Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2165~2177
e-ISSN: 2548-6861

Comparing Decision Tree and Optimized LightGBM for Attrition Prediction

Dhea Maharani[1], Farrikh Alzami[2*], MY. Teguh Sulistyono[3], Aris Nurhindarto[4], Dewi Agustini Santoso[5], Muslih[6], Henry Bastian[7]
[1234]Sistem Informasi, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
[567]Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
112202206900@mhs.dinus.ac.id [1], alzami@dsn.dinus.ac.id [2*], teguh.sulistyono@dsn.dinus.ac.id [3], dewi@dsn.dinus.ac.id [5], muslih@dsn.dinus.ac.id [6], henry@dsn.dinus.ac.id [7]

Article Info

Article history:

Received 2026-03-14
Revised 2026-04-22
Accepted 2026-05-05

Keyword:

Employee Attrition,
Feature Importance,
Hyperparameter Tuning,
LightGBM,
Machine Learning

ABSTRACT

Employee turnover poses a considerable challenge for organizations, impacting productivity and raising recruitment expenses. This research seeks to evaluate the effectiveness of Decision Tree and Light Gradient Boosting Machine (LightGBM) models in forecasting employee attrition. The study utilizes a quantitative experimental design, leveraging a secondary dataset sourced from Mendeley. Before model development, data preprocessing was performed, and model evaluation was carried out using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Each algorithm was assessed under three different configurations baseline, regularization, and hyperparameter tuning through GridSearchCV. The experimental findings indicate that the Decision Tree model is prone to overfitting and has limited capabilities in detecting attrition classes, even though optimization raises the ROC-AUC score to 0.80. In comparison, LightGBM demonstrates more reliable and consistent performance. The Tuned LightGBM model achieved the highest performance on the test dataset, with an Accuracy of 0.81, a Precision of 0.82, a *Recall* of 0.71, *F1-Score* of 0.76, and an ROC-AUC of 0.85. An analysis of feature importance reveals that job satisfaction, work-life balance, emotional commitment, work experience, and allowances are the key factors influencing attrition prediction. These results indicate that LightGBM not only performs exceptionally well, but it is also able to offer insights into the critical factors that are important for data-driven retention strategies.

I. INTRODUCTION

Human resources are the most crucial asset for organizational success, where employee retention has a strong positive correlation with the stability of the company's performance. [1]. However, high employee turnover rates have become a persistent global challenge. [2], [3], resulting in financial losses due to re-recruitment costs and loss of institutional knowledge [4]. In addition to financial consequences, high employee turnover also undermines team unity and hinders long-term strategic planning. To address this issue, data-driven approaches using Machine Learning (ML) are increasingly being adopted. [5]. Predictive analytics allows organizations to anticipate which employees may resign before turnover happens. Recent research has employed algorithms such as Random Forest, Support Vector Machine (SVM), and Deep Learning to predict attrition and support managerial decision-making. [1], [4]. These methods exhibit encouraging predictive capabilities and indicate that machine learning can improve evidence-based human resource management.

Despite the rapid growth of ML applications in HR management, a review of the current literature reveals significant fundamental challenges and research gaps. [6]. First, the majority of existing research still relies heavily on synthetic public datasets, such as IBM HR Analytics. [1]

Research by Alshiddy and Aljaber [4] and Govindarajan et al [7]. For example, use this IBM dataset to test ensemble and hybrid models. This use of synthetic data often does not reflect the complexity of real-world behavior, cultural nuances, or specific variables present in real-world labor markets, thus limiting its ecological validity. [7].

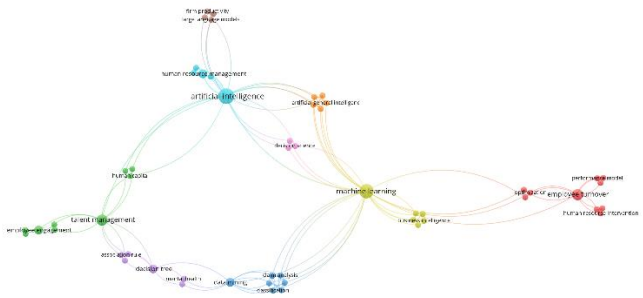

Figure 1. Bibliometric Network Visualization of Employee Turnover Research Trends

Based on Figure 1, it can be seen that the Machine Learning cluster is closely related to Employee Turnover. However, keywords related to 'Optimization' have not become a dominant node compared to other basic algorithms. This is in line with the findings of Fadel et al. [1]. and Govindarajan et al. [7]. This indicates that while many predictive models have been suggested, there is still a lack of thorough investigation into hyperparameter optimization methods aimed at improving model robustness and generalizability. This implies that optimization strategies are not sufficiently examined, even though they play a vital role in enhancing model stability and minimizing overfitting.

Second, many studies focus solely on maximizing accuracy using complex models while neglecting the interpretability aspect. [8]. A study by Fadel et al. [1] used the Stacking Ensemble method to improve accuracy in a financial services company, but did not discuss in depth the transparency of the model's decisions. These models are often "black boxes," where management finds it difficult to understand the specific reasons behind attrition predictions, such as whether salary or job satisfaction is more dominant. [9]. Third, modern boosting algorithms such as the Light Gradient Boosting Machine (LightGBM) have been shown to have high performance, but are very sensitive to hyperparameter settings. [10]. Utilizing default parameters frequently leads to overfitting or less-than-optimal models, making systematic optimization techniques essential. [11].

This study aims to bridge this gap by developing a turnover prediction model that is not only statistically precise but also explainable. This study conducts a comparative study between the Decision Tree algorithm (as a baseline) and LightGBM optimized using GridSearchCV on a primary survey dataset. The main contribution of this study lies in the application of Feature Importance analysis to overcome the black box problem, providing transparency regarding the ranking of the most dominant variables that influence model decisions. Thus, the results are expected to provide priority insights for management in designing targeted retention strategies based on real data.

II. METHOD

This study was carried out in multiple organized phases, beginning with the preprocessing of the dataset and concluding with the evaluation of the model. The cleaned dataset was subsequently split into training and testing sets. A classification model was developed utilizing Decision Tree as the foundational model and LightGBM as the ensemble learning model. Hyperparameter tuning was conducted to attain optimal performance and reduce overfitting. The model's performance was assessed using metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC, along with an analysis of feature importance. The flow of the research methodology is illustrated in Figure 2.

