Explainable AI for Employee Attrition Prediction

Galal Mohamed, Ibrahem Ali, Mohamed Hassan, Tasneem Mohammed
Zewail City
{s-galal.qassas, s-ibrahem.ali, s-mohamed-hassan, s-tasneem.ahmed}@zewailcity.edu.eg

***Abstract*—**Employee attrition is one of the biggest problems in HR Analytics. Companies invest a lot in the training of the employees keeping in mind the returns they would provide to the company in the future. Losing trained staff results in significant financial and knowledge losses for organizations. In this paper we used the IBM HR Analytics dataset which consist of 35 features to predict employee attrition using different machine learning models, including tree based models like Adaboost and Extra Tree Classifier (ETC), linear models like SVM and Linear Discriminant Analysis (LDA), Neural network like MLP and DNN, Probabilistic models like naive bayes and distance models like K-NN. Various explainability techniques were applied to ensure better interpretability of the predictions. Specifically, SHAP (SHapley Additive exPlanations) for global interpretation, LIME (Local Interpretable Model-agnostic Explanations) for local interpretations, PDP (Partial Dependence Plots) to visualize global feature effects, and ICE (Individual Conditional Plot) to visualize the relationship between features and model predictions for individual data points. By using these techniques across different machine learning models we found that "StockOptionLevel", "workload_score", "JobSatisifcation" and "JobLevel" are the most influential features of employee attrition prediction. With this knowledge we can develop employee retention techniques that lower employee attrition rate.

I. Introduction

Employee turnover remains a persistent challenge for organizations across industries, with fluctuation increasing each year. Each month, about 1.4% of an organization's employees leave, which equates to a turnover rate of 16.8% per year. This high turnover rate costs US companies $1 trillion annually [1]. Replacing an employee can cost from half to four times their salary, depending on their experience [2]. Besides costs, it affects employers in other ways, such as the loss of experience and knowledge that employees take away with them. According to Mobley [3], there is a strong relationship between job satisfaction and turnover rate. However, they also believe that the employee withdrawal decision process also has a few intermediate linkages. Although AI models offer powerful predictions, a significant challenge in AI-driven decision-making, particularly in sensitive areas like HR, is their "black box" nature, which can hinder interpretability, trust, and actionable insights.

In this study, we address the critical challenge of predicting employee attrition in the IBM HR Employee Dataset [4]. Although previous research has shown the effectiveness of single algorithms, we introduce a reproducible framework that uses an identical end-to-end process—exploratory data analysis, label encodinzg, ADYSAN resampling, and an 80/20 train/test division—for twelve machine learning classifiers from existing literature and applies explainability techniques to investigate the reasons behind each model prediction and identify the most effective features on the predictions.

Our work offers three main contributions. Initially, we create a unified benchmark to evaluate twelve machine learning models under the same preprocessing and evaluation conditions to compare their performance. Secondly, we provide a comprehensive evaluation that uses many evaluation methods to highlight the strengths and weaknesses of each model. Finally, we identify the best model that balances accuracy, and interpretability for HR analytics.

II. Literature Review

A. Motivation

The phenomenon of employees leaving their organization has been the focus of research for many years. Researchers have studied employee attrition topics from different perspectives. Different machine learning approaches have been used to predict employee attrition. In some studies, tree-based models were used because of their ability to handle numerical and categorical features effectively and capture complex, non-linear relationships in the dataset [5]- [9]. Logistic regression and Support Vector Machine (SVM) are also commonly used as they are known for their simplicity and efficiency in handling linear relationships [10], [11].

B. Tree-based Models

Tree models have been very effective in predicting employee attrition from the IBM HR dataset. XGBoost achieved 89.1% accuracy but with a large amount of tuning and enormous computational power [5]. An optimized Extra Trees Classifier (max depth = 300, criterion="gini") achieved the best accuracy of 93% [6]. Adaboost has also shown promise, with studies reporting 88% accuracy [7], achieved using 1000 estimators and a learning rate of 0.1. Another study used Random forest technique with n_estimators = 500 to achieve an accuracy of 85.12% [8], taking the advantage of many decision trees but with the cost of computations. [9] used the Employee dataset with 4,653 records to compare K-Nearest Neighbors (KNN), Decision Tree, and SVM, finding that the Decision Tree model achieved the best performance with 79% accuracy. Synthetic Minority Over-sampling Technique (SMOTE) was applied in class imbalance by generating synthetic data for the minority class to improve model performance.

C. Linear Models

Studies [10], [12] used linear models such as Logistic Regression and SVM have demonstrated their effectiveness in predicting employee attrition. In [10], 6 classifiers were applied to the IBM dataset. The SVM excelled with an accuracy of 89%. Similarly, [11] used the models: Ridge, Lasso, Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, SVM, and Linear Discriminant Analysis (LDA) on the same data. The LDA achieved the highest accuracy of them with 86.39%. Chen [12] used logistic regression on a dataset similar to IBM's, achieving an overall accuracy of 85.9%. They also were able to get 99.4% accuracy in predicting non-attrition cases.

D. DNN, MLP, NB & KNN

Other techniques for predicting employee attrition were used, each using a different approach to enhance the accuracy. A deep neural network model with an input layer of 53 neurons, a binary cross-entropy loss function, and the Adam optimizer (initial learning rate = 0.01) was used by [13] on the IBM HR dataset. Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN) was applied to the dataset to fix its imbalance and transform it into a balanced version, which increased the prediction accuracy from 91.16% to 94.16%. On the same dataset, a multi-layer perceptron (MLP) implemented in Apache Spark framework was used by [14] and achieved a high accuracy of 91.8%. Boosting techniques were applied to fix class imbalance, and with fine-tuning the model and using ensemble methods, the accuracy increased by 10%. While in a probabilistic approach, a Naïve Bayes classifier (NB) was trained on a proprietary dataset of 514 employee survey responses and achieved a prediction accuracy of 100% with no root mean squared error (RMSE) [15]. [16] used the KNN algorithm with one neighbor for its effectiveness in classifying employee attrition by analyzing proximity-based patterns in historical data.

E. Explainability

Recent XAI advancements have moved beyond simple feature importance, offering model-agnostic techniques like SHAP and LIME for more reliable and trustworthy explanations of complex models. Previous research has used various methods: [11] used information value, which provides a measure of how well a variable X can distinguish between binary responses (e.g., "stay" versus "leave") in some target variable y. The prioritized features, ranked by information value, were Job Role (0.4909), Overtime (0.4001), and Job Level (0.3841).

