Journal of Global Research in Electronics and Communication
Volume 2, No. 3, March 2026
Available Online at: www.jgrec.info
RESEARCHPAPER

# Predictive Analytics for Employee Retention Using Workday HCM Data

Dr. Chintal Kumar Patel
*Associate Professor, CSE*
*Geetanjali Institute of Technical Studies*
chintal.patel@gits.ac.in

***Abstract*—The retention of employees is an important aspect of organizational stability and performance because high rates of employee attrition mean that the organization incurs financial losses, decreased productivity, and institutional knowledge loss. Predictive analytics is a proactive way of comprehending and predicting employee turnover in using demographic, behavioral, and work-related information. A practical model for predicting employee retention is the goal of this research, which applies machine learning techniques to the 1,470-item Employee Attrition and Performance dataset from IBM HR Analytics. The dataset contains 35 different qualities. Data preprocessing such as the processing of missing values, outlier processing, label encoding, normalization, and feature selection were done comprehensively to guarantee data quality. The problem of the imbalance of classes was addressed by creating balanced training data with the SMOTE technique. The postulated model employs the Extra Trees Classifier (ETC) that is an ensemble learning algorithm that constructs numerous randomized decision trees to improve the predictive capability and strength. The ETC model was doing exceptionally well with an accuracy (ACC) of 99.1%, precision (PRE) of 98.6%, recall (REC) of 99.8% and F1-score(F1) of 99.2%. These results affirm the applicability and relevance of the model in employee retention prediction as it provides the organization with practical data of how to optimize HR practices and support businesses to make informed judgments concerning workforce management.**

***Keywords*—HR Analytics, Employee Retention Strategies, Predictive Turnover Analysis, Retention Decision-Making, Employee Performance, Predictive Analytics, Data-Driven HR, Strategic Decision-Making.***

## I. INTRODUCTION

One of the most pressing problems in modern management is retaining key employees. High employee turnover has both direct and indirect costs that involve the cost of recruitment, training and loss of productivity as well as organizational knowledge [1]. One of the biggest challenges, which industries cannot forget is the retention of employees. The adverse outcome of high attrition rates is that it influences the performance of the organization since it increases the expenses of recruitment, training and lost knowledge [2][3]. Employers across a wide range of sectors have the same challenge: keeping good staff. Huge expenses, such as those associated with recruiting, selecting, and training new personnel, as well as the potential loss of institutional knowledge and low morale among the surviving workforce, can result from high staff turnover rates [4]. These problems have made human resource (HR) professionals focus on employee turnover comprehension and prediction [5][6]. Traditional approaches in managing employee turnover, including satisfaction surveys and exit interview, are often ineffective in information and are mostly reactive, not proactive [7][8].

A comprehensive approach to understanding and anticipating employee attrition is predictive analytics, which involves the use of behavioral and demographic data [9]. Despite the fact that demographic data may be used in order to sub-categorize the workforce and determine the general risk factors, behavioral data gives a more detailed picture of what is occurring to some of the employees and possible risk of turnover. Organizations can also achieve more accurate and useful prediction models by combining more than one source of data [10][11][12]. HRM is a strategic role that can assist an organization to be more efficient, innovatable and competitive. Recruiting, retaining and maximizing on the performance of its employees to possess a productive and a motivated work force are some of its key functions. Companies that fail to adequately deal with such factors tend to spend a lot in terms of costs of turnover, poor productivity and failure to sustain sustained growth [13][14][15][16].

The development of predictive analytics and AI has transformed the human resources (HR) management strategy by enabling more rational and proactive attitude towards retaining employees [17][18][19]. ML and AI are new groundbreaking technologies that can be applied in HRM to make predictive analytics and automation on significant HR functions [20]. AI refers to both the theory and practice of creating and using computer systems to perform cognitive functions normally associated with humans, such as decision-making and pattern recognition. ML is a branch of AI that allows computers to learn new things on their own without human intervention or programming by drawing conclusions from existing data [21]. These technologies are being used to make plans using real-time data, predict staff attrition, and speed up the hiring process [22][23][24][25].

