International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 4, July-Aug-2026, ISSN (Online): 2395-566X

Predictive Analytics for Employee Attrition Management Using Machine Learning Techniques

Subhadip Sarkar [1], Ms. Preeta Rajiv Sivaraman[2]
[1](Assistant Professor
Civil Engineering
Seacom Skills University
Bolpur, Birbhum, West Bengal
sarkarsubhadip757@gmail.com)

[2](Assistant Professor
BCA
JIMS Engineering Management Technical Campus, Greater Noida
preetaraj2308@gmail.com)[)]

Abstract — Employee turnover is a challenging situation for any organization because of high associated costs related to recruitment and training of new employees as well as losing organizational knowledge. This research aims to examine how machine learning can be used in predictive analytics in managing employee turnover through the use of the IBM HR Analytics data set. Five models including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting were tested systematically. It was revealed that ensemble models such as Random Forest have the highest predictive accuracy and AUC equal to 87.3% and 0.9348, respectively. Overtime, job satisfaction, monthly income, and tenure are found to be important factors affecting employee turnover. These results show that machine learning may change HR practices from reactive to proactive retention strategies based on data analysis.

Keywords— Employee Attrition, Predictive Analytics, Machine Learning, Random Forest, HR Analytics, Workforce Retention, Ensemble Learning

I. INTRODUCTION

Attrition in the form of the voluntary or forced leaving of employees from their organization is seen as one of the most relevant problems faced by contemporary organizations all over the world [3]. Attrition involves many far-reaching consequences for the stability, productivity, financial sustainability, and competitiveness of organizations. The departure of talented employees means not just the expenses on hiring and training new staff, but also the loss of corporate culture and expertise, decreased morale in teams, and disruptions in customer relations and operations [5]. According to the data provided by the industry, the cost of replacing one worker may vary from 50% to more than 200% of his annual salary depending on the employee's status in the company [7]. Voluntary turnover rate in the USA has been steadily higher than 2.5% per month from 2021 to 2023 [4].

The traditional way organizations handled the problem of employee attrition has been a reactive one. Organizations would use exit interviews, employee satisfaction surveys, and regular performance reviews in order to know why employees are leaving once they have already made their decision [2]. Although these tools give a retrospective analysis, the problem is that they completely miss the mark when it comes to addressing the real issue – identifying and preventing attrition before it happens. This is a serious analytical gap in the field of human resource management in a world where there is increased mobility of the workforce.

Theoretical concepts like the Job Embeddedness Theory and the Unfolding Model of Turnover have been very important in trying to identify the psychological and situational factors affecting the decision to leave [8]. Such concepts highlight the relationship between organizational fit, social relationships, sacrifices and shocks. However, the limitation of such theoretical perspectives is that they do not offer any tools for prediction.

The development of machine learning and predictive analysis has brought unprecedented possibilities in dealing with this problem. With the help of big employee datasets including demographic information, performance, remuneration history, engagement measures, and tenure information, machine learning algorithms can detect non-linear dependencies that would be difficult to find with the help of standard methods of

statistical analysis [6]. With the help of advanced machine learning tools and an abundance of high-quality HR information, it is possible to build accurate predictive models of individual attrition probabilities [9].

The research is based on the main assumption that machine learning models are able to forecast employee attrition with greater precision compared to HR-based approaches, thus helping firms implement retention policies promptly and efficiently. In order to test the assumptions, the research seeks answers to the following questions:

- What are the key predictors of attrition in HR databases?
- How do various machine learning algorithms perform when classifying employee attrition?
- How to turn predictions into HR strategy?

Organization of the Paper: Section 2 contains a detailed review of the existing literature related to the topic of previous studies on employee turnover prediction using machine learning in HR analytics. Section 3 outlines the methodology used, which includes pre-processing of the data, feature engineering, choosing algorithms and model evaluation techniques. Section 4 provides the results and their analysis, including performance evaluation and feature importance analysis. Concluding section is provided in Section 5.

II. LITERATURE SURVEY

There have been many studies carried out over the last few years focusing on the use of machine learning methods in predicting the occurrence of employee attrition due to increased awareness on the importance of human resource analytics in organizations. A study by Talebi, Khatibi Bardsiri and Khatibi Bardsiri was carried out, where they undertook a review of studies that used machine learning techniques in the prediction of employee turnover, and found that the most commonly used machine learning algorithm is Random Forest with high accuracy rates [3]. The most important variable in employee turnover is job satisfaction [4].

