Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'24)
Barcelona, Spain - August 19 - 21, 2024
Paper No. CIST 174
DOI: 10.11159/cist24.174

Feature Selection and Classification Performance: A Multi-Dataset Comparative Analysis Using Boruta Algorithm and Random Forest

Ikhlass Boukrouh[1], Faouzi Tayalati[1], Abdellah Azmani[1]
[1]Intelligent Automation and BioMedGenomics Laboratory
FST of Tangier, Abdelmalek Essaâdi University, Tetouan, Morocco
ikhlass.boukrouh@etu.uae.ac.ma; faouzi.tayalati@etu.uae.ac.ma; a.azmani@uae.ac.ma

Abstract **-** Dimensionality reduction is crucial for managing high-dimensional datasets in machine learning, reducing complexity and overfitting. This study evaluates the efficiency of classification models without and with feature selection using the Boruta algorithm with Random Forest classifiers across three distinct datasets. Feature selection aims to improve model accuracy and interpretability by retaining only the most significant features. The three datasets were evaluated using full and reduced feature sets by comparing accuracy, precision, recall, and F1-score. Results show that feature selection significantly enhances model performance. For Dataset 1, accuracy improved by 1.06%, precision by 3.23%, recall by 3.46%, and F1-score by 3.36%. Dataset 2 saw increases in accuracy by 0.46%, precision by 2.36%, recall by 4.82%, and F1-score by 5.42%. Dataset 3 showed no significant changes, with both configurations yielding similar performance metrics. These findings confirm that the Boruta algorithm effectively enhances classification performance by reducing dataset dimensionality and retaining key features, especially in datasets with irrelevant features. However, when all features are relevant, the benefits of feature selection may be minimal.

Keywords**:** Dimensionality Reduction - Feature Selection - Boruta Algorithm - Random Forest - Classification Performance

1. Introduction

Artificial intelligence (AI) has revolutionized numerous industries, bringing transformative changes to sectors such as marketing [1], manufacturing [2], logistic [3], healthcare [4], and more. Within the realm of AI, machine learning (ML) stands out as a pivotal component, enabling the development of models capable of learning from data to generate predictions. A persistent challenge in ML is managing high-dimensional data, which can lead to increased computational demands, overfitting, and reduced interpretability of models.

To address these challenges, dimensionality reduction techniques are applied, with feature selection and feature extraction being the most prominent methods [5]. Feature selection aims to identify and retain only the most relevant features from the dataset, thereby enhancing model performance and interpretability. On the other hand, feature extraction involves transforming features into a lower-dimensional space while preserving essential information.

This study focuses on the application of the Boruta algorithm for feature selection in conjunction with Random Forest classifiers. The primary contributions of this study are as follows: (1) Assessing the impact of feature selection on the performance of Random Forest classifiers across three diverse datasets. (2) Comparing performance metrics with and without feature selection. (3) Demonstrating the advantages of using the Boruta algorithm to improve model performance by reducing dimensionality and eliminating irrelevant features.

The structure of this research is as follows: Section 2 provides an overview of feature selection techniques and the Boruta algorithm. Section 3 details the experimental setup, including data collection, preprocessing, model construction, and evaluation, and presents the results and discussion, underscoring key findings and their implications. Finally, Section 4 concludes the study and proposes directions for future research.

2. Feature Selection Overview

Handling high-dimensional data often necessitates dimensionality reduction techniques. Feature selection and feature extraction represent the two key approaches used for this purpose [5]. Feature selection aims to collect a subset of original

features that retain significant information, whereas feature extraction converts features into a lower-dimensional space [5]. Importantly, feature selection maintains the original features' physical meaning.

This research prioritizes feature selection to streamline high-dimensional data by removing irrelevant and redundant features, thereby preventing overfitting and boosting classifier accuracy [6], [7]. The Boruta algorithm is particularly effective for feature selection [8], [9], [10], [11], [12], [13]. Developed by Miron B. Kursa et al in [14], Boruta uses a wrapper method to assess feature importance by comparing them to random noise features [22]. It categorizes features as 'selected', 'tentative', or 'rejected' based on their importance [41]. To further enhance its efficacy, Boruta can be integrated with ensemble learning algorithms like Random Forest [10], Boruta-ERT [11], and XGBoost [12]. These combinations leverage Boruta's feature selection capabilities alongside the predictive power of ensemble methods, yielding superior results on complex datasets. The process involves extending the dataset by duplicating all independent variables to create hybrid features, generating shadow features by randomly shuffling the original features, and combining these with the original dataset. An ensemble learning algorithm is then initialized. The highest Z value among the shadow features, Z max, is identified to assess the importance of the original features. Features with a Z value greater than Z max are considered 'Important' and 'Selected', while those with a Z value less than Z max are deemed 'Unimportant' and 'Not selected'. This process is repeated until all features are either confirmed or rejected, or until the predefined iterations limits is achieved. By following this methodology, only the most relevant features are retained, enhancing both the performance and interpretability of predictive models. This approach is especially beneficial for high-dimensional datasets, enabling more accurate and reliable predictions.

