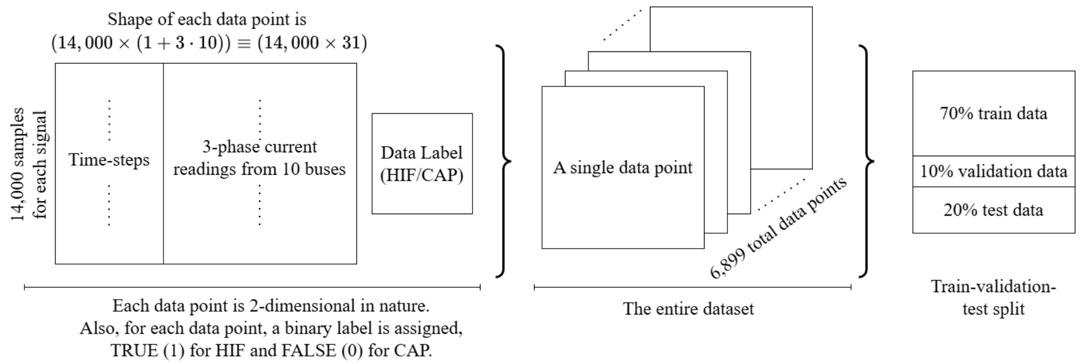
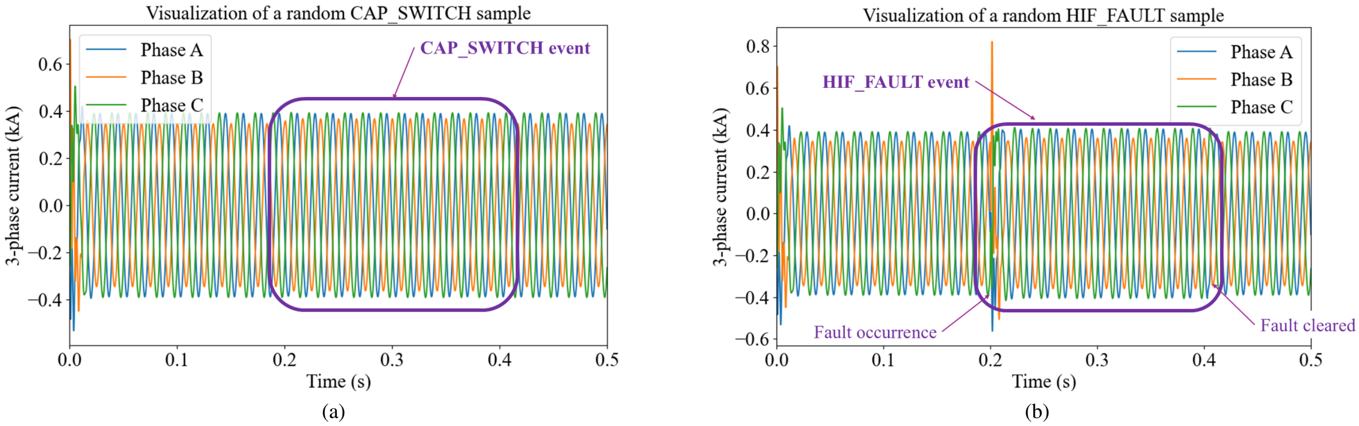


Fig. 1: Summary of processed dataset



(a)

(b)

Fig. 2: Side-by-side comparison of the first 0.5s of a (a) CAP_SWITCH event and a (b) HIF_FAULT event

library is used in Python. Besides, other ML-related tasks are conducted using SciKit-Learn.

All of the models are tuned on their best hyperparameters based on a grid search algorithm. A three-fold cross-validation is run upon the validation set to arrive at these hyperparameters. The most optimum hyperparameters obtained are described in the following subsections.

A. Support Vector Machines

SVMs create a hyperplane to separate classes in a classification problem by means of kernels. The theory behind SVMs can be found in [28]. The tuned hyperparameters for the SVM model for the HIF detection dataset are mentioned in Table I.