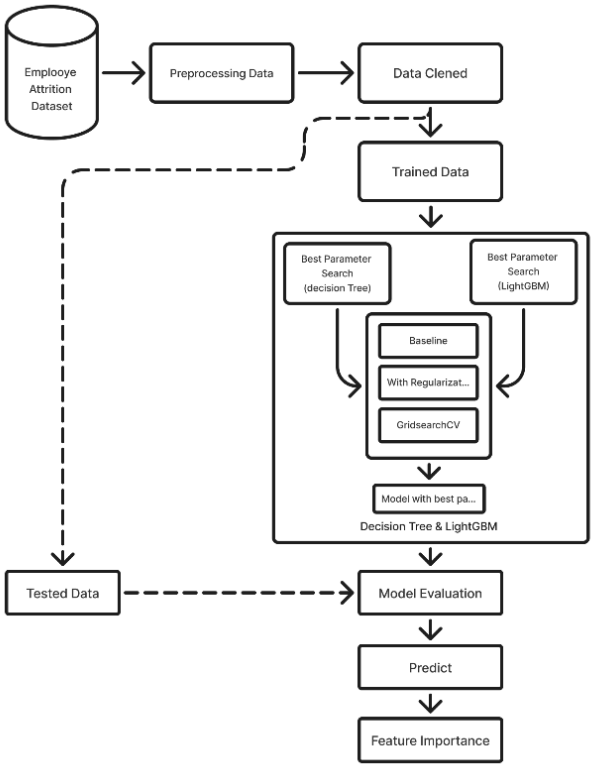

Figure 2. Methodology flow

A. Dataset

The dataset used in this study comes from the Mendeley Data repository via the official link https://data.mendeley.com/datasets/6z2hty8php/1, titled Saudi Employee Attrition Dataset, using the file "Original Dataset of Employee Attrition.xlsx." This dataset consists of 1,191 samples with a total of 34 features representing various aspects related to employee conditions. In contrast to previous studies that mostly used synthetic datasets from IBM HR Analytics [12]. The target variable in this study is attrition,

which indicates whether an employee leaves the company or not. Based on the data distribution, the class proportion shows that 56.8% of employees do not experience attrition, while 43.2% do. This distribution indicates that the dataset has a relatively good class balance for the modeling process [7].

The features in the dataset have diverse characteristics, including categorical and numeric variables. Categorical variables consist of several types, namely binary variables that have two values (yes/no) and are coded as numeric values 0 and 1, and ordinal variables that have certain levels, such as job satisfaction, work-life balance, and emotional commitment, that are coded based on their level order. In addition, there are nominal variables with more than two categories, such as marital status, employment sector, and department, that are processed using the one-hot encoding technique. This distribution indicates that the dataset has a relatively good class balance for the modeling process. [7].

Variables with a large number of categories, such as job title, were not used in the modeling process and were removed from the dataset to avoid excessive feature dimensionality and potential sparsity [4]. This dataset also includes numeric variables, such as allowances, which are used as continuous values. With these characteristics, the dataset used reflects a heterogeneous combination of features, requiring appropriate data preprocessing to ensure optimal use in machine learning modeling.

B. *Preprocessing*

Pre-processing steps are conducted to guarantee the quality and consistency of data before the modeling phase. The steps involved are

1. *Missing Value and Data Duplication Check.* The dataset was analyzed to identify data quality issues. Experimental results showed no missing values in the dataset. However, one duplicate data point was identified and removed to maintain data integrity [4]. This step was performed to ensure the model was built using clean data, thus reducing bias and improving predictive accuracy [3].
2. *Data Cleaning and Standardization.* At this stage, data cleaning and formatting are standardized. Column names are converted to lowercase, leading and trailing spaces are removed, and special characters such as non-breaking spaces are handled to ensure consistency. Furthermore, the data content of categorical features is normalized by converting the text to lowercase and removing unnecessary spaces. Irrelevant attributes, such as IDs, are removed as they contribute nothing to the modeling process [13].
3. *Categorical Data Encoding Process.* After the data cleaning process, categorical variables were transformed into numeric form so they could be processed by machine learning algorithms. Binary variables were encoded as 0 and 1, while ordinal variables such as job satisfaction and work-life balance were encoded based on their level to maintain hierarchical relationships between categories. Meanwhile, nominal variables such as marital status, department, and sector were transformed using one-hot encoding to avoid incorrect order assumptions [13]. This transformation approach ensured that all the encoding processes showed no null values (NaN), so the data was ready for use in the modeling stage. [4].
4. *Feature Separation and Target Variable*. Independent variables are separated from the dependent variable (attrition). Additionally, features with a high number of categories, such as job title, are removed to reduce data complexity and prevent sparsity, which can impact model performance. This stage aims to produce relevant and efficient features for model training. [5].
5. *Data Splitting.* The dataset was divided into training and test data in an 80:20 ratio using stratified sampling. This approach ensures a balanced class distribution across both datasets, resulting in more representative and reliable model evaluation results [7].

C. *Classification Model*

This study compares the performance of two machine learning algorithms

1. *Decision Tree* is used as the baseline model due to its high interpretability. This model builds a tree structure based on selecting the best features to maximize class separation [6]. However, this model is susceptible to overfitting on complex data.
2. *Light Gradient Boosting Machine (LightGBM),* here is a boosting-based ensemble algorithm that builds a tree incrementally using a leaf-wise growth approach [14]. His method enables the model to more effectively capture non-linear patterns and improves computational efficiency through histogram-based learning techniques. Studies by Zhang et al. [15]. and Yamini et al. [16]. show that this architecture significantly speeds up the training process while reducing memory usage without sacrificing accuracy.

D. *Model Optimization Strategy*

Model optimization in this study was carried out through three experimental scenarios, namely baseline, regularization, and hyperparameter tuning, which were applied to the Decision Tree and LightGBM algorithms. In the baseline scenario, each model is trained using the default parameters provided by the library without any adjustments. This stage aims to obtain an initial overview of the model's baseline performance on the dataset used.

Furthermore, in the regularization scenario, parameter adjustments are made to control model complexity and reduce the risk of overfitting. In the Decision Tree model, regularization is performed by limiting the tree depth (max_depth) and setting the minimum number of samples for the splitting process (min_samples_split) and the minimum number of samples at the leaf node (min_samples_leaf). Meanwhile, in the LightGBM model, regularization is performed by adding the reg_alpha and reg_lambda

parameters to penalize model complexity, as well as the subsample parameter to reduce variance by using a portion of the data at each training iteration.

The next stage is hyperparameter tuning, which is performed using the GridSearchCV method with a five-fold cross-validation technique. This process aims to find the best parameter combination that can optimally improve model performance. In this process, the primary evaluation metric used is ROC-AUC. [17]. The parameter search space used in the tuning process is shown in Table I.

TABLE I
HYPERPARAMETER SEARCH SPACE

Model	Parameter	Values
Decision tree	max_dept	[3,5,7,10,none]
Decision tree	min_samples_split	[2,10,20,50]
Decision tree	min_samples_leaf	[1,5,10,20]
LightGBM	n_estimators	[100,200]
LightGBM	learning_rate	[0.01,0.05,0.1]
LightGBM	num_leaves	[15,31]
LightGBM	max_depth	[3,5,7]
LightGBM	min_child_samples	[20,50]
LightGBM	reg_alpha	[0, 0.1, 0.5]
LightGBM	reg_lambda	[0, 0.5, 1]

E. Model Evaluation

Model evaluation is conducted to objectively measure the classification's ability to predict employee attrition. Given that the attrition problem is a binary classification problem, several evaluation metrics are used to provide a comprehensive assessment of model performance. [18].

1. *Confusion Matrix.* The confusion matrix is used as the basis for evaluation by comparing the model's prediction results to the actual labels. This matrix consists of four main components
   - *True Positive (TP): Employees correctly predicted to leave.*
   - *True Negative (TN):* Employees correctly predicted to retain
   - *False Positive (FP):* Retained employees incorrectly predicted to leave.
   - *False Negative (FN): E*mployees leaving are incorrectly predicted to retain.