Techniques like Partial Dependence Plots (PDP) were used to visualize the marginal effect of features like Job Satisfaction, Distance From Home, Monthly Income, and Overtime on attrition. The observation suggests that lower monthly income and job satisfaction are associated with a higher likelihood of attrition. Employees who are required to work overtime and those with longer commuting distances have a higher likelihood of attrition. ELI5 was used to determine global feature importance using permutation, identifying Overtime (weight = 0.0867), Monthly Income (0.0373), and Daily Rate (0.0089) as top factors.

SHAP (Shapley Additive Explanations) is used to explain the output of a machine learning model by assigning each feature an importance value for a particular prediction. The results showed that OverTime, Monthly Income, and Job Satisfaction have the most influence on the model. While SHAP offers detailed local and global insights and PDP visualizes marginal effects, literature often applies these in isolation. A comparative understanding of their specific strengths and limitations within the context of attrition is an area for further development.

F. Datasets

The prediction of employee attrition has been explored using different datasets. Several studies [5] – [8], [10] – [14], [16] have used the IBM HR dataset [18], which contains 35 features for 1470 employees. [15] collected their dataset of 514 survey responses, mainly focusing on factors such as job satisfaction, working hours, and job security to provide more context-specific analysis. [9] used a larger dataset of 4653 records from Kaggle that contains a wide range of employee-related factors like education, work history, and demographics. The IBM and Kaggle datasets are both imbalanced, so techniques like ADASYN and SMOTE were used to fix them and address bias. The survey collected by [15] didn't mention class imbalance, but their reported class distribution appears to be balanced. The differences between the datasets regarding their sizes, feature sets, and balance affect findings and interpretation across the studies.

G. Gap of Knowledge

While previous work explores various models and some XAI techniques, a comprehensive benchmark comparing a diverse set of models under identical preprocessing, coupled with a systematic application and comparison of multiple XAI techniques (SHAP, LIME, PDP, ICE) for both global and local interpretability in attrition prediction, is largely missing. Therefore, we aim to re-implement these models and conduct a comparative analysis, providing a more reliable interpretation of their results and the insights from different XAI methods.

III. Methodology

The methodology flow of our research study is illustrated in Figure 1. The IBM HR employee attrition dataset [18] was used for our research findings. To obtain useful insights, the employee attrition dataset and the factors that cause employee attrition were examined through Employee Exploratory Data Analysis (EEDA). Label encoding was applied during pre-processing, as the dataset contains mostly ordinal categorical features. The ADYSAN data resampling technique was applied to balance the dataset. The dataset was then split with a 80:20 ratio, and the proposed machine learning models were trained on 80% of the dataset and tested on the remaining 20%.

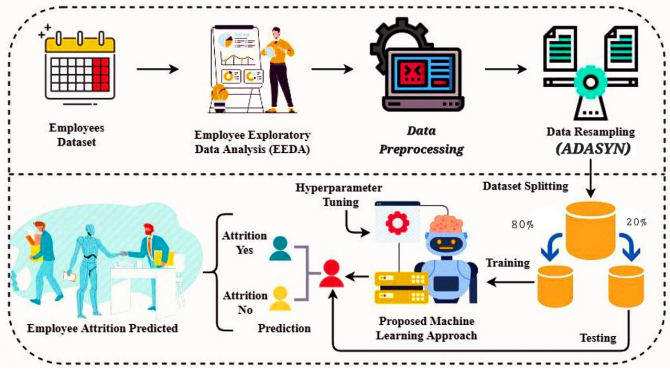

Fig. 1. The methodological analysis of our proposed research study for employee attrition prediction.

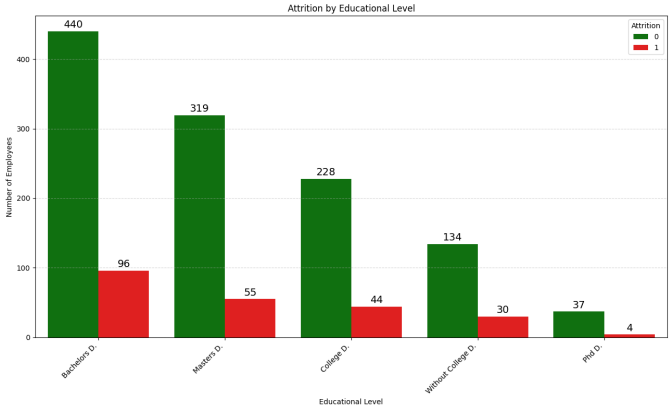

Fig. 2. Employee attrition distribution analysis by educational attainment level.

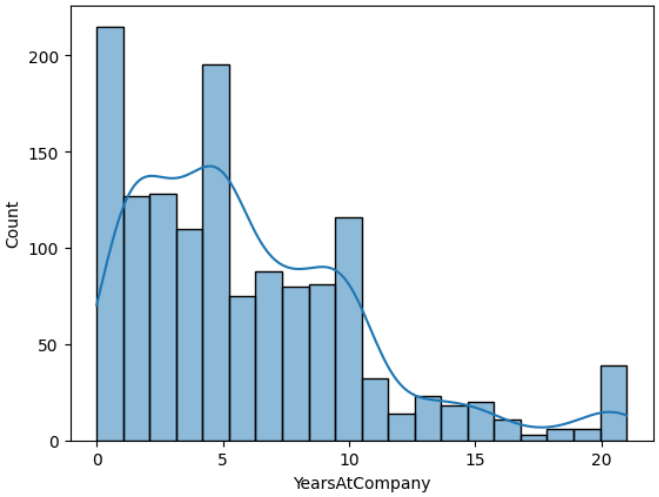

Fig. 3. Employee distribution by tenure at the company.

A. Dataset

The IBM HR Employee Attrition dataset [18] was used to analyze data and develop generalized machine learning models to predict employee attrition. The dataset, created by IBM data scientists, comprises 1,470 employee records and includes 35 features relevant to workforce analytics. It was employed to investigate key factors contributing to employee attrition.

B. Employee Exploratory Data Analysis

The Employee Exploratory Data Analysis (EEDA) was conducted to obtain useful insights from the HR employee attrition dataset. EEDA was used to critically examine the features and factors that are the major causes of employee attrition. We examined the features through various plots and analyses. The EEDA demonstrated data patterns and proved helpful for data factor analysis in the context of employee attrition.

The data distribution plot for the analysis of employee attrition across different educational levels is shown in Figure 2. The educational level categories are mapped on the x-axis, and the number of employees is mapped on the y-axis. The analysis demonstrated that employees with PhD qualifications show the lowest attrition rate, followed by those with a Master's degree. This suggests that higher education correlates with employee retention.

