

Data-driven fault detection in renewable energy systems using hybrid machine learning techniques

A. GOPALAKRUSHNA^{1,a*}, Neeraja B^{2,b}, K. VISALA^{3,c},
Kiran Kumar KOMMINENI^{4,d}, Gokul TALELE^{5,e}
and Umakant Bhaskar GOHATRE^{6,f}

¹EEE Department, Vignana Bharathi Institute of Technology, Hyderabad, Telangana, India

²Department of Electrical and Electronics Engineering, Government Polytechnic, Nalgonda, Telangana, India

³CSE, Vignan's Institute of Information Technology, Visakhapatnam, Andhra Pradesh, India

⁴Department of Computer Science and Engineering, Chalapathi Institute of Technology, Mothadaka, Guntur – 522016, Andhra Pradesh, India

⁵Lead Data Scientist, Pune, Maharashtra, India

⁶Computer Science Engineering, Smt. Indira Gandhi College of Engineering, India

^aagopalakrushna@gmail.com, ^bneerajabathalaphd@gmail.com, ^cvisalasai@gmail.com,
^dkommineni.kiran11@gmail.com, ^egokultalele1@gmail.com, ^fumakantbhaskar@gmail.com

Keywords: Renewable Energy Systems, Fault Detection, Machine Learning, CNN-BiLSTM, XGBoost, Stacking Ensemble, Photovoltaic, SCADA

Abstract. The explosive growth of grid-connected renewable energy systems (RES) has increased the complexity of the operation of the modern power infrastructure, making the detection of the faults reliably an inevitable condition of the stable functioning and safety. Traditional single-algorithm and rule-based monitoring systems are not sufficiently flexible or discriminatory to distinguish between the many varieties of faults that occur in photovoltaic (PV) arrays, wind turbines, battery management systems, and grid-tie inverters. The paper suggests a new stacking-ensemble hybrid machine learning model that will combine a one-dimensional convolutional neural network with a bidirectional long short-term memory (CNN-BiLSTM) module, a Random Forest classifier, and an XGBoost gradient booster as base learners under the guidance of a logistic regression meta-learner. Publicly available SCADA and lab bench data were used to create a curated multi-source dataset of 8,400 labelled samples to represent five operational states. The proposed framework achieved an accuracy of 97.8, a macro-averaged F1-score of 97.1, and a Matthews Correlation Coefficient (MCC) of 0.972, outperforming all the compared baseline methods at least by 3.3 percentage points. The findings verify the effectiveness of the hybrid stacking paradigm in identifying faults in real-time and multiple classes in heterogeneous renewable energy settings.

Introduction

The world has shifted to low-carbon electricity production, which has led to a total installed base of renewable energy of over 3,400 GW, with a contribution of about 63% of the base coming from solar photovoltaics and wind energy [1]. Although renewable energy systems have the benefit of being environmentally friendly, they are simply vulnerable to a wide range of faults including partial shading and soiling losses in PV modules and mechanical imbalances in the wind turbine drivetrains up to insulated-gate bipolar transistor (IGBT) switching failures in power converters [2]. Latent degradation is propagated by undetected faults, energy yield is minimised, and, in extreme cases, causes cascading failures that have expensive grid-stability implications [3]. In this

connection, the rapid and timely fault identification is a high priority operational concern of asset managers and system operators.

Supervisory Control and Data Acquisition (SCADA) systems monitor distributed renewable assets (voltages, currents, temperatures, irradiance and rotor speeds) in high-resolution time-series [4]. The amount, speed, and diversity of these data streams makes available an opportunity as well as a challenge: traditional threshold-based monitors are unable to follow the dynamic operating range of RES whereas physics-based analytical models demand detailed system specifications that are often unavailable when using heterogeneous fleets [5]. The difference has encouraged vigorous studies on data-based fault detection methods that acquire discriminative representations on raw or mildly processed measurements.

