Optimizing Recruitment: Harnessing Machine Learning for Predictive Hiring Decisions

Mukhesh Ravi
The Richard DeVos Graduate School of Management
Northwood University
Midland, Michigan, USA
ravim@northwood.edu

Harika Yenuga
The Richard DeVos Graduate School of Management
Northwood University
Midland, Michigan, USA
yenugah80@northwood.edu

Dr Itauma Itauma
The Richard DeVos Graduate School of Management
Northwood University
Midland, Michigan, USA
itaumai@northwood.edu

***Abstract* - This study investigates the application of machine learning models to optimize recruitment processes by predicting hiring decisions based on candidate profiles. The goal is to develop a predictive model that classifies candidates as "hired" or "not hired" using demographic data, qualifications, and recruitment scores. A dataset of 1,500 candidates was used, incorporating features such as age, gender, education level, work experience, and skill scores. Several machine learning algorithms, including Random Forest, Support Vector Machine, Logistic Regression, and ensemble methods (CatBoost, XGBoost), were evaluated to determine the best-performing model and identify the most influential features. Results show that CatBoost outperformed other models, achieving an accuracy of 95%, followed by Random Forest and XGBoost. Analysis of feature importance revealed that Recruitment Strategy, Education Level, and Personality Score were the top three factors influencing hiring decisions. These findings suggest that machine learning can enhance recruitment efficiency, reduce biases, and improve hiring outcomes.**

Keywords— Machine Learning, Predictive Hiring, Recruitment, Hyperparameter Tuning, Classification & Modelling

I. INTRODUCTION

The hiring process is crucial to an organization's success, as it determines the talent aligned with its strategic goals. However, traditional recruitment methods, including manual screening and interviews, are often time-consuming, resource-intensive, and prone to biases. These biases can lead to suboptimal hiring decisions, affecting diversity and organizational growth. As a result, there has been growing interest in leveraging machine learning (ML) models to improve recruitment efficiency, reduce biases, and enhance decision-making [1]. Despite their potential, existing predictive models in hiring face significant challenges, particularly in mitigating fairness issues.

A. Problem Statement

Although ML models have shown promise in improving recruitment accuracy, they are susceptible to issues like dataset imbalance, which can exacerbate biases in hiring predictions. For example, an imbalanced dataset where one gender or educational group predominates can lead to biased predictions favoring the overrepresented group, perpetuating existing disparities [2]. While existing studies have examined ensemble methods like CatBoost and XGBoost, this paper seeks to demonstrate their superiority over simpler models by highlighting their enhanced predictive performance and practical value in optimizing hiring decisions [3] [4] [5]. These models have demonstrated robust performance in handling imbalanced datasets and capturing complex, non-linear relationships, yet their application in recruitment prediction is underexplored [6].

B. *Objective*

This research aims to develop a predictive model to improve the accuracy and fairness of hiring decisions by addressing bias-related challenges in the recruitment process. It evaluates the performance of advanced ensemble models like CatBoost and XGBoost against traditional models such as Random Forest and Logistic Regression. These advanced models are chosen for their ability to handle imbalanced datasets and capture intricate patterns in candidate data [7].

C. *Significance*

This study contributes to the literature by providing a comparative analysis of advanced ensemble models for recruitment prediction. It emphasizes the importance of addressing dataset imbalance and bias mitigation to improve fairness in hiring decisions. By identifying key predictive features such as Recruitment Strategy, Education Level, and Personality Score, this research offers actionable insights for human resource (HR) professionals. Furthermore, it shows the potential of ML models to streamline recruitment, enhance decision accuracy, and reduce bias, thereby improving hiring outcomes and promoting diversity in the workplace.

D. *Research Question*

The central research question guiding this study is: **Which features are most predictive of hiring decisions, and how can machine learning models be leveraged to optimize recruitment strategies while mitigating bias?** To answer this, we examine features such as education, skill scores, and work experience, with the aim of developing models that support automated, efficient, and fair hiring processes.