Figure 3 presents an analysis of employee distribution by tenure at the company. The years at the company are mapped on the x-axis, and the number of employees is mapped on the y-axis. The analysis demonstrates that the company is good at hiring new employees, but needs to improve at retaining them, as fewer employees remain as time passes.

The relationship between monthly income and total working years compared to attrition status is examined in Figure 4. Blue points represent non-attrition employees, while red points represent attrition employees. Non-attrition is widely spread across various income ranges and working years, while attrition employees are concentrated at lower income levels and fewer working years, indicating that less experienced employees with lower incomes are more likely to leave the company.

The relationship between Job Satisfaction Level and attrition status is examined in Figure 5. Job satisfaction levels are mapped on the x-axis, and attrition values are mapped on the y-axis. The intensity of the color indicates the number of employees falling into each job satisfaction/attrition area. Employees with higher job satisfaction levels show lower attrition rates, while those with lower satisfaction levels have higher attrition rates.

C. Data Preprocessing

Outliers were removed since they affect employee attrition. To enhance model performance, three additional columns were created.

The first was **promotion_velocity**. It was calculated as:

$$\text{promotion\_velocity} = \frac{\text{YearsAtCompany}}{\text{YearsSinceLastPromotion} + \varepsilon}$$

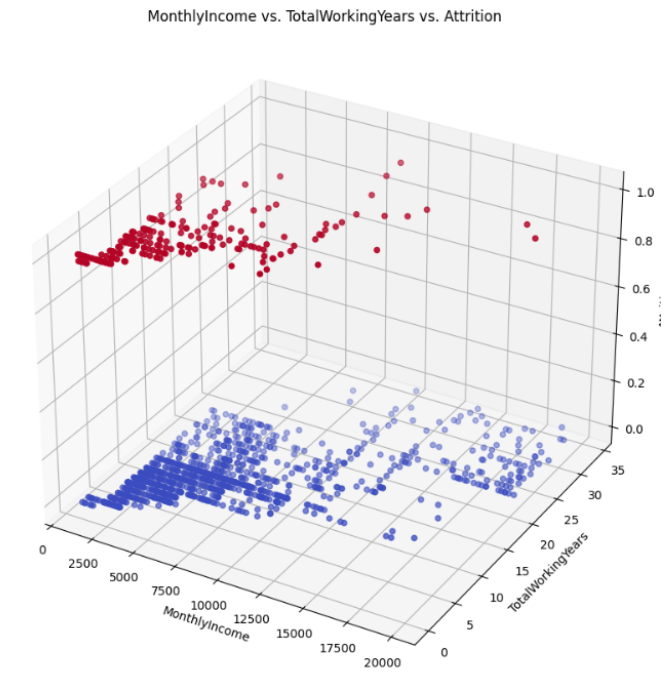

Fig. 4. Relationship between Monthly Income, Total Working Years, and Employee Attrition.

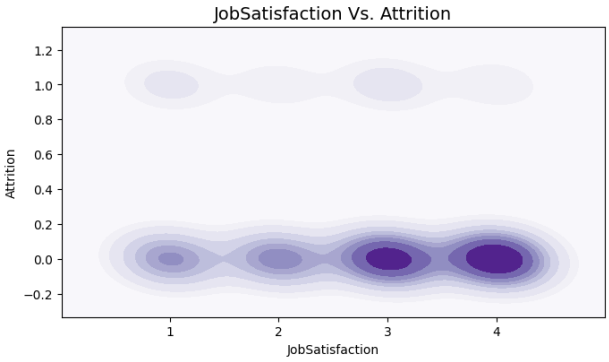

Fig. 5. Employee attrition by Job Satisfaction Level.

where $\varepsilon = 1 \times 10^{-5}$ was added to prevent division by zero.

The second was **workload_score**, intended to quantify job pressure by combining overtime hours and perceived work-life balance. It was computed as:

$$\text{workload\_score} = \text{OverTime} \times (5 - \text{WorkLifeBalance})$$

Finally, the third new column was **compensation_ratio**:

$$\text{compensation\_ratio} = \frac{\text{MonthlyIncome}}{\text{Median(MonthlyIncome by JobRole)}}$$

Categorical columns were processed using label encoding, as most contained ordinal data where the inherent order between categories was meaningful.

D. Data Resampling

ADYSAN (Adaptive Synthetic Sampling) resampling technique was utilized to balance the data. This technique adaptively focuses on minority instances that are harder to learn, thereby improving class distribution and enhancing the model's ability to generalize across both classes. The resampling was necessary to fix the bias in the model's predictions for the majority class.

E. Data Splitting

The generalization capability of the models was enhanced using an 80:20 split ratio. Specifically, 80% of the data was allocated for training, while the remaining 20% was reserved for testing and evaluation

F. Model Implementation and Setup

To evaluate performance across different machine learning models, we implemented twelve models using `scikit-learn` [20] and `TensorFlow` [21] libraries. The models include Decision Tree [10], Random Forest [8], Extra Trees [6], AdaBoost [7], XGBoost [5], K-Nearest Neighbors [16], Naive Bayes [15], Support Vector Machine [10], Logistic Regression [12], Linear Discriminant Analysis [11], Multi-Layer Perceptron [14], and Deep Neural Networks [13]. Each model's structure and setup were based on referenced literature.

G. Explainability Techniques

To provide comprehensive model interpretability, three complementary XAI approaches were implemented across all twelve ML models. Global explanations, local interpretations, and causal relationships. Specifically, we employed eleven techniques Shapely Additive Explanations (SHAP) for feature contribution via game theory, Local Interpretable Model-Agnostic Explanations (LIME) for local surrogate models, Partial Dependence Plots (PDPs) for the average effects of features to interpret the effects of specific instances, Individual Conditional Expectation (ICE), Correlation Coefficient for linear relationships, Accumulated Local Effects (ALE) for non-linear feature impacts, Chi-Square, Permutation Feature Importance (PFI), Information Gain, ANOVA, and Fisher Score.

H. Evaluation Metrics and Procedures

To ensure a fair comparison between implemented models, we apply the same set of quantitative metrics and plots to the held-out test set. Initially, we compute accuracy, precision, recall, and F1-score. Next, we create Receiver Operating Characteristic (ROC) curves and Precision–Recall (PR) curves, together with their corresponding areas under the curve (AUC), to evaluate discriminative effectiveness across thresholds, especially in the presence of class imbalance. We also present learning curves to compare training and validation

performance in relation to sample size or the number of epochs for neural network models. Finally, we create confusion matrices to detail true and false positive/negative rates. All these metrics are produced using the same 20% test split, ensuring consistency among models.