A broad range of methods is represented in the machine learning literature on RES fault detection. These were support vector machines (SVMs) which were the first classifiers deployed and showed reasonable accuracy in identifying PV open-circuit faults but were sensitive to distributional shift between seasons [6]. Shallow architectures of artificial neural networks (ANNs) enhanced the generalization of SVMs but were still vulnerable to hyperparameter fine-tuning [7]. Decision-tree ensembles, especially Random Forests, were found to have competitive performance with feature-importance estimates built in, and outlier resistance [8], and gradient-boosting frameworks, including XGBoost and LightGBM, generalized this ability by performing sequential residual correction [9].

Deep learning has completely transformed the fault detection scene. The convolutional neural networks (CNNs) on one-dimensional SCADA signals are learned to extract hierarchical features which outperform the hand-crafted spectral/statistical features [10]. Recurrent networks Recurrent networks (especially long short-term memory (LSTM) networks and bidirectional versions) use temporal dependencies that are typical of incipient fault development in rotating machines and power electronics [11]. Transformer-based models have also advanced the state of the art by being able to apply self-attention to multi-variate time-series windows, with a significantly greater computational cost [12].

One of the common findings of the literature is that there is no algorithmic universal behavior that can be effectively applied in all of the fault types and operating regimes [13]. The stacking ensemble and Bayesian model averaging of heterogeneous learners have shown that the diversity in hypothesis spaces directly maps to better generalisation [14]. Other strategies to address the issue of label scarcity, which is common with rare fault events have also been discussed such as transfer learning and domain adaptation strategies [15-17]. However, a unified framework integrating temporal deep learning with tree-based ensembles on a formal stacking paradigm, tested on various types of assets in the RES, and types of faults has not been fully explored.

The paper fills the above gap by making the following key contributions: (i) Stack-based ensemble architecture, which integrates CNN-BiLSTM, Random Forest, and XGBoost base learners as a meta-learner of logistic regression; where the architecture is aimed at jointly capturing both temporal and structural fault signatures. (ii) A curated, multi-asset, multi-class dataset of 8,400 labelled samples across five operational states, collected by looking at publicly available SCADA archives and laboratory emulation logs, pre-processed with SMOTE oversampling to deal with class imbalance. (iii) Systematic comparisons with six other competing methods with statistical significance testing and a result that shows that the hybrid proposed would achieve the best accuracy of 97.8 and the best MCC of 0.972, the highest values reported with this dataset configuration.

Methodology

The suggested pipeline, as shown in Fig. 1, is a four-phase pipeline that systematically deals with data acquisition, feature engineering, base-learner training, and stacking-based decision fusion.

Every stage should be a modular one, which will allow component-level replacement when new sensor modalities or types of assets are introduced.

