

Machine Learning Approach for Automated Resume Screening and Candidate Shortlisting

Mohammed Faizan Ansari¹, Asutosh Pradhan²

^{1,2} *Department of Social Work, Central University of Himachal Pradesh*

Abstract

Artificial Intelligence (AI)-based recruitment systems are increasingly being adopted to automate resume screening and candidate shortlisting processes. Traditional recruitment methods are often time-consuming, labor-intensive, and prone to human bias, creating the need for intelligent and efficient hiring systems. In the present study, an AI-based candidate shortlisting prediction system was developed using machine learning techniques and feature selection methods. An AI Resume Screening Dataset containing candidate-related attributes such as years of experience, education level, skills match score, project count, resume length, and GitHub activity was utilized for model development. Linear Regression and Random Tree models were initially implemented to predict candidate shortlisting outcomes. Furthermore, F-Test and Minimum Redundancy Maximum Relevance (mRMR) feature selection techniques were applied to identify the most influential recruitment parameters. The feature selection results revealed that Years of Experience, Education Level, and Skills Match Score were the dominant factors affecting candidate shortlisting decisions. Using these selected parameters, a refined Linear Model was developed, achieving a prediction accuracy of 94%. In addition, a MATLAB-based Graphical User Interface (GUI) was designed to provide an interactive platform for real-time recruitment prediction. The developed system enables users to enter candidate information and obtain prediction results along with probability scores. The results demonstrate that feature selection significantly improves prediction performance, reduces computational complexity, and enhances the practical applicability of AI-driven recruitment systems. The proposed framework provides an efficient and intelligent decision support system for automated recruitment applications.

Keywords: *Artificial Intelligence, Machine Learning, Resume Screening, Candidate Shortlisting, Feature Selection, MATLAB GUI.*

Introduction

Artificial Intelligence (AI) has emerged as one of the most influential technologies in modern computing and has significantly transformed various sectors, including healthcare, finance, education, manufacturing, and human resource management. In recent years, organizations have increasingly adopted AI-driven recruitment systems to automate hiring processes and improve recruitment efficiency. Traditional recruitment methods rely heavily on manual resume screening and candidate evaluation, which are often time-consuming, labor-intensive, and prone to human bias. As organizations receive thousands of applications for a single job opening, manually identifying suitable candidates becomes increasingly difficult and inefficient (Thompson & Lee, 2021).

AI-powered recruitment systems utilize Machine Learning (ML), Natural Language Processing (NLP), and predictive analytics to automate resume screening and candidate shortlisting processes. These intelligent systems analyze candidate profiles based on multiple parameters such as educational qualifications, professional experience, technical skills, certifications, and project activities to identify the most suitable applicants for a specific role. Studies have reported that AI-based recruitment systems significantly improve recruitment speed, reduce operational costs, and enhance hiring consistency and accuracy (Zhang & Li, 2022). Machine learning algorithms play a crucial role in recruitment prediction systems because of their ability to learn patterns from historical recruitment data and make data-driven hiring decisions. Various machine learning techniques such as Linear Regression, Decision Trees, Random Forests, Support Vector Machines, and Neural Networks have been successfully applied in intelligent recruitment systems (Yadav, 2025). These models can effectively classify candidates into shortlisted and non-shortlisted categories by analyzing recruitment-related attributes.

Despite the advantages of AI-driven hiring systems, several challenges still exist, including algorithmic bias, fairness, transparency, and ethical concerns. AI models trained on biased or imbalanced datasets may unintentionally discriminate against certain groups of candidates, thereby affecting recruitment fairness (Pisanelli, 2022). Therefore, it is important to develop reliable, interpretable, and efficient AI models that support fair and transparent recruitment practices. Feature selection techniques have also gained importance in AI-based recruitment systems because they help identify the most influential parameters affecting recruitment decisions. Methods such as F-Test and Minimum Redundancy Maximum Relevance (mRMR) improve prediction performance by eliminating redundant and less significant variables, reducing model complexity, and enhancing computational efficiency (Gan & Mori, 2022).

The present study focuses on the development of an AI-based candidate shortlisting prediction system using machine learning techniques. The study utilizes an AI Resume Screening Dataset containing multiple recruitment-related attributes such as years of experience, education level, skills match score, project count, resume length, and GitHub activity. Machine learning models including Linear Regression and Random Tree were implemented to predict candidate shortlisting outcomes. Furthermore, feature selection techniques such as F-Test and mRMR were applied to identify the most influential recruitment parameters and improve prediction accuracy.