   In the context of human resource management, false negative errors have a significant impact because companies fail to detect employees who are likely to leave.
2. *Accuracy, here* measures the ratio of total correct predictions (positive and negative) to the entire data sample.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

   While this metric provides a general overview of model performance, its interpretation should be approached with caution, as it does not necessarily reflect the specific positive class detection capability [1] [19].
3. *Precision*, here measures the model's confidence level when predicting the positive class.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

   A high precision value indicates that most employees predicted to be at risk of leaving actually do leave. This metric is important for ensuring the efficiency of a company's retention strategy [4].
4. *Recall*, here measures the model's ability to detect all employees who actually leave.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

   Recall is a crucial metric in this study because failure to detect employees at risk of leaving can lead to companies missing opportunities for early intervention. [1], [4]
5. *F1-Score*, here the harmonic mean between Precision and Recall

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (4)$$

   This metric provides a balance between model precision and sensitivity, making it suitable for use as a primary performance indicator. [4], [9].
6. *ROC-AUC*, here, the Receiver Operating Characteristic (ROC) curve describes the relationship between the True Positive Rate and False Positive Rate at various threshold values. The Area Under the Curve (AUC) value indicates the model's ability to distinguish between positive and negative classes. [3], [4] The closer the value is to 1, the better the model's discrimination ability. In this study, ROC-AUC was used as the primary metric in the hyperparameter optimization process using GridSearchCV.

F. Model Interpretation

Model interpretation is performed to increase the transparency of prediction results, particularly for complex ensemble algorithms such as the Light Gradient Boosting Machine (LightGBM) [14]. Although this model has high predictive performance, its internal structure is difficult to understand directly (black-box model) [20].

To address this issue, this study applies Feature Importance analysis to identify the relative contribution of each variable to the classification process. Feature importance is calculated based on the total gain, which is the accumulated reduction in the loss function generated each time a feature is used to separate nodes in a decision tree. [21].

This method is implemented on a global interpretation scale, enabling the recognition of the key factors that have the greatest impact on the overall prediction of employee

turnover. The feature exhibiting the highest gain value is understood as the variable that plays the most crucial role in differentiating between employees who depart and those who stay.

The outcomes of this analysis are anticipated to offer valuable insights for human resource management in crafting data-driven retention strategies, emphasizing the variables that significantly impact model decisions.

III. Result and Discussion

This section describes the results of the computational experiments, including data characteristic analysis, model performance evaluation before and after optimization, and interpretation of the dominant factors causing turnover

A. Data Dataset Description and Data Exploration

The Employee Attrition dataset used in this study consists of 1,191 sequentially arranged data entries with no missing values, as confirmed by the DataFrame.info() function, which shows 1,191 non-null entries across all columns. This dataset contains a total of 34 columns, encompassing 33 predictor features and one target variable, Attrition. This target variable is an object data type with binary categories ("Yes" and "No"), confirming that this study falls into the supervised binary classification category. The absence of missing values indicates excellent data quality, so the pre-processing stage does not require data imputation techniques and can instead focus directly on data format transformation.

```
RangeIndex: 1191 entries, 0 to 1190
Data columns (total 35 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   ID                                      1191 non-null   int64
 1   Attrition                               1191 non-null   object
 2   Gender                                  1191 non-null   object
 3   Age                                     1191 non-null   object
 4    Maritalstatus                          1191 non-null   object
 5    Academic_degree                        1191 non-null   object
 6   Years_Experience                        1191 non-null   object
 7   Years_experience_lastorganization       1191 non-null   object
 8   Sector                                  1191 non-null   object
 9   Department                              1191 non-null   object
 10  JobTitle                                1191 non-null   object
 11  MonthlySalary                           1191 non-null   object
 12  Allowances                              1191 non-null   int64
 13  MedicalInsurance                        1191 non-null   object
 14  Bonus                                   1191 non-null   object
 15  OverTime                                1191 non-null   object
 16  Payment_Overtime                        1191 non-null   object
 17  Rewards&Wages_Satisfaction              1191 non-null   object
 18  Get_ Deserved_Promotion                 1191 non-null   object
 19  Training_programs_ During_last_three_years  1191 non-null   object
 20  Useful_Training_Programs                1191 non-null   object
 21  Business_Travel                         1191 non-null   object
```

Figure 3. Feature characteristics

The results of Figure 3 can be seen in terms of feature characteristics. This dataset is dominated by categorical data types (objects) with 33 columns, with only one numeric (integer) column, namely allowances. The dominance of categorical features indicates that the dataset captures more qualitative dimensions and employee perceptions, such as job satisfaction, emotional commitment, and work-life balance, than purely quantitative metrics. Therefore, the main challenge in this modeling is not the completeness of the data, but rather the strategy of transforming categorical features into appropriate numeric representations through encoding techniques. This step is crucial so that non-numeric information can be optimally processed by decision tree-based and ensemble machine learning algorithms, such as LightGBM, applied in this study.

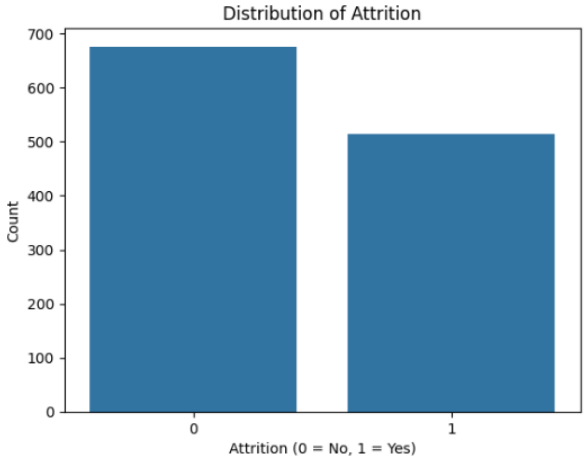

Figure 4. Distribution of target classes in the dataset

Based on Figure 4, the distribution visualization of the Attrition target variable shows a relatively balanced class proportion, with 676 employees retained (0) and 515 employees leaving (1). Proportionally, the retained class represents approximately 56.8% of the total data, while the leaving class represents 43.2%. This relatively balanced distribution condition is advantageous because it minimizes the risk of model bias towards the majority class, so that additional balancing techniques such as oversampling or undersampling are not required. To ensure an objective evaluation, the division of training and testing data was carried out using a stratified sampling technique to maintain consistency in class proportions in both subsets.

Initial data exploration was conducted to obtain a general overview of the statistical characteristics of the numeric features in the Employee Attrition dataset before modeling. This analysis sought to comprehend the distribution patterns of the data, the extent of variation, and the possibility of outliers that might affect the efficacy of the machine learning model.