TABLE I: Tuned Hyperparameters for SVM

Hyperparameter	Tuned Value
Regularization Parameter (C)	0.1
Kernel	Linear

B. Random Forest

Random forests (RF) formulate a classification task using multiple trees. The target classes remain at the leaves of the trees. RFs aggregate the decisions obtained from multiple trees through a voting mechanism, which provides the final decision. The mathematical basis of RF can be found in [29]. The hyperparameters considered for the RF classifier in these experiments are mentioned in Table II.

TABLE II: Tuned Hyperparameters for RF

Hyperparameter	Tuned Value
Estimator number	0.1
Maximum depth	None

C. Decision Tree

DTs can approximate several hyperplanes to generate the decision boundary, as opposed to a single hyperplane in SVMs. The unique benefit of DTs over DL is that the training time is significantly faster, and it utilizes fewer resources. The hyperparameters used for DT are given in Table III.

TABLE III: Tuned Hyperparameters for DT

Hyperparameter	Tuned Value
Criterion	Gini
Splitter	Best
Minimum number of samples at leaf	1
Minimum samples split	2

D. Convolutional Neural Networks

CNNs have been used extensively for image processing. However, the utility of this algorithm is not limited to pictorial fields. For example in [30], CNNs have been used for

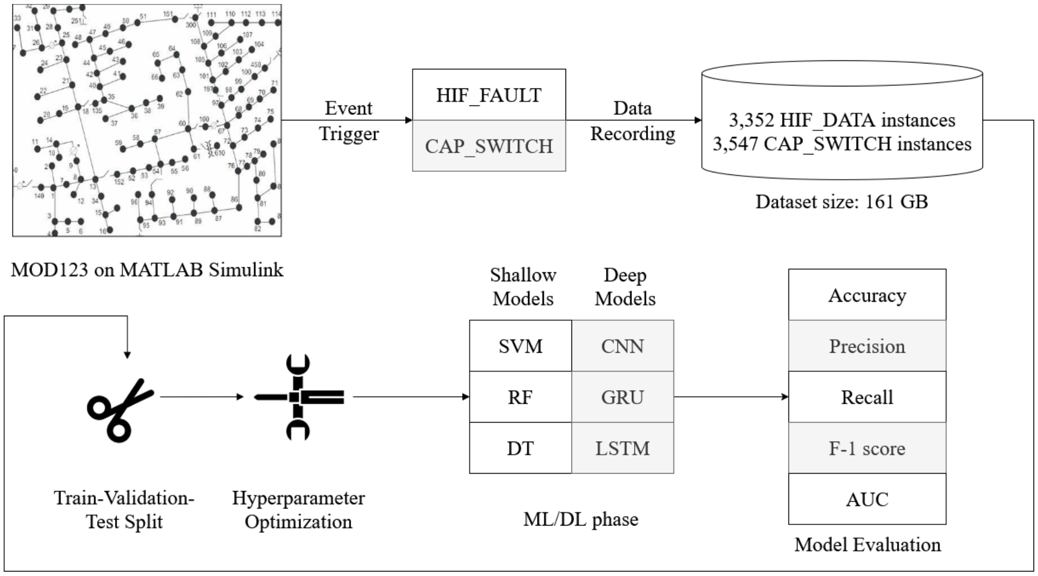


Fig. 3: Methodology adopted for the study. The details of MOD123 can be found in [26].

sentiment analysis. This is because, CNNs work well with datasets of higher dimensionality. Therefore, given that the dataset of this work is three-dimensional (3d) in nature, CNN is an optimal candidate. CNN can be considered to be a special kind of feed-forward neural network (FNN), which takes higher dimensional matrices as samples. The matrices get passed through convolutional layers that produce feature maps, and pooling layers, that reduce the feature space. Besides, filters reduce the number of parameters during the learning process. Filters are useful when mapping complex features [31]. Afterwards, the data is flattened and passed through fully-connected layers. Flattening converts all higher dimensional matrices to one-dimensional (1d) arrays. For this reason, CNNs are useful with data of higher dimensionality. At the end, a dense layer with the appropriate activation function is used according to the task at hand. CNNs are used mostly for classification problems [32].