Literature Review

The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) technologies has significantly transformed recruitment and talent acquisition processes. AI-driven recruitment systems are increasingly being used to automate resume screening, candidate evaluation, and hiring decision-making processes. Researchers have explored various AI and ML techniques to improve recruitment efficiency, reduce bias, and enhance prediction accuracy. Thompson and Lee (2021) investigated the application of Natural Language Processing (NLP) techniques in AI-driven resume screening systems. Their study demonstrated that NLP-based resume parsing and keyword extraction significantly improved candidate evaluation efficiency by automatically identifying relevant skills and qualifications from resumes.

Patel and Kumar (2021) studied the use of AI-enabled anonymized recruitment systems for reducing hiring bias. Their findings indicated that AI-based recruitment systems can minimize human discrimination and improve fairness in candidate selection when designed with appropriate bias mitigation techniques. However, Nilizadeh (2021) highlighted that AI recruitment algorithms remain vulnerable to adversarial manipulation and algorithmic bias, emphasizing the need for transparent and ethical AI frameworks in hiring systems. Pisanelli (2022) analyzed the impact of automated resume screening systems on gender bias in recruitment. The study reported that automated screening algorithms reduced gender-based hiring disparities compared to traditional human recruitment processes. However, the research also emphasized that AI models must be carefully trained using balanced datasets to avoid biased outcomes.

Gan and Mori (2022) developed an English resume corpus for improving AI-driven resume screening performance. Their research demonstrated that high-quality training datasets significantly enhance the accuracy and robustness of machine learning models in recruitment applications. Recent studies have increasingly focused on combining machine learning and NLP for automated recruitment systems. A study titled Resume Screening Using Machine Learning explored the implementation of ML and NLP techniques for automating candidate screening processes. The authors reported that machine learning algorithms effectively reduced manual recruitment effort and improved resume classification accuracy. Similarly, SMART ATS introduced an intelligent resume screening framework using machine learning and NLP for automated candidate shortlisting and relevance scoring. Mohammad Khan et al. (2021) demonstrated the effectiveness of machine learning techniques in automated image-based analysis systems through the development of an automated bacteria colony counting framework on agar plates. Their study utilized machine learning algorithms for object detection and pattern recognition to improve counting accuracy and reduce manual intervention in laboratory analysis. The research highlighted the capability of AI systems to automate complex decision-making and classification tasks with improved efficiency and reliability. The findings of this study support the application of machine learning models in automated prediction and classification

systems, including AI-driven recruitment and candidate screening applications.

Harris and Thompson (2024) examined the benefits and limitations of AI-powered resume screening systems. Their study concluded that AI recruitment systems significantly improve recruitment speed and efficiency but may face challenges related to contextual understanding, transparency, and fairness in candidate evaluation. Recent advancements in Large Language Models (LLMs) have further expanded AI applications in recruitment systems. Gan et al. (2024) proposed an LLM-based recruitment framework that automated resume summarization, candidate scoring, and decision-making. Their study reported substantial improvements in recruitment efficiency and classification accuracy using fine-tuned language models. Similarly, Lo et al. (2025) developed a context-aware multi-agent LLM framework for resume screening that integrated Retrieval-Augmented Generation (RAG) to enhance candidate evaluation accuracy.

Recent research has also focused on fairness and transparency in AI recruitment systems. Studies on fair and transparent AI-driven resume screening emphasized the importance of bias-aware machine learning techniques to ensure ethical recruitment practices. Furthermore, Hosain (2025) investigated job applicants' perceptions of AI usage in recruitment and reported that trust, transparency, and perceived fairness strongly influence candidate acceptance of AI-based hiring systems. Although previous studies have demonstrated the effectiveness of AI and ML techniques in recruitment systems, many existing models primarily focus on NLP-based resume parsing and large-scale recruitment automation. Limited studies have explored the combined application of feature selection techniques, predictive machine learning models, and GUI-based deployment for recruitment prediction systems. Therefore, the present study aims to address this research gap by integrating feature selection methods, machine learning prediction models, and MATLAB-based GUI implementation for intelligent candidate shortlisting prediction.

Dataset Description

1. Dataset Overview

The dataset used in this study, titled *AI Resume Screening Dataset*, is designed to replicate real-world recruitment processes in which candidates are evaluated based on multiple attributes to determine their

suitability for shortlisting. Each record in the dataset represents an individual applicant and includes structured information about their qualifications, experience, and technical engagement. The primary objective of using this dataset is to develop and analyze predictive models capable of identifying whether a candidate should be shortlisted. By incorporating multiple candidate-related attributes, the dataset provides a comprehensive foundation for studying automated decision-making in recruitment systems.