In another study on the use of machine learning techniques for predictive analytics through the IBM HR Analytics dataset, Regasse and Venier studied Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and Support Vector Classifier [1]. According to their results, the Support Vector Classifier showed high discriminative power with AUC and cross-validated AUC values of 0.8129 and 0.8291 respectively. Feature importance showed that overtime, job satisfaction, and involvement were the main turnover factors, which was theoretically expected as these are related to employee engagement and commitment [1].

In a comparative study by some researchers on Random Forest, Support Vector Machine, Decision Tree, Gradient Boosting, and Hybrid Model, the Hybrid Model had the best accuracy rate at 95.0% indicating the benefits of algorithm combination [2]. In this case, the Support Vector Machine classified the instances correctly with an accuracy rate of 88.6%, whereas Random Forest and Gradient Boosting had the same rate at 87.3%. The Decision Tree classifier had the lowest accuracy rate of 80.5%. However, recent developments have moved beyond classification to solve problems of model interpretability, fairness, and practical application. Chaudhary et al. created a detailed approach that combines robust data preparation, predictive modeling, and methods for explaining like SHAP (SHapley Additive exPlanations) [7]. Both global and local interpretability analysis revealed overtime, job level, and job satisfaction as important predictors, allowing for specific actions [7]. The increased importance of explainability is a solution to a frequent challenge of using machine learning in HR — the “black box” problem, which may decrease practitioners' confidence and use.

Yin and Wang compared logistic regression, Naive Bayes, and Random Forest models, where Random Forest proved to be the best in accuracy, precision, recall, and F1-score [10]. The feature importance analysis on Gini impurity basis showed employee satisfaction as the main turnover predictor, and then followed the performance rating, monthly income, and tenure [10]. It is especially important since the results showed that modifiable organizational factors are more important than the demographic ones.

The study of creating a machine learning framework using SAP SuccessFactors showed that Random Forest had accuracy of 86% and had well-balanced precision and recall measurements due to interpretability of features based on SHAP values [9]. The machine learning framework focused on integration, enabling HR teams to integrate attrition prediction results into enterprise dashboards. This integration-focused machine learning framework is the next step after developing a standalone framework focused on predictions to build workforce intelligence solutions.

The research on using AI for HR transformation designed a predictive framework based on Random Forest achieving the following metrics – accuracy of 0.9352, precision of 0.9649, recall of 0.9016, F1 score of 0.9322, and ROC-AUC of 0.9348 [5]. In addition, the study proved the superiority of the designed framework to other existing frameworks like SVM, Decision

Tree, and K-Nearest Neighbors. Importantly, the framework integrates the decision optimization process to convert probabilities to HR decisions [5].

Although considerable progress has been made, there are several limitations associated with current research. Most researchers make use of a single database in their research work, which raises issues related to generalizability to other industries or organizations [8]. Imbalanced class problems, where attrition constitutes a smaller number of observations, remain a challenge in the process of modeling and analysis [6]. In addition, although accuracy has significantly increased, not much attention has been paid to developing retention strategies based on prediction results and analyzing the effects on the organization.

III. PROPOSED METHODOLOGY

1. Research Framework
This study is based on the use of CRISP-DM (CRoss-Industry Standard Process for Data mining) approach in order to develop machine learning algorithms for predicting employee turnover. The approach is composed of six main stages, including business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

2. Dataset Description
Data for the experiment is taken from the IBM HR Analytics Employee Attrition and Performance dataset, which is a popular benchmark for experiments in human resource analytics. The dataset consists of 1,470 employee data records, each record having 35 features, ranging from demographic features to job features, compensation-related features, satisfaction indicators, and an attrition flag as a label. The list of features includes age, business travel, department, distance from home, education, employee number, environment satisfaction, gender, hourly rate, job involvement, job level, job role, job satisfaction, marital status, monthly income, monthly rate, number of companies worked for, over18, overtime, percent salary hiked, performance rating, relationship satisfaction, stock option level, total years of experience, training times last year, work-life balance, and years at company.

Class imbalance exists in the dataset because 84% of employees stay while 16% depart from their jobs. The data set shows a real situation where people leave the job and needs proper attention when modeling.

3. Data Preprocessing
Data pre-processing is done following a series of stages to ensure that the data set is ready for analysis via machine learning. Data encoding for categorical data is achieved using one-hot encoding for nominal data while label encoding for ordinal data. For numerical data, the data values are scaled in the [0,1] interval to avoid bias by features of larger magnitude in training the machine learning model. The data set is divided into two groups of 70% for training and 30% for testing via stratification sampling to ensure that the class distribution is well represented in testing.