3. Experiments and Results

To provide a clear overview of our approach, the methodology schema depicted in Fig. 1 outlines the key steps of our study. This includes data preparation, feature selection, construction and performance evaluation.

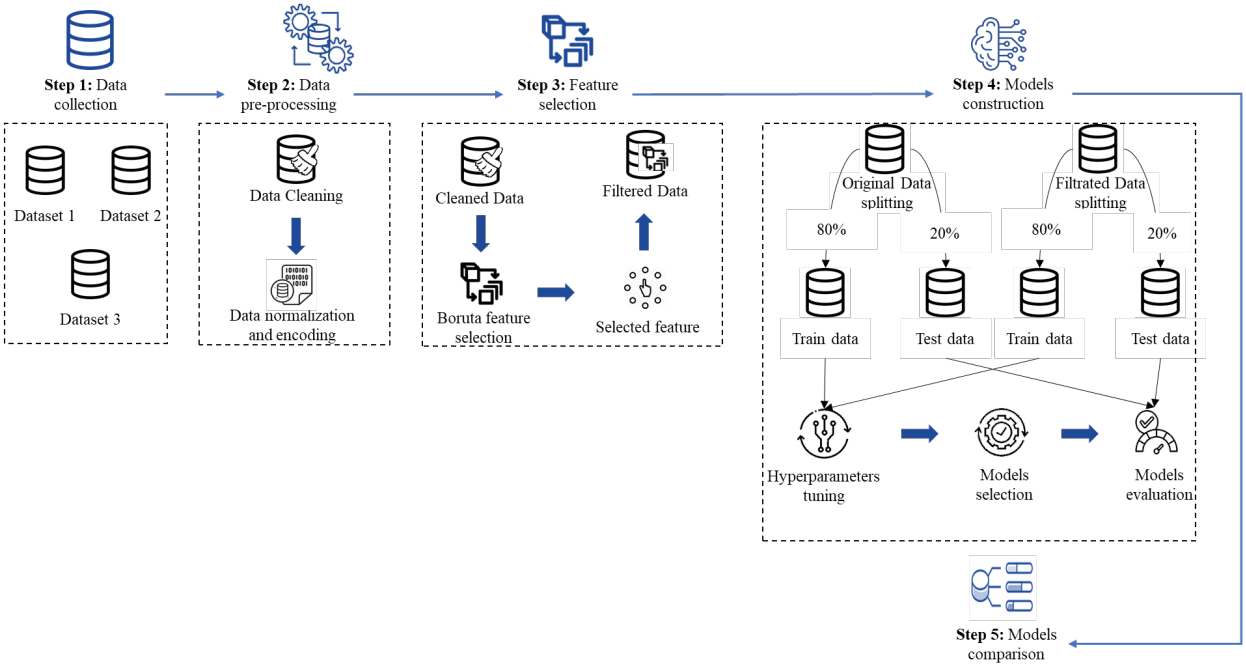

Fig. 1: Methodology schema.

2.1. Data Collection

The study uses three datasets with diverse features and classification targets. The first dataset, referred to as Dataset 1 (E-Commerce Dataset), includes features related to customer transactions and behaviour in an online retail environment, focusing on customer churn. The second dataset, referred to as Dataset 2 (Marketing Campaign Dataset), comprises data on customer responses to various marketing efforts and demographic information. The third dataset, referred to as Dataset 3 (American Bankruptcy Dataset), contains features pertaining to financial indicators and company information relevant to bankruptcy prediction. Each dataset undergoes rigorous pre-processing to ensure data quality and consistency. The details of each dataset, including the number of samples, number of features, target column, number of classes, and class distribution, are summarized in Table 1.

Table 1: Dataset Characteristics.