### *A. Motivation and Contribution*

The rising employee turnover rate and expenses has necessitated new workforce management approaches. The traditional HR practices, exit interviews and surveys, not probably provide any actionable data or make predictions on attrition. Using predictive analytics and machine learning, one might be able to predict employee retention more accurately by utilizing detailed HR data sets that comprises demographic and behavioral and performance data. Being aware of the risks of attrition that can be involved, as well as what can be the causes of the same, organizations can implement the required interventions in a timely manner to increase engagement and stability. Utilizing state-of-the-art AI techniques transforms the HR management system into a data-driven, proactive operation, thereby boosting organizations' efficiency,

productivity, and competitive advantages. This study has made a number of important contributions as follows:

- **Extra Trees Classifier:** It is used that can be utilized to predict employee retention and it can handle complex and high-dimensional HR data. The model is more generalization friendly and less susceptible to overfitting as compared to more traditional methods.
- **Preprocessing & Balancing:** Supports full preprocessing procedures such as feature selection, feature normalization, and class imbalance with SMOTE. These procedures would guarantee relevant, balanced and high-quality data to train models.
- **Pattern Recognition:** Extracts complicated relationships and behavior patterns in employee data. This allows proper determination of factors affecting retention and attrition.
- **HR Decision Support:** Offers viable recommendations to strategic human resource management. The predictive system also helps organizations in terms of planning and retaining workforce.

### *B. Novelty and Justification*

The originality of the study is that the Extra Trees Classifier (ETC) is applied as a predictor of employee retention that is more precise on complicated HR information than the customary and conventional deep learning frameworks. ETC reduces overfitting when constructing the trees and is robust and efficient. Issues of class imbalance have also been taken care of in the methodology of class imbalance using SMOTE and other preprocessing and feature selection to ensure that quality of the data is not compromised. A combination of these strategies enables the research to present a reliable, scalable, and viable predictive analytics tool to HR management, where organizations can understand employee behavior that can enable them to make effective decisions. Moreover, the ETC model can identify complicated relationships between employee attributes and offers greater information about the patterns of staff retention. The ability promotes HR planning strategy and provides the option to implement specific interventions to increase workforce stability and organizational development.

### *C. Organization of the Paper*

The paper is organized into the following sections: Survey the research on predicting employee retention in Section II. Section III provides a detailed description of the dataset, preprocessing, and proposed model. Section IV presents the findings from the experiments as well as a comparison of the various models. In Section V, the study concludes by summarizing the main findings and outlining new directions for research.

## II. LITERATURE REVIEW

A thorough review and analysis of significant research studies on employee retention prediction, as summarized in Table I, helps guide and strengthen the development of this study.

Ramya and Sanjay (2025) introduced a CNN-LSTM network hybrid deep learning model. The deep learning model can detect both structural and additional temporal patterns in the employee data; it uses CNN layers and is further improved using LSTM layers. In order to evaluate the DL model's predictive potential, used confusion matrices and performance metrics such as an ACC score of 0.87 and ROC-AUC of 0.54 [26].

Singh et al. (2025) supplied an HRMS that includes AI and blockchain technology. The system's smart contracts make record-keeping easier, and ML algorithms help with things like staff performance and retention forecasting. With an AUC-ROC of 0.95, an ACC of 92%, a PRE of 93%, a REC of 90%, and an F1 of 91%, the suggested system beats baseline models like Decision Trees and LR in important performance parameters. The integration of blockchain ensures the integrity of employee records, while AI models predict HR outcomes with higher PRE [27].

G et al. (2024) aim to Initially, an IBM-HR dataset is considered as the input for preprocessing, which is performed using Standard Scalar to preserve data distributions. Later, Gradient Boosting-Binary Logistic Regression (GB-BLR) is employed for classification to analyze and predict attrition of employees. The proposed GB-BLR method obtained better performance when compared with other existing Extra Trees Classifier (ETC) method in terms of ACC, PRE, REC, and f1 of 97.48%, 98.57%, 96.25%, and 98.77% respectively [28].

Ismail Al-Alawi and Ahmed Aljawder (2024) created a machine learning prediction classifier for staff promotions using a mixed ensemble approach. Developing a model that can accurately forecast employee promotions while reducing the impact of data discrepancies is the primary goal. Multiple blended ensemble models were trained using this meta-data. An area under the curve (AUC) of 80%, an accuracy rate of 97%, and an ACC rate of 94% were the best metrics for the blended ensemble Gradient boosting model (BEns_GradBoost) evaluation, according to the research [29].

Sharma and Sharma (2023) developed a more reliable and secure model by making use of several decision trees. Logistic regression is a statistical method that looks at how different independent factors (such gender, age, and salary) relate to a binary dependent variable (like employee turnover). The logistic regression model was beaten out by the RF classifier, which achieved an efficiency of 85%. In comparison to the logistic regression model, the random forest classifier model achieved better results in PRE, REC, and F1 [30].