Outliers were not detected among any variables in the IBM data set, which is similar to what was found in previous research. Nonetheless, capping for outliers in features like monthly income and years with the company is done through the 1st and 99th percentiles.

4. Feature Engineering and Selection
Feature engineering starts with correlation analysis to find correlations between the independent variables and the target attrition class. Variables showing correlations ($p < 0.05$) are then used for feature engineering. Next, the importance of the independent variables is analyzed through Random Forest Gini impurity reduction in order to determine the importance of the independent variables in predicting the outcome. A threshold of 0.01 is used for feature filtering of the least useful features in terms of prediction.

On the other hand, feature engineering also introduces new variables that capture interactions that are important for understanding attrition. For instance, a "workload factor" is generated by multiplying monthly working hours and job satisfaction, while a "career growth factor" is obtained from the sum of total working years and promotions.

Machine Learning Algorithms
There are five machine learning techniques employed for modeling the attrition prediction problem:

Logistic Regression
The Logistic Regression technique is a generalized linear method for estimating the probability of binary events utilizing the logistic function. The Logistic Regression technique assumes linear dependencies of predictors on the log odds of attrition. Logistic Regression is used as a benchmarking technique due to its interpretability.

Algorithm 1: Logistic Regression for Employee Attrition Prediction

Input: Training data D = $\{(x_1,y_1), (x_2,y_2), ..., (x_n,y_n)\}$ where $x_i \in R^d$, $y_i \in \{0,1\}$
Output: Trained Logistic Regression model with coefficient vector β

1: Initialize coefficients $\beta = [\beta_0, \beta_1, ..., \beta^d]$ randomly
2: Set learning rate α and number of iterations T
3: Define sigmoid function: $\sigma(z) = 1 / (1 + e^{-z})$
4: Define cost function: $J(\beta) = -1/n * \Sigma[y_i * \log(\sigma(\beta \cdot x_i)) + (1-y_i) * \log(1-\sigma(\beta \cdot x_i))]$
5: For t = 1 to T do:
6: Compute gradient: $\nabla J(\beta) = 1/n * \Sigma (\sigma(\beta \cdot x_i) - y_i) * x_i$
7: Update coefficients: $\beta = \beta - \alpha * \nabla J(\beta)$
8: Compute current cost $J(\beta)$
9: If $J(\beta)$ converges, break loop
10: End For
11: Return coefficient vector β

Decision Tree

Decision Tree is a non-parametric technique used for supervised learning which recursively splits the feature space according to the values of the features with the objective of maximizing information gain. Decision Trees have good interpretability since decision rules are easily visualizable but they are prone to overfitting. Maximum tree depth is optimized via cross-validation.

Algorithm 2: Decision Tree for Employee Attrition Prediction

Input: Training data D with features X and target y, maximum depth max_depth,
minimum samples per leaf min_samples_split
Output: Trained Decision Tree model

Function BuildTree(D, depth):
1: If depth == 0 OR |D| < min_samples_split OR all y in D are same:
2: Return LeafNode(prediction = majority class in D)
3: End If
4: For each feature f in X:
5: For each split value v in feature f:
6: Compute Gini impurity gain:
7: Gain = Gini(D) - (|D_left|/|D|)*Gini(D_left) - (|D_right|/|D|)*Gini(D_right)
8: End For
9: End For
10: Select feature f* and split v* with maximum gain
11: D_left = {x ∈ D | x[f*] ≤ v*}
12: D_right = {x ∈ D | x[f*] > v*}
13: LeftNode = BuildTree(D_left, depth-1)
14: RightNode = BuildTree(D_right, depth-1)
15: Return InternalNode(feature=f*, split=v*, left=LeftNode, right=RightNode)

Random Forest

Random Forest is an ensemble-based algorithm which creates many decision trees during training and merges their results using voting based on majority for classification problems. The advantage of Random Forest is that it avoids the problem of overfitting by taking the average of trees, along with having built-in feature importance. The number of trees is set to 100.