Dataset	Objective	Samples	Features	Target	Number of classes	Class distribution
1	Customer churn	5630	20	Churn	2	Class 0: 4682, Class 1: 948
2	Response to offer acceptance	2240	29	Response	2	Class 0: 1906, Class 1: 334
3	Campaign status	78682	21	Status	2	Class 0: 73462, Class 1: 5220

2.2. Preparing Data for Machine Learning Models

Data preprocessing involves several critical steps to ensure data quality and suitability for modeling. This includes addressing missing values, converting categorical data, standardizing numerical features, and partitioning the data into training and testing subsets. Missing categorical values are imputed with the mode, while missing numerical values are replaced with the mean. Categorical variables are transformed using label encoding, assigning a unique integer to each category. Numerical features are scaled to a range of [0,1] using MinMaxScaler, ensuring uniformity and improving model performance.

Post-preprocessing, the datasets are divided into training and testing sets with a 80/20 split ratio. This approach facilitates the training of the model on a substantial portion of the data while reserving a separate portion for unbiased evaluation. By splitting the data, the model's performance can be assessed on unseen data, preventing overfitting and ensuring generalizability. This systematic preprocessing ensures that the data is clean, consistent, and ready for effective model training and evaluation.

2.3. Feature Selection

The Boruta algorithm is used to identify and select the most relevant features for each dataset. This algorithm works by comparing the importance of original features to that of randomly generated shadow features. Through an iterative process, Boruta refines the selection, retaining only the features that demonstrate a significant importance over the shadow features. This ensures that only the most predictive and valuable features are included in the model, enhancing its performance and interpretability. After applying the Boruta algorithm, the number of features selected for each dataset are as follows: Dataset 1 retained 13 features, Dataset 2 retained 22 features, and Dataset 3 retained all features.

2.4. Models Construction

Random Forest classifiers are constructed using both the full and Boruta-reduced feature sets. Random Forest, developed by Adele Cutler and Leo Breiman [15], employs a unique splitting strategy for model construction. This method generates numerous decision trees, each trained by randomly selecting a subset of predictive attributes from the entire set. These trees grow to their maximum depth based on a specific subset of features [16]. While the accuracy of individual decision trees might be lower compared to a single tree trained on the full dataset [2], the overall performance of Random Forests improves

as the number of trees increases, leveraging their combined strengths for enhanced model reliability and predictive performance [2].

To construct the Random Forest model, it is important to select the optimal set of parameters, such as the number trees, tree depth, and the number of features considered at each split, to maximize model accuracy. Bayesian optimization employed to efficiently explore the hyperparameters space and find the best configuration, progressively refining its strategy by learning from the outcomes of previous iterations [17]. Consequently, Bayesian optimization helps to fine-the Random Forest classifiers, enhancing their overall performance.

2.5. Models Evaluation

To evaluate the effectiveness of classification models, a variety of metrics can be applied [18]. This study employs five specific metrics: the confusion matrix, accuracy, precision, recall, and F1-score. A confusion matrix, also known as an error matrix, visually represents a model's performance by displaying the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This matrix, as shown in Fig.1, provides a comprehensive overview of a classification model's effectiveness. Accuracy, precision, recall, and F1 score are among the metrics obtained from the confusion matrix.

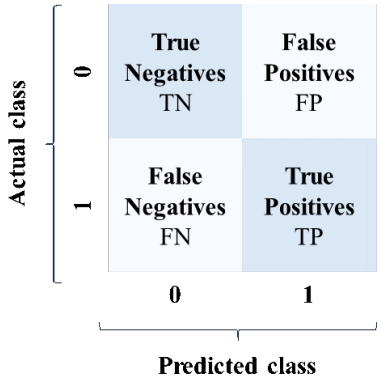

Fig. 2: Binary classification confusion matrix

Accuracy measures the overall correctness of the model's predictions and is calculated using Eq. 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision indicates the proportion of correctly predicted positive instances out of all instances predicted as positive. It is defined by the Eq.2.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall, also known as sensitivity, is the proportion of correctly predicted positive instances out of all actual positive instances, is given by Eq.3.

$$Recall = \frac{TP}{P} \tag{3}$$

The F1 score is a harmonic mean of precision and recall, providing a single metric that balances both. It assesses the model's accuracy by considering both precision and recall as shown in Eq.4.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

2.6. Results and Discussion

The performance of the Random Forest classifiers was evaluated on three datasets, both with and without feature selection, using key metrics such as accuracy, precision, recall, and F1-score. The results are summarized in Table 2. Additionally, the confusion matrices for models with all features and selected features are illustrated in Fig. 3 and Fig. 4, respectively.