IV. Results

The evaluation outcomes and explainability analyses of the twelve models were applied to the IBM HR Employee Attrition dataset. The results highlight the strengths and limitations of each model in predicting attrition. The evaluation includes standard evaluation techniques detailed in section 4.7, in addition to explainability techniques discussed in section 4.6 to interpret feature contributions and rationale behind decisions. The main visualization insights are presented below.

Our modeling revealed that tree-based models like XGBoost, Decision Tree, and Random Forest overfit the training data, while Naive Bayes and Support Vector Machines (SVM) underfit. In addition, all evaluated models showed bias towards predicting the majority no attrition class, which limits their effectiveness in identifying actual attrition cases.

A. Predictive Performance Analysis

To compare the classification performance of all twelve machine learning models, we evaluated them using accuracy, macro-averaged precision, recall, and F1-score, and ROC-AUC metrics.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
XGBoost	0.83	0.65	0.71	0.67	0.77
AdaBoost	0.81	0.73	0.69	0.71	0.79
KNN	0.86	0.57	0.80	0.59	0.70
DT	0.72	0.58	0.59	0.58	0.59
Random Forest	0.82	0.69	0.61	0.63	0.77
DNN	0.82	0.70	0.68	0.69	0.76
SVM	0.37	0.54	0.55	0.37	0.63
LogisticRegression	0.76	0.65	0.70	0.67	0.76
MLP	0.81	0.66	0.63	0.64	0.48
ETC	0.86	0.78	0.69	0.72	0.77
LDA	0.77	0.64	0.66	0.65	0.75
Naive Bayes	0.42	0.52	0.53	0.41	0.53

TABLE I
Macro-Averaged Performance Comparison of the Models on the Test Set

Table I summarizes the macro-averaged performance of the twelve models on the test set. Overall, ensemble methods such as Extra Trees Classifier (ETC), AdaBoost, and XGBoost achieved the most balanced results across all metrics. ETC achieved the top F1-score (0.72) and macro precision (0.78). ETC and KNN achieved the highest accuracy (0.86). Deep neural networks (DNN and MLP) and Random Forest showed competitive performance, highlighting the effectiveness of ensemble and deep models in handling complex patterns in the data. In contrast, simpler or more assumption-sensitive models such as Decision Tree, Naive Bayes, and SVM underperformed. Naive Bayes and SVM got notably low accuracy (0.42 and 0.37, respectively) and F1-scores below 0.42, possibly due to the strong independence assumption in Naive Bayes and poor margin separation in SVM. KNN achieved the highest accuracy (0.86), but its F1-score remained low (0.59), showing the bias towards the no attrition class. Logistic Regression and LDA performed moderately well with F1-scores of 0.67 and 0.65, respectively, showing stable behavior among linear models. These findings demonstrate that ensemble-based models generally provide robust performance for the employee attrition task. A comparison of model performance across accuracy and ROC-AUC is shown in Figure 6, while the confusion matrices for all models, illustrating their detailed classification outcomes, are presented in Figure 7.

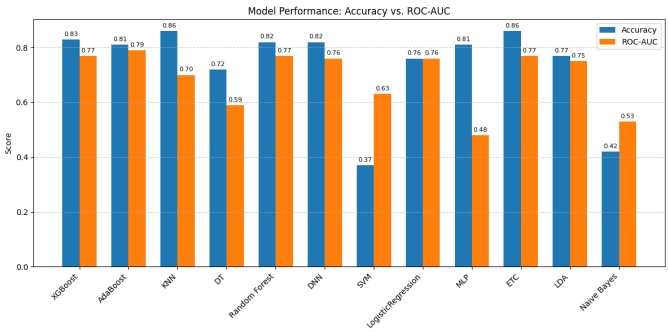

Fig. 6. Model performance comparison across accuracy and ROC-AUC

B. Explainability and Interpretability

XAI techniques used to analyze employee attrition and enhance the decision-making process. By examining the features and identifying which ones increased the attrition rate, we can develop employee retention strategies to address the employee retention problem.

1) Permutation Feature Importances: Permutation feature importance (PFI) is a technique used in machine learning to assess the importance of different features in a predictive model. The basic idea is to measure how much the model's performance deteriorates when the values of a particular feature are randomly shuffled or permuted while keeping other variables unchanged.

The most significant features are determined by consistently appearing as top-three attrition predictors across multiple machine learning models. Workload_score was the most frequently top-ranked feature across the models, followed by StockOptionLevel and JobSatisfaction.

2) Partial Dependence Plot: Partial dependence plots (PDP) analysis measures the marginal effect of a feature on the model's predictions by averaging outcomes while holding all other variables constant. This reveals the relationship between the feature and the predicted outcome, independent of interactions with other variables Fig 9.

Since the line is not flat, "job_satisfaction" is an influential feature. As the "job_satisfaction" increases, the predicted probability of the outcome decreases across all the models, with smoother decline in some models. We can conclude that "job_satisfaction" is a strong negative predictor for employee attrition.

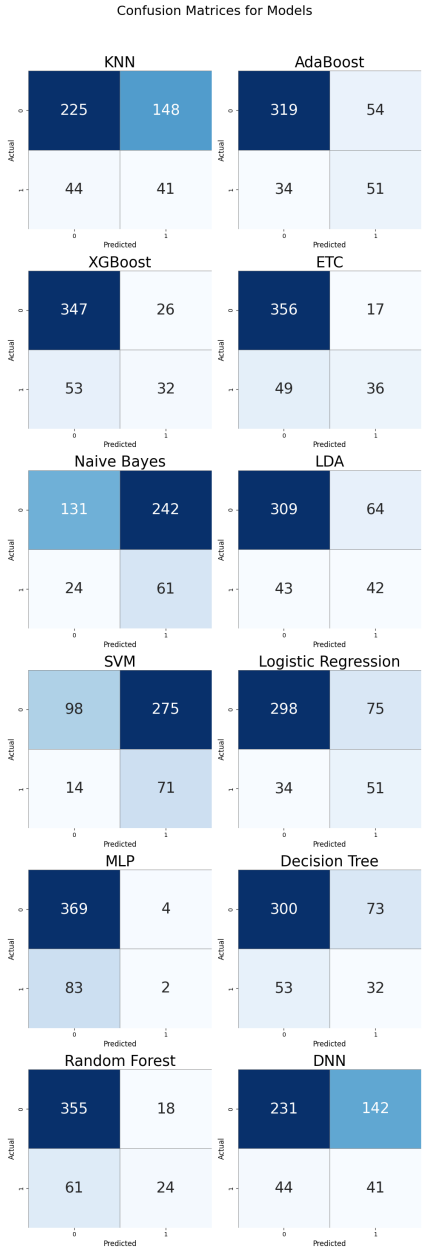

Fig. 7. Confusion Matrices of All Models

3) Individual Conditional Plot: Individual Conditional Expectation (ICE) plots display one line per instance that shows how the instance's prediction changes when a feature changes. An ICE plot visualizes the dependence of the prediction on a feature for each instance separately, resulting in one line per instance of a dataset [22].