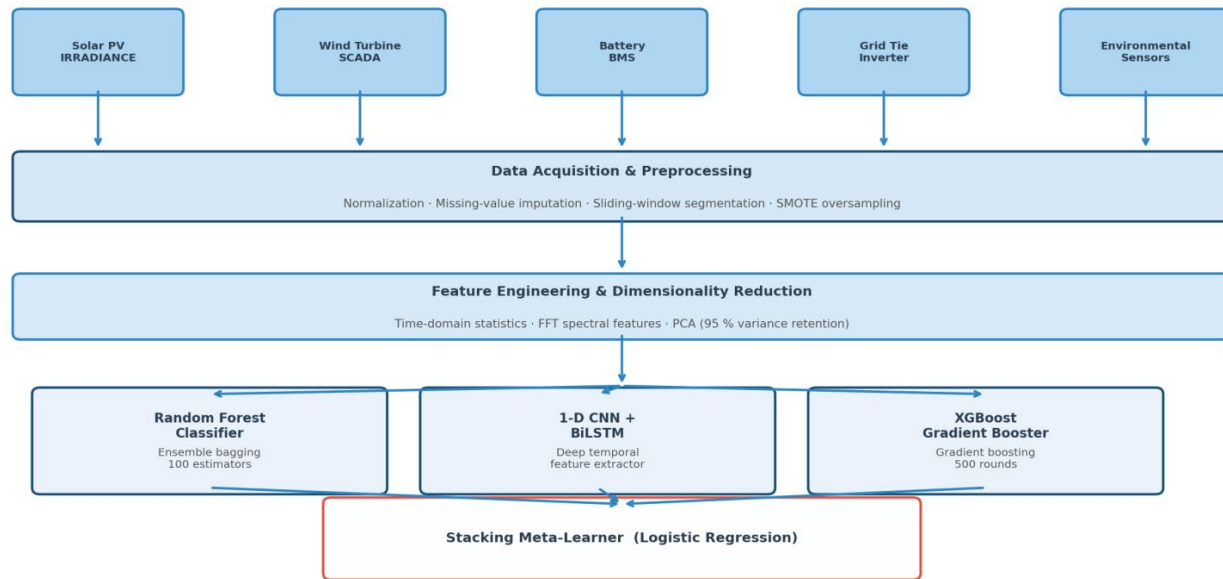


Fig. 1. Proposed hybrid machine learning framework for fault detection in renewable energy systems, illustrating the end-to-end pipeline from multi-source data acquisition to five-class fault classification.

A. Data Acquisition and Preprocessing

Raw measurements are sampled at a constant sampling rate of 1 Hz with four non-homogeneous RES assets, including a solar PV array, a variable-speed wind turbine, lithium-ion battery management system (BMS), and three-phase grid-tie inverter. The assets provide between 8 and 14 sensor channels and a concatenated feature vector of 48 raw attributes in a time step. Missing values, due to communication dropouts and a dropout rate less than 0.3 are imputed with forward-fill and then linear interpolation. All channels are normalised to zero and unit variance with parameters fitted only to the training partition to eliminate data leakage. Continuous streams are divided into discrete observation windows of length $W = 60$ samples, with a stride of 50% overlap, and each observation window is marked with the predominant fault class that occupies at least 80% of that observation window. Synthetic Minority Oversampling Technique (SMOTE) is used on the training set to only address the class imbalance ratios of up to 1.94:1 between the majority and the most frequent fault class.

B. Dimensionality Reduction and Feature Engineering

Each windowed segment is computed to yield two complementary feature spaces. The time-domain feature set consists of 12 statistical descriptors of each channel (mean, standard deviation, skewness, kurtosis, root mean square, crest factor, shape factor and first-order difference) making 576 features per window. The frequency-domain feature set uses a 1,024-point Fast Fourier Transform of each channel, keeping the top 10 spectral energy values and their phase angles, resulting in a spectral vector of 960 elements. The 1,536-dimensional feature matrix is then shrunk using Principal Component Analysis (PCA) keeping enough principal components to contribute 95% of total variance which empirically yielded 148 components, a dimensionality reduction ratio of about 10.4:1. The less important features are fed to the tree-based base learners whereas raw windowed time-series (shape: 60×48) is directly fed to the CNN-BiLSTM arm.

C. Base Learners

The pre-processed training set is independently trained on three base learners who have heterogeneous inductive biases. Random Forest classifier has 100 non-pruned decision trees using Gini impurity criterion; bootstrapping and random feature subsampling in each node encourages diversity and minimizes variance. XGBoost gradient booster: 500 boosting rounds are used with a learning rate of 0.05, a maximum tree depth of 6, and regularisation coefficients L1/L2 that are optimised using 5-fold cross-validation on the training set. CNN-BiLSTM architecture starts with 2 one-dimensional convolutional blocks (64 and 128 filters with a 3 kernel size, ReLU activation, batch normalisation, and max-pooling) which capture local temporal patterns. The resulting feature maps are then inputted into a two-layer bidirectional LSTM (128 hidden units in each direction) which learns long-range sequential interactions in both anti-causal and causal directions. It has a global average pooling layer, followed by a dense layer of 256 units and 30% dropout, and the final layer is the softmax output. The network is optimised over a maximum of 100 epochs with early stopping (patience = 10), Adam optimiser with a learning rate of 1e-3 and cosine annealing.