E. *Literature Review*

The application of machine learning in recruitment has gained traction in recent years. Studies have explored the potential of AI to reduce human biases and improve hiring efficiency. A study that investigated recruitment in the times of machine learning highlights how socio-economic changes and technological advancements can influence HR management, advocating for a balance between humanistic management practices and AI-driven recruitment tools [8]. However, it also cautions against over-reliance on technology, emphasizing the importance of maintaining human values in recruitment processes.

Peña et al. examined the challenges of bias in AI-driven recruitment systems, proposing fairness-aware methods to ensure transparent and unbiased decision-making. Their multimodal framework integrates structured data, image analysis (using ResNet-50), and text data (using BiLSTM models) to address biases and improve the fairness of recruitment algorithms [9]. Their work shows the necessity of fairness-aware ML techniques to mitigate inherent biases in AI recruitment tools.

Our study builds on these foundations by examining the effectiveness of advanced ensemble models like CatBoost and XGBoost. These models are particularly promising for handling imbalanced data and uncovering non-linear relationships between candidate features, which can improve both the accuracy and fairness of hiring predictions. Unlike previous studies, which focus primarily on traditional algorithms, this paper explores the unique advantages of these advanced models in recruitment scenarios, offering new insights into their practical application.

II. METHODOLOGIES

A. *Dataset Summary*

The dataset used in this study consists of 1,500 candidates and 11 features, capturing various candidate attributes including age, gender, education level, work experience, and skill scores. The target variable, **Hiring Decision**, indicates whether a candidate was hired (1) or not hired (0). A summary of the key features in the dataset is provided in **Table 1**.

TABLE 1: SUMMARY OF FEATURES IN THE DATASET

Feature	Description	Data Type
Age	Age of the candidate	Numerical
Gender	0 = Male, 1 = Female	Categorical
Education Level	1 = Bachelor’s (Type 1), 2 = Bachelor’s (Type 2), 3 = Master’s, 4 = PhD	Categorical
Experience Years	Years of professional experience	Numerical
Skill Score	Technical skill score (0-100)	Numerical
Personality Score	Personality fit score (0-100)	Numerical
Recruitment Strategy	1 = Aggressive, 2 = Moderate, 3 = Conservative	Categorical
Hiring Decision	0 = Not Hired, 1 = Hired	Categorical

B. Preprocessing Steps

1. Converting Numerical Variables to Categorical

In the preprocessing phase, some variables with discrete, limited values were recast as categorical variables to enhance model interpretability. For example, Gender, Education Level, and Recruitment Strategy were originally represented numerically (e.g., 0 for male, 1 for female) and were redefined as categorical labels for clearer classification during model training. This conversion helps improve the model's ability to distinguish between categories effectively and allows for better handling by certain machine learning algorithms (e.g., decision trees and ensemble methods).

2. Handling Missing Values, Duplicate Rows, and Feature Scaling

Initial data exploration revealed no missing values or duplicate entries in the dataset, ensuring the integrity of the dataset. Therefore, no imputation or duplicate removal was necessary. Given the lack of missing or redundant data, feature scaling was also not applied as the variables in the dataset were either categorical or already on comparable scales (e.g., skill scores and experience years). Feature scaling may be required for models like Support Vector Machines (SVM) or k-Nearest Neighbors (k-NN) if those models were used.

3. Summary Statistics

Table 2 provides an overview of summary statistics for the key features. The mean age of candidates is approximately 35.15 years with a standard deviation of 9.25 years, indicating a mix of mid-career professionals. Applicants have an average of 7.69 years of experience, reflecting a diverse range of experience levels. Other key statistics highlight the varied skill and personality scores among candidates, which can inform targeted recruitment strategies for optimizing organizational fit.