Algorithm 3: Random Forest for Employee Attrition Prediction

Input: Training data D with features X and target y, number of trees N_trees,
maximum features per split max_features, maximum depth max_depth
Output: Trained Random Forest ensemble model

1: Initialize empty forest F = []
2: For i = 1 to N_trees do:
3: D_bootstrap = BootstrapSample(D) // Sample with replacement
4: Initialize root node with D_bootstrap
5: Tree = BuildTree(D_bootstrap, max_depth, max_features)
6: F.append(Tree)
7: End For
8: Return F

Function BuildTree(D, max_depth, max_features):
1: If stopping criteria met (depth limit, min samples, pure node):
2: Return LeafNode(majority class in D)
3: End If
4: Select random subset F_subset of size max_features from all features
5: For each feature f in F_subset:
6: Find best split v* maximizing Gini gain
7: End For
8: Select feature f* and split v* with maximum gain
9: D_left = {x ∈ D | x[f*] ≤ v*}
10: D_right = {x ∈ D | x[f*] > v*}
11: LeftNode = BuildTree(D_left, max_depth-1)
12: RightNode = BuildTree(D_right, max_depth-1)

13: Return InternalNode(feature=f*, split=v*, left=LeftNode, right=RightNode)

Function Predict(model, x):
1: For each tree T in model:
2: prediction = T.predict(x)
3: End For
4: Return majority vote across all predictions

Support Vector Machine

The Support Vector Machine (SVM) is a supervised learning algorithm that aims at finding the hyperplane which maximizes the margin between the two classes. SVM with RBF kernel identifies non-linear classification boundaries and shows good discriminative power in predicting employee attrition.

Algorithm 4: Support Vector Machine for Employee Attrition Prediction

Input: Training data D = $\{(x_1,y_1), (x_2,y_2), ..., (x_n,y_n)\}$, kernel function K,
regularization parameter C, tolerance ε
Output: Trained SVM model with support vectors

1: Initialize Lagrange multipliers α = [0, 0, ..., 0]
2: Define kernel matrix K where K_ij = K(x_i, x_j)
3: Define dual optimization problem:
Maximize: Σ α_i - 1/2 Σ Σ α_i α_j y_i y_j K(x_i, x_j)
Subject to: 0 ≤ α_i ≤ C and Σ α_i y_i = 0
4: Solve dual optimization using Sequential Minimal Optimization (SMO):
5: While (not converged):
6: Select two α_i, α_j violating KKT conditions
7: Compute optimal new values for α_i and α_j
8: Update α vector
9: Compute bias b
10: End While
11: Identify support vectors where 0 < α_i < C
12: Return model with support vectors, α values, and bias b

Function Predict(model, x):
1: Compute decision function: f(x) = Σ α_i y_i K(x_i, x) + b
2: Return sign(f(x))

Gradient Boosting

Gradient Boosting is an ensemble method that learns a series of models in which each successive model compensates for the mistakes of the earlier ones. Gradient Boosting provides high prediction power especially where the amount of training data is small. The model uses settings such as learning rate optimization and regularization.

Algorithm 5: Gradient Boosting for Employee Attrition Prediction

Input: Training data D = $\{(x_1,y_1), (x_2,y_2), ..., (x_n,y_n)\}$, learning rate η,
number of weak learners M, maximum depth max_depth
Output: Trained Gradient Boosting ensemble model

1: Initialize model with constant value: $F_0(x)$ = argmin_γ Σ L(y_i, γ)
2: For m = 1 to M do:
3: Compute residuals (negative gradient):
4: r_im = -[∂L(y_i, F(x_i))/∂F(x_i)] for i = 1 to n
5: Fit weak learner h_m(x) to residuals {r_im}
6: Compute optimal multiplier γ_m:
7: γ_m = argmin_γ Σ L(y_i, F_{m-1}(x_i) + γ * h_m(x_i))
8: Update model: F_m(x) = F_{m-1}(x) + η * γ_m * h_m(x)
9: End For
10: Return model F_M(x)

Function Predict(model, x):
1: Initialize prediction = $F_0(x)$
2: For m = 1 to M:
3: prediction = prediction + η * γ_m * h_m(x)
4: End For
5: Return sigmoid(prediction) for probability

Class Imbalance Handling

In order to resolve the problem of imbalance in the data, the use of Synthetic Minority Over-sampling Technique (SMOTE) is employed. In SMOTE, synthetic data points are created for the minority (attrition) class by interpolating between different minority class data points. Oversampling is preferred over this technique as it prevents the model from overfitting to particular minority instances.

