Received: 19 Mar 2024 | Revised: 15 Sept 2024 | Accepted: 19 Sept 2024 | Published: 18 Oct 2024
Volume 2(2) Pages 42-54 DOI: https://doi.org/10.56279/jicts.v2i2.9028

JICTS
Journal of ICT Systems

Spark: A Statistical Comparison and Evaluation of Classification Algorithms for Fault Prediction in Electrical Secondary Distribution Networks

David T. Makota [a, 1], Naiman Shililiandumi [b], Hashim U. Iddi [b]

[a]*Department of Computer Science, Institute of Finance Management, Dar es Salaam, Tanzania*
[b]*Department of Electronics and Telecommunications Engineering, University of Dar es Salaam, Dar es Salaam, Tanzania*

[1]Corresponding author
Email: davetotty@gmail.com

Funding information

This work is the result of PhD research sponsored in full by the Institute of Finance Management.

Keywords

Apache Spark
Classification Algorithms
Electrical Secondary Distribution
Fault Prediction
Statistical Methods

Abstract

Managing faults in electrical secondary distribution networks is a challenging task given the nature, size, and complexity. Predicting faults before they occur helps in increasing the safety and reliability of the power distribution system. Various statistical and machine learning techniques are being used to predict different types of faults. This study applies classification algorithms available in the Apache Spark framework, through its python interface PySpark, to predict electrical secondary distribution network faults. The study evaluates and compares ten algorithms: Decision tree, Gradient-boosted tree, Binomial Logistic Regression, Multinomial Logistic Regression, Naïve Bayes, Multilayer perceptron, Random Forest, Linear Support Vector Machine, One-versus-rest, and Factorization machines. The research uses Friedman's test followed by Nemenyi post hoc test to find the significance of performance differences among the algorithms. The results show significant differences among the algorithms. Gradient-boosted tree and One-versus-rest with Gradient-boosted tree had the best performance for binary and multiclass classification, respectively, while Naïve Bayes had the worst performance. By identifying the most effective algorithms, this research provides a practical reference for selecting suitable models, aiding in fault prediction, reducing system downtime, and optimizing maintenance strategies. Additionally, the results can inform the selection of base models for ensemble methods, further improving prediction accuracy.

1. Introduction

Maintaining a safe and reliable Electrical Secondary Distribution Network (ESDN) with effective fault detection, identification and clearance have become difficult. This difficulty is attributed to the continuous increase in power demand coupled with ageing power distribution infrastructures that keep on increasing the size and complexity of the network [1-3]. Thus, effective fault prediction is of paramount importance to the management and control of ESDN for safety and reliability.

If failures are known in advance, appropriate measures for service restoration and optimization in the ESDN are taken. Zhang et al. [4] argue that taking proper steps earlier increases reliability and safety, reduces maintenance time and cost, and extends life of assets in the power system.

In recent years, machine learning techniques have demonstrated significant potential in predicting faults in electrical power distribution systems, offering a more proactive approach compared with traditional methods. The effectiveness of these models, however, varies based on the algorithms used and the specific characteristics of the datasets. A wide range of classification algorithms, from linear models and tree-based methods to ensemble models and artificial neural networks, has been applied for fault prediction in power distribution networks. Despite this, there is limited research specifically targeting ESDN with big data processing frameworks like Apache Spark. While a limited number of studies have explored algorithm comparisons in related tasks in the power distribution networks, such as load forecasting and assessing building energy efficiency, it remains unclear which algorithms are most effective for distributed and parallel processing of the growing ESDN datasets. This gap highlights the need for comprehensive evaluations of various algorithms within the context of big data frameworks to determine their performance in ESDN fault prediction.

Thus, this paper aims to address this gap by comparing and evaluating the classification algorithms supported by PySpark, the Python interface for the Apache Spark big data processing framework, to determine the most suitable algorithms for fault prediction in ESDN. This study builds on existing algorithm comparison frameworks and contributes to the field by evaluating the performance of multiple classification algorithms for fault prediction in ESDN. Statistical tests are used to determine significant performance differences, ensuring the robustness of the results. By filling the knowledge gap in algorithm performance for fault prediction in ESDN, this research provides actionable insights for improving system reliability and reducing operational risks.

Apache Spark [5] is among the prominent open-source big data distributed processing systems supporting machine learning, graph processing, real-time analytics, batch processing and interactive queries. It offers fast processing capability even on complex datasets due to its optimized query execution and in-memory caching mechanisms. Within the Apache Spark framework, PySpark offers a powerful combination of Spark's distributed computing capabilities and Python's vast machine learning ecosystem. While Scala and Java are also supported by Apache Spark, Python was preferred due to its widespread use in machine learning and its extensive library ecosystem such as Scikit-learn, TensorFlow, and Pandas, making it the most suitable choice for this study.