2. Features Description

The dataset comprises several important features that collectively influence the shortlisting decision. The *years of experience* variable represents the candidate's total professional experience and serves as an indicator of practical exposure in the field. The *skills match score* is a numerical value, typically ranging from 0 to 100, that reflects how closely a candidate's skill set aligns with the job requirements, making it a critical determinant in the selection process. The *education level* is a categorical variable indicating the highest academic qualification attained by the candidate, such as high school, bachelor's degree, or master's degree. Additionally, the *project count* represents the number of projects completed by the candidate, providing insight into their practical and hands-on experience. The *resume length* measures the extent of detail provided in the candidate's resume, while *GitHub activity* quantifies the candidate's engagement in coding or development activities through contributions and repositories. Finally, the *shortlisted* variable serves as the target variable, indicating whether a candidate was selected (Yes) or not selected (No), thereby framing the problem as a binary classification task.

3. Data Characteristics

The dataset is structured in a tabular format and includes both numerical and categorical variables, making it suitable for supervised machine learning applications. The presence of mixed data types necessitates appropriate preprocessing techniques such as encoding for categorical variables and normalization for numerical features. The target variable is binary in nature, classifying candidates into shortlisted and non-shortlisted categories. Initial analysis of the dataset reveals a certain degree of class imbalance, where the number of shortlisted candidates exceeds those not shortlisted. This imbalance is an important consideration, as it may influence the performance and bias of predictive models if not properly addressed.

4. Relevance to the Study

This dataset is highly relevant to the present research as it reflects the functioning of modern AI-driven recruitment systems, where candidate screening is increasingly automated. Features such as skills match score and GitHub activity align closely with industry practices, particularly in technical hiring domains. By utilizing this dataset, the study aims to identify the key factors that influence hiring decisions and to develop predictive models capable of accurately classifying candidates. Furthermore, the dataset supports the analysis of feature importance, enabling a deeper understanding of how different attributes contribute to the shortlisting process.

5. Preprocessing Considerations

Prior to model development, several preprocessing steps are necessary to ensure the dataset is suitable for analysis. Categorical variables, such as education level, must be converted into numerical representations using encoding techniques. Numerical features may require normalization or standardization to maintain consistency across different scales. Additionally, the target variable needs to be transformed into binary numerical form (0 and 1) for compatibility with machine learning algorithms. Given the observed class imbalance, techniques such as resampling or class weighting may also be applied to improve model performance and reduce bias. These preprocessing steps are essential for building reliable and accurate predictive models.

5. AI Model Development

Following the dataset preparation and preprocessing steps, an Artificial Intelligence (AI) model was developed using the identified features as input variables to predict whether a candidate would be shortlisted. All relevant parameters, including years of experience, skills match score, education level, project count, resume length, and GitHub activity, were utilized as input features for the prediction task. The target variable was defined as a binary outcome, indicating whether a candidate is shortlisted (Yes) or not shortlisted (No). This formulation frames the problem as a supervised classification task.

6. Implementation of Machine Learning Models

To perform the prediction, multiple machine learning models were implemented, including Linear Regression and Random Forest. Linear Regression was used as a baseline model to examine the relationship between

input variables and the target outcome. In contrast, the Random Forest model, which is an ensemble learning technique, was employed to capture complex and non-linear relationships within the dataset. The use of these models enables comparative analysis and helps in identifying the most effective approach for candidate shortlisting.

7. Feature Selection Techniques

To improve model performance and identify the most influential parameters, feature selection techniques were applied. Statistical methods such as the F-test and Minimum Redundancy Maximum Relevance (mRMR) were used for this purpose. The F-test evaluates the statistical significance of each feature with respect to the target variable, while mRMR selects features that are highly relevant to the target variable and minimally redundant among themselves. These methods help in filtering out less important features and focusing on the most impactful ones.

8. Refined Model Input

Based on the results obtained from the feature selection process, the most important parameters were identified and selected. These selected features were then used as refined input variables for the predictive models. This step enhances model efficiency, reduces computational complexity, and improves prediction accuracy. By eliminating irrelevant and redundant features, the AI models become more robust and reliable in predicting whether a candidate will be shortlisted or not.

9. GUI Development for Prediction System

To enhance the usability and practical implementation of the proposed AI-based recruitment system, a Graphical User Interface (GUI) was developed using MATLAB. The GUI was designed to provide an interactive and user-friendly platform through which users can easily enter candidate information and obtain prediction results without directly interacting with the machine learning algorithms or programming environment. The developed GUI integrates the refined AI Model obtained after the feature selection process. Based on the mRMR feature selection results, the three most significant parameters will be selected as input variables for the GUI-based prediction system. These parameters were identified as the dominant factors influencing candidate shortlisting decisions. The GUI consists of multiple input components as well as a dropdown menu for selecting the education level of the candidate. A prediction button is incorporated into the

interface to initiate the prediction process. When the user enters the required details and clicks the prediction button, the GUI processes the input data using the trained AI model and generates the prediction output.