In Fig 10, most models showed a consistent downgrade as job satisfaction increases, suggesting a strong negative relationship with employee attrition. Xgboost showed more variability as more instances are derived from the average line. The plot of the Decision Tree shows a sharp step-like pattern because it makes predictions by assigning a constant value to all data points falling into a specific node, resulting in abrupt changes at split points.

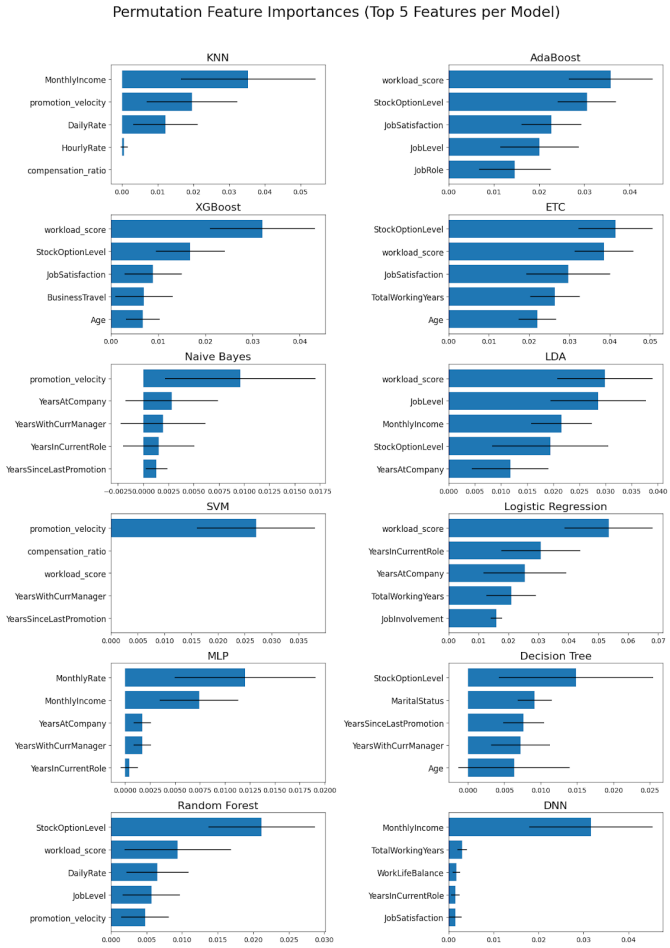

Fig. 8. Permutation Feature Importances (Top 5 Features per Model).

4) LIME: LIME is an algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model. Models from the same class performed similarly, so we will take samples from each class.

For the Linear Model (LDA) in Fig 11, the attrition risk is 37%. Green bars represent features that increase the likelihood of attrition, while red bars show features that decrease the risk. PerformanceRating is the most influential feature in pushing the employee towards leaving the company, while JobLevel reduces the probability of leaving.

For the Tree Model (AdaBoost)in Fig 12, the attrition risk is 58%. StockOptionLevel is the most influential factor that leads the employee to leave the company, while JobLevel is the most important feature to decrease the probability of leaving.

For the Probabilistic Model (Naive Bayes)in Fig 13, the attrition risk is 90%. MonthlyIncome, PromotionVelocity, and Age are the features that push the employee to leave the company, while HourlyRate and PerformanceRating decrease the probability of leaving. The irrelevant interpretation here indicates why Naive Bayes failed to predict employee attrition.

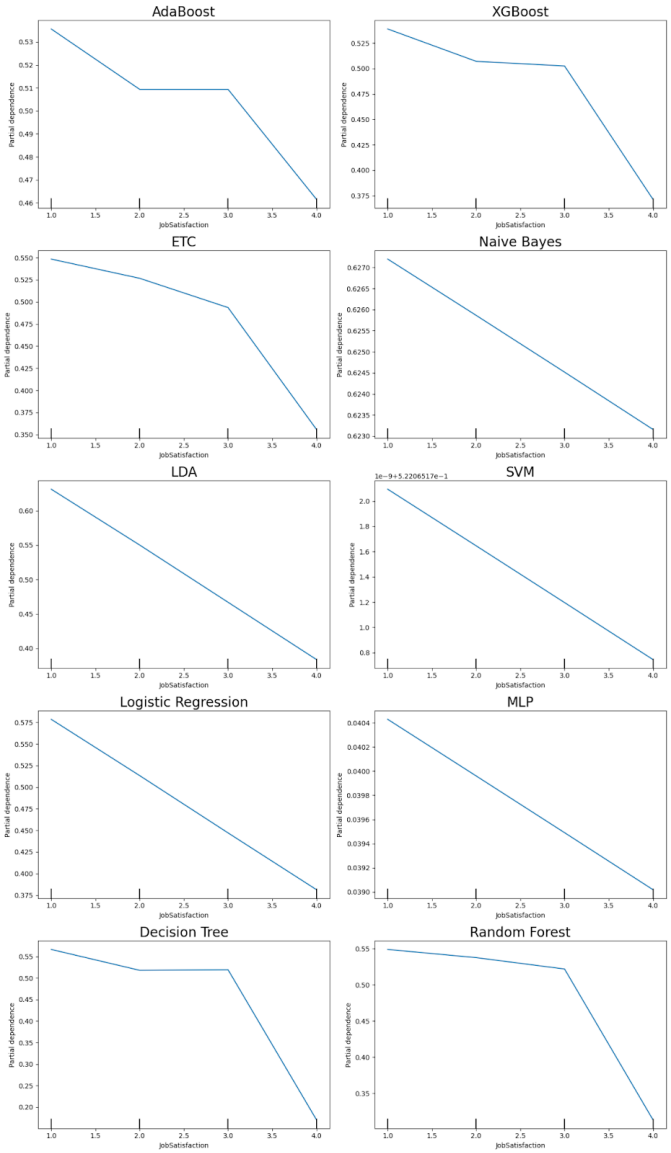

Fig. 9. Partial Dependence Plot for Job Satisfaction.