For the CNN architecture developed for the work endisclosed, the number of layers, neurons, filters, dropout rates, activation functions, optimizer and learning rate (lr) have been tuned. The architecture of the most optimum CNN model is given in Fig. 4 (a). The optimum optimizer was found to be root mean square propagation (RMSprop) with an lr of 0.0001, the number of epochs was found to be 10 and the batch size to be 512.

E. Gated Recurrent Units

Gated Recurrent Units (GRU) is a kind of Recurrent Neural Network (RNN). GRU can account for the temporal dependence of the dataset, thus making it specifically useful with time-series data. The mathematics behind GRUs are discussed in [33].

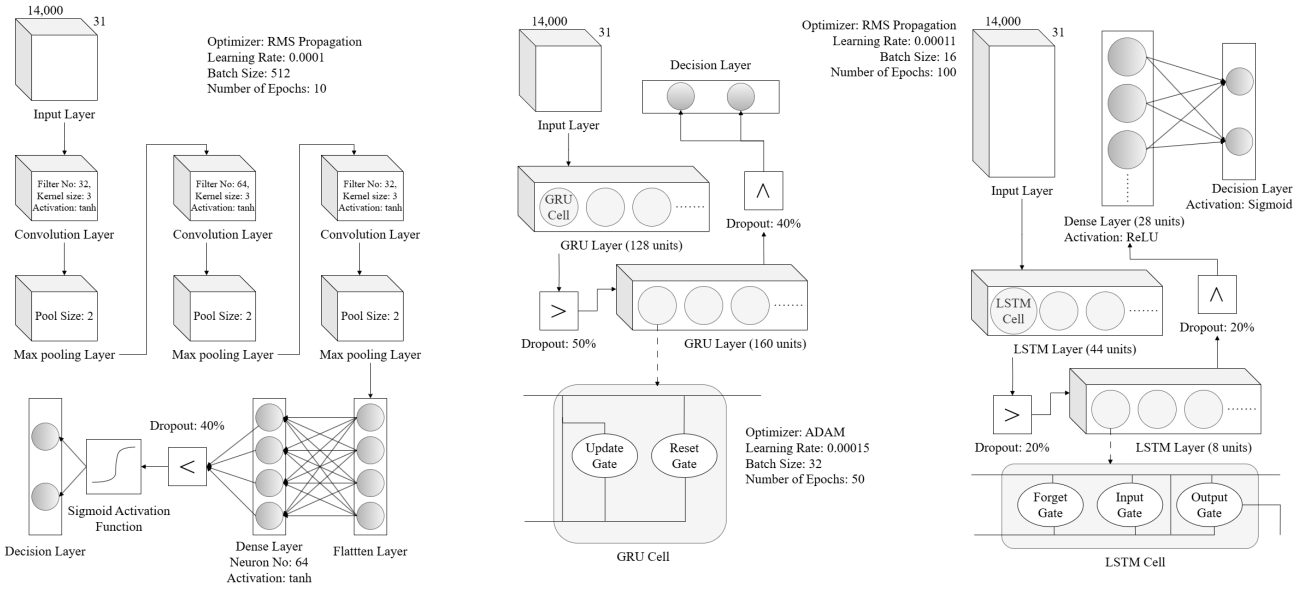
As of now, only one work has been done with GRU in the field of HIF detection [17]. This leaves untapped potential and therefore scope of research in the field. GRU is a more modern algorithm, and it uses the update gate and the reset gate to deal with the vanishing gradient problem of recurrent neural networks (RNN). Because GRU cells have fewer gates, they are faster than LSTM cells. A GRU model is trained after tuning the hyperparameters using a grid search process for this work. The number of neurons, number of layers, activation function, optimizer, lr, epoch no, batch size and dropout rate have been tuned. The final architecture after tuning is given in Fig. 4 (b). An Adaptive Moment Estimation (ADAM) optimizer is obtained with an lr of 0.00015 by means of hyperparameter optimization. The number of epochs was found to be 50 and the batch size 32.