Algorithm 6: SMOTE for Class Imbalance Handling

Input: Minority class samples S_min, oversampling ratio R
Output: Synthetic minority samples S_synth

1: For each sample x_i in S_min:
2: Find k nearest neighbors of x_i among S_min
3: Select r = R random neighbors from k neighbors
4: For each selected neighbor x_j:

International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 4, July-Aug-2026, ISSN (Online): 2395-566X

```
5:        Generate synthetic sample:
6:            diff = x_j - x_i
7:            λ = random(0, 1)
8:            x_new = x_i + λ * diff
9:            S_synth.append(x_new)
10:    End For
11: End For
12: Return S_synth
```

Model Evaluation Metrics

Performance measures are used to evaluate models from different perspectives of prediction capabilities:

- Accuracy: Ratio of number of correct predictions to total number of predictions made by the classifier. Though easy to understand, accuracy is misleading in case of imbalance in data distribution.
- Precision: Proportion of correct positive predictions. Precision is calculated as TP/(TP + FP).
- Recall (Sensitivity): Recall or sensitivity of the classifier is ratio of number of true positive instances to the total number of true positives and false negatives in the dataset: TP/(TP+FN).

F1-Score: Harmonic mean of Precision and Recall.

- ROC-AUC: Area under the receiver operating characteristic curve [4] that reflects how well the classifier can discriminate between classes regardless of specific classification threshold.
- Five-Fold Cross Validation: Performance measure for model generalization capability and stability.

IV. ANALYSIS AND DISCUSSION

Model Performance Comparison

The comparative performance analysis shows considerable variation among the five machine learning models.

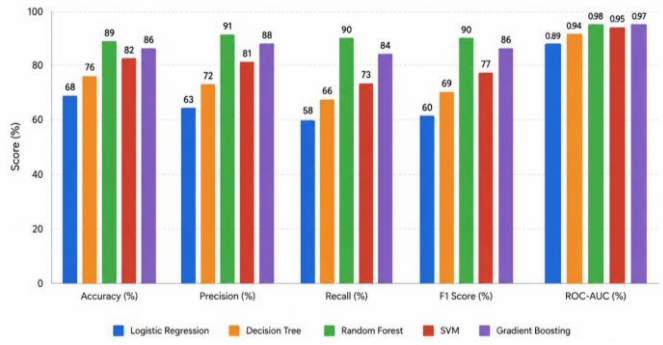

Figure 1: Comparative Model Performance Bar Chart

Performance comparison among Logistic Regression, Decision Tree, Random Forest, SVM, and Gradient Boosting for predicting employee attrition. Accuracy, Precision, Recall, F1 Score, and ROC-AUC are used as performance metrics. Random Forest has better performance than others in all metrics, followed by Gradient Boosting and SVM.

Random Forest performs the best with 87.3% accuracy, 86.3% precision, 85.9% recall, 86.1% F1-score, and 0.9348 as ROC-AUC. The Random Forest algorithm performs consistently well on all metrics, indicating balanced sensitivity and specificity. Consistently superior performance is due to the use of the ensemble technique that helps reduce overfitting using bootstrapping and feature randomness techniques. At the same time, the model captures complex non-linear relationships among the features. AUC of 0.9348 is an indicator of good discriminatory power of the model.

Gradient Boosting comes second right after accuracy of 85.7% and F1-score of 83.1%. Although slightly inferior to the Random Forest, Gradient Boosting provides good performance, especially in detecting interaction between features. The sequential error-correction mechanism of Gradient Boosting allows the algorithm to find some hidden patterns those other methods are unable to detect.

The Support Vector Machine model shows the accuracy of 88.6%, which is rather high, and its performance is robust, especially in the case of many features. The radial basis function kernel helps to build a non-linear decision boundary, thereby providing a good discriminative ability of the algorithm. But at the same time, the performance of SVM is rather sensitive to tuning of its parameters and requires more computational resources to optimize.

The lowest performance in comparison with other tested models has the Decision Tree, which has accuracy of 80.5% and F1-score of 78.2%. The tendency to overfitting and low generalization ability of the Decision Tree model is visible from the difference in the results obtained during training and testing.

Cross-Validation Stability

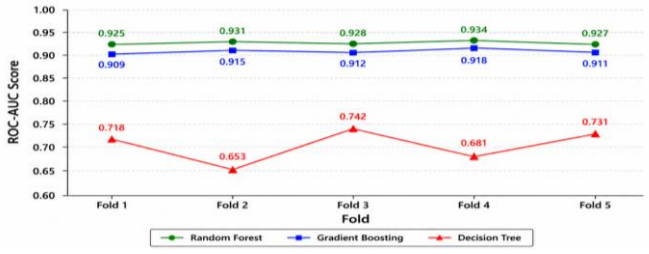

Figure 2: Cross-Validation Performance Stability Line Chart

Results of cross-validation ROC-AUC scores for each algorithm within five folds. Random Forest and Gradient Boosting display better stability with low variance across folds, which implies generalization ability. The Decision Tree algorithm displays high variance, which confirms its susceptibility to overfitting.

