Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 4, Agustus 2025, Page. 2153-2170
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4810

Interpretable Machine Learning for Employee Recruitment Prediction Using Boruta, CatBoost, Lasso, Logistic Regression, NLP, and RFE Feature Selection

Aswan Supriyadi Sunge*[1], Suzanna[2], Hamzah Muhammad Mardi Putra[3]

[1]Informatics Engineering Department, Faculty of Engineering, Pelita Bangsa University, Indonesia
[2]Informatics Information Systems Department, Binus Online Learning, Bina Nusantara University, Indonesia
[3]Management Department, Faculty of Economics and Business, Pelita Bangsa University, Indonesia

Email: [1]aswan.sunge@pelitabangsa.ac.id

Received: Jun 2, 2025; Revised: Jun 27, 2025; Accepted: Jul 20, 2025; Published: Aug 19, 2025

Abstract

Employee recruitment is one of the crucial processes in human resource management that has a direct impact on the performance and success of the company. In the digital era, the use of Machine Learning (ML) in candidate selection processes is increasingly prevalent due to its ability to enhance efficiency, accuracy, and transparency. This research is important because conventional recruitment methods often face issues such as subjective bias, slow processing times, and limitations in assessing a candidate's true potential. ML offers a more objective, data-driven, and faster approach, enabling companies to identify the best candidates more effectively. This study aims to identify the main features that influence recruitment decisions, as well as evaluate the effectiveness and interpretability of several ML models, namely Boruta, CatBoost, Lasso Regression, Logistic Regression, Natural Language Processing (NLP), and Recursive Feature Elimination (RFE). This study uses a dataset consisting of 1,501 samples with 10 features and one class variable (0 = Not Hired, 1 = Hired). The evaluation is carried out based on the ability of each model to identify the features that make the most significant contribution to the classification results. This study has several limitations, particularly the potential bias in the data, such as demographic bias that may be reflected in historical recruitment decisions. This could lead the ML models to replicate or even reinforce such biases. Additionally, the limited dataset size may affect the models' ability to generalize to new data. In the context of this study, the main parameter used to assess the superiority of the model is the most dominant feature or the highest feature produced by each method. The test results show that the Boruta model identifies Gender as the most influential feature, while the CatBoost, Lasso Regression, Logistic Regression, and NLP models consistently place Recruitment Strategy as the most significant feature in predicting candidate eligibility. Meanwhile, the RFE model produces Distance from the Company as the highest feature that influences recruitment decisions. The uniqueness of this study lies in its approach that integrates feature interpretability models within the real-world context of recruitment decision-making. This approach not only emphasizes prediction accuracy but also promotes transparency and a clear understanding of the rationale behind each decision. It supports the development of a fairer and more accountable selection process, particularly by minimizing unconscious bias in data-driven recruitment systems. From a practical standpoint, the findings are highly relevant for human resource professionals, as the identified key features can be used to design more objective selection strategies and enhance the efficiency of candidate evaluations. Therefore, this study makes a tangible contribution to the advancement of modern, technology-based recruitment systems that prioritize fairness and decision-making efficiency. Additionally, the selection of evaluation metrics could be further elaborated to strengthen the analysis, for example by presenting the overall accuracy of each model or comparing them with alternative approaches to provide a more comprehensive view of the models' performance.

Keywords: *Employee Selection, Feature Selection, Interpretable Models, Recruitment Prediction.*

1. INTRODUCTION

Employee recruitment is one of the processes of searching for human resources that includes stages to select and obtain the most suitable or appropriate candidates for the company [1]. Stages include compiling job criteria and specifications, with the dissemination of information through job advertisements, social media, or through colleagues [2]. It is very important to align the message with the company's vision and values so that the company succeeds in achieving its goals [3].

After the vacancy announcement, the next step is screening through application files, interviews, and tests. This process is intended to see the abilities and suitability that match the criteria sought [4]. Thus, good recruitment helps companies build productive teams that help achieve the company's long-term planning goals [5],[6].

Some of the problems that often occur in terms of employee recruitment are the gap between the company's needs and the quality of existing candidates. As a result, many companies do not get candidates who do not match the required qualifications, especially when those who will work are not by their portion, even if there are sometimes candidates who have to be trained or taught again which has a long impact on the process to be truly competent in working [7].

One of the biggest challenges in the recruitment process is the uncertainty or lack of clarity regarding the characteristics or criteria that should form the basis for hiring decisions. In addition, companies aim to allocate candidates who possess not only work experience or technical skills but also strong communication and teamwork abilities. This issue has a tangible impact on companies and the industry, as poor hiring decisions can lead to decreased productivity, increased training costs, and high employee turnover rates, ultimately affecting business competitiveness and sustainability.

This problem is further complicated by the fact that the field of work is constantly changing, and employees must be very flexible. In other words, companies must identify the individual needs of each position more carefully and evaluate the prospects of prospective candidates. In this way, companies identify the highest features that are relevant to each criterion and recruitment becomes more effective and efficient.

One potential solution to address recruitment challenges is the utilization of ML technology, particularly through top feature selection techniques. This approach allows companies to analyze historical data from employees who were either successful or unsuccessful in previous roles, in order to identify which features consistently have the most significant impact on recruitment outcomes. By applying specific algorithms, ML helps organizations uncover important patterns that may not be easily detected through manual evaluation. ML plays a strategic role in tackling modern recruitment challenges, such as handling a large volume of applicants, limited time for candidate evaluation, and the need for objective and unbiased decision-making. In addition to automating the initial screening process, ML provides data-driven insights into the characteristics of candidates that best match the company’s requirements. The dataset used in this study reflects a broad spectrum of recruitment needs, consisting of 1,501 samples with ten diverse candidate features, including recruitment strategy, educational background, work experience, distance from the company, and location preferences. This diversity offers a representative overview of real-world hiring dynamics and enables the evaluation of ML models in a realistic and practically applicable recruitment context.

Previous studies have extensively explored the application of ML in employee recruitment, highlighting its potential to automate candidate screening, extract predictive features, and improve the quality of selection decisions. ML has proven effective in accelerating recruitment processes that are otherwise complex and time-consuming when conducted manually, particularly during the initial document review and candidate evaluation stages [8],[9]. It also offers cost efficiency and improved predictive accuracy [10]. Various models have been employed, including Random Forest, Logistic Regression, K-Nearest Neighbors (K-NN), and Natural Language Processing (NLP) using classification techniques [11],[12],[13]. Other studies have utilized models such as CatBoost, KD-Trees, and ensemble learning approaches to achieve more objective and efficient outcomes [14],[15],[16]. However, as the volume and complexity of recruitment data increase—encompassing aspects such as cultural diversity, ethics, talent, and privacy—more sophisticated approaches like Neural Networks are required to capture nonlinear patterns and handle large-scale data effectively [17],[18],[19],[20]. Nonetheless, many ML applications in recruitment remain overly focused on predictive accuracy, often overlooking transparency and interpretability. This leads to the so-called "black box" problem, where model decisions are difficult to understand or explain to non-technical users [21],[22],[23]. To address this issue, feature selection has emerged as a more interpretable and informative ML approach, allowing the identification of the most relevant attributes in the employee recruitment and selection process [24],[25].

By leveraging ML, companies can enhance the process of determining employee selection criteria, making it more detailed, efficient, and data-driven. One of the key advantages is the ability to analyze historical performance data to identify and prioritize the most relevant features. This approach integrates various pertinent parameters, enabling not only a reduction in the time required to identify suitable candidates but also promoting a more transparent, objective, and error-resistant selection process. The feature selection model employed in this study represents an application of established ML techniques within a novel context—namely, prediction and interpretability in employee recruitment. Ultimately, more accurate and informed decision-making facilitates the selection of high-performing teams, contributing to the development of a work culture that fosters innovation and strengthens competitive advantage.