Kaur and Dogra (2022) positively impacts a company's success. A high retention rate is a key indicator of a successful organization. The machine learning model was built using methods like SVM, Ensemble with Boosted Tree (KNN), and DT. Obtaining a well-trained model requires manually adjusting the dataset's feature value types according to the model's standards. The resulting model provided a 98% ACC rate [31].

Yahia, Hlel and Colomo-Palacios, (2021) created a plan to use people analytics to forecast employee turnover that puts an emphasis on data quality rather than quantity, thereby putting deep data ahead of big data. The second thing is that this approach to turnover prediction uses ML, DL, and EL to its foundation. It has been evaluated on a large-scale simulated HR dataset, a medium-sized one, and a real-life one with four hundred and fifty responses. The approach outperforms the prior solutions on all three datasets, with accuracies of 0.96, 0.98, and 0.99 respectively [32].

Marvin, Jackson, and Alam (2021) presents models that successfully forecast the possibility of candidate retention before training and evaluates various machine learning

 

classifiers that could generate these kinds of predictions. To describe the outcomes of the algorithms, traditional metrics are used. When it came to ACC percentages, the RF Classifier was on top. It achieved 99.1% on the training dataset, 84.6% on the testing dataset, and 91.8% on the entire dataset [33].

Although various ML and DL models have improved employee analytics, gaps remain in interpretability, handling imbalanced or heterogeneous data, and real-time adaptability. Few studies address the integration of AI with blockchain for data integrity and ethical HR management. Most research relies on specific datasets, limiting generalizability, and lacks standardized comparisons of models for predictive power, efficiency, and scalability. There is a need for frameworks that combine high ACC with interpretability, fairness, and practical applicability across diverse HR contexts.

TABLE I. RECENT STUDIES ON EMPLOYEE RETENTION PREDICTION USING WORKDAY HCM DATA

| Author | Dataset used | Proposed Work | Results | Key Findings | Limitations & Future Work |
|---|---|---|---|---|---|
| Ramya & Sanjay (2025) | Employee dataset | Hybrid deep learning model integrating CNN and LSTM layers | Accuracy: 0.87, ROC-AUC: 0.54 | CNN-LSTM captures both structural and temporal patterns in employee data | Moderate ROC-AUC; further improvement in predictive power needed |
| Singh et al. (2025) | HR dataset | AI and blockchain-based HR management system with ML models & smart contracts | Accuracy: 92%, Precision: 93%, Recall: 90%, F1-Score: 91%, AUC-ROC: 0.95 | AI improves retention and performance prediction; blockchain ensures data integrity | High system complexity; integration with other HR systems can be explored |
| G et al. (2024) | IBM-HR dataset | Gradient Boosting-Binary Logistic Regression (GB-BLR) is a technique for forecasting staff turnover. | Accuracy: 97.48%, Precision: 98.57%, Recall: 96.25%, F1-score: 98.77% | GB-BLR outperforms Extra Trees Classifier | Limited dataset diversity; further testing on larger datasets suggested |
| Ismail Al-Alawi & Ahmed Aljawder (2024) | Employee promotion metadata | Blended ensemble ML models, including BEns_GradBoost | Accuracy: 94%, Precision: 97%, AUC: 0.80 | Gradient boosting ensemble effective for promotion prediction | Imbalanced data may still affect results; more robust handling needed |
| Sharma & Sharma (2023) | Employee attrition dataset | Random Forest classifier and comparison with Logistic Regression | Accuracy: 87% (RF) vs 85% (LR) | Random Forest outperforms Logistic Regression in all metrics | Dataset may not cover all HR scenarios; further generalization needed |
| Kaur & Dogra (2022) | Employee dataset | ML models: Decision Tree, Boosted Ensemble, KNN, SVM | Accuracy: 98% | High retention prediction accuracy using multiple ML algorithms | Dataset feature engineering is manual; automation could improve efficiency |
| Yahia, Hlel & Colomo-Palacios (2021) | Simulated HR datasets & small real dataset (450 responses) | Machine, deep, and ensemble learning for attrition prediction | Accuracy: 0.96, 0.98, 0.99 (large, medium, small datasets) | Focus on data quality improves prediction accuracy | Small real dataset limits generalization; larger real datasets needed |
| Marvin, Jackson & Alam (2021) | Candidate HR dataset | ML classifiers to predict retention probability | Accuracy: 99.1% (training), 84.6% (testing), 91.8% (overall) | Random Forest effective in retention prediction | Testing accuracy lower than training; risk of overfitting |