The classification algorithms that were applied in this study are Binomial Logistic Regression (BLR), Multinomial Logistic Regression (MLR), Decision Trees (DT), Random Forest (RF), Gradient-Boosted Tree (GBT), Multilayer

Perceptron (MP), Linear Support Vector Machine (LSVM), One-vs-Rest (OVR), Naïve Bayes (NB) and Factorization Machines (FM). These are all classification algorithms available in PySpark 3.0.3. We were particularly interested in finding out the usefulness of OVR, BLR, MLR, MP, NB, and FM since a limited number of studies previously applied them to predict faults in the electrical distribution network [6-12].

The performance metrics that were used to evaluate the classification algorithms are accuracy, recall, precision. and F1-Score. Additionally, binary classification algorithms were also evaluated using Area under Precision-Recall Curve (AUPR) and Area under Receiver Operator Characteristic Curve (AUROC) which are visually presented using Receiver Operator Characteristic (ROC) curve and Precision-Recall (PR) curve.

The statistical inferences from the observed differences in accuracy and AUPR for multiclass classification and binary classification were drawn based on the approach proposed by Vázquez et al. [13] and Pizarro et al. [14] for comparing multiple algorithms on a single dataset. Since parametric conditions were not met, the accuracy and AUPR measures were ranked using the non-parametric Friedman's test. After that, Nemenyi post hoc test was used to determine whether there is a statistical significance in rank differences. Finally, the results were visually presented using Demšar significance diagrams.

The results of this study have practical implications for improving fault prediction in electrical distribution networks. By identifying the best-performing algorithms (Gradient-boosted tree and One-versus-rest with Gradient-boosted tree), these algorithms offer enhanced fault prediction accuracy and can serve as strong base models for ensemble learning, potentially improving predictive performance in complex scenarios. Integrating these techniques can reduce downtime, optimize maintenance, and enhance safety. Additionally, leveraging Apache Spark's scalability advances big data applications in power system management for more efficient, real-time fault prediction.

2. Related Work

Roland and Eseosa [15] established a classification model by applying Artificial Neural Network (ANN) to predict incipient transformer faults in the distribution network. The established model was trained and validated using Dissolved Gas Analysis (DGA) data. Similarly, Sayar and Yüksel [16] used the same approach to predict power outages in the electrical distribution network. The established model uses hourly outages and meteorological data that influence power system failures and network health conditions. The model yielded performance results of 70.59% accuracy.

Huang et al. [17] proposed a model for fault prediction in the distribution network using support vector machines (SVMs). The model was constructed and validated using historical datasets from distribution network management and monitoring systems. It demonstrated superior performance compared with both ANN and C4.5 decision tree models. Similar results were obtained by Wang et al. [18] when predicting cascade failures in the smart grid. It was reported that the model produced prediction accuracy close to 100% during validation with real-time data.

Lin et al. [19] presented a fault prediction model for a smart grid distribution system based on Voted Random Forest Algorithm (VRF). The model was trained and validated using distribution network faults logs and meteorological data. The authors then compared the performance of the model with ANN, Random Forest (RF), and SVM algorithms. The results indicate that the VRF model surpassed the performance of the other models.

Cai et al. [20] built a model for feeder fault prediction in a power distribution network. The proposed model was based on XGBoost, utilizing datasets from Fujian Electric Power in China collected over 17 years. The results showed that the proposed model is valid and efficient with an AUC of 0.8899.

Stefenon et al. [21] implemented a hybrid time series model for fault prediction in distribution insulators. The model was implemented using the wavelet technique and long short-term memory. The authors then compared their model with non-linear Auto-Regressive (NAR) and eXogenous input (NARX) models. The results showed that wavelet LSTM outperformed both NAR and NARX.

Hou et al. [22] used random forest (RF) to predict power outage faults caused by typhoon disasters. They built their model using a dataset with 14 features from multiple sources, including geographical data, power grid data and meteorological data. The model was validated through a case study from typhoon "Mijiage" of 2015 and obtained an accuracy of up to 92.44%.

Yang [23] developed a fault prediction model for the distribution system based on RF and Multi-classification Support Vector Machines (MSVM) using original data from the distribution network coupled with meteorological data. The results showed that the proposed method has a practical application value with accuracy reaching 95%.

Relatively few studies have compared machine learning algorithms for various tasks in power distribution networks. Moradzadeh et al. [24] assessed machine learning methods for predicting heating and cooling loads in residential buildings, but their comparison involved only two algorithms: Multilayer Perceptron (MLP) and Support Vector Regression (SVR). Markovics and Mayer [25] performed a comparative analysis of machine learning methods for forecasting photovoltaic power using numerical weather predictions, evaluating and comparing a total of 24 algorithms.

Ullah et al. [26] conducted a comparative analysis of machine learning algorithms for predicting electric vehicle energy consumption. They employed advanced algorithms such as Light Gradient Boosting Machine and Extreme Gradient Boosting, and compared their performance to conventional algorithms such as Multiple Linear Regression and Artificial Neural Networks. Egwim et al. [27] carried out a comparative analysis of machine learning algorithms for assessing building energy efficiency through big data analytics, exploring commonly used algorithms for developing models to assess energy efficiency in buildings. Additionally, Luo et al. [28] conducted a performance comparison of three machine learning algorithms (Support Vector Regression, Long-Short-Term Memory Neural Network, and Artificial Neural Network) for predicting Building Integrated Photovoltaic (BIPV) power production and multiple building energy loads simultaneously.