Therefore, the contribution of this research is:

1. Model interpretation is also presented in the form of a bar chart which provides information on the highest importance of features and reduces uncertainty in determining the most relevant and objective selection criteria.
2. Provides literature in the comparison of six models in the selection of the highest features, and also provides an in-depth understanding of the advantages and limitations of each model.
3. Development of new models in evaluation using models based on candidate selection through the highest or most influential feature search method to better understand the dynamics of company needs.

2. SIMILAR PREVIOUS RESEARCH

One of the main advantages of various ML tasks is its ability to perform feature selection automatically, quickly, and easily [26],[27]. This process allows to identify and select the most relevant variables for analysis, thereby reducing data problems and increasing model efficiency [28]. With feature selection, ML can eliminate unimportant data, which in turn helps reduce overfitting, improve model accuracy, and speed up training time [29]. Another advantage is that the resulting model becomes simpler and easier to interpret, so it can be used to make more informed decisions based on relevant data [30],[31].

In terms of employee recruitment, feature selection methods are instrumental and important, especially considering the diversity of candidate data that must be analyzed to determine their suitability for job placement. Using the right and appropriate feature model is the key to filtering relevant information from the available data so that it can form a more accurate predictive model in assessing prospective employees. In addition, not only in recruitment but also factors that can contribute to the success of candidates in the work environment, such as cultural fit and individual motivation [32].

There are various ML models that are often used for feature selection, each with different approaches and techniques, thus providing many options to choose the one that best suits the needs of data analysis. The model that is often used is Boruta. Several studies with this model are selecting or searching for influential features in predicting heart disease [33], diabetes [34], prostate cancer [35], stock price prediction [36].

In addition, other models CatBoost, and some studies in feature search such as cervical cancer prediction [37], battery health [38], and soil resistance structure [39]. In addition, the Least Absolute Shrinkage and Selection Operator (Lasso Regression) is one of the models used to determine Bitcoin transaction fees [40], credit risk evaluation [41], and genomic cancer feature search [42]. Another model in feature selection is Logistic Regression and several studies use this model such as the selection of colorectal cancer features [43], and breast cancer [44]. Meanwhile, the Recursive Feature Elimination model abbreviated as RFE, and several others are used in research in feature selection such as the selection of colorectal cancer features [45], Parkinson's [46] and liver [47]. Finally, the model that is often used is Natural Language Processing known as NLP, although it is often used in sentiment analysis, it can be used in feature selection such as inappropriate content [48] and Arabic dialects [49].

This study presents a comprehensive approach to feature selection and predictive modeling for employee recruitment using a comparative analysis of several machine learning techniques. Employing multiple models enables the identification of the most influential features, thereby improving both predictive performance and interpretability. The Boruta algorithm, built upon the Random Forest ensemble, effectively selects all relevant features by leveraging the robustness of multiple decision trees. CatBoost demonstrates superior performance in handling categorical variables, offering a powerful mechanism for identifying significant predictors in structured datasets. Lasso Regression introduces regularization to eliminate irrelevant or redundant features, enhancing the model’s generalizability and interpretability. Logistic Regression serves as a reliable and interpretable baseline model for binary classification tasks within recruitment contexts. Recursive Feature Elimination (RFE) iteratively refines feature subsets by evaluating their contributions to model performance, facilitating the selection of an optimal feature set. Additionally, Natural Language Processing (NLP) techniques are employed to extract meaningful features from unstructured text data such as resumes, cover letters, and job descriptions, enriching the overall feature space with context-sensitive linguistic information. The integration of these methodologies supports the development of a robust, accurate, and interpretable recruitment prediction framework, capable of assisting decision-makers in enhancing the efficiency and fairness of the candidate selection process. Compared to standalone methods like Random Forest or interpretability tools such as SHAP, this integrated approach offers a more holistic and practical solution. While Random Forest provides high predictive power, it often lacks transparency in feature importance without further interpretation tools. Similarly, SHAP values—although powerful in explaining model outputs—can be computationally intensive and complex to implement across multiple models. By combining diverse models and selection techniques, this study ensures both performance and interpretability without over-reliance on a single method, offering more balanced and actionable insights for real-world recruitment decisions.

3. RESEARCH METHODOLOGY

This study is an experimental research employing a quantitative approach, grounded in a systematic review of relevant literature and theoretical frameworks. The primary focus of this research is to implement ML models to identify the most significant features influencing candidate selection processes. The objective is to provide meaningful scientific contributions to the development of data-driven recruitment systems and to address research questions related to model performance comparison and the selection of the most relevant features for successful recruitment, as outlined in Figure 1.

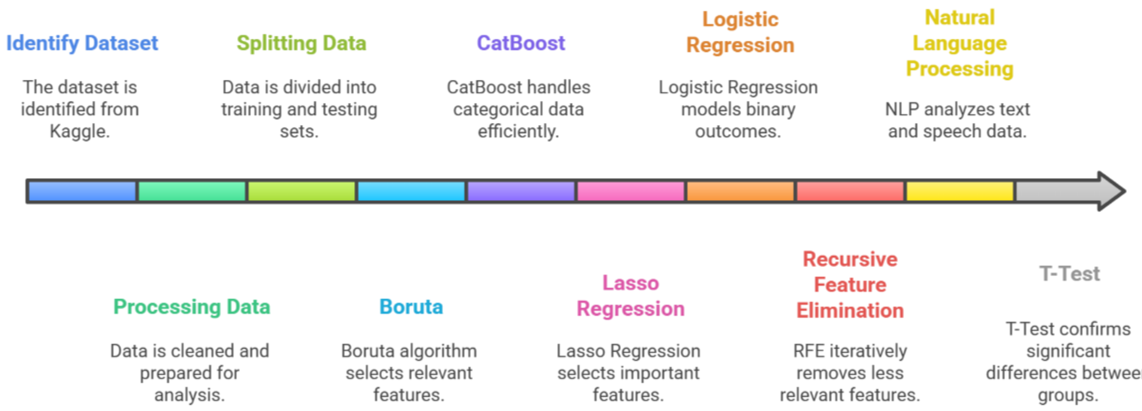
Identify Dataset
The dataset is identified from Kaggle.
Splitting Data
Data is divided into training and testing sets.
CatBoost
CatBoost handles categorical data efficiently.
Logistic Regression
Logistic Regression models binary outcomes.
Natural Language Processing
NLP analyzes text and speech data.
Processing Data
Data is cleaned and prepared for analysis.
Boruta
Boruta algorithm selects relevant features.
Lasso Regression
Lasso Regression selects important features.
Recursive Feature Elimination
RFE iteratively removes less relevant features.
T-Test
T-Test confirms significant differences between groups.

Figure 1. Framework Model

3.1. Datasets

This section aims to identify the data used in this study and ensure that the data can be tested effectively and relevantly. Table 1 shows the dataset used in this study, which was obtained from the Kaggle ML Repository and is licensed under CC BY 4.0. The dataset is accessible via *https://www.kaggle.com/datasets/rabieelkharoua/predicting-hiring-decisions-in-recruitment-data,* and consists of data contributed by the machine learning community for analysis using various algorithms. This dataset is open to the public and is updated regularly. The dataset consists of 1501 samples, with 10 features and 1 class (0 = Not Hired, 1 = Hired). The purpose of this dataset is to identify important factors to help predict hiring decisions.