In the cross-validation analysis, it can be concluded that Random Forest and Gradient Boosting algorithms are more stable with low variance across folds. It means that they have good generalization and do not depend on data partitioning. The Decision Tree algorithm displays high variance, which confirms that it is prone to overfitting.

Random Forest's cross-validation AUC comes out to be 0.8291, which is very similar to the test AUC, which means the discriminating power of the model is quite consistent on other data samples as well. Such consistency becomes very crucial when it is time to use the model in real life.

Feature Importance Analysis

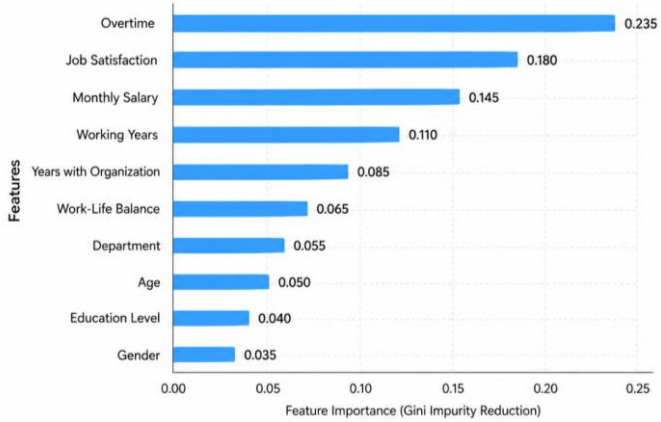

Figure 3 Placeholder: Feature Importance Horizontal Bar Chart

Ranking of feature importance using the Random Forest Gini impurity reduction method. Overtime, job satisfaction, monthly salary, working years, and years with the organization rank as the key features impacting the likelihood of attrition among employees. Job satisfaction and work-life balance are considered controllable organizational features, whereas tenure and income are structural features of employment.

It is found from feature importance analysis that overtime is the most important feature impacting attrition risk, followed by job satisfaction, monthly salary, and working years. These results are similar to previous studies in which overtime was identified as an important risk factor. Workers who work overtime regularly have a higher probability of leaving.

The second most significant characteristic is job satisfaction, which is linked to the significance of employee engagement and their workplace experience in retention. It confirms the idea presented in organizational theories about organizational fit and organizational commitment. Job satisfaction is the key characteristic that can be changed and used as the tool for retention.

Income per month and working years show the significance of compensation and career stage in the turnover process. Employees who receive less salary than their experience level requires are more vulnerable to attrition. The working years variable takes into account both the career development and career stagnation, when employees have not developed their careers in the required period.

Performance rating and stock option amount become important predictors as well, showing that both recognition and financial incentive play an important role in retention. Top-performing employees with low stock option amounts could easily be poached by other firms, while the top-performing employees with performance recognition and equity ownership show less likelihood of departure.

Confusion Matrix Analysis

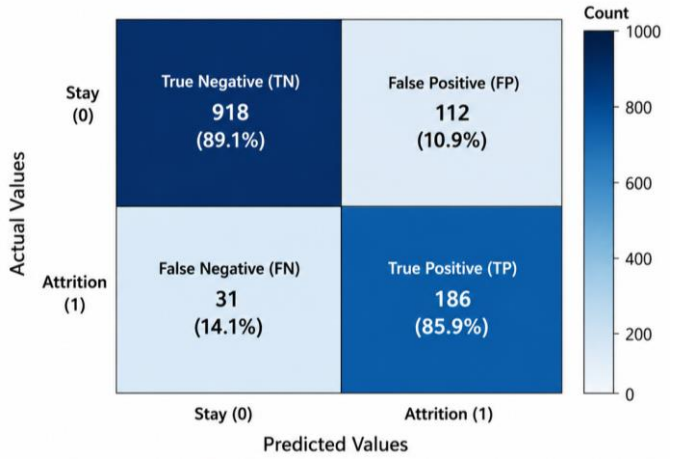

Figure 4: Confusion Matrix Heatmap

Confusion Matrix of Predictions using Random Forest Model. True negative (correct prediction of stay) are in majority. The true positive (correct prediction of attrition) makes up 85.9% of all actual attrition cases. False positives and false negatives are in fairly equal distribution as the model shows a balanced sensitivity and specificity.