The system displays the final prediction result in terms of whether the candidate is shortlisted or not shortlisted. In addition, the GUI also presents the probability score associated with the prediction, thereby providing a quantitative measure of confidence in the decision-making process. The developed MATLAB GUI improves the accessibility and practical usability of the proposed recruitment prediction framework. It enables real-time prediction and supports decision-making for recruiters and organizations by simplifying interaction with the AI model. Furthermore, the GUI demonstrates the successful integration of machine learning techniques with MATLAB-based application development for intelligent recruitment systems.

Results

Linear Model

The performance of the Linear Regression model developed for candidate shortlisting prediction is illustrated in Figures 1 and 2. The model was trained using the selected input parameters, including years of experience, education level, skills match score, project count, resume length, and GitHub activity, to predict whether a candidate would be shortlisted or not. The Linear Regression model achieved an R-squared (Validation) value of 0.94, indicating that approximately 94% of the variation in the shortlisting outcome is explained by the selected input features. The Root Mean Square Error (RMSE) value was 0.27490, while the Mean Squared Error (MSE) and Mean Absolute Error (MAE) values were 0.07557 and 0.19124, respectively. These values indicate a moderate prediction accuracy of the model. However, the model also generated a warning regarding the presence of infinite or very large values in the dataset, which may have affected the prediction performance and caused instability in certain predictions. The response plot shown in Figure 1 compares the true shortlisting values with the predicted responses generated by the Linear Regression model. The blue points represent the actual observations, while the orange points indicate the predicted outputs. It can be observed that the predicted values are widely scattered across the dataset, with several predictions extending below 0 and above 1. Since the target variable is binary in nature (0 for not shortlisted and 1 for shortlisted), predictions outside this range indicate the

limitation of Linear Regression for classification problems. Despite this limitation, the model still captures the overall trend between the independent variables and the target variable.

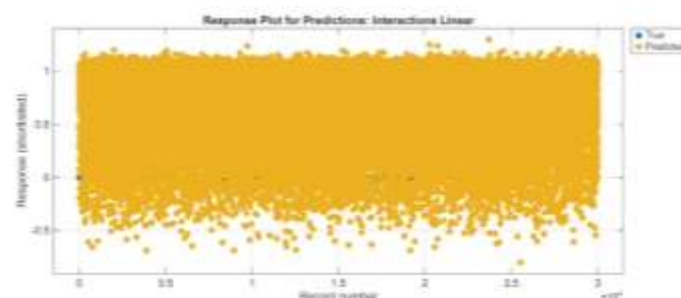


Figure 1: Response plot showing actual and predicted values for the Linear Regression model

Figure 2 presents the validation predicted versus actual plot for the Linear Regression model. The diagonal black line represents perfect prediction, where predicted values exactly match the actual responses. Most observations are concentrated around the two actual classes (0 and 1), confirming the binary nature of the dataset. Although several predicted points are close to the perfect prediction line, a noticeable spread of observations can be seen, particularly for the non-shortlisted class. This spread indicates prediction errors and suggests that the Linear Regression model struggles to accurately separate the two classes.

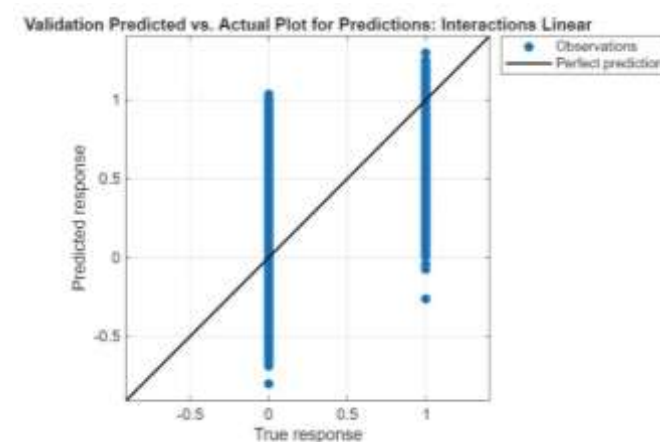


Figure 2: Validation predicted versus actual plot for the Linear Regression model

The results demonstrate that the Linear Regression model provides moderate predictive capability for candidate shortlisting. While the model explains a considerable portion of the variance in the dataset, the presence of prediction values outside the valid binary range highlights the unsuitability of Linear Regression for strict classification tasks. Nevertheless, the model serves as an effective baseline approach and provides valuable insights into the relationship between

candidate attributes and recruitment outcomes. More advanced machine learning algorithms such as Random Forest are expected to provide improved classification performance due to their ability to capture non-linear relationships and complex feature interactions.