## III. RESEARCH METHODOLOGY

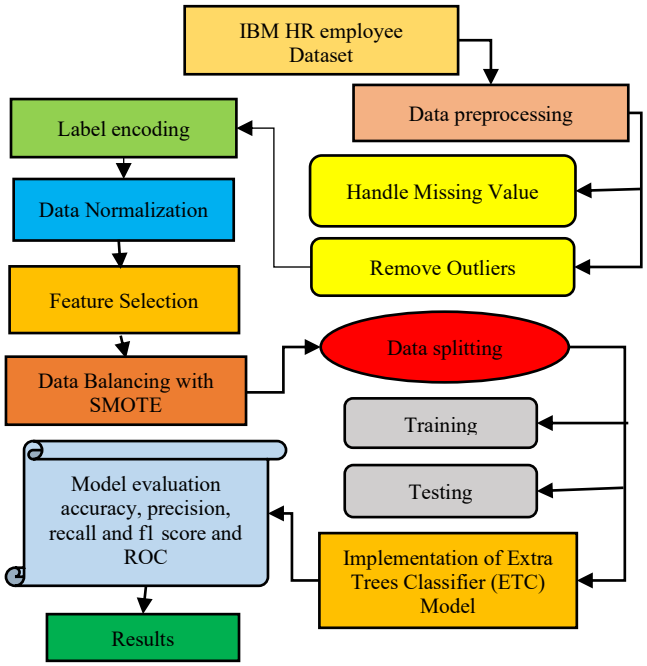


Fig. 1. Proposed flowchart for Employee Retention Prediction

In this study, the IBM HR Analytics Employee Attrition & Performance dataset was used to conduct predictive analytics on employee retention using data from Workday HCM (As shown in Fig. 1). Addressing missing values, removing outliers, labelling categorical features, and normalizing numerical variables using min-max scaling were all part of the data preprocessing that was necessary to optimize the model's performance. To achieve a balanced dataset, used feature selection to keep the most important qualities and the SMOTE technique to create synthetic minority class examples to fix the class imbalance. A training set and a testing set were created from the stratified data in order to keep the proportions of each class. The Extra Trees Classifier (ETC) utilized in the prediction model was subjected to a number of performance measures for evaluation, including ACC, PRE, REC, F1, confusion matrix, and ROC analysis. A comprehensive assessment of the model's ability to predict employee retention and attrition might be conducted as a result.

Here is a detailed explanation of each step in the proposed plan:

### *A. Data Gathering and Analysis*

A big number of employee records with various attributes are contained in the IBM HR Analytics Employee Attrition & Performance dataset, which is used in this study. There are 1,470 records in the collection, and each one has 35 unique

properties that cover things like demographics, job-related variables, and satisfaction measures. Data visualizations such as bar plots and heatmaps were used to examine attack distribution, feature correlations etc., are given below:

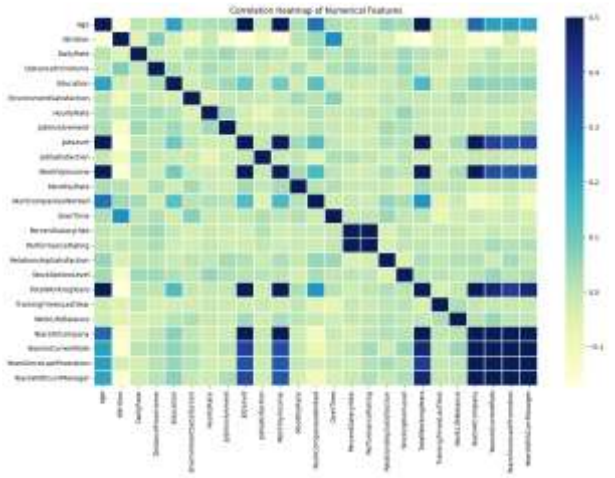

Fig. 2. Correlation Heatmap of Numerical Features

Fig. 2 presents a correlation heatmap, a matrix plot that visualizes the pairwise correlation coefficients between all numerical features in a dataset. The degree and direction of a link between two variables are shown by the color intensity and shade of each square. A weak or nonexistent correlation is shown by a lighter color, such white or yellow, whereas a strong positive correlation is denoted by darker hues of blue and a negative correlation by darker shades of green-blue. With a perfect positive correlation of 1.0, the diagonal line—the darkest blue—represents the correlation of each feature with itself. This plot is a useful method to rapidly establish features with very correlated features, which may be critical to feature selection and the interpretation of the data structure. As an example, the dark squares indicate that the pairs such as the MonthlyIncome and JobLevel and also the TotalWorkingYears and YearsAtCompany are strongly correlated and thus tend to move together.