This information reflects the challenges in the employee selection process due to the variety of factors and attributes used to evaluate candidates. As methods and strategies in this field evolve, so does the use of new tools and techniques to assess candidates. However, there are some limitations, such as the lack of current data that includes changes in the features of the selection process, and the reliance on variables that may not always be relevant or reflect the company’s current needs. Additionally, this information may not fully reflect practices across industries or regions, as much of the data comes from a single source. Nevertheless, the analysis performed still provides valuable insights to improve employee hiring decision predictions.

Table 1. Features Name, and Values

No	Training	Testing
1	Age	Numerical
2	Gender	Categorical
3	Education Level	Categorical
4	Experience Years	Numerical
5	Previous Companies	Numerical
6	Distance From Company	Numerical
7	Interview Score	Numerical
8	Skill Score	Numerical
9	Personality Score	Numerical
10	Recruitment Strategy	Numerical
11	Hiring Decision	Categorical

3.2. Preprocessing Dataset

The data processing process begins with the pre-processing stage, which aims to prepare the data before testing so that it can reduce errors and support the ML data analysis process. The steps of this stage vary depending on the purpose and model used, but the data quality must be thoroughly checked before testing is carried out. Common techniques used include checking and cleaning the data, such as identifying unrecognized symbols, duplicates, typos, or empty data. The main goal is to ensure that the data is filled in completely, there are no blanks and is free from errors. To check for missing or invalid data, the syntax *data.isnull().sum()* is often used to display the number of errors or missing data, which can be seen in Table 2.

Table 2. Checking and Cleaning Data Results

No	Features Name	Results
1	Age	0
2	Gender	0
3	Education Level	0
4	Experience Years	0
5	Previous Companies	0
6	Distance From Company	0
7	Interview Score	0
8	Skill Score	0
9	Personality Score	0
10	Recruitment Strategy	0
11	Hiring Decision	0

3.3. Splitting Dataset

At this stage, the dataset is divided into two main parts, namely 80% of the dataset is used for training data, and the remaining 20% is used for testing data. This division is chosen to ensure that the model has enough data to train, while still providing sufficient data for an accurate evaluation process. The purpose of this division is to build an effective ML model and evaluate its performance objectively.

3.4. Boruta Model

It is a feature selection algorithm that uses a Random Forest classifier to determine which features are most relevant to the dependent variable in a dataset [50]. Unlike other feature selection methods, this model focuses on selecting features that are truly important to the prediction model, rather than simply reducing the number of features. This algorithm compares the importance of the original features with shadow features and is capable of handling non-linear relationships and interactions between features. The specific steps of how it works are as follows [51];

1. Feature duplication, for each original feature (e.g. features X1, X2, ..., Xn), create a random copy of that feature, called a shadow feature.
2. Model training uses Random Forest to build a model based on existing data, including original features and shadow features.
3. Assessing features, after the model is built, each feature (both original and shadow) is assessed based on the model performance. Important features have higher scores compared to shadow features. If the original feature is more important than the best shadow feature, then the feature is considered relevant and is retained.

4. Repeating the process, to assess the consistency of the importance of features, then classifying which features are important, unimportant, and not necessarily relevant or not.
5. The final result, important features are retained, and unimportant or similar shadow features are removed from the model.

3.5. CatBoost Model

Short for Categorical Boosting, is an algorithm that uses a gradient boosting approach and is designed to process categorical data efficiently. Developed by Yandex in 2017, this algorithm is known for its excellent ability to handle categorical data [52]. This algorithm relies on symmetric trees as the basic structure, applies preprocessing techniques for feature separation, and combines a sorted boosting strategy to reduce gradient bias and prediction offset problems that are common in gradient boosting algorithms [53]. In general, the formula used is gradient enhancement, which is explained below:

1. Gradient Boosting Model, every m-iterations, the model updates the prediction by adding a new decision tree to correct the errors of the previous model.

$$F_m(x) = F_{m-1}(x) + \eta . h_m(x) \qquad (1)$$

Where:

- $F_m(x)$ is the model prediction at the mth iteration for input *x*
- $F_{m-1}(x)$ is the model prediction at the previous iteration (m-1)
- η is the learning rate that controls how much the decision tree contributes.
- $h_m(x)$ is the mth decision tree learned at the m-iterations.

2. Loss Function, which means trying to minimize the loss function, depending on the type of problem to be solved, whether classification, regression or others.
3. Ordered Boosting, which means it is designed to reduce the overfitting problem that usually occurs in other boosting algorithms. With this method, the data is specifically sorted to calculate the contribution of each tree more carefully.
4. Converting categorical data to numeric representation in a more sophisticated way. One technique used is combinatorial hashing or mean encoding, which converts categories to numbers based on the average distribution of the target for each category.

3.6. Lasso Regression Model

Least Absolute Shrinkage and Selection Operator or known as Lasso Regression, is a regression method designed to improve model accuracy by reducing its complexity through automatic feature selection [54]. This model is very effective in situations with many features because it automatically selects features that have a large contribution to the prediction [55].

3.7. Logistic Regression Model

A model is one of the statistical techniques used to model the relationship between a categorical dependent variable (usually binary) and one or more independent variables [56]. Although the term "regression" is in its name, Logistic Regression is more often used for classification tasks, especially when the target variable has two categories, such as "yes" or "no", "true" or "false", or 1 and 0 [57].

1. Logistic Regression Function:

$$P(Y = 1|X = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_n X_n)}} \qquad (2)$$

Here β_1, β_2,…..,β_n are coefficients indicating the contribution of each feature X_1,X_2,…..,X_n to the prediction probability.

2. In feature selection using logistic regression with the steps of feature coefficients, feature elimination, and gradient selection.

3.8. Recursive Feature Elimination Model

Known as RFE, it is a feature selection method designed to improve model performance by gradually removing less relevant features through a recursive process [58]. This technique is very effective in overcoming high-dimensional data problems because it can reduce the computational complexity of the model while increasing predictive capabilities [59]. In general, RFE does not have a special formula, it still follows an iterative process in removing features, for example, X= [X_1, X_2, X_3, …..X_n] is the initial feature set, and Y is the target variable.

3.9. Natural Language Processing Model

This model is known as NLP, which is a part of artificial intelligence that focuses on the interaction between computers and human language [60]. This technology allows machines to understand, analyze, and generate natural language effectively so that it can be applied in various fields such as sentiment analysis, machine translation, and chatbots [61]. NLP includes a variety of techniques and models used to process text and speech in various formats, including unstructured text such as articles, conversations, or documents [62]. In general, RFE does not have a single formula because NLP includes various approaches and algorithms for processing and analyzing human language.

3.10. T-Test

The statistical analysis in this study was conducted using the independent samples t-test, which is commonly used to determine whether there is a statistically significant difference in the means of two independent groups [63]. This method was chosen because it is one of the most familiar statistical analysis techniques, offering simplicity in calculation and efficiency in handling samples. The t-test is also well-regarded for its ability to assess mean differences while accounting for sample size and variance under the assumption of approximate normality [64]. In this study, the resulting p-value was significantly lower than the conventional threshold of 0.05, leading to the rejection of the null hypothesis [65]. This outcome statistically confirms the presence of a real difference between the two groups and supports the study’s overall conclusions.

4. RESULT

This section presents an analysis of the Boruta, CatBoost, Lasso Regression, Logistic Regression, RFE, and NLP algorithm models with the aim of seeing the highest features of each model and comparing each model in handling the complexity and characteristics of the available data.

From the Boruta model testing in Figure 2, the highest features in employee recruitment prediction are Gender, because there may be historical patterns in the data where certain genders are more likely to be accepted, and Previous Companies, as an indicator of work experience also stands out because it shows how much exposure a candidate has had to previous work environments, an important consideration for recruiters.