For Dataset 1, out of the original 20 features, 7 were selected by the Boruta algorithm. The model without feature selection achieved an accuracy of 95.32%, precision of 88.89%, recall of 81.45%, and F1-score of 85.01%. With feature selection, the model's performance improved, achieving an accuracy of 96.39%, precision of 92.13%, recall of 85.09%, and F1-score of 88.47%. This improvement indicates that removing irrelevant features can enhance model performance by reducing noise and overfitting.

For Dataset 2, out of the original 29 features, 15 were selected by the Boruta algorithm. The model without feature selection showed an accuracy of 87.05%, precision of 65.38%, recall of 17.89%, and F1-score of 28.10%. After applying feature selection, the accuracy slightly increased to 87.50%, with improvements in precision (67.74%) and recall (22.11%), resulting in a higher F1-score of 33.33%. These results suggest that feature selection can significantly improve the model's ability to correctly identify relevant patterns, particularly in datasets with many features.

For Dataset 3, all 21 features were retained by the Boruta algorithm. The model's performance remained consistent with and without feature selection, with both scenarios resulting in an accuracy of 94.31%, precision of 97.87%, recall of 17.10%, and F1-score of 29.11%. This consistency indicates that all features in this dataset were relevant, and feature selection did not provide additional benefits.

Table 1: Performance Metrics Comparison.

Dataset1	Accuracy	Precision	Recall	F1-score
Without feature selection	0.953226	0.888888	0.814545	0.850095
With feature selection	0.963884	0.921259	0.850909	0.884688
Dataset2	Accuracy	Precision	Recall	F1-score
Without feature selection	0.870535	0.653846	0.178947	0.280992
With feature selection	0.875	0.677419	0.221052	0.333333
Dataset3	Accuracy	Precision	Recall	F1-score
Without feature selection	0.943062	0.978723	0.171003	0.291139
With feature selection	0.943062	0.978723	0.171003	0.291139

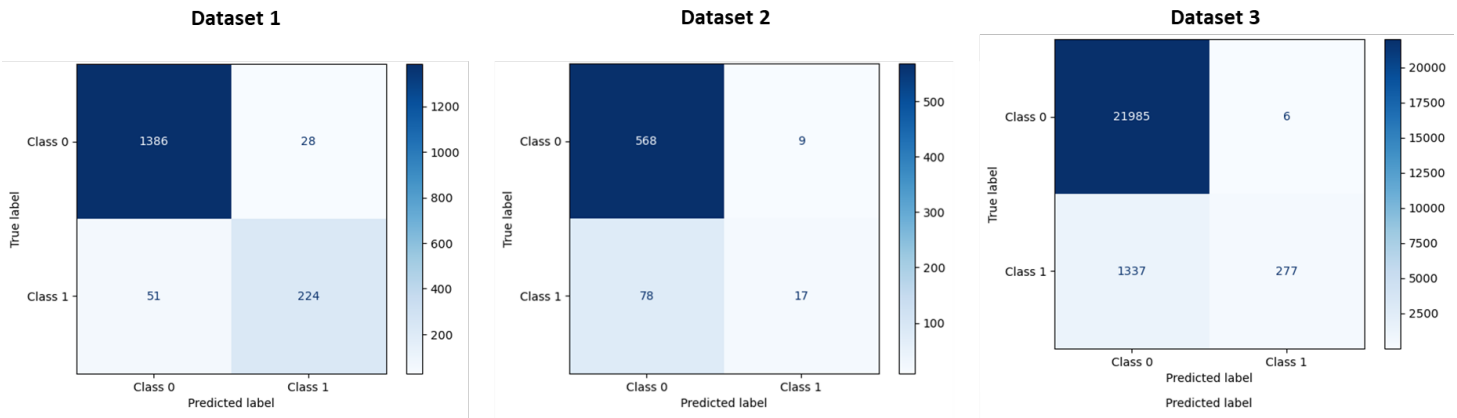

Fig. 3: Confusion matrix with all features.