TABLE 2: SUMMARY STATISTICS OF KEY VARIABLES

Variable	Mean	SD
Age	35.15 years	9.25
Experience Years	7.69 years	4.64
Skill Score	51.12	29.35
Personality Score	49.39	29.35

4. Correlation Matrix

Figure 1 presents the correlation heatmap for the dataset's features. SkillScore, PersonalityScore and InterviewScore have the highest positive correlation with the target feature, HiringDecision, indicating that higher scores in these areas are associated with a higher likelihood of a positive hiring decision.

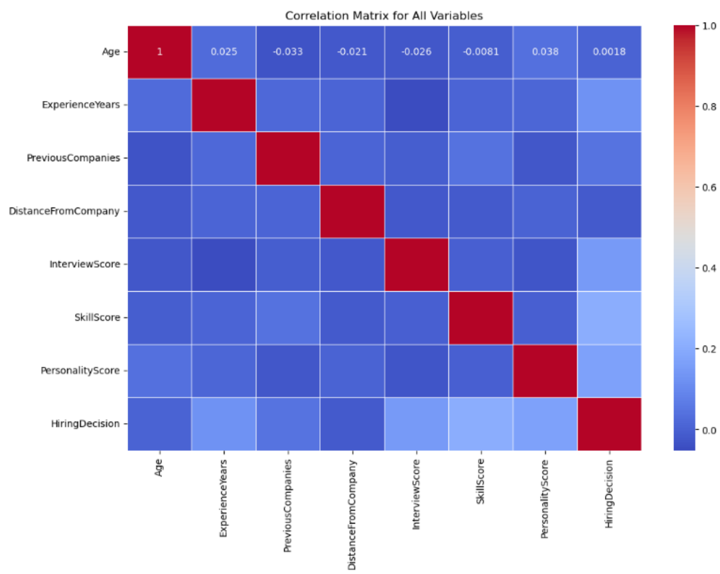

Fig. 1 Correlation Heatmap

5. Gender Distribution in the Dataset

Figure 2 shows a near-equal gender distribution in the dataset, with males comprising 50.8% of the data and females making up 49.2%. This balanced gender representation is important for ensuring fairness in predictive modeling and reducing gender biases in hiring predictions.

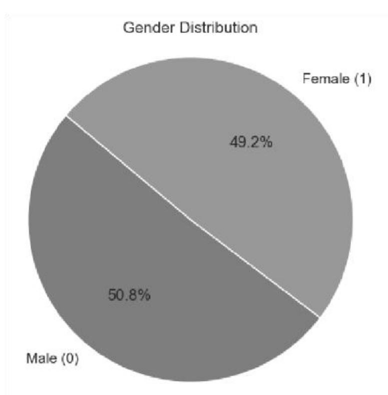

Fig. 2 Gender Distribution

C. *Machine Learning Models*

To ensure a thorough evaluation of the factors influencing hiring decisions, this study employed a range of machine learning algorithms. These models were selected for their ability to handle complex, structured datasets and to mitigate overfitting, as follows:

1) Random Forest (RF): Random Forest was chosen for its robustness in handling high-dimensional data and its ability to mitigate overfitting by averaging multiple decision trees. It is particularly useful in handling complex relationships between features and can provide feature importance rankings [10].

2) Support Vector Machine (SVM): SVM was employed to test the model's ability to classify candidates based on a linear or non-linear decision boundary. SVM is well-suited for datasets where the decision boundary is not easily separable and can perform well with high-dimensional data.

3) Logistic Regression (LR): Logistic Regression was used as a baseline model due to its simplicity and interpretability. As a linear classifier, LR serves as a useful benchmark to compare the performance of more complex models in binary classification tasks.

4) Gradient Boosting Algorithms (XGBoost, LightGBM, CatBoost): These ensemble methods were included due to their strong predictive power and efficiency in handling large datasets. Gradient boosting algorithms iteratively improve model predictions by correcting errors from previous iterations. XGBoost, LightGBM, and CatBoost were specifically chosen for their ability to handle imbalanced data and capture complex non-linear relationships [11][12].