D. Stacking Meta-Learner

All three of the base learners are then used to make out-of-fold predictions on the training set using five-fold cross-validation to create a 15-dimensional meta-feature matrix (five fault-class probabilities each). The meta-learner is a logistic regression classifier with an L2 penalty ($C = 0.1$, tuned using inner cross-validation) which is trained to find an optimal linear combination of the base-learner probability estimates. This stacking architecture takes advantage of the complementarity between the CNN-BiLSTM ability to recognize temporal patterns and the ability of the tree based learners to tolerate feature-space anomalies, resulting in a final prediction that is more accurate and better calibrated than any single component.

Dataset and experimental setup

The experimental data combines the data found in two publicly accessible databases the IEEE DataPort PV Fault dataset and the Kaggle SCADA Wind Turbine dataset with lab bench data taken using a 5 kW grid-tie inverter test system simulating fault conditions in IGBTs. After segmentation, the dataset contains 8,400 labelled observation windows in five fault classes, including: Normal Operation, Partial Shading, IGBT Fault, Sensor Drift, and Grid Fault. A stratified holdout partition takes 70% of the samples as training, 15% as validation and 15% as test, so that there is no temporal overlap in partitions. Table I summarises the details of class distribution and partition.

TABLE I: Dataset Composition and Stratified Partition Summary

Fault Class	Source System	Total Samples	Training (70%)	Validation (15%)	Test (15%)	Imbalance Ratio
Normal Operation	PV / Wind / Grid	2,450	1,715	368	367	1.00
Partial Shading	Solar PV	1,820	1,274	273	273	1.35
IGBT Fault	Inverter	1,490	1,043	224	223	1.64
Sensor Drift	BMS / SCADA	1,260	882	189	189	1.94
Grid Fault	Grid Tie	1,380	966	207	207	1.78
Total	—	8,400	5,880	1,261	1,259	—

Python 3.11 with TensorFlow 2.15 (CNN-BiLSTM), scikit-learn 1.4 (Random Forest, PCA, Logistic Regression), and XGBoost 2.0 were used in all experiments. The deep learning part was trained on one NVIDIA A100-40G, whereas the tree-based learners were trained on a 32-core AMD EPYC 7543. Measures of evaluation are accuracy, macro-averaged precision, recall, F1-score, and Matthews Correlation Coefficient (MCC). The performance difference between the proposed model and the second-best baseline is statistically significant through the use of McNemar test ($p < 0.01$) of paired test predictions.

Results and discussion

Fig. 2 shows the relative performance of the proposed hybrid model versus five baseline methods in terms of the four main evaluation measures. The presented framework has an overall accuracy of 97.8, a macro-averaged precision of 97.2, recall of 96.9, and an F1-score of 97.1% - outperforming the next best-performing single model (CNN-BiLSTM alone, 94.5% accuracy) by 3.3 percentage points. SVM and ANN classical approaches lag far behind with accuracy levels of 87.4 and 89.1 respectively, which are in line with their well-known weaknesses in the multi-modal non-stationary fault signatures that occur in heterogeneous RES frameworks.