From the results of the CatBoost model test in Figure 3, the highest features are Recruitment Strategy, because because it is able to evaluate the interactions between features and recognize that recruitment strategies (e.g., internal, external, referral channels) can have a strong influence on candidate success, and Education Level, because educational background often plays a role in screening candidates in many companies.

Vol. 6, No. 4, Agustus 2025, Page. 2153-2170
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4810

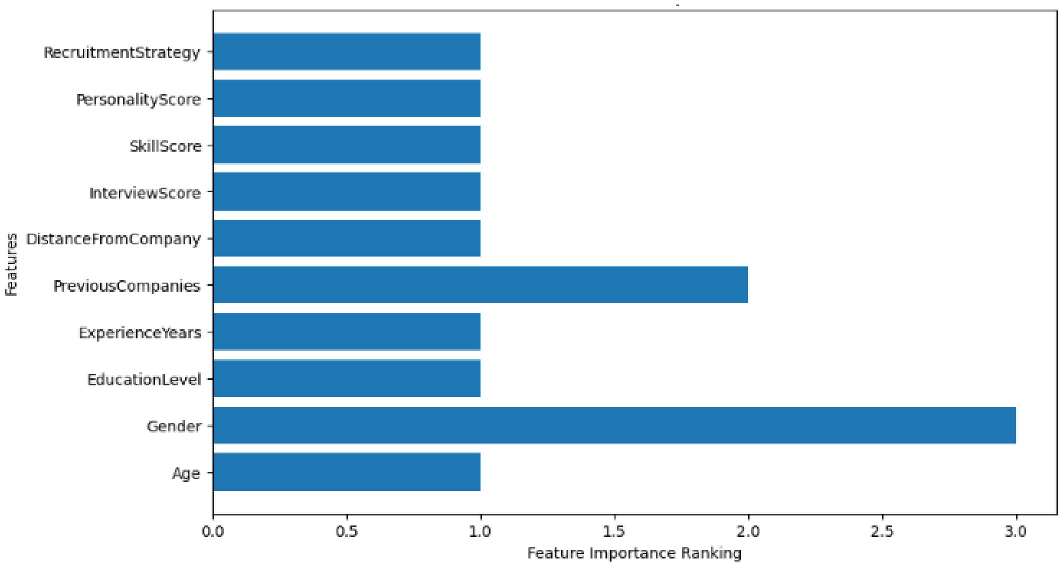

Figure 2. Top Features of Boruta

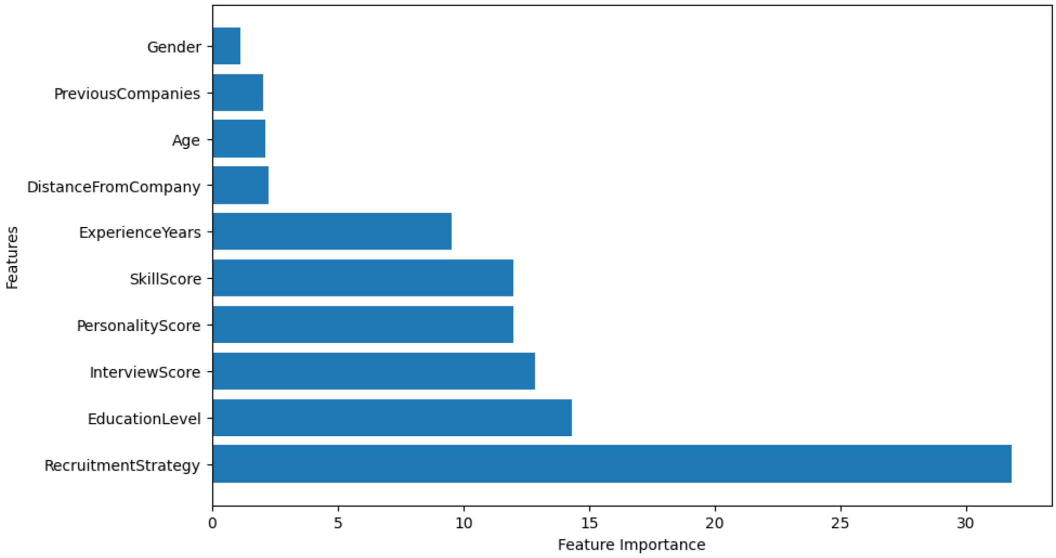

Figure 3. Top Features of CatBoost

From the results of the Lasso model test in Figure 4, the highest features are Recruitment Strategy, because persists because it has a strong linear correlation with acceptance status, and Experience Years, selected to retain features with the most stable predictive contribution to the target variable, especially when these features show a clear linear relationship.

From the test results with the Logistic Regression model in Figure 5, the highest features are Recruitment Strategy, and Experience Years, because shows that both features have a significant contribution in the linear model to the probability of a candidate being accepted.

From the test results with the RFE model in Figure 6, the highest features are Recruitment Strategy, because it is a key feature because it provides a stable contribution to model accuracy, and Skill Score, because in its iterative process this feature consistently improves model performance when other features are eliminated.

From the test results with the NLP model in Figure 7, the highest features are Distance from Company, because because it is an important feature because it reflects geographic proximity that may affect punctuality, work commitment, or loyalty, and Age, because in numerical representation, age can be closely related to experience or work readiness.