F. Long Short Term Memory

LSTMs are advanced RNNs that are particularly effective for analyzing time series data due to their ability to capture spatio-temporal dynamics [2]—an advantage over CNNs and other FNNs. This capability is facilitated by their unique memory units, which consist of an input gate, a forget gate, and an output gate. These help dealing with the vanishing gradient problem. Unlike standard RNNs, LSTMs incorporate loops within their neurons, enabling them to maintain a memory of past information [34]. The most optimum LSTM architecture upon hyperparameter tuning is illustrated in Fig. 4 (c). The number of layers, neurons, the activation function, dropout rate, number of epochs and batch size have been tuned. The RMSprop optimizer is used with an lr of 0.00011. The batch size is kept at 16 and the number of epochs is set to 100.

IV. EXPERIMENTAL RESULTS

The problem is formulated to be a binary classification problem. The models are tuned on the validation set and the performances are evaluated on the test set. Five evaluation metrics are used to compare the six models under discussion, and these are accuracy (ACC), precision (DR), recall (RE),



(a) (b) (c)

Fig. 4: Hyperparameter-tuned DL architectures: (a) CNN, (b) GRU, (c) LSTM

F1 score [35], [36] and area under curve (AUC) [37]. The performance of each of the models for HIF detection on MOD123 is given in Table IV.

TABLE IV: Comparative analysis among the models

Model	SVM	RF	DT	CNN	GRU	LSTM
ACC	84.000%	99.634%	100%	95.306%	97.074%	97.074%
DR	99.569%	100%	100%	94.828%	100%	100%
RE	99.338%	99.282%	100%	94.908%	94.253%	94.253%
F1	86.087%	99.640%	100%	97.354%	97.041%	97.041%
AUC	94.933%	99.640%	100%	99.714%	95.689%	97.126%

A. Performance of Shallow Models

It can be seen here that the shallow models (SVM, RF and DT) outperform the deeper models (CNN, GRU, and LSTM). DT exhibits a near-perfect score. This can be attributed to DT's pattern recognition skills. Although it is difficult for the naked eye to distinguish between other switching events and HIF, there clearly is an underlying pattern to it, which DTs can capture. It is noteworthy here that DTs have also been the fastest models to train.

The dataset needed flattening before training the shallow models because they don't adapt well to such data complexity. This discards certain correlations among the data, which is a flaw of shallow models.

B. Performance of Deep Models

With the availability of data, deeper models tend to perform better. Here comes the need for a trade-off. Available power systems data is disorganized and unlabeled. The real question is, given the results that have been achieved upon the available data, would the small boost in performance for the deeper models with the integration of more data be significant?

LSTMs outperform GRUs only slightly, with it having a 1.502% higher AUC. This is because LSTMs perform better with larger datasets. However, the RNN algorithms not outperforming CNN too well (0.989% increase in accuracy) shows that temporality is not very important for the case of HIF detection. The time-steps recorded during HIF simulations do not represent a global clock, meaning that the entire data-point may appear at any time for the duration of the recorded time, and this information is not significant in detecting HIFs.

C. Performance Remarks

The false alarm rate (FAR) has been calculated but not presented here, because it was found mostly to be 0 across all models. This means that none of the models is misclassifying CAP_SWITCH events. This means learning CAP_SWITCH events is too easy. The same experiments can, therefore, be done with fewer CAP_SWITCH instances.