The models were evaluated based on accuracy, precision, recall, F1-score, and AUC-ROC. These metrics help assess model performance from both an overall accuracy and class-specific perspective, which is crucial in the context of imbalanced datasets where class distribution may not be uniform.

III. RESULTS

A. *Model Accuracy Scores*

Figure 3 presents the accuracy scores across all models evaluated in this study. CatBoost outperforms the other models with an accuracy of 95%, closely followed by Random Forest and XGBoost, both of which exhibit accuracy scores around 94%. In contrast, Logistic Regression shows the lowest accuracy, highlighting its limitations for this dataset.

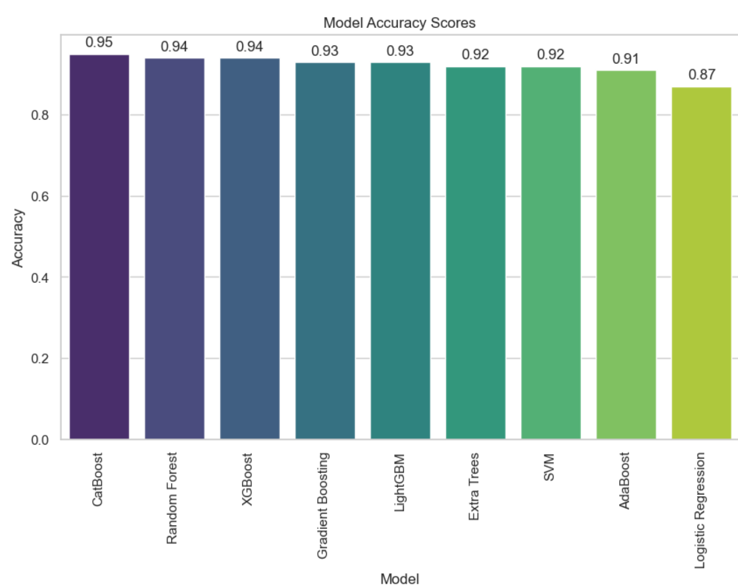

Fig. 3 Accuracy Scores across all models

B. Model Comparison & Evaluation

1) CatBoost: Table 3 shows the classification report for CatBoost. This model achieves impressive results with a high precision of 0.96 for class 0 (Not Hired) and 0.93 for class 1 (Hired). The recall for class 1 (0.91) and class 0 (0.97) suggests that CatBoost performs well in both identifying hires and non-hires, with an overall accuracy of 0.95.

TABLE 3: CLASSIFICATION REPORT FOR CATBOOST

	Precision	Recall	F1 Score	Support
Class 0	0.96	0.97	0.97	215
Class 1	0.93	0.91	0.92	85
Accuracy			0.95	300

2) Random Forest: Table 4 shows the classification report for Random Forest. The model exhibits a high precision of 0.95 for class 0 and 0.94 for class 1, but the recall for class 1 is slightly lower (0.86), suggesting it may miss some hires. Despite this, it maintains an accuracy of 0.94, indicating strong overall performance.

TABLE 4: CLASSIFICATION REPORT FOR RANDOM FOREST

	Precision	Recall	F1 Score	Support
Class 0	0.95	0.98	0.96	215
Class 1	0.94	0.86	0.90	85
Accuracy			0.94	300

3) XGBoost: Table 5 shows the classification report for XGBoost. This model exhibits strong precision and recall for class 0 (0.95 and 0.96, respectively). However, its performance for class 1 (hired) is slightly lower with precision of 0.90 and recall of 0.88, suggesting it may miss more positive instances. Nevertheless, its overall accuracy remains 0.94.