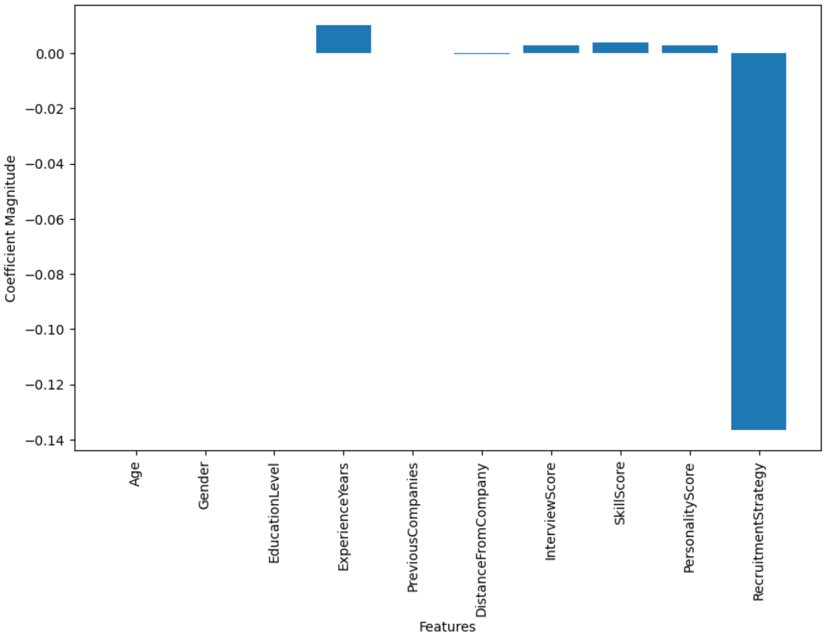

Figure 4. Top Features of Lasso

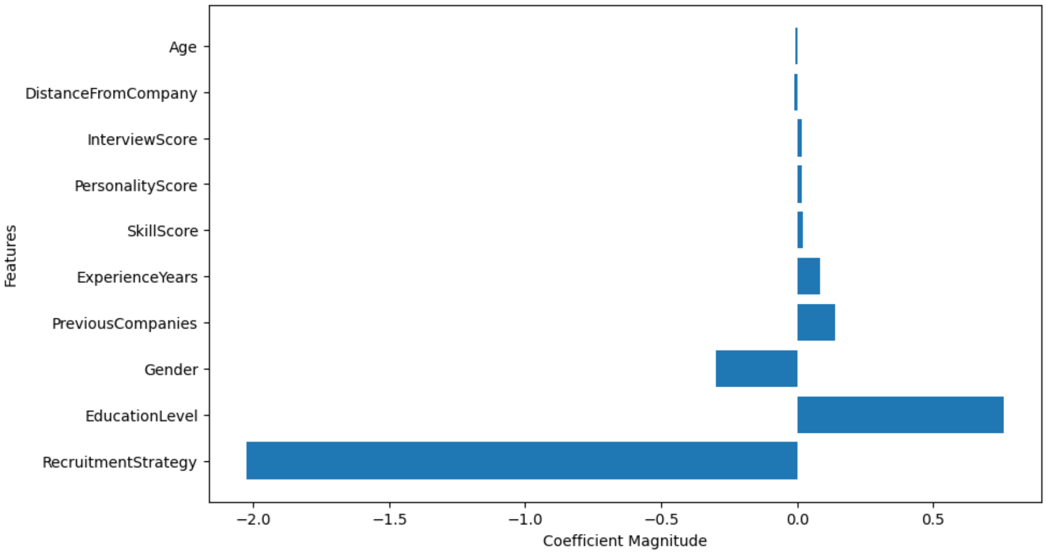

Figure 5. Top Features of Logistic Regression

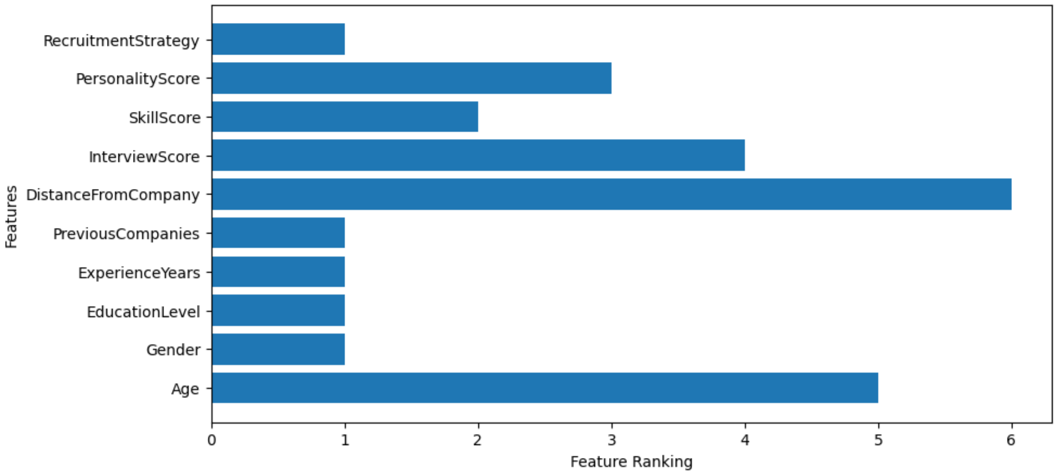

Figure 6. Top Features of RFE

DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4810

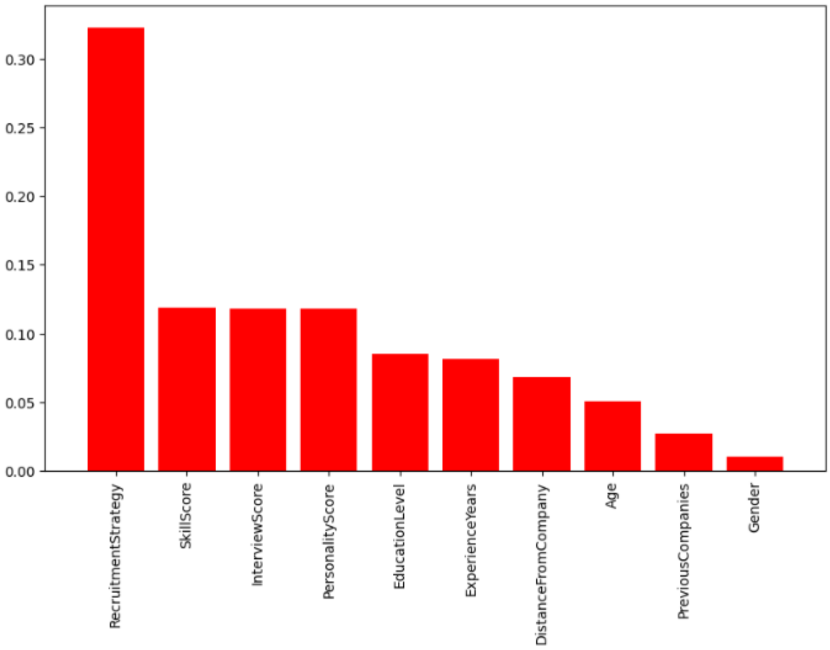

Figure 7. Top Features of NLP

Based on the results of all tests, Figure 8, shows the top two features identified by each model.

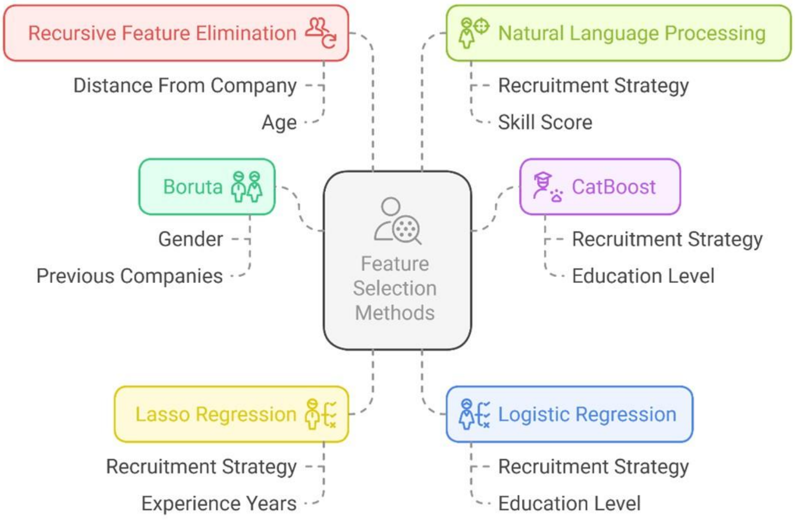

Figure 8. Top Features of Each Model

5. DISCUSSIONS

The testing results revealed variations in the key features identified by each model, although some models produced similar outcomes. One of the most consistent findings was that Recruitment Strategy emerged as the most dominant feature across the six evaluated models, indicating its high relevance and significant predictive power in recruitment success. These differences in selected features can be attributed to the distinct algorithmic approaches used in each feature selection method. For instance, Boruta applies a Random Forest-based method to compare original features with randomly permuted ones, selecting those with the most statistical influence. CatBoost, as a boosting model, effectively captures complex interactions between features, especially categorical data. Lasso Regression uses L1 regularization to shrink the coefficients of less relevant features to zero, whereas Logistic Regression relies on the strength of linear relationships between features and the target variable. RFE (Recursive

Feature Elimination) gradually removes the least contributing features to retain only those that significantly impact model performance. Meanwhile, NLP-based approaches utilize techniques such as TF-IDF or word embeddings to extract relevant textual features by accounting for linguistic and semantic context.

These methodological differences highlight that each model evaluates feature importance and relevance through its unique computational lens, leading to diverse feature selection outputs. Nevertheless, several top-ranked features—such as Recruitment Strategy, Experience Years, Education Level, Skill Score, and Distance from Company—demonstrate strong potential for practical application in modern recruitment systems. For example, organizations can prioritize recruitment strategies that have been empirically proven effective, such as employee referrals or internal hiring, to enhance recruitment outcomes. Additionally, Experience Years and Education Level can serve as initial filters in digital recruitment platforms to efficiently screen candidates who meet baseline requirements. Skill Score may be integrated into competency-based assessments, while Distance from Company could offer insights into a candidate's likelihood of long-term retention or job stability. By incorporating these features into data-driven decision support systems, companies can streamline the recruitment process, making it more efficient, transparent, and accurate in identifying the most suitable candidates.