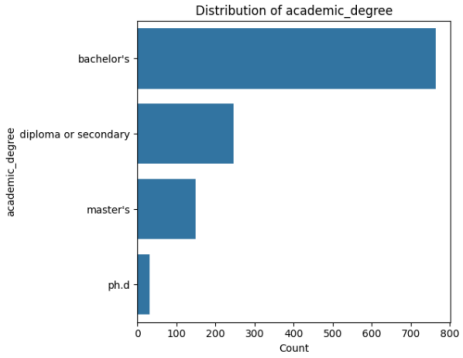

Figure 6. Visualization of categorical distribution

Based on figure 6, The distribution of the academic_degree feature shows that most employees have a bachelor's degree, with lower numbers in the diploma or secondary, master's, and PhD categories, reflecting the general educational structure of the workforce

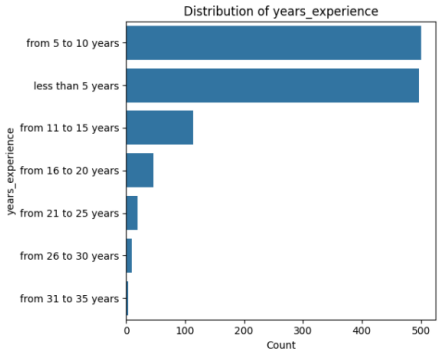

Figure 7. Visualization of categorical distribution

Based on figure 7, In the years_experience feature, the majority of employees have work experience ranging from less than 5 years and from 5 to 10 years, both for total experience and in the years_experience_lastorganization feature. This indicates that most respondents are in the early to middle phase of their careers

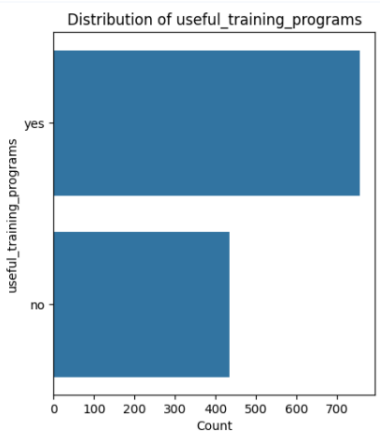

Figure 8. Visualization of categorical distribution

Based on figure 8, The distribution of the useful_training_programs feature shows that most employees perceive the training programs provided by the organization as beneficial (the "yes" category), with a lower but still notable number of employees indicating that the programs were not useful (the "no" category), reflecting the general perception of workforce development effectiveness within the organization.

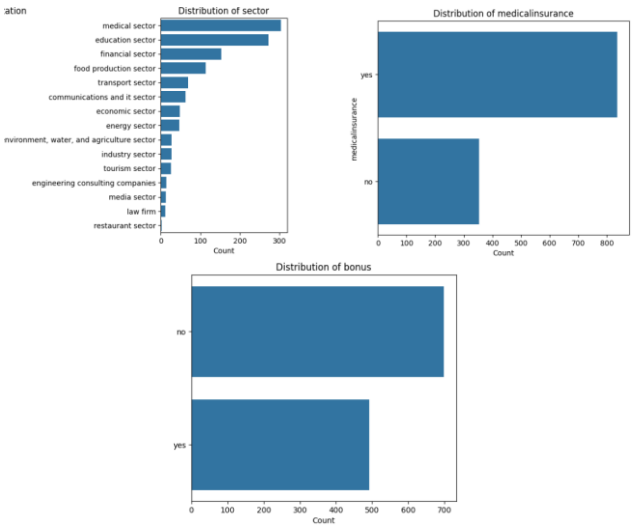

Figure 9. Visualization of categorical distribution

The distribution of employment sectors is dominated by the medical, education, and financial sectors, while other sectors have a relatively smaller number of observations. The distribution of department features shows that the administration, teaching, and patient affairs categories have the highest number of employees compared to other departments. Overall, the distribution of categorical features is not uniform but still represents real-world organizational conditions. This variation in the number of observations between categories is a consideration in the preprocessing and encoding stages, but it is not a major obstacle because the tree-based algorithm used in this study is able to handle the imbalanced distribution of categories well.

B. Data Pre-Processing Result

The data pre-processing phase was carried out to guarantee that all features in the employee attrition dataset were formatted appropriately for machine learning modeling. According to the analysis results from the earlier stage, the dataset was of high quality, with no missing values detected and only one duplicate entry eliminated. Consequently, the pre-processing efforts concentrated on cleaning categorical data, eliminating irrelevant features, and encoding without the need for imputation or numerical normalization.

The initial pre-processing step involved cleaning and standardizing the values of the categorical features. All text values were standardized to avoid differences in technical representations that have the same meaning, such as capitalization and excess spaces. Furthermore, identity columns that did not contribute to the attrition prediction process were removed to avoid introducing noise into the model's learning process.

Subsequently, an encoding procedure is executed to transform categorical features into numerical formats. This study employs a multifaceted encoding approach, customized to the unique attributes of each feature. Binary encoding is utilized for features that consist of only two categories,

preserving the relationship between classes while avoiding an increase in data dimensionality. Ordinal encoding is applied to features that possess a defined level or order, thereby retaining the hierarchical information among categories in a numerical representation.

Meanwhile, one-hot encoding is applied to nominal categorical features that have no order, to prevent the assumption of ordinal relationships that are inconsistent with the meaning of the data. After the encoding process is complete, the attrition target variable and the predictor features are separated. At this stage, further feature selection is also performed by removing attributes with very high cardinality or containing redundant information, such as job title, to reduce model complexity and the risk of multicollinearity.

```
Index: 1190 entries, 0 to 1190
Data columns (total 65 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   gender                                      1190 non-null   int64
 1   age                                         1190 non-null   int64
 2   academic_degree                             1190 non-null   int64
 3   years_experience                            1190 non-null   int64
 4   years_experience_lastorganization           1190 non-null   int64
 5   monthlysalary                               1190 non-null   int64
 6   allowances                                  1190 non-null   float64
 7   medicalinsurance                            1190 non-null   int64
 8   bonus                                       1190 non-null   int64
 9   overtime                                    1190 non-null   int64
 10  payment_overtime                            1190 non-null   int64
 11  rewards&wages_satisfaction                  1190 non-null   int64
 12  get_ deserved_promotion                     1190 non-null   int64
 13  training_programs_ during_last_three_years  1190 non-null   int64
 14  useful_training_programs                    1190 non-null   int64
 15  business_travel                             1190 non-null   int64
 16  job_support                                 1190 non-null   int64
 17  recognition                                 1190 non-null   int64
 18  emotional_commitment                        1190 non-null   int64
 19  job_engagement                              1190 non-null   int64
 20  distance_to_work                            1190 non-null   int64
 21  work_live_balance                           1190 non-null   int64
 22  physical_stress                             1190 non-null   int64
 23  psychological_exhaustion                    1190 non-null   int64
 24  job_stability                               1190 non-null   int64
 25  health_issues                               1190 non-null   int64
 26  environment_satisfaction                    1190 non-null   int64
 27  job_satisfaction                            1190 non-null   int64
 28  job_opportunities                           1190 non-null   int64
 29  maritalstatus_married                       1190 non-null   bool
```

Figure 10. Data cleaned after preprocessing

The final result of this stage is a transformed dataset in which all features have been converted into a uniform numeric format. Unlike conventional approaches, this study stipulates that data partitioning occurs after the entire preprocessing sequence is complete to ensure feature consistency between the training and test data.

Next, the dataset is partitioned into two independent subsets with a ratio of 80:20, where 80% of the data is allocated for model training and 20% for testing. Given the class imbalance in the target variable, the division was performed using the Stratified Sampling technique. As suggested by Shabbir et al. and Zhang et al., this technique ensures that the proportion of the Attrition class (Yes/No) in both subsets remains representative of the original population, thus preventing evaluation bias that often occurs in imbalanced datasets.