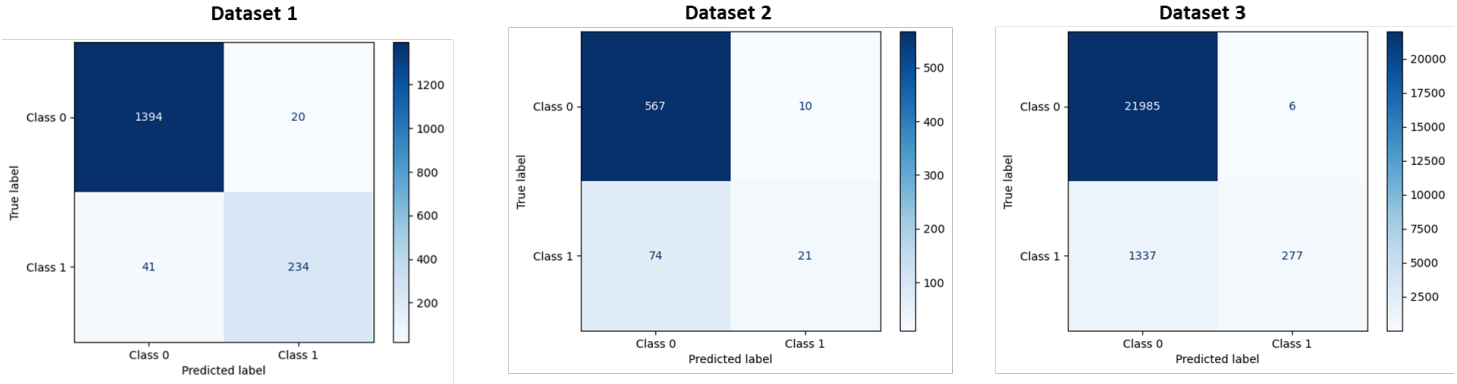

Fig. 4: Confusion matrix with selected features.

These results demonstrate that feature selection generally enhances the performance of the Random Forest classifiers, particularly in terms of precision and F1-score. However, the impact of feature selection can vary depending on the dataset. In cases where irrelevant features are present, feature selection can lead to significant improvements. Conversely, when all features are relevant, the benefits of feature selection may be negligible.

The findings from this study have several important implications for the field of machine learning. The effectiveness of the Boruta algorithm in feature selection underscores its value in improving model performance by retaining relevant features, simplifying models, and making them more interpretable and computationally efficient.

4. Conclusion

This study investigated the impact of feature selection on the performance of Random Forest classifiers across three diverse datasets. The Boruta algorithm was utilized for feature selection with the aim of enhancing model accuracy, precision, recall, and F1-score. Our findings indicate that feature selection generally improves the performance of classification models, especially in terms of precision and F1-score.

For Dataset 1, the application of feature selection led to a significant improvement in model performance, as evidenced by higher accuracy and better precision and recall metrics. Similarly, for Dataset 2, feature selection resulted in notable enhancements in precision and F1-score, demonstrating the algorithm's effectiveness in identifying relevant features. Conversely, for Dataset 3, where all features were deemed relevant, the impact of feature selection was minimal, suggesting that the benefits of feature selection are contingent on the dataset's characteristics.

Confusion matrices were used to provide a comprehensive evaluation of the models, facilitating a detailed analysis of their classification capabilities. These matrices, along with performance metrics such as accuracy, precision, recall,

and F1-score, underscored the importance of appropriate feature selection and hyperparameters tuning in achieving optimal model performance.

Future research could extend this work by applying feature selection methods like Boruta to other ML models, such as as Decision Tree, Gradient Boosting Machines, Support Vector Machines, and Neural Networks, to evaluate their effectiveness across different models. Investigating the synergy between feature selection algorithms and ensemble learning learning could further enhance predictive models. Additionally, implementing automated machine learning (AutoML) frameworks with integrated feature selection and hyperparameters tuning could streamline the model development process, making advanced techniques more accessible and efficient.

Acknowledgements

The support for this research is provided by the Ministry of Higher Education, Scientific Research, and Innovation, as well as the Digital Development Agency (DDA) and the National Center for Scientific and Technical Research (CNRST) of Morocco, under the Smart DLSP Project - AL KHAWARIZMI AI PROGRAM.

References

[1] I. Boukrouh and A. Azmani, 'ARTIFICIAL INTELLIGENCE APPLICATIONS IN E-COMMERCE: A BIBLIOMETRIC STUDY FROM 1995 TO 2023 USING MERGED DATA SOURCES', *Int. J. Prof. Bus. Rev.*, vol. 9, no. 4, p. e4537, Apr. 2024, doi: 10.26668/businessreview/2024.v9i4.4537.

[2] F. Tayalati, A. Azmani, and M. Azmani, 'Application of supervised machine learning methods in injection molding process for initial parameters setting: prediction of the cooling time parameter', *Prog. Artif. Intell.*, Apr. 2024, doi: 10.1007/s13748-024-00318-z.