Then a test was conducted using the independent two-sample t-test method, which is a statistical method used to determine whether there is a significant difference between the averages of two independent groups. In this case, the groups being compared are:

1. Group 1: Unaccepted individuals (label 0)
2. Group 2: Accepted individuals (label 1).

The focus of the test is directed at the Recruitment Strategy feature, which was previously identified as the most dominant feature by several interpretability models. The purpose of this test is to determine whether there is a difference in the average Recruitment Strategy score between the two groups.

The test results show that the T-statistic value is 25.8810, which indicates that the difference between the two groups is quite large when compared to the variation of the data. The very small p-value (3.55×10^{-128}) is far below the significance limit of $\alpha = 0.05$, so the statistical decision is to reject the null hypothesis (H_0) which states that there is no difference in the average between the groups.

Thus, it can be concluded that there is a statistically significant difference in the Recruitment Strategy scores between accepted and rejected individuals. This means that the Recruitment Strategy feature is significantly related to the acceptance decision, and can be considered as one of the key factors in the recruitment prediction process.

6. CONCLUSION AND SUGGESTIONS FOR THE FUTURE

Based on the test results, it can be concluded that each feature selection model produces different highest features. In the Boruta model, the highest features selected are Gender and Previous Companies, while in the CatBoost model, the most influential features are Recruitment Strategy, Education Level, and Interview Score. The Lasso model selects Recruitment Strategy, Experience Years, and Skill Score as the most important features, while the Logistic Regression model identifies Recruitment Strategy, Education Level, and Gender as the main features. The NLP model selects Recruitment Strategy, Skill Score, and Education Level as the most important features, while the RFE model prioritizes Distance from Company, Age, and Interview Score. These differences in results occur because each model uses a different approach in assessing the relevance of features to predictions, with different ways of handling data, measuring feature contributions, and adjusting to the objectives or algorithms used.

Although the dataset used in this study is balanced between the classes of hired and non-hired candidates, several limitations should still be acknowledged. One key limitation is the relatively small sample size compared to the number of features analyzed. While the dataset is sufficiently representative for initial analysis, it could be expanded to capture more complex variations in real-world recruitment scenarios. Additionally, in practical implementation within corporate environments, several challenges may arise, including the need to safeguard candidate data privacy, regularly update models to maintain relevance, and potentially incorporate Deep Learning approaches to better adapt to evolving trends in human resource management.

Suggestions for future research are to further explore the use of a combination of feature selection models that can optimize feature selection results by considering more complex data characteristics, as well as conducting a more in-depth evaluation of the performance of each model in various data conditions to improve the accuracy and generalization of prediction results. In addition, further research can focus on testing models with more sophisticated or hybrid feature selection techniques that combine the advantages of each approach, such as the XGBoost, LIME, SHAP, or Elastic Net models, which can handle large and complex data more efficiently. The findings of this study are expected to be directly applicable in corporate recruitment processes, particularly as part of a decision support system based on historical data. By leveraging the key features identified, companies can design a selection process that is more accurate in evaluating candidate eligibility, more equitable by reducing subjectivity, and more consistent across different evaluation stages.

The main contribution of this study lies in the application of a comparative approach to various interpretable feature selection methods. This approach not only identifies which features have the most significant influence on recruitment outcomes but also explains why those features are important in the context of decision-making. Consequently, the proposed method can assist organizations in developing recruitment mechanisms that are not only technically efficient but also uphold ethical standards, fairness, and equity in workforce selection.

REFERENCES

[1] G. D. Byrd and X. C. Simcock, "Human Resources Management: From Recruitment to Retention to Pitfalls," *Hand Clinics*, vol. 40, no. 4, pp. 467-476, 2024, doi: 10.1016/j.hcl.2024.06.002

[2] A. Mohammad, "A review of recruitment and selection process," *Journal of Business and Management*, vol. 22, no. 5, pp. 28-34, 2020, doi: 10.9790/487X-2205012834.

[3] L. T. V. Ha, P. N. Linh, D. D. Thanh, T.-H. Nguyen, D. V. Nguyen, L.-A. T. Nguyen, and P.-H. Nguyen, "The impact of corporate vision, customer orientation, and core values with experience as a moderator – insights from Vietnamese enterprises," *Journal of Open Innovation: Technology, Market, and Complexity.*, vol. 11, no. 1, p. 100460, 2025, doi: 10.1016/j.joitmc.2024.100460.

[4] D. M. Truxillo, T. N. Bauer, and B. Erdogan, "Selection and recruitment: An organizational perspective," *Organizational Psychology Review*, vol. 12, no. 3, pp. 199-225, 2022, doi: 10.1177/20413866211005417.

[5] I. Farida and D. Setiawan, "Business strategies and competitive advantage: The role of performance and innovation," *Journal of Open Innovation: Technology, Market, and Complexity.*, vol. 8, no. 3, p. 163, 2022, doi: 10.3390/joitmc8030163.

[6] M. Rožman, P. Tominc, and T. Štrukelj, "Competitiveness through development of strategic talent management and agile management ecosystems," *Global Journal of Flexible Systems Management*, vol. 24, no. 3, pp. 373–393, Jun. 2023, doi: 10.1007/s40171-023-00344-1.

[7] F. L. Schmidt and J. E. Hunter, "The impact of selection methods on organizational success: A meta-analytic review," *Journal of Applied Psychology*, vol. 106, no. 3, pp. 450-463, 2021. doi: 10.1037/apl0000846.

[8] D. Sam, M. Ganesan, S. Ilavarasan and T. J. Victor, "Hiring and Recruitment Process Using Machine Learning," *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, Chennai, India, 2023, pp. 1-4, doi: 10.1109/ICECONF57129.2023.10084133.

[9] D. N.H.A.S, W. E.J.K.D, C. S.M.A, W. K.W.M, S. Thelijjagoda and N. Giguruwa, "AI Bot to Increase the Accuracy and Efficiency of Hiring Process of Business Organizations," *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, Chennai, India, 2024, pp. 1-6, doi: 10.1109/ICSES63760.2024.10910737.

[10] B. Meraliyev, B. Alibekova and I. Bekturganova, "Machine Learning as an effective tool for Human Resource management in recruiting process in the higher educational field," *2023 17th International Conference on Electronics Computer and Computation (ICECCO)*, Kaskelen, Kazakhstan, 2023, pp. 1-5, doi: 10.1109/ICECCO58239.2023.10147133.

[11] P. B, S. Fahimuddin, A. H. S, H. B, K. P and L. A. A, "Advanced Recruitment Strategies in Business Intelligence Systems: A Comparative Study of Machine Learning Models," *2025 8th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2025, pp. 1628-1631, doi: 10.1109/ICOEI65986.2025.11013060.

[12] R. Dugyala, V. K. Gaddam, H. Eroju, M. V. Dantuluri and M. Ch, "Smart Recruitment System," *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10725202.

[13] V. S. Pendyala, N. Atrey, T. Aggarwal and S. Goyal, "Enhanced Algorithmic Job Matching based on a Comprehensive Candidate Profile using NLP and Machine Learning," *2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService)*, Newark, CA, USA, 2022, pp. 183-184, doi: 10.1109/BigDataService55688.2022.00040.