TABLE 5: CLASSIFICATION REPORT FOR XGBOOST

	Precision	Recall	F1 Score	Support
Class 0	0.95	0.96	0.96	215
Class 1	0.90	0.88	0.89	85
Accuracy			0.94	300

C. ROC-AUC Curves

Figure 4 displays the ROC-AUC curves for all models. CatBoost, with the highest ROC-AUC score of 0.9292, demonstrates superior performance in distinguishing between hired and non-hired candidates. Random Forest follows closely, achieving an AUC of 0.9309. XGBoost, while still strong, has a slightly lower ROC-AUC score of 0.9262. The ROC-AUC b curve provides a more nuanced evaluation of model performance, particularly when the dataset is imbalanced.

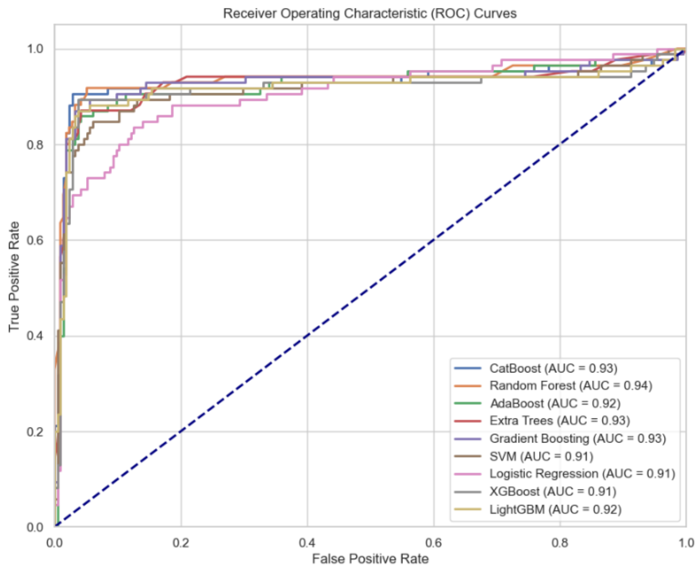

Fig. 4 ROC-AUC Curves for All Models

D. *Hyperparameter Tuning*

Hyperparameter tuning was performed using GridSearchCV and RandomSearchCV to optimize model performance. Table 6 summarizes the results before and after tuning for all four models. CatBoost and XGBoost showed stable performance after tuning, indicating their default parameters were well-suited for this dataset. Random Forest exhibited a slight decrease in accuracy, suggesting limited gains from tuning.

TABLE 6: Model Accuracy Pre & Post Hyperparameter Tuning

Model	Before Tuning	After Tuning (GridSearchCV)	After Tuning (RandomSeachCV)
CatBoost	0.9533	0.9542	0.9542
XGBoost	0.9400	0.9292	0.9292
Random Forest	0.9400	0.9242	0.9225
Logistic Regression	0.8667	0.8657	0.8558

IV. DISCUSSION

A. *Model Accuracy Scores*

CatBoost achieves the highest accuracy of approximately 95%, outperforming Random Forest and XGBoost, which exhibit accuracy scores of 94%. Logistic Regression shows the lowest accuracy, confirming that more complex models perform better for this dataset, particularly when capturing non-linear relationships.

B. *Model Comparison & Evaluation*

CatBoost is the most effective model, balancing precision and recall well, with a strong F1-Score indicating reliable performance for both hires and non-hires. Random Forest has strong precision for class 0 but lower recall for class 1, potentially missing some hires. XGBoost performs well overall but also has lower recall for class 1. While CatBoost is the best based on accuracy and F1-Score, all three models are viable depending on the importance of precision versus recall in hiring.

C. ROC-AUC Curves

The ROC-AUC scores provide a better understanding of model performance beyond accuracy. CatBoost leads with the highest AUC, followed by Random Forest and XGBoost, demonstrating that these models are quite capable of distinguishing between hires and non-hires. The AUC scores further confirm that all three models are well-suited for the task of predicting hiring outcomes.

D. Hyperparameter Tuning

The stability of CatBoost and XGBoost after tuning demonstrates their robustness, making them reliable options for recruitment predictions. In contrast, the sensitivity of Random Forest to parameter changes highlights the need for careful tuning to avoid overfitting.