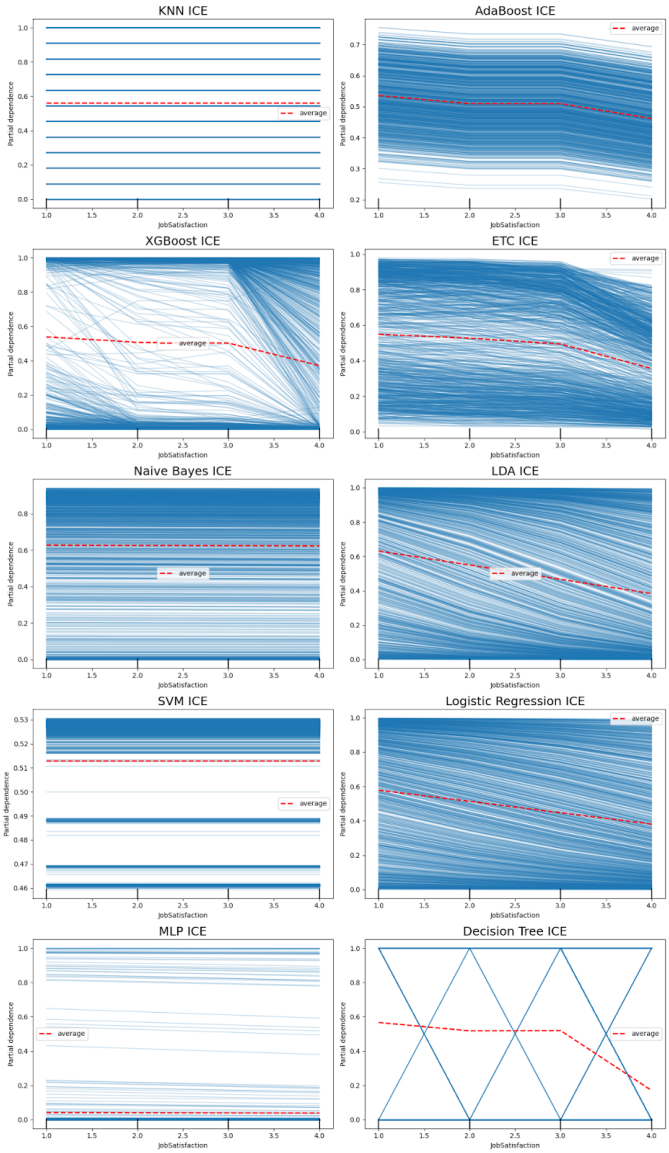

Fig. 10. Individual Conditional Expectation (ICE) Plots for Age across all Models.

5) SHAP: SHAP (SHapley Additive exPlanations) is a game-theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions. Models from the same class performed similarly, so we will take samples from each class. SHAP's general interpretation is as follows: each dot represents a person, and its color shows whether their value for this feature is high (red) or low (blue). The features are ranked by their importance. For this specific task, the dots on the right, whether high or low, increase the attrition value, while those on the left decrease it.

For the Linear Model (LDA) Fig 14, lower JobLevel, JobSatisfaction, JobInvolvement, and YearsInCurrentRole, as well as higher workload_score, monthly income, and YearsAtCompany, increase the attrition rate. Conversely, the opposite trends for these variables decrease the attrition rate.

For the Tree Model (Extra Trees Classifier) in Fig 15, lower StockOptionLevel, JobSatisfaction, JobLevel, and higher workload_score and OverTime increase employee attrition. Conversely, the opposite trends for these variables decrease the attrition rate.

For the Probabilistic Model (Naive Bayes) in Fig 16, lower MonthlyIncome and PromotionVelocity increase employee attrition, while the opposite trends for these variables decrease it. This SHAP graph demonstrates that the Naive Bayes model didn't learn the pattern of the data, as it is a very simple model.