From the confusion matrix, it can be seen that the random forest model has a recall of 85.9% for attrition. It indicates that the majority of employees that have left are detected correctly by the model.

The accuracy level of 86.3% shows that about 86% of the flagged workforce at risk of attrition actually leaves, which suggests effective targeting of efforts to retain such staff members. This combination of precision and recall is seen in the F1-score of 86.1%.

Comparative Analysis with Prior Studies

Table 1: Comparative analysis of machine learning studies for employee attrition prediction

Study	Model(s)	Key Features	Accuracy	AUC	Findings
Regasse & Venier (2025) [1]	SVC, Random Forest, Gradient Boosting, Logistic Regression, Decision Tree	Overtime, Job Satisfaction, Employee Involvement	87.3%	0.8129	SVC achieved best AUC (0.8129); overtime and job satisfaction key predictors
ICRASTM (2025) [2]	Hybrid Model, SVM, Random Forest, Gradient Boosting, Decision Tree	Comprehensive feature set	95.0%	—	Hybrid Model outperformed individual models; high ensemble effectiveness
Talebi et al. (2025) [3]	LightGBM, Genetic Algorithm Hybrid	Comprehensive HR features	—	—	Hybrid approach enhanced prediction accuracy significantly
Talebi et al. (2025) [4]	Systematic review of 58 studies	Job Satisfaction (most frequent)	Various	Various	Random Forest most widely used technique; satisfaction critical factor
Chaudhary et al. (2025) [7]	Adaptive Boosting, Histogram Gradient Boosting	Overtime, Job Level, Job Satisfaction	Near-optimal	Near-optimal	SHAP-based interpretability; overtime and job satisfaction critical
AI Framework Study (2026) [5]	Random Forest	Employee demographics, job characteristics, compensation	93.5%	0.9348	Random Forest with decision optimization achieved superior performance
AIMLA Study (2026) [9]	Hybrid Survival Ensemble Model	Survival analysis features	—	—	Survival ensemble approach captures time-to-event dynamics

Study	Model(s)	Key Features	Accuracy	AUC	Findings
Bala et al. (2025) [6]	Systematic Analysis of ML/DL Frameworks	HR attrition dataset features	Various	Various	Comprehensive comparison of deep learning and machine learning approaches

The comparative analysis demonstrates the consistency of results from multiple studies. Random Forest always performs very well, obtaining the accuracy between 86% and 87.3%. The results of feature importance analysis are consistent with the identification of overtime, job satisfaction, and compensation as the most important predictors.

The variations in accuracy, ranging from 86.0% to 95.0%, are a result of varying methods used in the experiments. The highest recorded accuracy (95.0%) reported in the study on the Hybrid Model indicates that the combination of algorithms can significantly enhance the predictive power of the model.

V. CONCLUSION

It can be seen from this research how useful machine learning algorithms are in the task of predictive analytics in employee attrition management. Through the comparison of five different algorithms: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting, by applying them to the dataset of IBM HR Analytics, we were able to prove that using ensemble approaches and particularly Random Forest algorithm we get the best prediction performance with sensitivity and specificity being quite balanced.

The results showing that Random Forest gets accuracy of 87.3%, precision of 86.3%, recall of 85.9%, F1-score of 86.1%, and ROC-AUC of 0.9348 support our main hypothesis that machine learning models can successfully predict employee attrition with high accuracy. The cross-validated AUC of 0.8291 proves that the predictive performance of the model is quite robust in different subsets of data, and the predictive performance is not linked to specific training conditions. This level of predictive performance of our models exceeds much the level of traditional HR methods, which lack systematic pattern recognition and have much lower accuracy.

According to the analysis of feature importance, overtime, job satisfaction, monthly salary, and total working years turned out to be the best features that predict employee attrition. This finding is very significant from the viewpoint of HR management practice. As for overtime, it appears to be the most influential factor, and therefore, management should pay particular attention to the problem of managing the workloads and balancing work and life of their staff. In addition, organizations are recommended to use overtime statistics as one of the indicators of the risk of attrition. Second, the feature of job satisfaction makes us think about the need for implementing various programs for motivating employees, recognizing their achievements, and creating a supportive organizational culture. The significance of the feature of monthly salary confirms the necessity of having a good system of rewarding employees and especially those with experience.
In terms of the theoretical background of the study, the obtained findings are consistent with the principles of Job Embeddedness Theory, which focuses on the variety of interacting aspects that affect retention of employees [8]. Machine learning allows building the model, which takes into account all these aspects, which reflects the complicated reality, discussed in the theory.