Although HIF events are only slightly differentiable from switching events to the naked eye, history-based artificially intelligent (AI) algorithms can capture the required discrepancies, thus learning key trends. Shallower models perform better than deeper models because the data of the same classes contain similar trends, which are indistinguishable to humans but are too simple to distinguish for machines. The more complex algorithms learn the trends a bit too well, thus leading to overfitting. The complexity of the HIF fault data does not appear to necessitate DL models. Shallow models seem to be free from these overfitting tendencies, thus fulfilling the goal of this work.

V. CONCLUSION

The purpose of this paper is to formulate the detection of HIF as a binary classification task and solve the issue of HIF detection. A dataset comprising waveform readings of both HIF and capacitor switching events was synthesized in the laboratory. The waveforms were used to train three shallow models, namely, SVM, RF and DT, and three deep models, namely, CNN, GRU and LSTM. In the end, the prediction accuracies of the models were analyzed. Although most models in consideration portray decent performance, shallower models may be good candidates for the formulation of a standard for HIF detection. Among SVM, RF and DT, DT showed the best results. As can be seen from the evaluation metrics, all models perform well but only subjectively, depending on what the criterion for judgment is. That is to say, the utility of each model depends on the specific needs of the researcher.

ACKNOWLEDGMENT

We express our deepest gratitude to Chashi Mahiul Islam, who was instrumental in the success of the LSTM model.

REFERENCES

[1] K. Rai *et al.*, "Deep learning for high-impedance fault detection and classification: transformer-CNN," *Neural Computing and Applications*, vol. 34, no. 16, pp. 14 067–14 084, 2022.

[2] H. Kumar *et al.*, "High resistance fault detection in DC microgrid using Hilbert Huang transform and vector-based ensemble optimized LSTM networks," *Scientific Reports*, vol. 15, no. 1, pp. 1–26, 2025.

[3] R. V. Pandikuddy *et al.*, "Performance evaluation of MODWT families for high impedance fault detection in a mesh network using Bi-LSTM," in *AIP Conference Proceedings*, vol. 3311, no. 1. AIP Publishing LLC, 2025, p. 070002.

[4] A. Sultan and G. Swift, "Security testing of high impedance fault detectors," in *[Proceedings] WESCANEX'91*. IEEE, 1991, pp. 191–197.

[5] M. M. Islam, "Application of artificial intelligence for detecting distribution system faults and power quality disturbances," PhD thesis, Florida State University, Tallahassee, FL, December 2023.

[6] S. S. Khan and H. Wen, "A comprehensive review of fault diagnosis and tolerant control in DC-DC converters for DC microgrids," *IEEE Access*, vol. 9, pp. 80 100–80 127, 2021.

[7] A. Emanuel *et al.*, "High impedance fault arcing on sandy soil in 15 kV distribution feeders: contributions to the evaluation of the low frequency spectrum," *IEEE Transactions on Power Delivery*, vol. 5, no. 2, pp. 676–686, 1990.

[8] J. Carr, "Detection of high impedance faults on multi-grounded primary distribution systems," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-100, no. 4, pp. 2008–2016, 1981.

[9] S. Chakraborty and S. Das, "Application of smart meters in high impedance fault detection on distribution systems," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3465–3473, 2019.

[10] B. Wang *et al.*, "High-impedance fault detection based on nonlinear voltage–current characteristic profile identification," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3783–3791, 2018.

[11] S. Gautam and S. M. Brahma, "Detection of high impedance fault in power distribution systems using mathematical morphology," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1226–1234, 2013.

[12] A. Meier *et al.*, "Synchrophasor monitoring for distribution systems - technical foundations and applications," 01 2018.

[13] H. Saadat *et al.*, *Power system analysis*. McGraw-hill, 1999, vol. 2.

[14] Y. Sheng and S. Rovnyak, "Decision tree-based methodology for high impedance fault detection," *IEEE Transactions on Power Delivery*, vol. 19, no. 2, pp. 533–536, 2004.