In previous studies, authors primarily used conventional machine learning methods to predict faults in electrical distribution networks with small datasets. While some research has compared machine learning algorithms for various tasks in power distribution networks, there is a lack of comprehensive evaluations focused on ESDN, especially using big data frameworks like Apache Spark. Many studies overlook the advantages of these frameworks, which are essential for processing the large datasets generated in ESDN. In this paper, we evaluated and compared the performance of all the classification algorithms available in PySpark, the Python interface for the well-known big data framework Apache Spark. The evaluation focused on fault prediction in ESDN datasets, ranking the algorithms based on their performance.

3 Methods and Materials

3.1 Data Collection

The dataset used to assess the performance of the classification algorithms in this study was collected from the electrical secondary distribution network using Automatic Meter Reading (AMR). Additionally, the corresponding weather data of temperature and rainfall were obtained and aggregated. The dataset comprises of 105,118 instances with a resolution of 20 minutes collected from January 2015 to December 2018. Table 1 displays the description of ESDN dataset.

3.2 Evaluation of Classification Algorithms

The metrics used in the evaluation of performance for each of the classification algorithms are accuracy, recall, precision and F1-Score as presented in equations (1) through (4). These metrics come from the four fundamental parameters of all the classification results, namely True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). TP indicates that the positive label is predicted positive, while TN means that the negative label is predicted negative. On the other hand, FP shows that the negative label is predicted positive, and FN shows that the positive label is predicted negative.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1-Score} = \frac{2 \times TP}{(2 \times TP) + FP + FN} \quad (4)$$

Other metrics are area under the precision-recall curve (AUPR) and area under the receiver operator characteristic curve (AUROC). AUROC provides a performance measure that shows the true positives against the false positives. Unlike accuracy, AUROC is threshold independent because it implicitly compares the base error rates between classifiers. Thus, AUROC is considered more reliable in the presence of high false-positive rates [29, 30]. AUPR is another threshold independent metric obtained from the plot of precision against recall. In the case of the highly imbalanced ESDN dataset, AUPR is said to be more effective and informative than AUROC [31].

Table 1. Description of ESDN Dataset.

S/N	Attribute	Description
1	Timestamp	Timestamp measurements were recorded
2	VA	Voltage measurement of Phase A
3	VB	Voltage measurement of Phase B
4	VC	Voltage measurement of Phase C
5	CA	Current measurement of Phase A
6	CB	Current measurement of Phase B
7	CC	Current measurement of Phase C
8	kVAh	Apparent power measurement
9	kVArh	Reactive power measurement
10	kWh	Active power measurement
11	Temperature	Temperature measurement
12	Precipitation	Rainfall measurement
13	Fault	Target variable

3.3 Statistical Comparison of Classification Algorithms

In this research, the hypothesis was developed to evaluate the performance of different classification algorithms available in PySpark for fault prediction in ESDN. The specific hypotheses tested are as follows:

H_0: There is no significant difference in the performance of the classification algorithms provided by PySpark for fault prediction in ESDN

H_1: At least one of the classification algorithms performs significantly better than the others for fault prediction in ESDN

The rationale for this hypothesis is based on several factors. Firstly, the existing literature shows that machine learning algorithms often perform differently depending on the dataset and computational framework used. ESDN presents complex data challenges, which can influence algorithm performance, making it essential to test whether some algorithms outperform others. Additionally, not all machine learning algorithms are equally suited for distributed processing in big data environments like PySpark, and the hypothesis aims to identify the most effective algorithms for fault prediction in this context. By incorporating statistical tests, the study seeks to determine whether these differences in performance are statistically significant, providing practical insights for real-world deployment in ESDN to enhance reliability and safety.

Our motivation is based on the interest in comparing more than two classification algorithms over a single data set. Studies conducted by Vázquez et al. [13] and Pizarro et al. [14] recommend ANOVA if the parametric conditions are met and Friedman's test when the parametric conditions are not met.

In this study, we follow the procedure provided by Pizarro et al. [14]. Mean AUPR and mean accuracy are used to compare the performance of the algorithms for binary-class classification and multiclass classification, respectively. 10-fold walk-forward cross-validation is performed with the threshold p=0.95 (95% confidence level) to check significance ($p<0.95$). We selected the non-parametric Friedman's test to test the null hypothesis because the Levene test on the algorithms performance results scored p-values of 0.0028 for accuracy and less than 0.0001 for AUPR ($\alpha=0.05$), rejecting the homoscedasticity assumption for ANOVA.