E. Feature Importance

In all three models, RecruitmentStrategy is identified as the most important feature in predicting hiring outcomes. This is not surprising, as the recruitment strategy likely encompasses critical factors such as the methods and channels through which candidates are sourced. EducationLevel and PersonalityScore are also crucial features, indicating that both academic qualifications and personal traits play an important role in predicting hiring success. Interestingly, Random Forest places more importance on SkillScore and PersonalityScore, suggesting a model preference for assessing personal qualities alongside educational background and experience.

F. Challenges Faced

1) *High Accuracy Scores:* With models like CatBoost, Random Forest, and XGBoost all showing high accuracy, it becomes difficult to choose the best performer. To address this, evaluating additional metrics such as F1-Score, recall, and ROC-AUC is crucial. In scenarios with imbalanced data, precision and recall become particularly important for balancing the trade-offs between false positives and false negatives.

2) *Hyperparameter Tuning:* Hyperparameter tuning is a critical but time-consuming process. While automated techniques like GridSearchCV and RandomSearchCV offer solutions, they can still be computationally expensive and time-consuming for complex models.

G. Real-World Implications

1) *Streamlined Recruitment:* Using machine learning models can significantly reduce time and costs in recruitment processes. By automating tasks such as resume screening, the model helps HR professionals focus on more strategic and value-added activities.

2) *Bias Reduction:* The model's data-driven approach can minimize human biases, leading to more equitable hiring decisions. By evaluating candidates based on objective criteria like experience and skills, the model can promote diversity and inclusion within the workplace.

3) *Enhanced Decision-Making:* Insights from the model can help HR professionals refine their recruitment strategies and identify candidates more likely to succeed in their roles. The ability to make evidence-based decisions can improve hiring outcomes and align the recruitment process with organizational goals.

H. Limitations of this Study

The study's dataset is relatively small (1,500 candidates), which may limit the generalizability of the findings. Furthermore, it relies primarily on demographic and qualification-based features, potentially overlooking other factors like cultural fit or soft skills, which may influence real-world hiring decisions. The black-box nature of machine learning models, particularly in terms of interpretability, could be a challenge for HR professionals who seek to understand and trust model predictions fully [13]. Finally, despite efforts to mitigate bias, the use of historical data means the model could unintentionally perpetuate past biases present in the hiring process.

Conclusion

This study identified CatBoost as the top-performing model, achieving 95% accuracy with balanced precision and recall for both non-hires and hires, making it the most effective at identifying candidates across all categories. Random Forest and XGBoost also demonstrated strong performance but showed a slight bias toward the majority class, indicating areas for improvement in identifying minority class instances. Among the evaluated features, Recruitment Strategy emerged as the most influential predictor of hiring outcomes, followed by Education Level and Personality Score. These findings highlight the importance of strategic recruitment methods, formal education, and personal traits in driving hiring success.

The study also found that hyperparameter tuning had minimal impact on CatBoost's performance, demonstrating its robustness and stability for this dataset, while slight improvements were observed for Random Forest and XGBoost. Overall, the results emphasize the potential of machine learning models, particularly CatBoost, to enhance recruitment processes by automating hiring decisions, reducing bias, and promoting fairness. These insights can help HR professionals make data-driven decisions, streamline recruitment workflows, and implement strategies that foster diversity and improve hiring efficiency.

Areas for Future Research or Improvements

Future research should explore additional machine learning algorithms and ensemble methods, including deep learning, to improve prediction accuracy and handle complex data. Expanding feature engineering to include new candidate attributes, such as social media activity or psychometric assessments, can enhance predictive power. Addressing biases with fairness-aware techniques will ensure equitable treatment and support diversity. Real-world validation through practical recruitment scenarios is crucial to assessing the model's impact on efficiency and diversity. Furthermore, enhancing model interpretability is essential to make machine learning models more transparent and trustworthy for HR professionals. Using larger and more diverse datasets will also improve the generalizability of the findings and ensure the models are robust across different contexts.