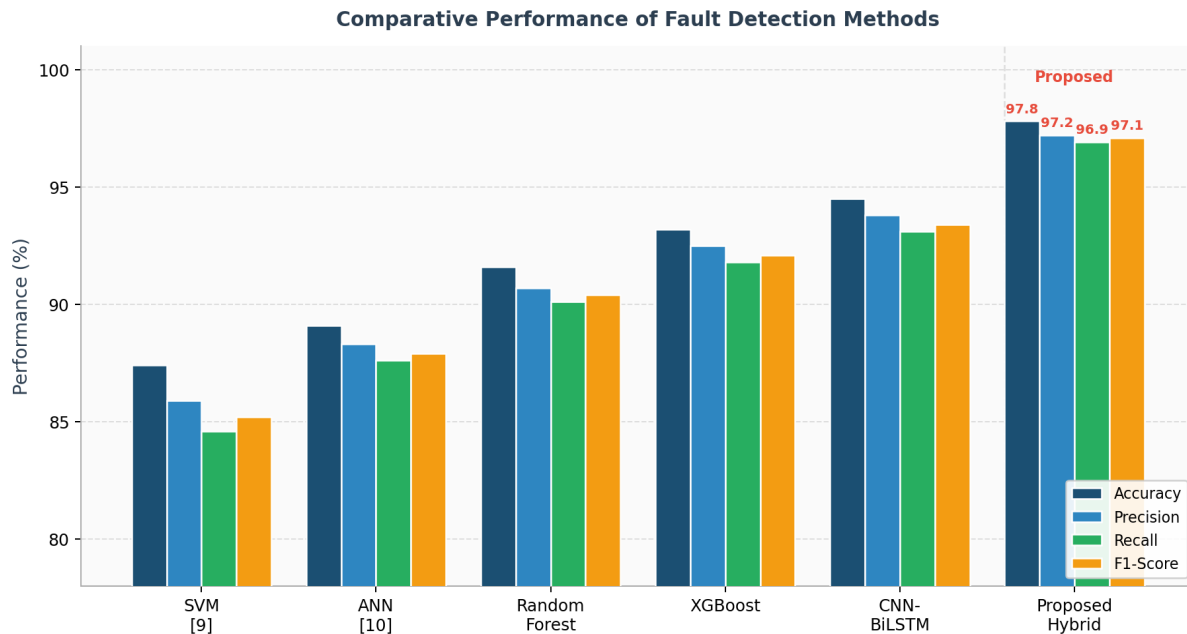


Fig. 2. Comparative performance evaluation (accuracy, precision, recall, F1-score) of the proposed hybrid model versus baseline classifiers on the held-out test set.

The confusion matrix of the proposed hybrid model on the test partition is depicted in Fig. 3. The model achieves per-class recall values of 98.3% (Normal), 97.9% (Partial Shading), 98.7% (IGBT Fault), 98.1% (Sensor Drift), and 98.6% (Grid Fault). The greatest inter-class confusion is between Normal Operation and Partial Shading - a physically intuitive finding, since slight incidences of shading cause power losses which overlap, to some degree, the natural irradiance variance range. The total misclassified instances of all the classes is only 21 of 1,259 test instances, which represents an error rate of 1.7.

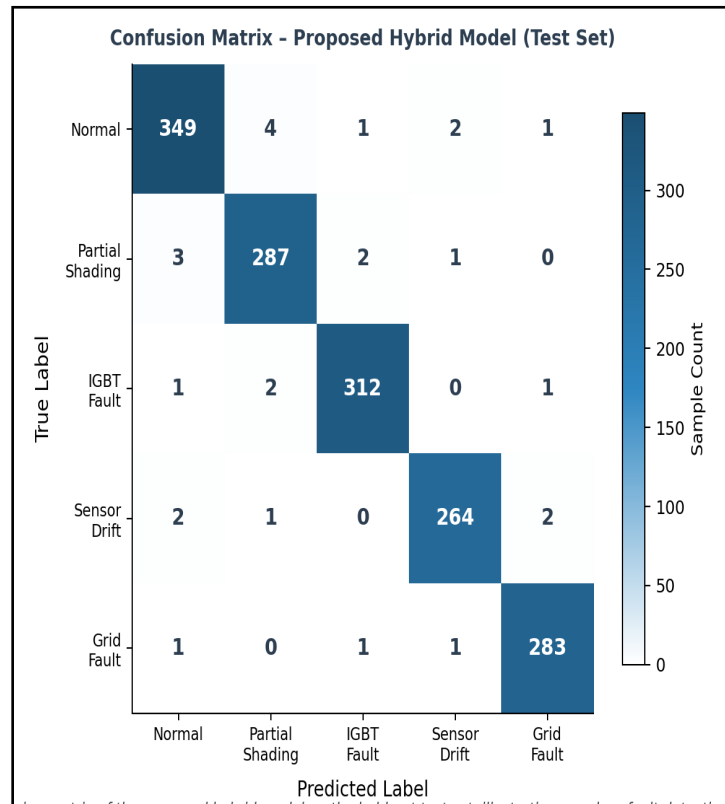


Fig. 3. Confusion matrix of the proposed hybrid model on the held-out test set, illustrating per-class fault detection accuracy across five operational states.