Friedman's test compares the classification algorithms based on the average ranks of their performances to show whether a statistical difference exists among the classification algorithms. The observed value of Friedman's test statistic is given by

$$X_F^2 = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1), \qquad (5)$$

where n is the number of blocks, k is the number of treatment levels, and R_j is the sum of the ranks for sample j. X_F^2 is compared with a chi-square distribution using $k-1$ degrees of freedom, and when it is large enough, the null hypothesis is rejected [32, 33].

When the null hypothesis is rejected, the post hoc Nemenyi test is used to compare the classification algorithms to identify the significant differences in their performances. The Nemenyi test uses a critical difference (CD) calculated using

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \qquad (6)$$

where q_α is the critical value and k is the number of treatment levels. If the average ranks of two algorithms differ by a value larger than CD, their performance is significantly different [34].

Finally, the statistical comparison results obtained from Friedman's and Nemenyi post hoc test are visually presented using Demšar significance diagrams [35]. In these diagrams, algorithms with no significant difference in performance are connected by horizontal bold

lines. In the significance diagram, the best-performing algorithms are positioned on the right side, and the critical difference is shown above the same significance diagram.

4 Results and Discussion

This section shows the results produced by applying all the evaluated algorithms. Firstly, we present the performance evaluation results of the algorithms in terms of accuracy, recall, precision F1-Score, AUROC and AUPR for each classifier. Then, statistical tests to determine whether the results are significantly different or not are presented. Finally, the results are discussed in detail, highlighting key insights and the broader implications of the study.

4.1 Performance of the Algorithms

The performance of all evaluated algorithms for binary classification is presented in Table 2. The Gradient Boosted Tree algorithm exhibited the best performance among all the classifiers in all the evaluated metrics, followed by Decision Tree, Random Forest, Binomial Logistic Regression, Linear Support Vector Machine and Multilayer Perceptron. Conversely, Naïve Bayes and Factorization Machine demonstrated the worst performance of all the classifiers. Figure 1 and Figure 2 show the graphical representation of the binary classification evaluation using the ROC curve and PR curves, respectively. With the ROC curve, best performing algorithms tend to produce curves nearer to the top-left corner and those which come closer to the dashed diagonal of the ROC space indicate poor performance. The algorithms producing curves nearer to the top-right corner in the PR curve diagram indicate a better performance.

Table 3 shows the evaluation results of the algorithms for multiclass classification. The OVR with GBT (OVR-GBT) model exhibited the best performance among all the classifiers in all the evaluated metrics. In contrast, Naïve Bayes, OVR with FM (OVR-FM) and Multilayer Perceptron demonstrated the worst performance of all the classifiers.

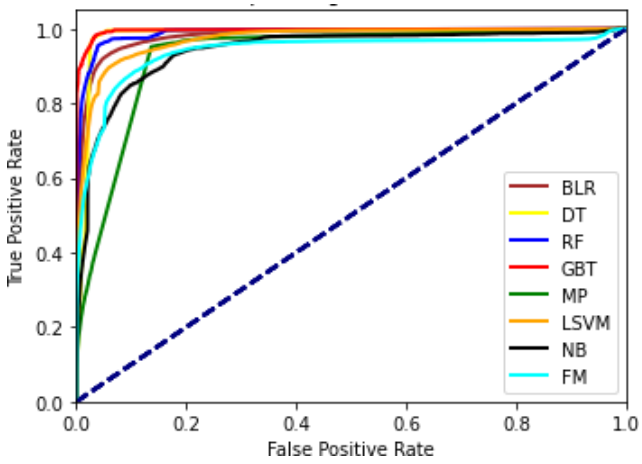

Figure 1. ROC Curves for Binary Classification Performance Results.

Table 2. Binary Classification Performance Results.

Algorithm	Evaluation Criteria					
	Accuracy	AUROC	AUPR	Precision	Recall	F1-Score
Binomial Logistic Regression	94.67	98.11	94.19	94.70	94.67	0.9464
Decision Tree	97.40	98.60	96.68	97.45	97.40	0.9741
Random Forest	95.99	99.09	96.44	96.15	95.99	0.9603
Gradient Boosted Tree	97.96	99.61	98.99	97.96	97.96	0.9796
Multilayer Perceptron	88.60	90.17	81.77	90.70	88.60	0.8894
Linear Support Vector Machines	94.42	97.16	90.65	94.43	94.42	0.9442
Naïve Bayes	63.74	74.75	35.30	83.09	63.74	0.6694
Factorization Machine	88.34	79.84	70.05	89.74	88.34	0.8732

Table 3. Multiclass Classification Performance Results.