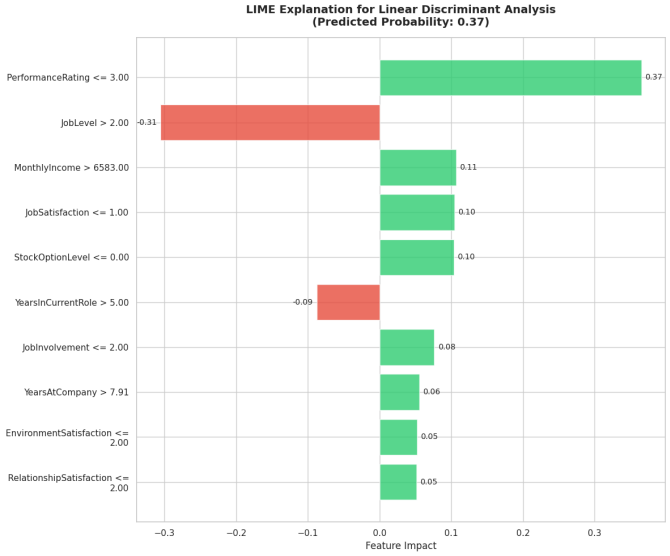

Fig. 11. LIME Explanation for Linear Model LDA

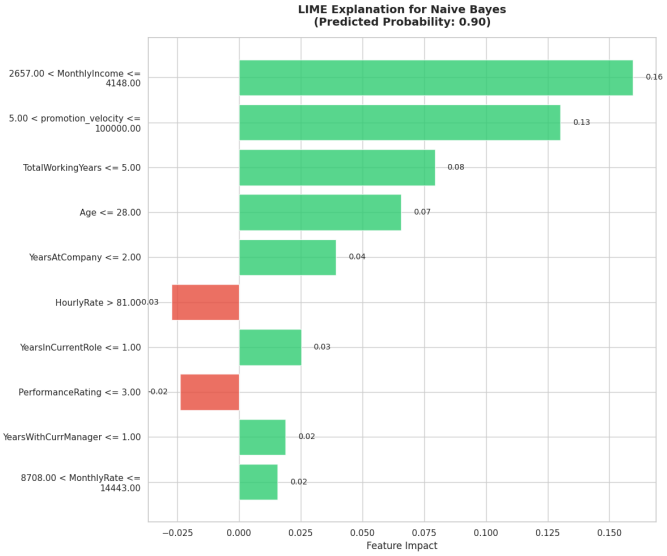

Fig. 13. LIME Explanation for Probabilistic Model Naive Bayes

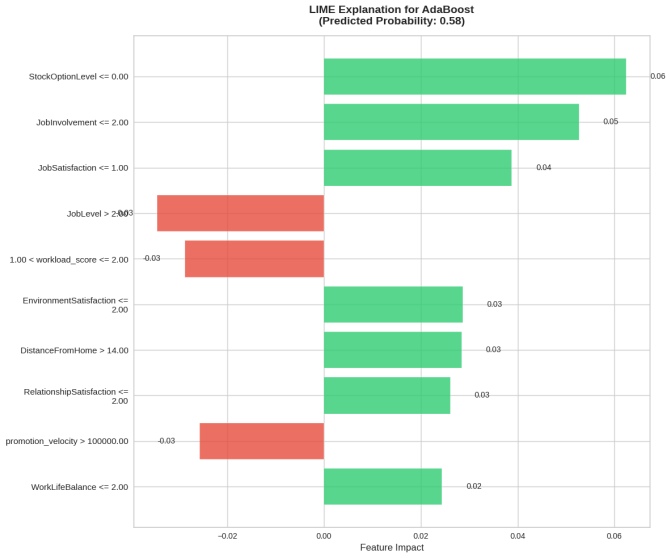

Fig. 12. LIME Explanation for Tree Model AdaBoost

V. Discussion

A. Key Takeaways and Analysis

The results of our experiments showed several insights in the effectiveness of different machine learning approaches for predicting employee attrition and provide interpretable insights into how different features affect the prediction. The evaluation metrics showed that Ensemble-based models such as Extra Trees Classifier, AdaBoost, and XGBoost achieved better performance than other models. This shows that ensemble methods are effective in realizing complex patterns in data. Simpler models such as Decision Tree and Naive Bayes showed poor performance.

Through applying different explainability techniques, it was observed that StockOptionLevel, workload_Score, JobSatisfaction, and JobLevel features have the highest effect on the prediction of the models. These insights help organizations reduce attrition risk and create a suitable work environment that reduces attrition.

B. Challenges and Limitations

We have encountered several challenges during the project, the main problem was the small dataset, which consisted of only 1470 rows. Class imbalance was another main problem, which caused the models to be biased towards the attrition class. The preprocessing was uniform for all the models, which can lead to irrelevant results. More specifically, MLP and DNN would perform better using one-hot encoding instead of label encoding because it introduces an artificial ordinal relationship that likely doesn't exist.

C. Comparison with Prior Work

The unified reimplementation of the twelve models from prior studies [5], [16] enhanced the interpretation of them. Random Forest achieved 82% accuracy, matching [8] 85% closely. In contrast, SVM's 37% accuracy was significantly lower than [10]'s 89%, due to the absence of the feature selection method RFE in [10].

D. Practical Implications

These findings provide important insights for the features that predict employee attrition. With the help of XAI techniques we identified the most valuable features that contributed to employee attrition. Using these insights, we can develop retention strategies to address the employee attrition problem.

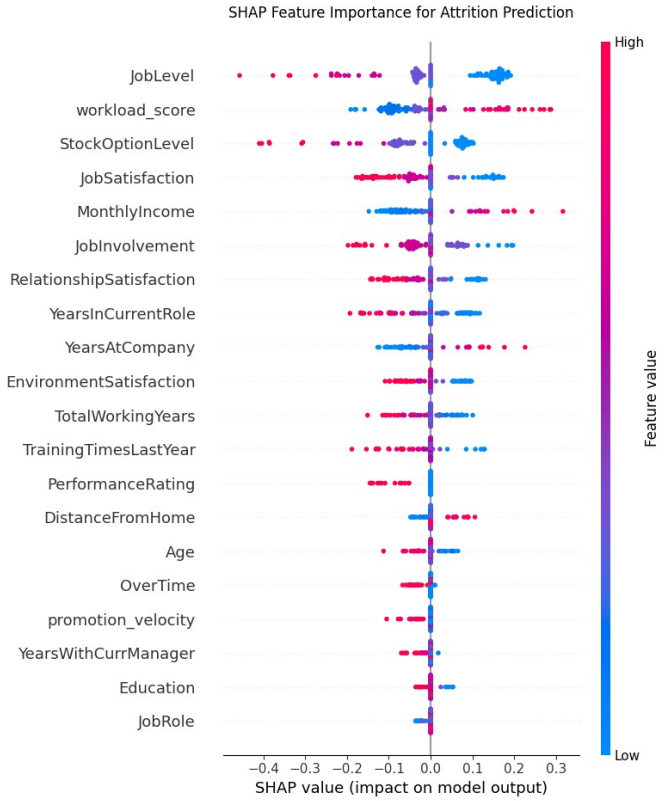

Fig. 14. SHAP Explanation for Linear Model LDA

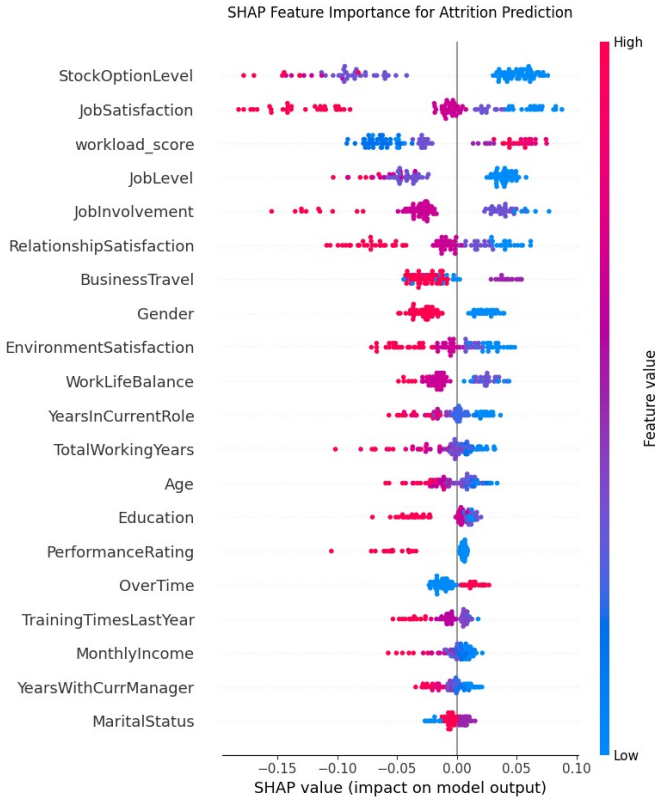

Fig. 15. SHAP Explanation for Tree Model Extra Trees Classifier

VI. Conclusion

In this paper, employee attrition was predicted using the IBM HR Analytics dataset, comprising 35. The focus was to compare 12 different machine learning algorithms with various explainability techniques to uncover the most significant key features that influenced employee attrition. By evaluating these models, AdaBoost performed the best among the other models with ROC-AUC equal to 0.79. This score enables AdaBoost to distinguish between the two classes effectively. The most influential features for this decision were StockOptionLevel, workload_score, JobSatisfaction, and JobLevel.