[3] Y. Riahi, T. Saikouk, A. Gunasekaran, and I. Badraoui, 'Artificial intelligence applications in supply chain: A descriptive bibliometric analysis and future research directions', *Expert Syst. Appl.*, vol. 173, p. 114702, 2021.

[4] Y. Guo, Z. Hao, S. Zhao, J. Gong, and F. Yang, 'Artificial intelligence in health care: bibliometric analysis', *J. Med. Internet Res.*, vol. 22, no. 7, p. e18228, 2020.

[5] B. Beirami and M. Mokhtarzade, 'Supervised and Unsupervised Clustering Based Dimensionality Reduction of Hyperspectral Data', *Int. J. Eng.*, vol. 34, no. 6, pp. 1407–1412, 2021, doi: 10.5829/ije.2021.34.06c.03.

[6] M. Biglari, F. Mirzaei, and H. Hassanpour, 'Feature Selection for Small Sample Sets with High Dimensional Data Using Heuristic Hybrid Approach', *Int. J. Eng.*, vol. 33, no. 2, pp. 213–220, 2020, doi: 10.5829/ije.2020.33.02b.05.

[7] P. Kalpana and K. Mani, 'A New Hybrid Framework for Filter based Feature Selection using Information Gain and Symmetric Uncertainty (TECHNICAL NOTE)', *Int. J. Eng.*, vol. 30, no. 5, pp. 659–667, 2017, doi: 10.5829/idosi.ije.2017.30.05b.05.

[8] G. Manikandan, B. Pragadeesh, V. Manojkumar, A. L. Karthikeyan, R. Manikandan, and A. H. Gandomi, 'Classification models combined with Boruta feature selection for heart disease prediction', *Inform. Med. Unlocked*, vol. 44, p. 101442, 2024, doi: 10.1016/j.imu.2023.101442.

[9] J. Li, Y. Liu, H. Gong, and X. Huang, 'Stock price series forecasting using multi-scale modeling with boruta feature selection and adaptive denoising', *Appl. Soft Comput.*, vol. 154, p. 111365, 2024, doi: https://doi.org/10.1016/j.asoc.2024.111365.

[10] N. Farhana, A. Firdaus, M. F. Darmawan, and M. F. Ab Razak, 'Evaluation of Boruta algorithm in DDoS detection', *Egypt. Inform. J.*, vol. 24, no. 1, pp. 27–42, Mar. 2023, doi: 10.1016/j.eij.2022.10.005.

[11] L. Wang, S. He, Z. Zhao, and X. Zhang, 'Prediction of hot-rolled strip crown based on Boruta and extremely randomized trees algorithms', *J. Iron Steel Res. Int.*, vol. 30, no. 5, pp. 1022–1031, May 2023, doi: 10.1007/s42243-023-00964-y.

[12] Yuan, X., Chen, F., Xia, Z., Zhuang, L., Jiao, K., Peng, Z., ... & Hou, Z., 'A novel feature susceptibility approach for a PEMFC control system based on an improved XGBoost-Boruta algorithm', *Energy AI*, vol. 12, p. 100229, 2023, doi: https://doi.org/10.1016/j.egyai.2023.100229.

[13] D. Du, B. He, X. Luo, S. Ma, Y. Song, and W. Yang, 'Spatio-Temporal Variation Analysis of Soil Salinization in the Ougan-Kuqa River Oasis of China', *Sustainability*, vol. 16, no. 7, 2024, doi: 10.3390/su16072706.

[14] M. B. Kursa and W. R. Rudnicki, 'Feature Selection with the Boruta Package', *J. Stat. Softw.*, vol. 36, no. 11, pp. 1–13, 2010, doi: 10.18637/jss.v036.i11.

[15] L. Breiman, 'Random Forests', *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[16] I. Ullah, H. Hussain, I. Ali, and A. Liaquat, 'Churn Prediction in Banking System using K-Means, LOF, and CBLOF', in *1st International Conference on Electrical, Communication and Computer Engineering, ICECCE 2019*, Institute of Electrical and Electronics Engineers Inc., 2019. doi: 10.1109/ICECCE47252.2019.8940667.

[17] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, 'Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis', *Informatics*, vol. 8, no. 4, p. 79, Nov. 2021, doi: 10.3390/informatics8040079.

[18] Ž. Đ. Vujovic, 'Classification Model Evaluation Metrics', *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, 2021, doi: 10.14569/IJACSA.2021.0120670.