C. Decision Tree and LightGBM Algorithm Modeling Results

In this stage, two classification algorithms were implemented Decision Tree and Light Gradient Boosting Machine (LightGBM). These two algorithms were chosen because of their ability to handle encoded categorical data and their ease of interpretation in tree-based models. The modeling process was carried out in three main scenarios :

- Baseline model (without optimization)
- Model with regularization
- Model with hyperparameter optimization using GridSearchCV

Model evaluation was performed using Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics. The use of ROC-AUC as the primary metric aims to measure the model's ability to distinguish between attrition and non-attrition classes at various classification thresholds.

1. Decision Tree Model Evaluation

The Decision Tree model was first implemented using default parameters as a baseline. Evaluation results showed that the model performed reasonably well in classification, but still showed indications of overfitting due to the complexity of the unconstrained tree structure.

Indications of overfitting in the Decision Tree algorithm are identified by comparing the performance of the training data with the testing data. In the baseline Decision Tree model without regularization, it was found that the model was able to achieve near-perfect accuracy on the training data, but experienced a significant drop in performance when tested on the testing data. This large gap indicates that the model tends to 'memorize' the noise in the training data rather than learning general patterns, thus failing to generalize to new data. This phenomenon is the main reason for optimization using GridSearchCV with a 5-fold cross-validation and the application of regularization parameters (such as max_depth and min_samples_split) to limit the tree's complexity and produce a more stable model with better generalization capabilities for the organization.

TABLE 3
DECISION TREE MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
DT Baseline	0.72	0.68	0.66	0.67	0.71
DT Regularization	0.72	0.68	0.66	0.67	0.78
DT GridSearchCV	0.74	0.78	0.55	0.65	0.80

These results demonstrate that limiting tree complexity can reduce the risk of overfitting without significantly reducing model performance. The increased ROC-AUC value also

indicates that the model is more stable and has better generalization capabilities.

The Decision Tree model in baseline conditions showed indications of overfitting due to uncontrolled tree complexity. After limiting the parameters max_depth, min_samples_split, and min_samples_leaf, model performance improved, with the ROC-AUC value increasing from 0.71 to 0.78, and reaching 0.80 after hyperparameter tuning. This indicates that limiting complexity can improve the model's stability and generalization capabilities.

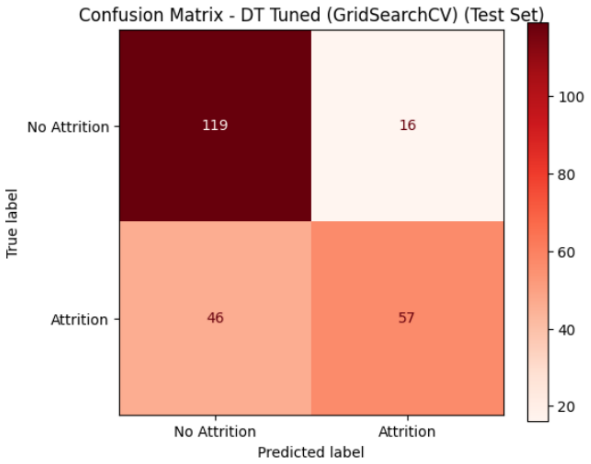

Figure 11. Confusion Matrix with the best model in the decision tree

Based on Figure 11, the Decision Tree model produced an accuracy of 73.9%. The DT model had a higher number of false negatives (46 vs. 30), meaning it more often failed to detect employees who were actually at risk of leaving. This indicates that the DT model may be less sensitive than the LGBM for this attrition case.

2. *LightGBM Model Evaluation*

The LightGBM model is implemented as a boosting-based ensemble method that gradually builds multiple decision trees to minimize residual error. In the baseline model, initial parameters such as n_estimators, learning_rate, max_depth, and num_leaves are used as the initial configuration. Next, regularization is applied to control model complexity through parameters such as reg_alpha, reg_lambda, and subsample.

Hyperparameter tuning is then performed using GridSearchCV with a search space that includes parameters such as n_estimators, learning_rate, num_leaves, max_depth, and min_child_samples. The optimization process is performed using 5-fold cross-validation using the ROC-AUC metric.

TABLE 4
LIGHTGBM MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
LGBM Baseline	0.79	0.77	0.74	0.75	0.87
LGBM Regularization	0.78	0.78	0.69	0.73	0.88
Tuned LGBM GridSearchCV	0.81	0.82	0.71	0.76	0.85

Experimental results show that the LightGBM model, after optimization, performs better than the Decision Tree model, particularly in the ROC-AUC and Recall metrics for detecting employee attrition. This indicates that the boosting approach is more effective in capturing complex patterns and interactions between features than a single decision tree-based model.

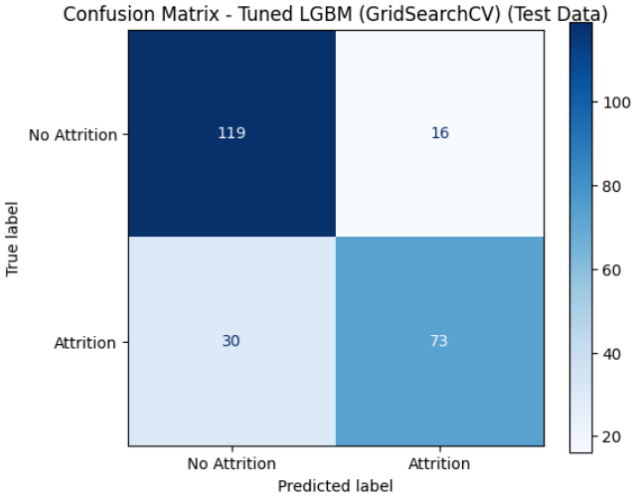

Figure 12. Confusion Matrix with the best model in the LightGBM

From Figure 12, the model shows quite good performance with an accuracy level of 80.6%. The main focus of the attrition model is usually on the Recall value, where the model successfully identified around 70.8% of the total employees who actually left. The low False Positive value (16) indicates that the model is quite selective in assigning the "Attrition" label, thus minimizing unnecessary intervention from management.

To deepen the analysis of the influence of hyperparameters, this study used the Optuna approach as an additional analysis on the LightGBM model. Tuning results using Optuna showed that several parameter combinations were able to produce ROC-AUC values of up to 0.91 in the cross-validation stage. However, this value was not used as the final model because this study determined the best model based on GridSearchCV results on the test data with an ROC-AUC value of 0.85.

Based on the optimization results, the LightGBM model obtained the best ROC-AUC value of 0.9136. The optimal

parameters produced include n_estimators of 212, learning_rate of 0.0949, num_leaves of 60, max_depth of 3, and min_child_samples of 40. In addition, the subsample and colsample_bytree parameters were 0.9081 and 0.7698, respectively, while the regularization parameters reg_alpha and reg_lambda were 0.1972 and 0.0205, respectively.