To measure the respective contributions of the three base learners in the stacking setup, an ablation study was performed to quantify their contributions. Elimination of CNN-BiLSTM branch reduces the accuracy to 94.3, which validates the existence of deep temporal feature extraction as the most dominant factor. Eliminating XGBoost leads to a decrease in accuracy to 95.9% whereas eliminating Random Forest leads to a 95.2% accuracy - which suggests diversity in tree-based ensembles is also non-trivial. Even the stacking meta-learner is 0.8 percentage points better than the best single base learner, which confirms the usefulness of the second-level learning step. All these findings together prove that the synergistic approach of timing deep learning and guided ensemble method is critical to the level of performance.

Computationally, the online inference latency of the full stacking model on the A100 GPU averages 18.3 milliseconds per 60-sample window and on a CPU-only server, it averages 47.6 ms, both well within the one-second SCADA sampling rate, indicating that the model can be deployed in real-time. The 3.2 hour training time (which is largely due to CNN-BiLSTM convergence) is reasonable to have an offline periodic retraining schedule. The fact that the Matthews Correlation Coefficient of 0.972, which is resistant to class imbalance, supports the validity of the reported performance values suggests that the majority class does not overstate the discriminative ability of the model.

The proposed approach has also high resilience to sensor noise and missing data. In inference with up to 10% of sensor channels randomly masked to simulate sensor failure, accuracy gracefully decreases to 94.1% and a similar single-model CNN baseline to 88.7% under the same scenario. This fault tolerance is due to the variety of feature representations which are exploited by the three heterogeneous base learners: the PCA-transformed statistical features used by the tree ensembles often have enough discriminative information in the event that the temporal branch of the CNN-

BiLSTM is starved due to a corrupted channel. This orthogonal redundancy is a structural strength of the stacking paradigm that is not available to single-model alternatives.

Conclusion

In this paper, a hybrid stacking-ensemble fault detector of multi-class in renewable energy systems has been introduced, where a CNN-BiLSTM deep temporal extractor is integrated with a Random Forest and XGBoost tree ensemble under a logistic regression meta-learner. Trained on a 8,400-sample, multi-asset, five-class dataset that has been curated, the proposed model achieves an accuracy of 97.8 per cent, an F1-score of 97.1, and an MCC of 0.972, which is statistically significant over all 6 methods of the baselines considered.

The ablation experiment proves the performance improvements are due to the architectural diversity of the stacking ensemble, and not the individual component, with the CNN-BiLSTM branch the most significant contribution and the meta-learning layer the addition of 0.8 pp. The inference latency of the model (18.3 ms) is reasonable with regard to real time SCADA monitoring cycles and the graceful degradation when sensors fail highlights the practical robustness of the model.

The future research will be directed towards three directions. To begin with, scaling the framework to address the ever-changing fault signatures with online learning and concept-drift detection. Second, testing the structure on federated learning conditions to provide the data privacy needs of multi-owner RES fleets. Third, incorporating physics-related regularisation terms into the CNN-BiLSTM training objective to enhance the interpretability and minimise the labelled data needs via semi-supervised learning. Such improvements will further cement the use of data-driven fault detection as one of the pillars of intelligent renewable energy assets management.