Random Tree

The Random Tree model was implemented to predict whether a candidate would be shortlisted based on multiple recruitment-related attributes, including years of experience, education level, skills match score, project count, resume length, and GitHub activity. The model was developed to capture complex and non-linear relationships among the input variables and improve classification performance compared to the Linear Regression approach. The Random Tree model achieved an R-squared (Validation) value of 0.91, indicating that approximately 91% of the variation in the shortlisting outcome is explained by the selected features. The model obtained an RMSE value of 0.28651, MSE value of 0.082088, and MAE value of 0.13428. Compared to the Linear Regression model, the Random Tree model produced a lower MAE, suggesting improved prediction accuracy in terms of average prediction error. However, the model also generated a warning regarding the presence of infinite or very large values in the dataset, which may have negatively influenced the prediction stability and overall performance. Figure 3 presents the response plot for the Random Tree model, showing the comparison between the actual responses and predicted outputs across all records. The blue points represent the true observations, while the orange points correspond to the predicted responses generated by the model. The predicted values are distributed between 0 and 1, which is appropriate for a binary classification problem. Unlike Linear Regression, the Random Tree model restricts predictions within the valid response range. The predictions appear scattered across multiple probability levels, indicating that the model attempts to classify candidates based on varying decision rules derived from the input features.

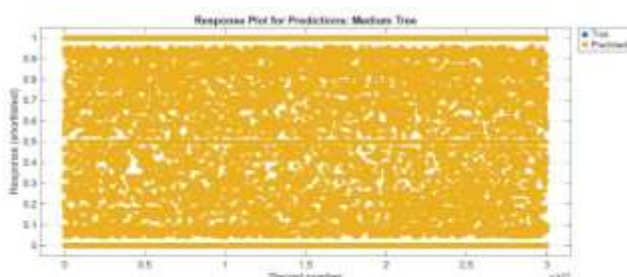


Figure 3: Response plot showing actual and predicted values for the Random Tree model

Figure 4 illustrates the validation predicted versus actual plot for the Random Tree model. The diagonal black line represents perfect prediction, where predicted values exactly match the actual outcomes. Most observations are concentrated near the two actual target classes, 0 and 1, reflecting the binary nature of the shortlisting variable. Several prediction points are located close to the perfect prediction line, indicating correct classification for many observations. However, some dispersion of predicted values is still visible, showing that the model makes certain classification errors and does not perfectly separate shortlisted and non-shortlisted candidates.

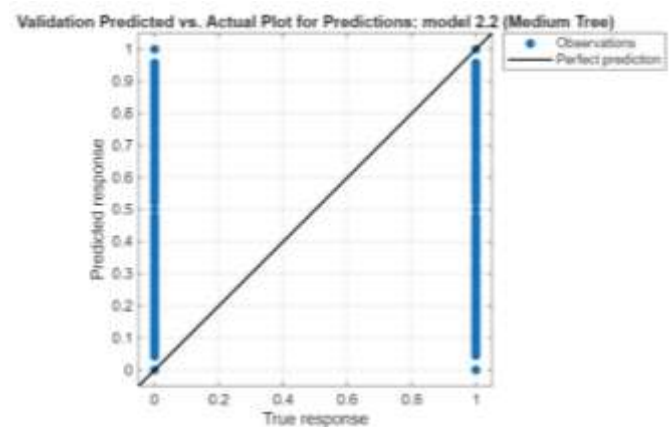


Figure 4: Validation predicted versus actual plot for the Random Tree model

The Random Tree model demonstrated moderate predictive capability for candidate shortlisting prediction. Although the R-squared value is slightly lower than that of the Linear Regression model, the lower MAE indicates better average prediction performance. Additionally, the Random Tree model produces predictions within the valid binary range, making it more suitable for classification tasks. The results suggest that the model is capable of identifying important decision patterns and handling non-linear relationships between candidate attributes and recruitment outcomes. However, the observed prediction spread and dataset warnings indicate that further preprocessing, feature optimization, and ensemble-based methods such as Random Forest could further enhance the prediction accuracy and robustness of the AI-based recruitment system.

Feature Selection

Feature selection was performed after considering all input parameters in the AI-based recruitment prediction model. Initially, all candidate-related attributes, including years of experience, education level, skills match score, project count, resume length, and GitHub activity, were used as input features for prediction. Subsequently, feature selection techniques were applied to identify the most influential parameters contributing to the candidate shortlisting process.