THE AUTHORS

Mukhesh Ravi, MSBA, is a dedicated professional currently pursuing a Master of Science in Business Analytics at Northwood University. With a focus on leveraging data-driven insights to solve complex business problems, Mukhesh is well-positioned to make a significant impact in the field of business analytics. Prior to their academic pursuits, he gained valuable experience working at Muthu Pillay Enterprise, where they contributed to various initiatives and honed their skills in business development, operations management, and customer relationship management. His commitment to continuous learning and their passion for data analytics make them a promising candidate for future opportunities in the field.

Harika Yenuga, MSBA, is a committed student working towards a Master of Science in Business Analytics at Northwood University. She is enthusiastic about leveraging data insights to tackle intricate business issues and drive strategic decisions. Before starting her academic journey, Harika acquired valuable experience at Tata Consultancy Services and Infosys, where she worked on various data-driven projects. This allowed her to develop her skills in machine learning, data engineering, automation, statistical analysis, and data visualization. Her solid academic foundation and hands-on industry experience make her a strong candidate for future roles in business analytics.

Itauma Itauma, Ph.D., is an experienced academic and influential thinker, currently serving as the Division Chair and Assistant Professor at Northwood University's DeVos Graduate School of Management. With a strong background in statistics and business analytics, Dr. Itauma brings a wealth of knowledge and experience to his role. Prior to his current position, he held various academic and leadership roles at Northwood University and the Harvard Business Analytics

Program. His commitment to progressive education, research excellence, and strategic thinking are evident in his teaching and research endeavors, making him a valuable asset to the field of education and business analytics.

REFERENCES

[1] Kalusivalingam, A. K., Sharma, A., Patel, N., & Singh, V. (2020). Optimizing Workforce Planning with AI: Leveraging Machine Learning Algorithms and Predictive Analytics for Enhanced Decision-Making. *International Journal of AI and ML*, *1*(3).

[2] Smith, A., Johnson, T., & Lee, M. (2022). Bias and fairness in predictive hiring models: A review of current approaches. *Journal of Artificial Intelligence Research*.

[3] Uppal, A., Awasthi, Y., & Srivastava, A. (2024). Machine learning-based approaches for enhancing human resource management using automated employee performance prediction systems. *International Journal of Organizational Analysis*.

[4] Begum, S. A., Nammalwar, S., Vijayalakshmi, J., & Vallabha, M. (2024, July). Predictive HR: Ensemble & Deep Learning Methods for Strategic Employee Retention. In *2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)* (pp. 1-6). IEEE.

[5] Hossain, M. S., Hossain, M. R., Hasan, S. M., Akter, F., Srizon, A. Y., Faruk, M. F., & Islam, H. (2023, December). Leveraging AI-Driven Strategies to Mitigate Employee Turnover in Commercial Industries. In *2023 26th International Conference on Computer and Information Technology (ICCIT)* (pp. 1-6). IEEE.

[6] Liu, X., Chen, Y., & Zhang, H. (2024). Ensemble learning techniques for predictive hiring: A comparative study of traditional and advanced models. *Proceedings of the IEEE Conference on Machine Learning and Applications*.

[7] Jones, D., & Patel, R. (2023). The role of machine learning in modern recruitment: Challenges and opportunities. *International Journal of Data Science and Analytics*.

[8] Rąb-Kettler, K., & Lehnervp, B. (2019). Recruitment in the times of machine learning. *Management Systems in Production Engineering.*

[9] Peña, A., Serna, I., Morales, A., Fierrez, J., Ortega, A., Herrarte, A., Alcantara, M., & Ortega-Garcia, J. (2023). Human-centric multimodal machine learning: Recent advances and testbed on AI-based recruitment.

[10] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[11] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system.

[12] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232

[13] Suresh, H., & Guttag, J. V. (2021). A framework for understanding unintended consequences of machine learning. Communications of the ACM, 64(11), 62-71