Algorithm	Evaluation Criteria			
	Accuracy	Precision	Recall	F1-Score
Multinomial Logistic Regression	94.16	94.46	94.16	94.02
Decision Tree	96.78	96.58	96.78	96.61
Random Forest	95.33	95.07	95.33	95.12
Multilayer Perceptron	86.60	86.85	86.60	85.07
Naïve Bayes	64.37	87.13	64.37	68.52
OVR-BLR	94.40	94.55	94.40	94.23
OVR-FM	85.38	88.21	85.38	84.30
OVR-GBT	97.78	97.69	97.78	97.70
OVR-LSVM	94.17	93.84	94.17	93.92

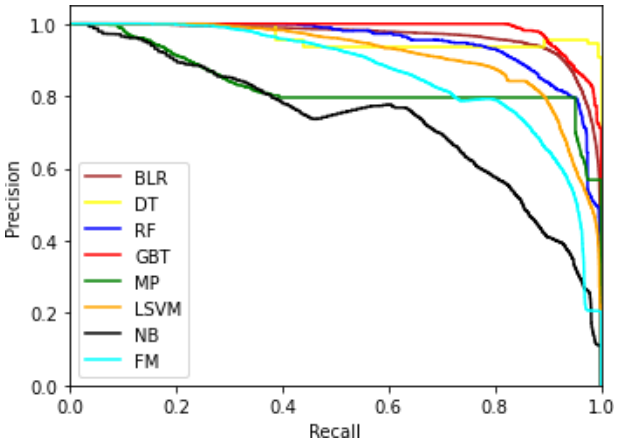

Figure 2. PR Curves for binary classification performance results.

4.2 Statistical Analysis

After evaluating the performance of the classification algorithms as illustrated in the previous section, a statistical analysis was conducted to test whether the differences in performances of the algorithms are significant. This research has compared the algorithms only for the classification evaluation measures of accuracy and AUPR for multiclass classification and binary classification, respectively. Results show that parametric conditions for ANOVA were not met. The Levene test rejected the homogeneity hypothesis of all the evaluation results with p-values of 0.0028 for accuracy and less than 0.0001 for AUPR. Thus, a non-parametric method, the Friedman test, was used to compare the evaluation results.

Table 4 and Table 5 report Friedman's test rankings of all the evaluated algorithms on the ESDN dataset for AUPR and accuracy. The results indicate that Gradient Boosted Tree has the highest Friedman score for binary classification, and OVR-GBT scored high for multiclass classification. Oppositely, Naïve Bayes demonstrated the worst average ranking than all classifiers in both cases.

The Friedman test statistics for accuracy and AUPR observed using equation (5) were 75.73 (greater than $x^2_{0.05,7} = 15.507$) and 66.13 (greater than $x^2_{0.05,7} = 14.067$), respectively. In all cases, the corresponding p-value for the Friedman test was less than 0.001 ($\alpha = 0.05$). These results reject the null hypothesis that all the evaluated algorithms perform similarly. Then, we applied the post hoc Nemenyi test to each evaluation criteria.

The significance diagrams, Figure 3 and Figure 4, illustrate the Nemenyi post hoc test results for AUPR and accuracy, respectively. Horizontal bold lines connect algorithms whose performances are not significantly different, and the best performing algorithms are presented to the right of the diagram. The observed critical difference values using (6) for accuracy and AUPR are 3.7988 and 3.3202, respectively.

Table 4. Rankings from AUPR for binary classification.

Algorithm	AUPR
Binomial Logistic Regression	3.9
Decision Tree	2.7
Random Forest	2.6
Gradient Boosted Tree	1.0
Multilayer Perceptron	5.9
Linear Support Vector Machines	5.0
Naïve Bayes	8.0
Factorization Machine	6.9

4.3 Discussions

The results of the statistical analysis affirm the observed performance trends and provide deeper insights into the strengths and weaknesses of the evaluated algorithms. GBT emerged as the best-performing algorithms for binary classification, with OVR-GBT showing similar dominance in multiclass classification. The high rankings of these algorithms, as confirmed by the Friedman test, demonstrate their superior fault detection capabilities in ESDN. GBT's success can be attributed to its ability to enhance model accuracy by sequentially correcting errors, which proves highly effective for complex prediction tasks with imbalanced data.

Table 5. Rankings from accuracy for multiclass classification.

Algorithm	Accuracy
Multinomial Logistic Regression	4.3
Decision Tree	1.9
Random Forest	3.5
Multilayer Perceptron	7.5
Naïve Bayes	9.0
OVR-BLR	4.5
OVR-FM	7.5
OVR-GBT	1.1
OVR-LSVM	5.7

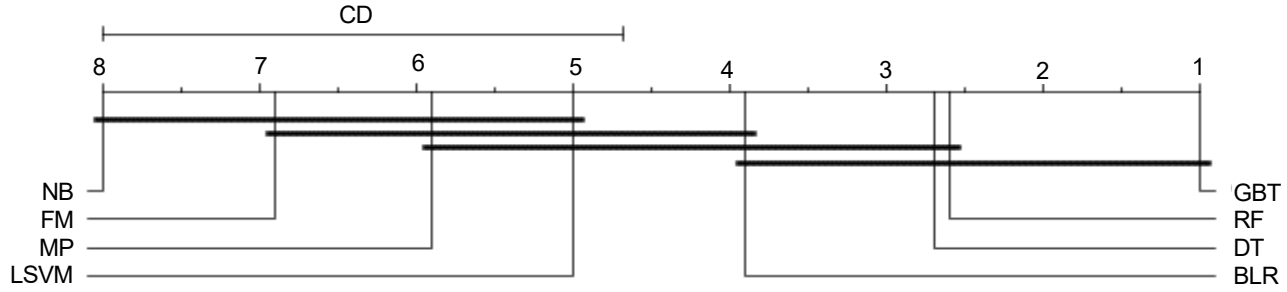

Figure 3. Significance diagram for AUPR.