[15] B. K. Chaitanya *et al.*, "An intelligent detection of high-impedance faults for distribution lines integrated with distributed generators," *IEEE Systems Journal*, vol. 14, no. 1, pp. 870–879, 2020.

[16] H. Lala and S. Karmakar, "Detection and experimental validation of high impedance arc fault in distribution system using empirical mode decomposition," *IEEE Systems Journal*, vol. 14, no. 3, pp. 3494–3505, 2020.

[17] M. M. Irfan *et al.*, "Bi-directional gated recurrent unit approach for detecting and classifying high impedance fault in power distribution," in *2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*. IEEE, 2024, pp. 1–5.

[18] M. Bhatnagar *et al.*, "LSTM-based low-impedance fault and high-impedance fault detection and classification," *Electrical Engineering*, vol. 106, no. 5, pp. 6589–6613, 2024.

[19] I. Abasi-obot *et al.*, "High impedance fault detection and localization using fully-connected convolutional neural network: A deep learning approach," *Nigerian Journal of Technological Development*, vol. 20, no. 4, pp. 62–71, 2023.

[20] J. P. Nelson, "High impedance ground fault detection for a safer petrochemical industry," in *2021 IEEE IAS Petroleum and Chemical Industry Technical Conference (PCIC)*. IEEE, 2021, pp. 1–8.

[21] M.-F. Guo *et al.*, "High-impedance fault detection methodology using time–frequency spectrum and transfer convolutional neural network in distribution networks," *IEEE Systems Journal*, vol. 17, no. 3, pp. 4002–4013, 2023.

[22] G. A. Rakib *et al.*, "Innereye: A tale on images filtered using instagram filters-how do we interact with them and how can we automatically identify the extent of filtering?" in *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*. Springer, 2022, pp. 494–514.

[23] C. M. Islam *et al.*, "Malicious path manipulations via exploitation of representation vulnerabilities of vision-language navigation systems," *arXiv preprint arXiv:2407.07392*, 2024.

[24] M. M. Islam *et al.*, "Unsupervised clustering of disturbances in power systems via deep convolutional autoencoders," in *2023 IEEE Power Energy Society General Meeting (PESGM)*, 2023, pp. 1–5.

[25] Y. Cheng *et al.*, "Online power system event detection via bidirectional generative adversarial networks," *IEEE Transactions on Power Systems*, vol. 37, no. 6, pp. 4807–4818, 2022.

[26] N. Samaan *et al.*, "Combined transmission and distribution test system to study high penetration of distributed solar generation," 04 2018.

[27] M. Jamei *et al.*, "Phasor measurement units optimal placement and performance limits for fault localization," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 1, pp. 180–192, 2019.

[28] R. Sharda, S. Voß, and S. Suthaharan, "Support vector machine," *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pp. 207–235, 2016.

[29] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, 2017.

[30] M. R. Kabir *et al.*, "Procuring MFCCs from Crema-D Dataset for Sentiment Analysis using Deep Learning Models with Hyperparameter Tuning," in *2021 IEEE International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON)*, 2021, pp. 50–55.

[31] S. Min *et al.*, "Deep learning in bioinformatics," *Briefings in bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017.

[32] S. Patel and A. Patel, "Deep Leaning Architectures and its Applications: A Survey," *International Journal of Computer Sciences and Engineering*, vol. 6, pp. 1177–1183, 06 2018.

[33] J. Chung *et al.*, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[35] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, no. 1, p. 6086, 2024.

[36] B. J. Erickson and F. Kitamura, "Performance metrics for machine learning models," *Radiology: Artificial Intelligence*, vol. 3, no. 3, p. e200126, 2021.

[37] G. Naidu, T. Zuva, and E. M. Sibanda, "A review of evaluation metrics in machine learning algorithms," in *Computer science on-line conference*. Springer, 2023, pp. 15–25.