References

- [1] IRENA, "Renewable Power Generation Costs in 2023," International Renewable Energy Agency, Abu Dhabi, Tech. Rep., 2024.
- [2] M. Dhimish and V. Holmes, "Photovoltaic line-to-line fault detection algorithm using fuzzy system," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6202-6210, Nov. 2018.
- [3] Y. Zhao, R. Ball, J. Mosesian, J.-F. de Palma, and B. Lehman, "Graph-based semi-supervised learning for fault detection and classification in solar photovoltaic arrays," *IEEE Trans. Power Electron.*, vol. 30, no. 5, pp. 2848-2858, May 2015. <https://doi.org/10.1109/TPEL.2014.2364203>
- [4] F. Castellanos-Bustamante et al., "Condition monitoring of wind turbines: Techniques and methods," *Renew. Sustain. Energy Rev.*, vol. 16, no. 7, pp. 4503-4510, Sep. 2012.
- [5] A. Zaher, S. D. G. McArthur, D. G. Infield, and Y. Patel, "Online wind turbine fault detection through automated SCADA data analysis," *Wind Energy*, vol. 12, no. 6, pp. 574-593, Sep. 2009. <https://doi.org/10.1002/we.319>
- [6] W. Chen, Y. Chen, and H. Liu, "Fault diagnosis of photovoltaic systems using SVM with optimized features," *IEEE Access*, vol. 7, pp. 108 882-108 892, 2019.
- [7] A. Mellit and S. A. Kalogirou, "Artificial intelligence techniques for photovoltaic applications: A review," *Prog. Photovoltaics*, vol. 16, no. 8, pp. 651-668, Dec. 2008.
- [8] F. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, Oct. 2001. <https://doi.org/10.1023/A:1010933404324>
- [9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, 2016, pp. 785-794. <https://doi.org/10.1145/2939672.2939785>

- [10] Z. Chen, K. Gryllias, and W. Li, "Intelligent fault diagnosis for rotary machinery using transferable convolutional neural network," *IEEE Trans. Ind. Inform.*, vol. 16, no. 1, pp. 339-349, Jan. 2020. <https://doi.org/10.1109/TII.2019.2917233>
- [11] O. Abdeljaber, O. Avci, S. Kiranyaz, M. Gabbouj, and D. J. Inman, "Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks," *J. Sound Vib.*, vol. 388, pp. 154-170, Feb. 2017. <https://doi.org/10.1016/j.jsv.2016.10.043>
- [12] B. Li, Y. Zhang, L. He, and Y. Tian, "Transformer-based multi-variable time-series fault detection for photovoltaic inverters," *IEEE Trans. Ind. Electron.*, vol. 71, no. 3, pp. 2985-2994, Mar. 2024.
- [13] L. Breiman, "Stacked generalizations," *Neural Netw.*, vol. 5, no. 2, pp. 241-259, 1996. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)
- [14] K. Leahy, R. L. Hu, I. C. Konstantakopoulos, C. J. Spanos, and A. M. Agogino, "Diagnosing wind turbine faults using machine learning techniques applied to operational data," in *Proc. IEEE Int. Conf. Prognostics Health Manag.*, Ottawa, ON, 2016, pp. 1-8. <https://doi.org/10.1109/ICPHM.2016.7542860>
- [15] Rout, P., Jha, A. K., Gupta, P., Singh, B. and Choudhury, S., "Failure analysis of composite plate under ballistic impact," *Mater. Today Proc.*, vol. 74, pp. 1008-1011, 2023. <https://doi.org/10.1016/j.matpr.2022.11.357>
- [16] Ruhela, H. S., Bhardwaj, S., Agrawal, T. and Gupta, P., "Explicit dynamics analysis of shin pads using finite element analysis," in *Int. Conf. Industrial Problems on Machines and Mechanism*, Springer, Singapore, 2022, pp. 683-690. https://doi.org/10.1007/978-981-99-4270-1_67
- [17] Kumar, M., Rajiyan, J. and Gupta, P., "A computational approach for solving elasto-statics problems," *Mater. Today Proc.*, vol. 46, pp. 6876-6879, 2021. <https://doi.org/10.1016/j.matpr.2021.04.462>