TABLE 5
OPTUNE BEST PARAMETERS

Hyperparameter	Best value	Importance
n_estimators	212.0	0.061
learning_rate	0.094	0.141
num_leaves	60.00	0.010
max_depth	3.000	0.015
min_child_samples	40.00	0.129
subsample	0.908	0.024
colsample_bytree	0.769	0.025
reg_alpha	0.197	0.556
reg_lambda	0.020	0.033

Next, a parameter importance analysis was performed to measure the influence of each hyperparameter on model performance. The analysis results show that the reg_alpha parameter is the most dominant parameter with a contribution of 55.62%, followed by learning_rate at 14.17%, and min_child_samples at 12.98%. Meanwhile, other parameters such as n_estimators (6.18%), reg_lambda (3.36%), colsample_bytree (2.60%), subsample (2.48%), max_depth (1.54%), and num_leaves (1.07%) have a relatively smaller contribution to model performance.

The dominance of the reg_alpha parameter indicates that L1 regularization plays a crucial role in controlling model complexity and reducing the risk of overfitting in LightGBM. An optimal reg_alpha value helps the model perform feature selection indirectly by penalizing less relevant parameters. Furthermore, the learning_rate parameter plays a crucial role in the boosting process, where an appropriate value allows the model to learn gradually and steadily. The min_child_samples parameter also contributes to controlling the complexity of the decision tree, making the model less sensitive to noise in the data.

Thus, the results of this analysis indicate that, in addition to learning parameters such as learning_rate, regularization aspects, especially reg_alpha, have a more significant influence on improving model performance. These findings complement the results from GridSearchCV by providing additional insights into the factors that influence the performance of the LightGBM model in predicting employee resignation.

3. Model Performance Comparison

Based on experimental findings, there was a significant difference in performance between the baseline and optimized models. Although the application of regularization and optimization to the Decision Tree improved the ROC-AUC score to 0.80, its ability to identify minority classes (attrition) remained limited, with the lowest recall value being 0.55.

To strengthen the comparative analysis, a Random Forest (GridSearchCV) model was added as a representative bagging method. This model provided more stable performance than the Decision Tree, with an accuracy of 0.75 and an ROC-AUC of 0.83. However, these results were still below those achieved by Tuned LightGBM.

TABLE 6
MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
DT Baseline	0.72	0.68	0.66	0.67	0.71
DT Regularization	0.72	0.68	0.66	0.67	0.78
DT GridSearch CV	0.74	0.78	0.55	0.65	0.80
Random Forest GridsearchCV	0.75	0.74	0.66	0.70	0.83
LGBM Baseline	0.79	0.77	0.74	0.75	0.87
LGBM Regularization	0.78	0.78	0.69	0.73	0.88
Tuned LGBM GridSearch CV	0.81	0.82	0.71	0.76	0.85

In contrast, the LightGBM model demonstrated the most stable and reliable performance. In the baseline setting, LightGBM outperformed all Decision Tree and Random Forest variants. Further optimization using GridSearchCV resulted in the Tuned LightGBM model with the highest overall performance, achieving an accuracy of 0.81, a precision of 0.82, and an F1-score of 0.76 on the test data. These results demonstrate that the optimized ensemble boosting approach provides the best balance in detecting employee resignation risk compared to other methods.

In addition to using the quantitative metrics summarized in the previous table, the model's effectiveness in distinguishing between retained and departing employees was further tested using a Receiver Operating Characteristic (ROC) curve. This curve plots the relationship between True Positive Rate and False Positive Rate at various classification thresholds to provide a comprehensive overview of each algorithm's discriminatory ability.

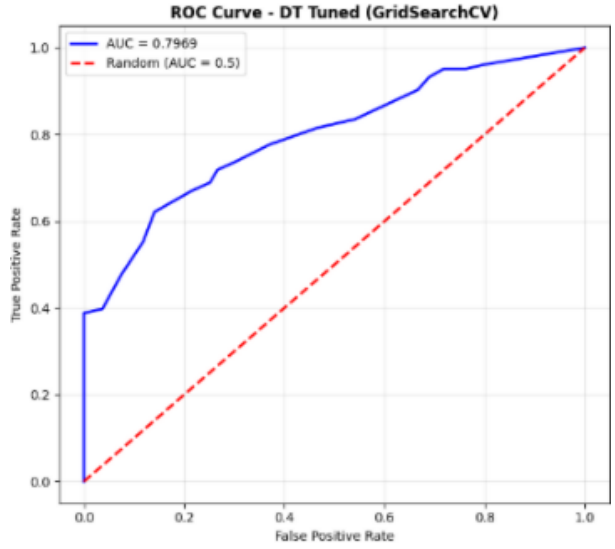

Figure 13. ROC graph of Decision tree hyperparameter tuning gridseachCV

Figure 13 shows that the optimized Decision Tree model produces an AUC value of 0.7969. Although the hyperparameter tuning via GridSearchCV has successfully stabilized the model and reduced the previous severe overfitting, its overall discriminative ability remains the lowest among the three evaluated models.

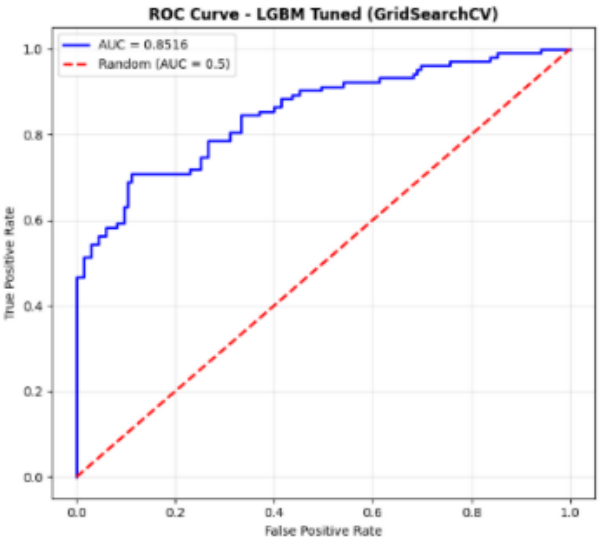

Figure 14. ROC graph of LGBM hyperparameter tuning gridseachCV

Ultimately, Figure 14 shows that the highest and most optimal result is consistently achieved by the Tuned LightGBM model, yielding an AUC value of 0.8516. Visually, its curve appears closest to the upper left corner, confirming that the boosting-based ensemble learning approach has a significant advantage in identifying subtle and non-linear employee attrition patterns. Although certain manual regularization trials might show slightly higher standalone AUC numbers, Tuned LightGBM via GridSearchCV was chosen as the final best model due to its superior capacity to provide the most reliable balance across all other essential evaluation metrics—such as accuracy (0.81), precision (0.82), and recall (0.71)—which are vital for risk-free strategic implementation in human resource departments

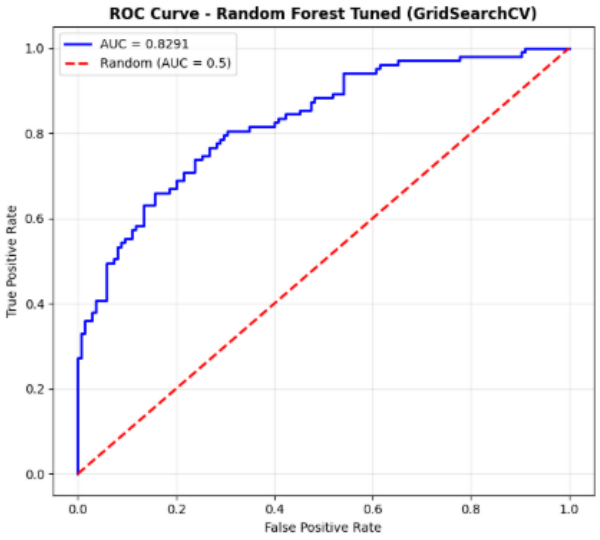

Figure 15. ROC graph of random forest hyperparameter tuning gridseachCV

As a secondary baseline representing the bagging method, the optimized Random Forest model demonstrates an enhanced performance. As illustrated in Figure 15, this model achieves an improved AUC value of 0.8291, proving that aggregating multiple independent trees successfully captures more complex interactions within the data compared to a single decision tree.