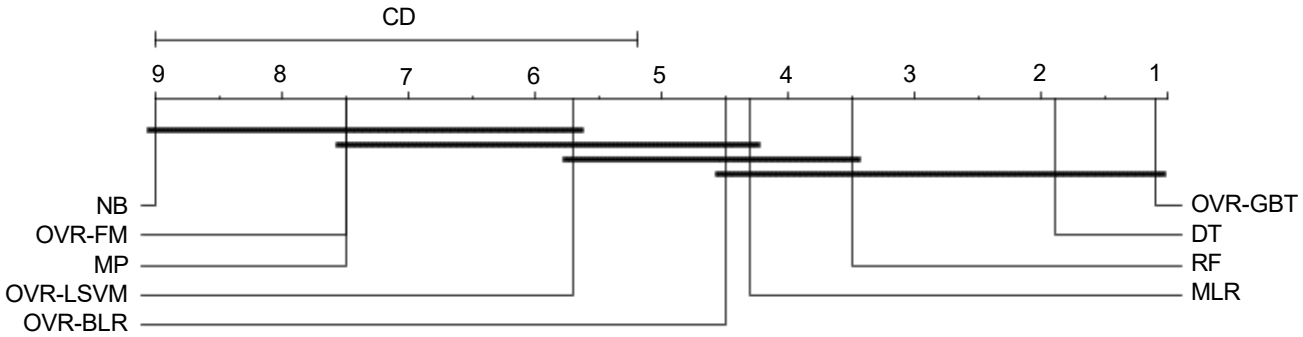

Figure 4. Significance diagram for accuracy.

Similarly, the strong performance of RF and DT algorithms reflects the advantages of ensemble-based models, which aggregate multiple decision paths to increase predictive accuracy and robustness. These algorithms' consistent performance across different metrics and their statistically significant ranking makes them viable alternatives where model simplicity or interpretability is preferred.

The poor performance of NB and FM can be explained by their assumptions and limitations. NB relies on feature independence, which is often violated in complex datasets, leading to suboptimal results. Similarly, FM, which excels in recommendation systems, failed to capture the complexities of the fault prediction task in this study.

The post hoc Nemenyi test further clarifies that, while GBT, DT, RF, and OVR-GBT are statistically similar in performance, algorithms like NB, FM, and MP consistently perform worse. These insights are critical for selecting appropriate algorithms for real-time fault prediction systems, where accuracy and computational efficiency are paramount.

The findings of this study have significant practical implications for real-time fault prediction in ESDN. By identifying GBT and OVR-GBT as the top-performing algorithms, operators can confidently deploy these algorithms for reliable and scalable fault prediction, enhancing system reliability and safety. The findings can also guide the selection of strong candidates for base models in ensemble learning, which could further improve predictive performance in more complex scenarios. Conversely, the poor performance of NB indicates its unsuitability for this application, and it should be avoided. Additionally, leveraging Apache Spark's scalability ensures that these algorithms can handle real-time processing in large networks, optimizing maintenance strategies and reducing downtime, thus advancing big data applications in power system management.

Despite the promising results, this study has several limitations. First, the algorithms were evaluated on a specific dataset from ESDN measurements obtained from the AMR system, which may limit the generalizability of the findings to other types of networks or domains. Additionally, while the study focused on the accuracy and performance of various algorithms, other important factors such as computational complexity, and real-time scalability under different conditions were not fully explored. Furthermore, the results rely on the performance of Apache Spark's implementation, which may vary across different big data frameworks. Future research could investigate the adaptability of these algorithms to other frameworks and scenarios, as well as explore hybrid approaches that combine multiple algorithms for enhanced performance.

5 Conclusion

This study evaluated the performance of ten classification algorithms in predicting faults within the ESDNs using data from AMR system. Among the evaluated algorithms, the GBT and OVR-GBT consistently outperformed other algorithms for binary and multiclass classification tasks, respectively. These algorithms demonstrated superior accuracy, recall, precision, and overall robustness, making them strong candidates for deployment in real-time fault prediction systems. Conversely, algorithms such as NB and FM exhibited the poorest performance and are less suited for this application.

The use of Apache Spark framework enabled scalable and efficient processing of large datasets, showcasing the potential of big data frameworks in managing complex power system tasks. Furthermore, the statistical analyses confirmed the significance of performance differences among the

algorithms, validating the robustness of the top-performing algorithms.