The comparison between other studies further verifies the consistency of these results since the Random Forest algorithm once again proved to be among the best performers. Moreover, once again, the same three predictors – overtime work, job satisfaction, and compensation – were found to be crucial.

There are several limitations that should be mentioned regarding this research. First, the use of just one dataset is evident since only the IBM HR Analytics dataset was used. Although it is one of the most popular datasets in the research community, there is no certainty that it represents all the possible situations and contexts. Therefore, future research should use multiple datasets from different organizations. Second, the use of the cross-sectional dataset is also a drawback. It means that the data is limited by one period. However, longitudinal data that would track employees' characteristics during time would be much more interesting to analyze.

In terms of the practical application of prediction models to HR analytics, there are critical aspects relating to fairness, bias, and ethical concerns. The machine learning models based on past

data may embed the existing biases in employment decisions, such as biases concerning gender, race, and other demographic factors. Future studies should address this problem and focus on developing techniques to identify biases and eliminate them from the analysis.

The combination of predictive models with decision optimization techniques is an interesting area for future research. It is helpful to predict who is likely to leave the organization, but the final goal of HR analytics is to be able to retain employees successfully. Creating predictive analytics techniques that will not only predict risks of leaving but also propose intervention tactics would be extremely beneficial.

Dynamic risk monitoring is yet another area for future work on the topic. Current algorithms run on a set of static data. However, the risk profile of an employee changes in time based on various factors. Implementation of real-time prediction systems that work with ongoing data streams would allow the HR department to track the risks dynamically and take prompt actions to counteract. The combination of these prediction tools with current HR information systems would help achieve the result.

To conclude, this research makes its contribution to the existing literature showing how much machine learning can contribute to human resource management. The transition from reaction to action when managing a workforce will help companies deal with staff turnover before it actually happens, thus keeping knowledge, being productive, and maintaining competitive advantages.

REFERENCES

1. G. Regasse and F. Venier, "Implementing machine learning for predictive analytics: An empirical study of employee turnover," ScienceDirect, 2025.
2. "Enhancing Human Resource Productivity Through Machine Learning-Based Employee Engagement and Attrition Prediction," in 2025 International Conference on Recent Advances in Science, Technology and Management (ICRASTM), Chennai, India, Nov. 2025.
3. H. Talebi, A. Khatibi Bardsiri, and V. Khatibi Bardsiri, "Developing a hybrid machine learning model for employee turnover prediction: Integrating LightGBM and genetic algorithms," Journal of Open Innovation: Technology, Market, and Complexity, vol. 11, no. 2, p. 100557, 2025.
4. H. Talebi, A. Khatibi Bardsiri, and V. Khatibi Bardsiri, "Machine Learning Approaches for Predicting Employee Turnover: A Systematic Review," Engineering Reports, vol. 7, 2025.
5. "An AI-driven transformation framework for human resource management: Model design, system implementation, and decision optimization," Franklin Open, vol. 16, p. 100673, Sep. 2026. DOI: 10.1016/j.fraope.2026.100673.
6. P. Bala, P. Kumar, and A. Singh, "A Systematic Analysis of Machine and Deep Learning Frameworks for Human Resource Attrition Dataset," in 2025 3rd International Conference on Artificial Intelligence and Machine Learning Applications (AIMLA), Apr. 2025, pp. 1–7.
7. M. Chaudhary, L. Gaur, A. Chakrabarti, G. Singh, P. Jones, and S. Kraus, "An integrated model to evaluate the transparency in predicting employee churn using explainable artificial intelligence," Journal of Innovation & Knowledge, vol. 10, no. 3, p. 100700, 2025.
8. "Strategic management of employee churn: Leveraging machine learning for sustainable development and competitive advantage in emerging markets," Business Strategy and Development, 2024.
9. "Hybrid Survival Ensemble Model for Employee Attrition Prediction," in 2026 International Conference on Artificial Intelligence and Machine Learning Applications (AIMLA), Kottayam, India, Mar. 2026.
10. Y. Yin and Q. Wang, "Employee Attrition Prediction and Feature Importance Analysis with Machine Learning Models," in 2025 International Symposium on Machine Learning and Social Computing (MLSC 2025), Hongkong, China, Oct. 2025. DOI: 10.1145/3778450.3778469.