This visually confirms that the boosting-based ensemble learning approach in LightGBM has a greater advantage in identifying attrition patterns compared to single tree-based models or bagging. Although there are models with AUC values that may be slightly higher in certain regulatory scenarios, Tuned LightGBM was chosen as the optimal model because it is able to provide the best balance across all other evaluation metrics, such as accuracy, precision, and recall, which are crucial for strategic implementation in human resource departments.

D. Model Interpretation and Feature Importance Analysis

Although LightGBM provides the best performance, this model is complex and difficult to interpret directly. Therefore, feature importance analysis was performed to identify the variables that contribute most to attrition prediction. Feature importance is calculated based on the total gain obtained by each feature during the tree formation process in the boosting algorithm, where the gain value represents the feature's contribution to reducing classification errors.

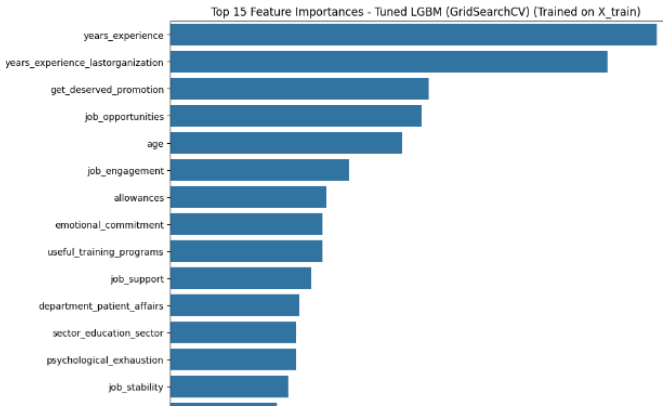

Figure 12. Feature Importances

Visualization results in Figure 12, the variables that have the most dominant contribution are as follows.

- Years of Experience, This is the feature with the highest contribution. In the context of industrial psychology, this variable relates to an individual's career stage. Employees with very long or very short work experience often have different motives for deciding to change organizations.
- Years of Experience in Last Organization. This feature indicates an employee's historical level of stability. A short tenure at a previous organization may indicate a recurring pattern of job-hopping behavior or a cultural mismatch.
- Get Deserved Promotion. The high gain score on this feature indicates that organizational justice plays a crucial role. Employees who feel their contributions are not recognized through equal career advancement have a strong psychological drive to seek professional validation outside the company.
- Job Opportunities, These external factors indicate that the decision to leave is not only influenced by the company's internal conditions, but also by the availability of options in the labor market that offer more attractive positions or compensation.
- Age, These demographic variables often serve as proxies for life stability and personal priorities, where certain age groups may be more willing to take the risk of changing jobs for career growth.

Contextually, the dominance of features related to tenure and promotion opportunities suggests that attrition in this dataset is heavily influenced by career trajectory and professional recognition. When an employee feels they have reached a plateau in their tenure without perceived fair promotion progress, the perceived benefits of remaining with the company decline. This is exacerbated if job opportunities outside the organization appear more promising, increasing the probability of attrition.

Furthermore, the presence of features such as job engagement and emotional commitment in the list of important features confirms that the relationship between employees and the organization is not only transactional (financial) but also involves a psychological bond. Psychological exhaustion and job support, which also emerged in this analysis, indicate that work environment stress and the lack of a support system in the workplace are significant factors that drive employees to terminate their employment.

E. Discussion

This study was designed to answer three main research questions a comparison of classification algorithm performance, the effect of model optimization on predictive performance, and the identification of dominant factors influencing employee attrition.

Regarding the first research question, the experimental results show that the ensemble learning approach, specifically Tuned LightGBM, consistently produces superior classification performance compared to Decision Tree, Random Forest, and Logistic Regression. This superiority is clearly evident in key evaluation metrics such as ROC-AUC and F1-Score, which reflect the model's ability to distinguish employees at risk of leaving from those who are likely to remain. These findings confirm that attrition patterns are complex and involve non-linear interactions between variables, thus the boosting method proves more effective than single tree models, bagging methods, and linear models. The main strength of these findings lies in the empirical evidence that the ensemble boosting approach provides more stable generalization to attrition data dominated by categorical features.

Answering the second research question, this study reveals that the application of hyperparameter regularization and tuning through GridSearchCV significantly improves the model's stability and generalization ability. Through the 5-fold cross-validation process integrated into GridSearchCV, the optimized model demonstrated more consistent performance compared to the baseline model while reducing the risk of overfitting. This strengthens the methodological argument that the optimization process is not merely an additional step, but rather a crucial element in building a reliable prediction system. The contribution of this research lies in its systematic and transparent optimization approach, thereby increasing the replicability and validity of the experiment.

Meanwhile, the third research question was answered through feature importance analysis, which identified the dominant variables predicting attrition. Substantially, the analysis results showed that employees' decisions to leave a company were significantly influenced by factors such as career trajectory and organizational fairness. The dominance of variables such as total tenure (years of experience) and perceptions of promotions (deserved promotion) indicates that employees tend to evaluate their long-term value at the

company based on professional recognition. When career aspirations are not met, or there is a perception of unfairness in the promotion process, the psychological urge to seek validation outside the organization increases dramatically, especially when supported by the availability of job opportunities in the external market.

In addition to career factors, psychological aspects and the work environment also play a significant role. Variables such as job engagement, emotional commitment, and psychological exhaustion indicate that resignation decisions are often the result of a combination of workload and a lack of job support. Unaddressed psychological exhaustion can sever employees' emotional bonds, transforming what was once a professional interaction into a desire to escape a stressful work environment. These findings strengthen the literature on the determinants of turnover intention through machine learning-based validation on real-world data.

Overall, this study makes two main contributions (1) a methodological contribution by demonstrating the effectiveness of ensemble learning and hyperparameter optimization in the context of attrition prediction, and (2) a practical contribution by providing a predictive framework that can support data-driven decision-making in HR departments. Future research could focus on model testing on cross-industry datasets to improve model generalizability, as well as the application of explainable AI (XAI) methods such as SHAP for deeper interpretation at the individual level.

IV. CONCLUSION

While this study demonstrates that hyperparameter optimization improves overall model performance, it is important to acknowledge several limitations. First, the performance differences between the training and testing datasets indicate that model generalization is still influenced by data characteristics. While the use of 5-fold cross-validation integrated with GridSearchCV has improved model stability compared to the baseline model, further development is needed to strengthen the model's robustness to broader applications. Furthermore, the lack of a specific data balancing technique in this study is a technical limitation that needs to be considered when interpreting the evaluation metrics for minority classes.