The findings from this study provide actionable insights for improving the reliability and safety of ESDN systems by selecting high-performing algorithms for fault prediction. These insights can also guide the selection of base models for ensemble learning, offering the potential to further improve predictive performance in more complex scenarios. However, the study's limitations suggest that future research should focus on testing these algorithms in different domains, exploring hybrid approaches, and assessing computational efficiency under real-time operational conditions.

ACKNOWLEDGEMENT

This research was carried out as part of the iGrid-Project at the University of Dar es Salaam (UDSM) under the sponsorship of Swedish International Development Agency (SIDA). The authors appreciate TANESCO for the collaboration provided during the research.

CONTRIBUTIONS OF CO-AUTHORS

David T. Makota	[ORCID: 0000-0003-2151-3190]	Conceived the idea, conducted experiments, analyzed results and wrote the paper.
Naiman Shililiandumi	[ORCID: 0000-0002-8499-7543]	Reviewed and improved the paper.
Hashim U. Iddi	[ORCID: 0000-0002-4025-9653]	Reviewed and improved the paper.

REFERENCES

[1] V. Zapata Castillo, H. De Boer, R. Maícas Muñoz, D. E. Gernaat, R. Benders, and D. van Vuuren, *Future Global Electricity Demand Load Curves*, Harmen Maícas Muñoz Raúl Gernaat David EHJ Benders René Van Vuuren Detlef Future Glob. Electr. Demand Load Curves, Accessed: Sep. 14, 2024. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3935492

[2] S. Bandara, P. Rajeev, and E. Gad, *A review on condition assessment technologies for power distribution network infrastructure*, Struct. Infrastruct. Eng., vol. 20, no. 11, pp. 1834–1851, Nov. 2024, doi: 10.1080/15732479.2023.2177680.

[3] M. Meliani, A. E. Barkany, I. E. Abbassi, A. M. Darcherif, and M. Mahmoudi, *Energy management in the smart grid: State-of-the-art and future trends*, Int. J. Eng. Bus. Manag., vol. 13, p. 184797902110329, Jan. 2021, doi: 10.1177/18479790211032920.

[4] S. Zhang, Y. Wang, M. Liu, and Z. Bao, *Data-based line trip fault prediction in power systems using LSTM networks and SVM*, IEEE Access, vol. 6, pp. 7675–7686, 2017.

[5] M. Zaharia *et al.*, *Apache spark: a unified engine for big data processing*, Commun. ACM, vol. 59, no. 11, pp. 56–65, 2016.

[6] M. Tang, Z. Kuang, Q. Zhao, H. Wu, and X. Yang, *Fault Detection of Wind Turbine Pitch System Based on Multiclass Optimal Margin Distribution Machine*, Math. Probl. Eng., vol. 2020, pp. 1–10, Aug. 2020, doi: 10.1155/2020/2091382.

[7] C. Wei, H. Jingshan, L. Zheng, W. Cong, and W. Jiahao, *Research on transmission line trip prediction based on logistic regression algorithm under icing condition*, in 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), IEEE, 2021, pp. 565–569. Accessed: Sep. 14, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9482230/

[8] L. Wang, T. Littler, and X. Liu, *Gaussian process multi-class classification for transformer fault diagnosis using dissolved gas analysis*, IEEE Trans. Dielectr. Electr. Insul., vol. 28, no. 5, pp. 1703–1712, 2021.

[9] A. Moradzadeh, B. Mohammadi-Ivatloo, K. Pourhossein, and A. Anvari-Moghaddam, *Data mining applications to fault diagnosis in power electronic systems: A systematic review*, IEEE Trans. Power Electron., vol. 37, no. 5, pp. 6026–6050, 2021.

[10] I. B. Taha and D. A. Mansour, *Novel power transformer fault diagnosis using optimized machine learning methods*, Intell. Autom. Soft Comput., vol. 28, no. 3, pp. 739–752, 2021.

[11] Z. Zhao, D. Moscovitz, L. Du, and X. Fan, *Factorization Machine Learning for Disaggregation of Transmission Load Profiles with High Penetration of Behind-the-Meter Solar*, in 2023 IEEE Energy Conversion Congress and Exposition (ECCE), IEEE, 2023, pp. 1278–1282. Accessed: Sep. 14, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10362108/

[12] Y. Zhang, *Electrical line fault prediction using a novel grey wolf optimization algorithm based on multilayer perceptron*, Adv. Control Appl., vol. 6, no. 3, p. e213, Sep. 2024, doi: 10.1002/adc2.213.

[13] E. G. Vázquez, A. Yañez Escolano, P. Galindo Riaño, and J. Pizarro Junquera, *Repeated Measures Multiple Comparison Procedures Applied to Model Selection in Neural Networks*, in Bio-Inspired Applications of Connectionism, J. Mira and A. Prieto, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2001, pp. 88–95. doi: 10.1007/3-540-45723-2_10.