These results demonstrate how combining explainability techniques with machine learning can yield valuable insights into employee behavior. This helps provide researchers with a baseline for next research on employee attrition prediction, and it is crucial to the industry, especially for HR departments who need to understand why their employees leave or stay.

Looking ahead, exploring larger and more varied datasets could show more patterns in the employee behavior. It would also be valuable to investigate newer explainability methods as they emerge, and perhaps even develop ways to directly test how well our insights translate into real-world retention strategies. It will only make us better at creating supportive workplaces.

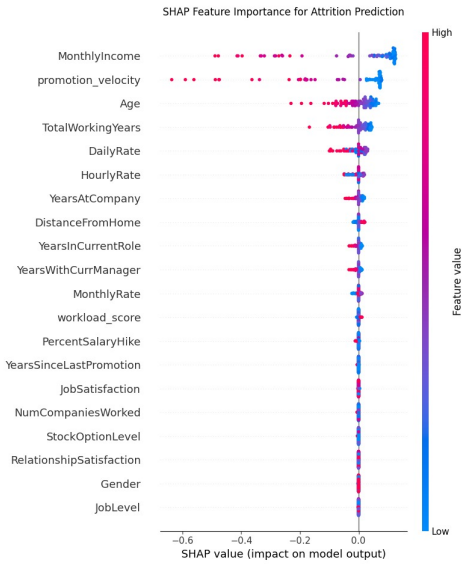

Fig. 16. SHAP Explanation for Probabilistic Model Naive Bayes

References

[1] B. Peng, "Statistical analysis of employee retention," in *Proc. Int. Conf. Statistics, Applied Mathematics, and Computing Science (CSAMCS)*, Apr. 2022, p. 199. [Online]. Available: https://doi.org/10.1117/12.2628107

[2] Bureau of Labor Statistics, "JOLTS home," Nov. 4, 2022. [Online]. Available: https://www.bls.gov/jlt/home.htm

[3] U.S. Bureau of Labor Statistics, "Job openings and labor turnover - January 2025," USDL-25-0331, Mar. 11, 2025. [Online]. Available: https://www.bls.gov/news.release/pdf/jolts.pdf

[4] IBM Developer, "Data science life cycle in action to solve employee attrition problem." [Online]. Available: https://developer.ibm.com/patterns/data-science-life-cycle-in-action-to-solve-employee-attrition-problem/

[5] Business Management Daily Editors, "SHRM survey: Average cost per hire is $4,129," *Business Management Daily*, Jun. 11, 2019. [Online]. Available: https://www.businessmanagementdaily.com/46997/shrm-survey-average-cost-per-hire-is-4129/

[6] R. Jain and A. Nayyar, "Predicting employee attrition using XGBoost machine learning approach," in *Proc. Int. Conf. System Modeling & Advancement in Research Trends (SMART)*, Moradabad, India, 2018, pp. 113-120. [Online]. Available: https://sci-hub.se/downloads/2019-10-22/8e/jain2018.pdf

[7] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting employee attrition using machine learning approaches," *Appl. Sci*, vol. 12, no. 13, p. 6424, Jun. 2022. [Online]. Available: https://doi.org/10.3390/app12136424

[8] M. Karimi and K. S. Viliyani, "Employee turnover analysis using machine learning algorithms," *arXiv Preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2402.03905

[9] M. Pratt, M. Boudhane, and S. Cakula, "Employee attrition estimation using random forest algorithm," *Baltic J. Mod. Comput.*, vol. 9, no. 1, pp. 49–66, 2021. [Online]. Available: https://doi.org/10.22364/bjmc.2021.9.1.04

[10] A. E. Kanuto, "Identifying patterns and predicting employee turnover using machine learning approaches," *Int. J. Sci. Bus.*, vol. 36, no. 1, pp. 20–35, 2024. [Online]. Available: https://ijsab.com/wp-content/uploads/2373.pdf

[11] ResearchGate, "An approach for predicting employee churn by using data mining," Aug. 1, 2022. [Online]. Available: https://www.researchgate.net/publication/320298197_An_Approach_for_Predicting_Employee_Churn_by_Using_Data_Mining

[12] K. Bhuvai and K. Srivastava, "Comparative study of machine learning techniques for predicting employee attrition," *Int. J. Res. Anal. Rev. (IJRAR)*, vol. 5, no. 3, pp. 568–576, Jul. 2018. [Online]. Available: https://www.ijrar.org/papers/IJRAR1903202.pdf

[13] ResearchGate, "Factors of employee attrition: A logistic regression approach," May 1, 2023. [Online]. Available: https://www.researchgate.net/publication/373896134_Factors_of_Employee_Attrition_A_Logistic_Regression_Approach

[14] S. Al-Darraji *et al.*, "Employee attrition prediction using deep neural networks," *Computers*, vol. 10, no. 11, p. 141, Nov. 2021. [Online]. Available: https://doi.org/10.3390/computers10110141

[15] E. Ramalakshmi and S. R. Kamidi, "Prediction of employee attrition and analyzing reasons: Using multi-layer perceptron in Spark," in *Springer eBooks*, 2019, pp. 183–192. [Online]. Available: https://doi.org/10.1007/978-981-13-8461-5_20

[16] I. L. Emmanuel-Okereke and S. O. Anigbogu, "Predicting the perceived employee tendency of leaving an organization using SVM and Naïve Bayes techniques," *OALib*, vol. 09, no. 03, pp. 1–15, Jan. 2022. [Online]. Available: https://doi.org/10.4236/oalib.1108497

[17] K. S. Pokkuluri and S. U. D. N, "Employee attrition prediction using KNN machine learning algorithm," *SSRN Electron. J.*, Jan. 2023. [Online]. Available: https://dx.doi.org/10.2139/ssrn.4452350

[18] G. M. Díaz *et al.*, "Analyzing employee attrition using explainable AI for strategic HR decision-making," *Mathematics*, vol. 11, no. 22, p. 4677, Nov. 2023. [Online]. Available: https://doi.org/10.3390/math11224677

[19] Kaggle, "Employee dataset," Sep. 6, 2023. [Online]. Available: https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset

[20] Scikit-learn: Machine Learning in Python. *scikit-learn 1.6.1 Documentation*. Available: https://scikit-learn.org/stable/

[21] TensorFlow Keras Documentation. *TensorFlow Docs*. Available: https://www.tensorflow.org/api_docs/python/tf/keras

[22] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation," *J. Comput. Graph. Stat.*, vol. 24, no. 1, pp. 44–65, Mar. 2015, doi: 10.1080/10618600.2014.907095.