[14] N. D. Dogiparthy, R. D and V. S. K. Devi, "Optimizing Hiring Practices: A Machine Learning Approach for Candidate Selection," *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, Bengaluru, India, 2025, pp. 1498-1504, doi: 10.1109/IDCIOT64235.2025.10914989.

[15] J. Brito, J. Ferro, D. Costa, E. Costa, R. Lopes and J. Fechine, "A ranking between attributes selection models using data from NCAA Basketball players to determine their tendency to reach the NBA," *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, Aveiro, Portugal, 2023, pp. 1-6, doi: 10.23919/CISTI58278.2023.10211486.

[16] D. Jagan Mohan Reddy, S. Regella and S. R. Seelam, "Recruitment Prediction using Machine Learning," *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, Patna, India, 2020, pp. 1-4, doi: 10.1109/ICCCS49678.2020.9276955.

[17] R. Khurana, M. Yadav, M. Quttainah, A. P. Srivastava, A. Balodi and P. K. Singh, "Neural Networks in Recruitment: Trends and Future Directions," *2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE)*, Ballari, India, 2023, pp. 1-5, doi: 10.1109/AIKIIE60097.2023.10390017.

[18] P. N. Mwaro, K. Ogada and W. Cheruiyot, "Neural Network Model for Talent Recruitment and Management for Employee Development and Retention," *2021 IEEE AFRICON*, Arusha, Tanzania, United Republic of, 2021, pp. 1-6, doi: 10.1109/AFRICON51333.2021.9571014.

[19] C. Qin *et al.*, "Towards Automatic Job Description Generation With Capability-Aware Neural Networks," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 5, pp. 5341-5355, 1 May 2023, doi: 10.1109/TKDE.2022.3145396.

[20] M. K. Shaw, P. Dey, S. Chowdhury and T. Ghosh, "Job Candidate Eligibility Prediction using Convolutional Neural Network," *2024 4th International Conference on Intelligent Technologies (CONIT)*, Bangalore, India, 2024, pp. 1-6, doi: 10.1109/CONIT61985.2024.10625937.

[21] M. Arboleda, C. Vieira and J. L. Chiu, "Opening the Machine Learning Black Box for Multidisciplinary Students: Scaffolding from GUI to Coding," *2023 IEEE Frontiers in Education Conference (FIE)*, College Station, TX, USA, 2023, pp. 1-5, doi: 10.1109/FIE58773.2023.10343043.

[22] N. Khan, M. Nauman, A. S. Almadhor, N. Akhtar, A. Alghuried and A. Alhudhaif, "Guaranteeing Correctness in Black-Box Machine Learning: A Fusion of Explainable AI and Formal Methods for Healthcare Decision-Making," in *IEEE Access*, vol. 12, pp. 90299-90316, 2024, doi: 10.1109/ACCESS.2024.3420415.

[23] M. R. Karim *et al.*, "Interpreting Black-box Machine Learning Models for High Dimensional

Datasets," *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, Thessaloniki, Greece, 2023, pp. 1-10, doi: 10.1109/DSAA60987.2023.10302562.

[24] S. Bala and K. Arora, "Interpretable Investigation of Feature Relevance and Sparsity of IoT Datasets," *2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, Goathgaun, Nepal, 2025, pp. 374-379, doi: 10.1109/ICMCSI64620.2025.10883378.

[25] T. R. N and R. Gupta, "Feature Selection Techniques and its Importance in Machine Learning: A Survey," *2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS)*, Bhopal, India, 2020, pp. 1-6, doi: 10.1109/SCEECS48394.2020.189.

[26] K. Liu, Y. Fu, L. Wu, X. Li, C. Aggarwal and H. Xiong, "Automated Feature Selection: A Reinforcement Learning Perspective," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 3, pp. 2272-2284, 1 March 2023, doi: 10.1109/TKDE.2021.3115477.

[27] A. Pandit, A. Gupta, M. Bhatia, and S. C. Gupta, "Filter based feature selection anticipation of automobile price prediction in Azure Machine Learning," in 2022 *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, 2022, pp. 256–262. doi: 10.1109/COM-IT-CON54601.2022.9850615.

[28] L. Quan, T. Gong, and K. Jiang, "Denying Evolution Resampling: An Improved Method for Feature Selection on Imbalanced Data," *Electronics*, vol. 12, no. 15, p. 3212, 2023. doi: 10.3390/electronics12153212.

[29] A. M. Dallo and A. J. Humaidi, "Optimizing machine learning models with data-level approximate computing: The role of diverse sampling, precision scaling, quantization, and feature selection strategies," *Results in Engineering*, vol. 24, p. 103451, 2024. doi: 10.1016/j.rineng.2024.103451.

[30] H. M. Farghaly and T. Abd El-Hafeez, "A high-quality feature selection method based on frequent and correlated items for text classification," *Soft Computing*, vol. 27, pp. 11259–11274, 2023. doi: 10.1007/s00500-023-08587-x.

[31] D. Theng and K. K. Bhoyar, "Feature selection techniques for machine learning: A survey of more than two decades of research," *Knowledge and Information Systems*, vol. 66, pp. 1575-1637, 2024. doi: 10.1007/s10115-023-02010-5.

[32] M. H. Khan and J. Jin, "The relationship between ethnocentric behaviour and workforce localisation success: The mediating role of knowledge sharing tendency," *European Research on Management and Business Economics*, vol. 30, no. 2, p. 100245, 2024. doi: 10.1016/j.iedeen.2024.100245.

[33] 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," *Informatics in Medicine Unlocked*, vol. 44, p. 101442, 2024. doi: 10.1016/j.imu.2024.101442.

[34] H. Zhou, Y. Xin, and S. Li, "A diabetes prediction model based on Boruta feature selection and ensemble learning," *BMC Bioinformatics*, vol. 24, p. 224, 2023. doi: 10.1186/s12859-023-05300-5.

[35] E. Mylona, D. I. Zaridis, C. N. Kalantzopoulos, et al., "Optimizing radiomics for prostate cancer diagnosis: Feature selection strategies, machine learning classifiers, and MRI sequences," *Insights Imaging*, vol. 15, p. 265, 2024. doi: 10.1186/s13244-024-01783-9.

[36] J. Li, Y. Liu, H. Gong, and X. Huang, "Stock price series forecasting using multi-scale modeling with Boruta feature selection and adaptive denoising," *Applied Soft Computing*, vol. 154, p. 111365, 2024. doi: 10.1016/j.asoc.2024.111365.

[37] J. Dhar and S. Roy, "Identification and diagnosis of cervical cancer using a hybrid feature selection approach with the Bayesian optimization-based optimized CatBoost classification algorithm," *Journal of Ambient Intelligence and Humanized Computing*, vol. 15, pp. 3459–3477, 2024. doi: 10.1007/s12652-024-04825-8.

[38] Y. Zhou, S. Wang, Y. Xie, J. Zeng, and C. Fernandez, "Remaining useful life prediction and state of health diagnosis of lithium-ion batteries with multiscale health features based on optimized CatBoost algorithm," *Energy*, vol. 300, p. 131575, 2024. doi: 10.1016/j.energy.2024.131575.

[39] X. Huang, W. Liu, Q. Guo, and J. Tan, "Prediction method for the dynamic response of expressway lateritic soil subgrades on the basis of Bayesian optimization CatBoost," *Soil Dynamics and Earthquake Engineering*, vol. 186, p. 108943, 2024. doi: 10.1016/j.soildyn.2024.108943.

[40] Guenther G. Pavankumar, J. Velmurugan and S. Padmakala, "Real-Time Bitcoin Cost Identification to Improve Efficiency Using Lasso Regression in Comparison with Decision Tree," *2024 IEEE Wireless Antenna and Microwave Symposium (WAMS)*, Visakhapatnam, India, 2024, pp. 1-5, doi: 10.1109/WAMS59642.2024.10528033.