Second, the determination of the best-performing model (Tuned LightGBM) was not based solely on the highest ROC-AUC value, but rather through a comprehensive evaluation of various metrics such as accuracy, precision, recall, and F1-score. This approach was taken to ensure the model's practical relevance in assessing employee resignation risk. However, the recall value of 0.71 indicates a business risk where approximately 29% of employees who actually planned to leave were not detected by the model. Therefore, these prediction results should be positioned as a decision support system combined with qualitative evaluation from the HR department to close this detection gap.

Third, this study used a single dataset with specific organizational characteristics in Saudi Arabia, which limits the generalizability of the results to different cultural or industrial contexts. Furthermore, model interpretation is limited to global feature importance analysis and does not include individual-level interpretation techniques. Comparisons between Decision Tree and Random Forest models confirm that, although the ensemble method yields superior results, external factors and organizational justice remain significant influences on attrition behavior.

Consequently, further research is recommended to validate this approach on more diverse cross-industry datasets, apply class imbalance handling techniques for more precise results, and explore advanced interpretation methods to strengthen practical applications in organizations in a more personalized manner.

REFERENCES

[1] M. Fadel, K. Kanasfi, Z. Arifin, and G. Triyono, "Application Of Ensemble Method For Employee Turnover Predictions In Financial Services Company," *J. Tek. Inform. Jutif*, vol. 5, no. 3, pp. 767–775, May 2024, doi: 10.52436/1.jutif.2024.5.3.1871.

[2] J. Park, Y. Feng, and S.-P. Jeong, "Developing an advanced prediction model for new employee turnover intention utilizing machine learning techniques," *Sci. Rep.*, vol. 14, no. 1, p. 1221, Jan. 2024, doi: 10.1038/s41598-023-50593-4.

[3] A. Nurhindarto, E. W. Andriansyah, F. Alzami, P. Purwanto, M. A. Soeleman, and D. P. Prabowo, "Employee Attrition and Performance Prediction using Univariate ROC feature selection and Random Forest," *Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control*, Nov. 2021, doi: 10.22219/kinetik.v6i4.1345.

[4] M. S. Alshiddy and B. N. Aljaber, "Employee Attrition Prediction using Nested Ensemble Learning Techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 7, 2023, doi: 10.14569/IJACSA.2023.01407101.

[5] "Employee Turnover Prediction Research of Human Resource Management on Machine Learning Algorithms and Big Data Analysis," *J. Organ. End User Comput.*, vol. 38, no. 1, Jan. 2026, doi: 10.4018/JOEUC.399146.

[6] X. Wang and N. Huang, "Application of data visualization technology in human resource management and employee resignation prediction," *Syst. Soft Comput.*, vol. 7, p. 200355, Dec. 2025, doi: 10.1016/j.sasc.2025.200355.

[7] R. Govindarajan, N. K. Kumar, S. R. P, S. P. E, D. B, and P. K. G, "Predicting Employee Attrition: A Comparative Analysis of Machine Learning Models Using the IBM Human Resource Analytics Dataset," *Procedia Comput. Sci.*, vol. 258, pp. 4084–4093, 2025, doi: https://doi.org/10.1016/j.procs.2025.04.659.

[8] S. Rawat, A. Rawat, D. Kumar, and A. S. Sabitha, "Application of machine learning and data visualization techniques for decision support in the insurance sector," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 2, p. 100012, Nov. 2021, doi: 10.1016/j.jjimei.2021.100012.

[9] "Grey clustering machine learning model for predicting voluntary employee turnover," *Grey Syst. Theory Appl.*, vol. 15, no. 4, pp. 771–791, Aug. 2025, doi: 10.1108/GS-02-2025-0020.

[10] R. H. M. Aly, A. I. Hussein, and K. H. Rahouma, "Grasshopper KUWAHARA and Gradient Boosting Tree for Optimal Features Classifications," *Comput. Mater. Contin.*, vol. 72, no. 2, pp. 3985–3997, 2022, doi: https://doi.org/10.32604/cmc.2022.025862.

[11] J. C. M. Bustillo, "Optimization-based Techniques Prediction Model in Determining Employee Turnover," *Procedia Comput. Sci.*, vol. 252, pp. 440–449, Jan. 2025, doi: 10.1016/j.procs.2025.01.003.

[12] H.-C. Chen, J.-Y. Wang, Y.-C. Lee, and S.-Y. Yang, "Examining the Predictors of Turnover Behavior in Newly Employed Certified Nurse Aides: A Prospective Cohort Study," *Saf. Health Work*, vol. 14, no.

2, pp. 185–192, 2023, doi: https://doi.org/10.1016/j.shaw.2023.04.003.

[13] E. Ahmed and M. Omer, "Predicting Employee Attrition Using Artificial Neural Networks: A Comparative Study of Machine Learning Models and Imbalanced Data Handling Techniques," 2025, *SSRN*. doi: 10.2139/ssrn.5105905.

[14] K.-T. Nguyen, T.-N. Tran, and H.-T. Nguyen, "Research on the Influence of Hyperparameters on the LightGBM Model in Load Forecasting," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 5, pp. 17005–17010, Oct. 2024, doi: 10.48084/etasr.8266.

[15] H. Zhang, Y. Wang, Z. Li, and X. Wang, "Machine Learning Models for Bank Customer Churn Prediction: A Comparative Study of LightGBM, CatBoost, and XGBoost," in *Proceedings of the 2025 International Conference on Big Data, Artificial Intelligence and Digital Economy*, Kunming China: ACM, Jul. 2025, pp. 6–16. doi: 10.1145/3767052.3767054.

[16] S. A. Alteer and A. Alariyibi, "Customer Churn Prediction Using Machine Learning: A Case Study of Libyan Internet Service Provider Company," in *2024 IEEE 4th International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA)*, May 2024, pp. 605–612. doi: 10.1109/MI-STA61267.2024.10599671.

[17] G. Vijh, N. Sharma, S. Tiwari, S. Vijh, and A. Sao, "Predicting Accurate Employee Performance: An Evaluation of Regression Models," *Procedia Comput. Sci.*, vol. 259, pp. 433–442, 2025, doi: https://doi.org/10.1016/j.procs.2025.03.345.

[18] Z. Liu and T. Kong, "Evaluation of Enterprise Internal Control Based on Artificial Intelligence," *Procedia Comput. Sci.*, vol. 262, pp. 1217–1227, 2025, doi: https://doi.org/10.1016/j.procs.2025.05.163.

[19] G. Vijh, N. Sharma, S. Tiwari, S. Vijh, and A. Sao, "Predicting Accurate Employee Performance: An Evaluation of Regression Models," *Procedia Comput. Sci.*, vol. 259, pp. 433–442, 2025, doi: 10.1016/j.procs.2025.03.345.

[20] M. Kang and H. Yim, "Unveiling employee perspectives: A comparative analysis of online reviews on Korean SMEs and large corporations," *Int. J. Inf. Manag. Data Insights*, vol. 4, no. 2, p. 100268, Nov. 2024, doi: 10.1016/j.jjimei.2024.100268.

[21] M. Madanchian, H. Taherdoost, and N. Mohamed, "AI-Based Human Resource Management Tools and Techniques; A Systematic Literature Review," *Procedia Comput. Sci.*, vol. 229, pp. 367–377, Jan. 2023, doi: 10.1016/j.procs.2023.12.039.