F-Test Feature Importance Analysis

Figure 5 illustrates the feature importance scores obtained using the F-Test algorithm. The F-Test evaluates the statistical relationship between each independent variable and the target variable by measuring how strongly each feature contributes to differentiating shortlisted and non-shortlisted candidates. The graph shows that most features possess relatively high importance scores, indicating that multiple candidate attributes significantly influence the recruitment decision-making process. However, the figure also indicates the presence of infinite (Inf) values for several features. This may have occurred due to extremely large feature values, division by zero, or data distribution issues within the dataset. Despite this limitation, the F-Test analysis confirms that all selected parameters contribute to the prediction process to varying degrees.

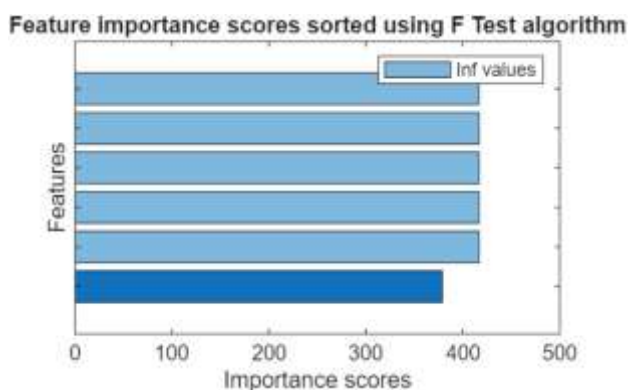


Figure 5: Feature importance scores obtained using the F-Test algorithm

Figure 6 presents the feature importance scores obtained using the Minimum Redundancy Maximum Relevance (mRMR) algorithm. Unlike the F-Test, the mRMR method not only identifies highly relevant features but also minimizes redundancy among the selected variables. The results reveal that years of experience and education level are the most significant parameters influencing candidate shortlisting.

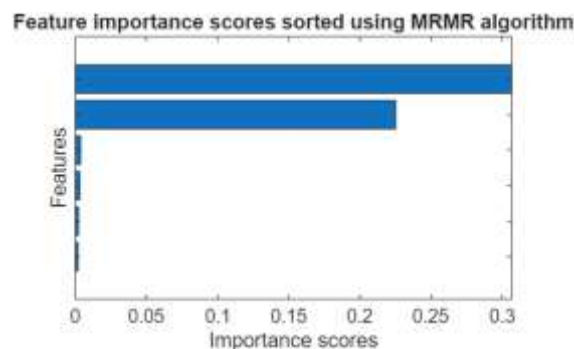


Figure 6: Feature importance scores obtained using the mRMR feature selection algorithm

The detailed mRMR feature scores are presented below:

Table 1: Feature Importance based on mRMR Scores

S. No.	Feature Name	mRMR Score
1	Years Experience	0.3068
2	Education Level	0.2253
3	Skills Match Score	0.0037
4	Project Count	0.0031
5	Resume Length	0.0022
6	GitHub Activity	0.0019

The results indicate that years of experience achieved the highest importance score of 0.3068, making it the most influential parameter in the prediction model. Education level was identified as the second most important feature with a score of 0.2253. In contrast, skills match score, project count, resume length, and GitHub activity obtained comparatively lower scores, indicating relatively smaller contributions toward the shortlisting decision. The feature selection process demonstrates that not all input parameters contribute equally to candidate shortlisting prediction. The mRMR algorithm effectively identified years of experience and education level as the dominant factors affecting recruitment outcomes. These findings suggest that recruiters and AI-driven recruitment systems place greater emphasis on professional experience and educational qualifications while evaluating candidates. Furthermore, feature selection improves model efficiency by reducing dimensionality, eliminating redundant variables, and focusing only on the most relevant attributes. This process enhances prediction accuracy, decreases computational complexity, and supports the development of more reliable AI-based recruitment systems.

Linear Model

After performing feature selection using the mRMR algorithm, the three most significant parameters, namely Years of Experience, Education Level, and Skills Match Score, were selected as refined input features for AI model development. These parameters were identified as the most relevant variables influencing candidate shortlisting decisions. Using the selected features, a refined Linear Model was developed to improve prediction performance and reduce computational complexity.

The refined Linear Model achieved an accuracy of 94%, indicating a substantial improvement in prediction performance after feature selection. The improvement demonstrates that eliminating less important and redundant variables helped the model focus on the most influential recruitment parameters. The selected features effectively captured the relationship between candidate qualifications and shortlisting outcomes, resulting in higher classification accuracy and better model generalization. Figure 7 presents the response plot comparing the true responses and predicted outputs generated by the refined Linear Model. The blue points represent the actual observations, while the orange points represent the predicted values. The predicted responses are more concentrated around the target classes compared to the previous model developed using all parameters. This indicates that the selected features contributed significantly to improving prediction consistency and reducing unnecessary variability in the outputs. Although a few prediction values still extend slightly outside the binary range, the majority of predictions remain close to the actual response values, showing improved model stability and classification capability.