[41] X. Li, Z. Zhang, and L. Li, "Combining feature selection and classification using LASSO-based MCO classifier for credit risk evaluation," *Computational Economics*, vol. 64, pp. 2641-2662, 2024. doi: 10.1007/s10614-023-10535-8.

[42] C. Ai, "A method for cancer genomics feature selection based on LASSO-RFE," *Iranian Journal of Science and Technology, Transactions of Science*, vol. 46, pp. 731–738, 2022. doi: 10.1007/s40995-022-01292-8.

[43] C. H. Feng, M. L. Disis, C. Cheng, and L. Zhang, "Multimetric feature selection for analyzing multicategory outcomes of colorectal cancer: random forest and multinomial logistic regression models," *Laboratory Investigation*, vol. 102, no. 3, pp. 236-244, 2022. doi: 10.1038/s41374-021-00662-x.

[44] Z. Khandezamin, M. Naderan, and M. J. Rashti, "Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier," *Journal of Biomedical Informatics*, vol. 111, p. 103591, 2020. doi: 10.1016/j.jbi.2020.103591.

[45] F. Deng, L. Zhao, N. Yu, Y. Lin, and L. Zhang, "Union with recursive feature elimination: A feature selection framework to improve the classification performance of multicategory causes of death in colorectal cancer," *Laboratory Investigation*, vol. 104, no. 3, p. 100320, 2024. doi: 10.1016/j.labinv.2023.100320.

[46] P. K. Chawla, M. S. Nair, D. G. Malkhede, and S. P. Narwaria, “Parkinson’s disease classification using nature inspired feature selection and recursive feature elimination,” *Multimedia Tools and Applications*, vol. 83, pp. 35197–35220, 2024, doi: 10.1007/s11042-023-16804-w.

[47] P. Theerthagiri and S. Devarayapattana Siddalingaiah, "RG-SVM: Recursive Gaussian Support Vector Machine based feature selection algorithm for liver disease classification," *Multimedia Tools and Applications,* vol. 83, pp. 59021-59042, 2024. doi: 10.1007/s11042-023-17825-1.

[48] M. Anand, K. B. Sahay, M. A. Ahmed, D. Sultan, R. R. Chandan, and B. Singh, "Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques," *Theoretical Computer Science*, vol. 943, pp. 203-218, 2023. doi: 10.1016/j.tcs.2022.06.020.

[49] A. E. Abdellah, H. Ouahi, E. M. Cherrat, and A. Haqiq, “Exploring advanced feature selection techniques: An application to dialectal Arabic data,” *International Journal of Information Technology*, vol. 16, pp. 4637–4649, 2024, doi: 10.1007/s41870-024-01974-z.

[50] H. Zhou, Y. Xin, and S. Li, "A diabetes prediction model based on Boruta feature selection and ensemble learning," *BMC Bioinformatics*, vol. 24, Art. no. 224, 2023. doi: 10.1186/s12859-023-05300-5.

[51] M. Al Fatih Abil Fida, T. Ahmad and M. Ntahobari, "Variance Threshold as Early Screening to Boruta Feature Selection for Intrusion Detection System," *2021 13th International Conference on Information & Communication Technology and System (ICTS)*, Surabaya, Indonesia, 2021, pp. 46-50, doi: 10.1109/ICTS52701.2021.9608852.

[52] Y. Wang, R. Wang, J. Wang, and X. Zhang, “A rock mass strength prediction method integrating wave velocity and operational parameters based on the Bayesian optimization CatBoost algorithm,” *KSCE Journal of Civil Engineering*, vol. 27, pp. 3148–3162, 2023, doi: 10.1007/s12205-023-2475-9.

[53] Y. Cai, Y. Yuan, and A. Zhou, "A predictive slope stability early warning model based on CatBoost," *Scientific Reports*, vol. 14, Art. no. 25727, 2024. doi: 10.1038/s41598-024-77058-6.

[54] Y. Zhao, Y. Zhao, H. Liao, S. Pan, and Y. Zheng, "Interpreting LASSO regression model by feature space matching analysis for spatio-temporal correlation-based wind power forecasting,"

Applied Energy, vol. 380, p. 124954, 2024. doi: 10.1016/j.apenergy.2023.124954.

[55] P. Y. Ng, E. Aruchunan, F. Furuoka, S. A. Abdul Karim, J. V. L. Chew, and M. K. M. Ali, "Intelligent LASSO Regression Modelling for Seaweed Drying Analysis," in *Intelligent Systems Modeling and Simulation III*, S. A. Abdul Karim, Ed. Cham, Switzerland: Springer, 2024, vol. 553, pp. 103-122. doi: 10.1007/978-3-031-67317-7_8.

[56] J. Nyholm, A. N. Ghazi, S. N. Ghazi, dan J. Sanmartin Berglund, "Prediction of dementia based on older adults' sleep disturbances using machine learning," *Computers in Biology and Medicine*, vol. 171, hal. 108126, 2024, doi: 10.1016/j.compbiomed.2024.108126.

[57] C.-C. Huang, W.-Y. Kuo, Y.-T. Shen, C.-J. Chen, H.-J. Lin, C.-C. Hsu, C.-F. Liu, and C.-C. Huang, "Artificial intelligence prediction of in-hospital mortality in patients with dementia: A multi-center study," *International Journal of Medical Informatics*, vol. 191, p. 105590, 2024, doi: 10.1016/j.ijmedinf.2024.105590.

[58] D. Roman, S. Saxena, V. Robu, "Machine learning pipeline for battery state-of-health estimation," *Nature Machine Intelligence*., vol. 3, pp. 447–456, 2021, doi: 10.1038/s42256-021-00312-3.

[59] F. Tian, S. Chen, X. Ji, J. Xu, M. Yang, and R. Xiong, "Robust lithium-ion battery state of health estimation based on recursive feature elimination-deep bidirectional long short-term memory model using partial charging data," *International Journal of Electrochemical Science*., vol. 20, no. 1, p. 100891, 2024, doi: 10.1016/j.ijoes.2024.100891.

[60] N. Ahmed, A. K. Saha, M. A. Al Noman, J. R. Jim, M. F. Mridha, and M. M. Kabir, "Deep learning-based natural language processing in human–agent interaction: Applications, advancements, and challenges," *Natural Language Processing Journal*., vol. 9, p. 100112, 2024, doi: 10.1016/j.nlp.2024.100112.

[61] M. Levis, J. Levy, M. DiMambro, V. DuFort, D. J. Ludmer, M. Goldberg, and B. Shiner, "Using natural language processing to evaluate temporal patterns in suicide risk variation among high-risk veterans," *Psychiatry Research*., vol. 339, p. 116097, 2024, doi: 10.1016/j.psychres.2024.116097.

[62] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools and Applications*, vol. 82, no.3, pp. 3713–3744, 2023, doi: 10.1007/s11042-022-13428-4.

[63] Y. Tian and N. Cao, "Case Study on the Application of Information Technology in Physical Education Teaching Based on Independent Sample T test," *2023 3rd International Conference on Information Technology and Contemporary Sports (TCS)*, Guangzhou, China, 2023, pp. 6-10, doi: 10.1109/TCS59553.2023.10455452.

[64] J. Li, "Finite sample t-tests for high-dimensional means," *Journal of Multivariate Analysis*, vol. 196, p. 105183, 2023, doi: 10.1016/j.jmva.2023.105183.

[65] G. Di Leo and F. Sardanelli, "Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach," *European Radiology Experimental.*, vol. 4, p. 18, 2020, doi: 10.1186/s41747-020-0145-y.