[14] J. Pizarro, E. Guerrero, and P. L. Galindo, *Multiple comparison procedures applied to model selection*, Neurocomputing, vol. 48, no. 1, pp. 155–173, Oct. 2002, doi: 10.1016/S0925-2312(01)00653-1.

[15] U. Roland and O. Eseosa, *Artificial neural network approach to distribution transformers maintenance*, Int. J. Sci. Res. Eng. Technol., vol. 1, no. 4, pp. 62–70, 2015.

[16] M. Sayar and H. Yüksel, *Real-Time Prediction of Electricity Distribution Network Status Using Artificial Neural Network Model: A Case Study in Salihli (Manisa, Turkey)*, Celal Bayar Univ. J. Sci., vol. 16, no. 3, pp. 307–321, 2020.

[17]W.-S. Huang, X. LU, Y. LIU, Q. CHEN, M.-H. QI, and H.-J. GAO, *Fault Prediction of Distribution Network Based on Support Vector Machine*, DEStech Trans. Eng. Technol. Res., no. amee, 2019.
[18]Y. Wang, Y. Li, H. Liang, X. Weng, and M. Huang, *An Active Power Failure Early Warning Probability Model Based on Support Vector Machine Algorithm*, in IOP Conference Series: Earth and Environmental Science, IOP Publishing, 2021, p. 042042.
[19]R. Lin, Z. Pei, Z. Ye, B. Wu, and G. Yang, *A voted based random forests algorithm for smart grid distribution network faults prediction*, Enterp. Inf. Syst., pp. 1–19, 2019.
[20]J. Cai, Y. Cai, H. Cai, S. Shi, Y. Lin, and M. Xie, *Feeder Fault Warning of Distribution Network Based on XGBoost*, in Journal of Physics: Conference Series, IOP Publishing, 2020, p. 012037.
[21]S. F. Stefenon *et al.*, *Fault detection in insulators based on ultrasonic signal processing using a hybrid deep learning technique*, IET Sci. Meas. Technol., vol. 14, no. 10, pp. 953–961, 2021.
[22]H. Hou *et al.*, *Spatial distribution assessment of power outage under typhoon disasters*, Int. J. Electr. Power Energy Syst., vol. 132, p. 107169, Nov. 2021, doi: 10.1016/j.ijepes.2021.107169.
[23]X. Yang, *Power Grid Fault Prediction Method Based On Feature Selection And Classification Algorithm*, Int J Electron. Eng. Appl., vol. 9, no. 2, 2021.
[24]A. Moradzadeh, A. Mansour-Saatloo, B. Mohammadi-Ivatloo, and A. Anvari-Moghaddam, *Performance evaluation of two machine learning techniques in heating and cooling loads forecasting of residential buildings*, Appl. Sci., vol. 10, no. 11, p. 3829, 2020.
[25]D. Markovics and M. J. Mayer, *Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction*, Renew. Sustain. Energy Rev., vol. 161, p. 112364, 2022.
[26]I. Ullah, K. Liu, T. Yamamoto, R. E. Al Mamlook, and A. Jamal, *A comparative performance of machine learning algorithm to predict electric vehicles energy consumption: A path towards sustainability*, Energy Environ., vol. 33, no. 8, pp. 1583–1612, Dec. 2022, doi: 10.1177/0958305X211044998.
[27]C. N. Egwim, H. Alaka, O. O. Egunjobi, A. Gomes, and I. Mporas, *Comparison of machine learning algorithms for evaluating building energy efficiency using big data analytics*, J. Eng. Des. Technol., vol. 22, no. 4, pp. 1325–1350, 2024.
[28]X. J. Luo, L. O. Oyedele, A. O. Ajayi, and O. O. Akinade, *Comparative study of machine learning-based multi-objective prediction framework for multiple building energy loads*, Sustain. Cities Soc., vol. 61, p. 102283, 2020.
[29]C. X. Ling, J. Huang, and H. Zhang, *AUC: a statistically consistent and more discriminating measure than accuracy*, in Ijcai, 2003, pp. 519–524.
[30]C. Sibona and J. Brickey, *A Statistical Comparison of Classification Algorithms on a Single Data Set*, AMCIS 2012 Proc., Jul. 2012, [Online]. Available: https://aisel.aisnet.org/amcis2012/proceedings/ResearchMethods/2
[31]T. Saito and M. Rehmsmeier, *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*, PLoS ONE, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.
[32]M. Friedman, *The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance*, J. Am. Stat. Assoc., vol. 32, no. 200, pp. 675–701, Dec. 1937, doi: 10.1080/01621459.1937.10503522.
[33]M. Friedman, *A Comparison of Alternative Tests of Significance for the Problem of m Rankings*, Ann. Math. Stat., vol. 11, no. 1, pp. 86–92, 1940.
[34]P. B. Nemenyi, *Distribution-free multiple comparisons*, Princeton University, 1963.
[35]J. Demšar, *Statistical comparisons of classifiers over multiple data sets*, J. Mach. Learn. Res., vol. 7, pp. 1–30, 2006.