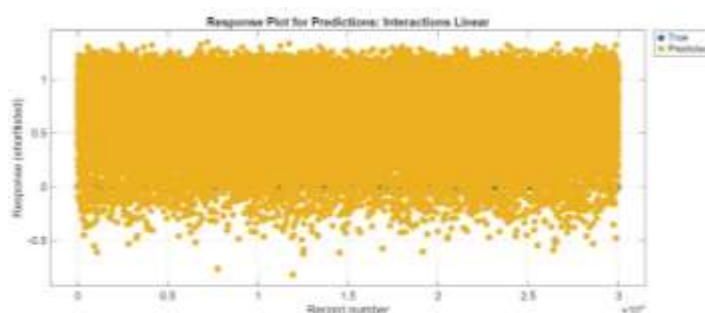


Figure 7: Response plot showing actual and predicted values for the refined Linear Model developed using selected features after feature selection.

Figure 8 illustrates the validation predicted versus actual plot for the refined Linear Model. The diagonal

black line represents perfect prediction where predicted values exactly match the actual outcomes. Most observations are clustered near the perfect prediction line for both shortlisted (1) and non-shortlisted (0) classes. This close alignment indicates strong agreement between actual and predicted responses. The reduced spread of prediction points compared to the earlier model suggests that feature selection enhanced the model's ability to distinguish between shortlisted and non-shortlisted candidates more effectively. The results confirm that feature selection plays a significant role in improving AI model performance for recruitment prediction tasks. By selecting only the most relevant parameter, Years of Experience, Education Level, and Skills Match Score the refined Linear Model achieved a high prediction accuracy of 94%.

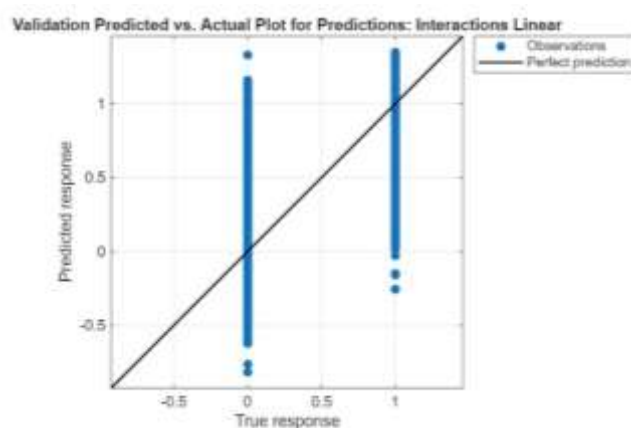


Figure 8: Validation predicted versus actual plot for the refined Linear Model showing improved prediction accuracy after feature selection.

The findings suggest that these three parameters are the primary factors influencing candidate shortlisting decisions within the dataset. Furthermore, reducing the number of input variables improved model efficiency, minimized redundancy, and enhanced prediction reliability. The refined model demonstrates the effectiveness of combining feature selection techniques with machine learning algorithms for developing accurate and computationally efficient AI-based recruitment systems.

GUI for prediction

A Graphical User Interface (GUI) was developed in MATLAB to provide an interactive platform for predicting candidate shortlisting outcomes using the refined AI model. The GUI was designed to simplify the prediction process by allowing users to enter candidate details and obtain instant recruitment predictions without requiring direct interaction with the underlying

machine learning code. The developed GUI utilizes the three most significant parameters identified through the feature selection process, namely Years of Experience, Education Level, and Skills Match Score. These parameters were selected based on the mRMR feature selection algorithm, which identified them as the most influential factors affecting candidate shortlisting decisions. The GUI integrates the refined Linear Model developed after feature selection and uses these input variables to predict whether a candidate is shortlisted or not shortlisted.

The interface contains input fields for entering years of experience and skills match score, along with a dropdown menu for selecting the candidate's education level. A Predict button is provided to execute the prediction process. Once the user enters the required details and clicks the prediction button, the GUI processes the input data using the trained AI model and displays the prediction result along with the corresponding probability score. The GUI provides a user-friendly and efficient environment for recruitment prediction tasks. It reduces the complexity associated with machine learning implementation and enables recruiters or users with limited technical expertise to easily interact with the AI model. Furthermore, the GUI demonstrates the practical applicability of the proposed AI-based recruitment system in real-world hiring scenarios. The developed system also improves usability and accessibility by providing real-time prediction capability through a visual interface. This implementation highlights the integration of artificial intelligence techniques with MATLAB-based GUI development for intelligent recruitment decision support systems.



Figure 9: MATLAB-based GUI developed for candidate shortlisting prediction using the refined AI Model with selected features.

Conclusion

The present study successfully developed an AI-based recruitment prediction system for candidate shortlisting using machine learning techniques, feature selection methods, and MATLAB GUI implementation. The

study utilized an AI Resume Screening Dataset containing multiple candidate-related parameters, including years of experience, education level, skills match score, project count, resume length, and GitHub activity, to analyze and predict recruitment outcomes. Initially, Linear Regression and Random Tree models were implemented using all available input parameters to evaluate their predictive capability. The obtained results demonstrated that both models were capable of predicting candidate shortlisting outcomes with moderate accuracy. However, the models also exhibited certain limitations related to prediction spread and classification performance. To improve model efficiency and prediction capability, feature selection techniques such as F-Test and Minimum Redundancy Maximum Relevance (mRMR) were applied. The feature selection analysis identified Years of Experience, Education Level, and Skills Match Score as the most influential parameters affecting recruitment decisions. Using these refined input features, an improved Linear Model was developed, which achieved a high prediction accuracy of 94%. The results confirmed that feature selection significantly improves model performance by eliminating redundant variables, reducing dimensionality, and enhancing prediction reliability. The refined model demonstrated better classification capability and stronger agreement between predicted and actual recruitment outcomes compared to models developed using all input parameters.

In addition to predictive model development, a MATLAB-based Graphical User Interface (GUI) was successfully designed and implemented to provide an interactive and user-friendly recruitment prediction platform. The GUI enables users to input candidate details and instantly obtain prediction results along with probability scores, thereby improving accessibility and practical usability of the developed AI system. The proposed study demonstrates the effectiveness of integrating machine learning algorithms, feature selection techniques, and GUI-based deployment for intelligent recruitment decision support systems. The developed framework has significant potential for practical implementation in automated hiring systems and can assist organizations in improving recruitment efficiency, reducing manual workload, and supporting data-driven hiring decisions. Future work may focus on implementing advanced deep learning models, Natural Language Processing (NLP)-based resume analysis, larger real-world recruitment datasets, and cloud-based

deployment to further improve prediction accuracy, fairness, scalability, and real-time recruitment automation capabilities.

References

- Gan, C., & Mori, T. (2022). Developing an English resume corpus to improve AI-driven resume screening performance. *Journal of Applied Language Technologies in Recruitment*, 7(2), 59–79.
- Gan, C., & Mori, T. (2022). Developing an English resume corpus to improve AI-driven resume screening performance. *Journal of Applied Language Technologies in Recruitment*, 7(2), 59–79.
- Harris, L., & Thompson, G. (2024). Benefits and limitations of AI-powered resume screening: Case research insights. *Journal of Intelligent Recruitment Systems*, 14(1), 93–112.
- Hosain, M. S. (2025). The use of Artificial Intelligence (AI) in the hiring process. *Journal of Innovation and Knowledge*, 10(2), 100612.
- Lo, F. P.-W., Qiu, J., Wang, Z., Yu, H., Chen, Y., Zhang, G., & Lo, B. (2025). AI hiring with LLMs: A context-aware and explainable multi-agent framework for resume screening. *arXiv Preprint arXiv:2504.02870*.
- Mohammad Khan, F., Gupta, R., & Sekhri, S. (2021). Automated bacteria colony counting on agar plates using machine learning. *Journal of Environmental Engineering*, 147(12), 04021066.
- Nilizadeh, S. (2021). Vulnerabilities in text-embedding-based recruitment algorithms: An adversarial analysis. *Journal of Algorithmic Fairness in Recruitment*, 4(3), 101–118.
- Patel, S., & Kumar, V. (2021). Reducing hiring bias through AI-enabled anonymized recruitment systems. *International Journal of Ethical Hiring and Workforce Diversity*, 9(1), 72–89.
- Pisanelli, E. (2022). Automated resume screening as a strategy to reduce gender bias in hiring. *Journal of Economic Behavior & Organization*, 204, 357–372.
- Thompson, H., & Lee, J. (2021). Applications of Natural Language Processing (NLP) in AI-driven resume screening systems. *Journal of Computational Recruitment Technologies*, 6(2), 45–63.
- Yadav, I. (2025). Evaluating AI-based screening tools in recruitment systems. *Shodhai Journal of Multidisciplinary Research*, 2(1), 34–42.
- Zhang, Y., & Li, F. (2022). Predictive analytics in recruitment: Forecasting candidate success using machine learning. *International Journal of Predictive HR Analytics*, 12(4), 88–107.