Fig. 3. Data Visualization

Fig. 3 shows a row of histograms which are the distributions of various features of a dataset, probably concerning employee data. These grid structures can allow one to have a rapid visual analysis of the type of data. As an illustration, some variables like Age, MonthlyIncome and TotalWorkingYears are heavily skewed towards the right or even slightly normal, with the majority of the values concentrated at the lower end of the value range. Quite the contrary discrete variables or categorical variables such as: Education, JobLevel and JobSatisfaction are in the shape of bar charts showing frequency of each category. Other variables, such as "EmployeeCount" and "StandardHours" seem to be fixed values, since their histograms are represented by a single bar only. Important for initial data exploration and understanding, the charts provide a full explanation of the data structure and distribution of feature values.

### *B. Data Pre-Processing*

Data preparation utilized the IBM HR Employee dataset, which included concatenation, cleaning of the data and feature engineering. Preprocessing was done to handle missing values, remove outliers, and standardized and normalized data. The most important preprocessing steps are as follows:

- **Handle missing value:** Data analysis and machine learning models rely on accurate and reliable handling of missing variables. The main reasons are: Better Model ACC: The missing values are tackled to prevent wrong predictions and enhance the model performance.
- **Remove Outliers:** The term "data lopping" refers to the practice of finding outliers in data and either deleting or changing them so they don't impact statistical analysis or ML model performance.
- **Label Encoding:** Label Encoding is a machine learning data preprocessing method of converting categorical data into a numerical form. This is needed as most machine learning algorithms take numerical input and are not capable of directly dealing with non-numeric and categorical features.

### *C. Min-Max Normalization*

The min-max technique was used to normalize the records and to make the values to lie within a range of 0 and 1. This was done with the view to optimizing the performance of the used classifiers and preventing the influence of outliers. According to the mathematical Equation (1) normalization was carried out as follows:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

The feature's initial value X, its normalized value $X'$ , its minimum value $X_{min}$ and its highest value ($X_{max}$

### *D. Feature Selection*

Feature selection is a method for improving machine learning system performance, streamlining the model, and making it easier to understand by identifying and using the most important characteristics from a dataset. Feature selection, by excluding irrelevant or unnecessary features, can lessen the likelihood of overfitting, lower computing costs, and, in certain instances, provide more efficient and accurate predictions. Data preprocessing, also known as feature selection, is the first step in building a machine learning model from a dataset. It involves finding and selecting a subset of the most relevant features, which are variables or properties.

### *E. Data balancing using SMOTE*

Data balancing techniques remove the issue of unbalanced class distribution, which tends to cause the majority class to be well-learned, and the minority class to end up poor in terms

of classification ACC. SMOTE is a popular approach to the imbalance in classes of a dataset especially in machine learning. It operates by producing artificial examples of the minority group, as opposed to merely cloning the examples that are available, in order to get a more equal distribution of classes.

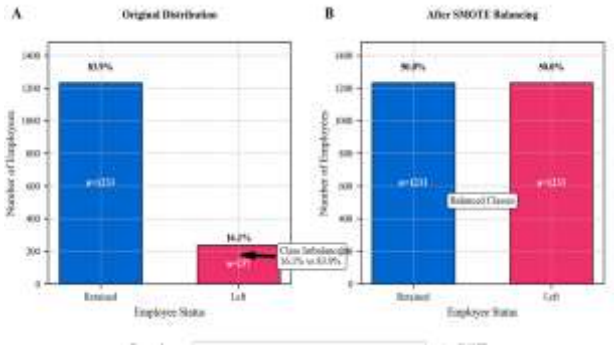


Fig. 4. Class Distribution and SMOTE Balancing Effect

Fig. 4 shows the distribution of employee status before and after the implementation of the SMOTE balancing technique. Class imbalance is evident in original dataset (Panel A) as the proportion of those retained (n=1233) is higher than those who left (n=237) at 83.9% and 16.1% respectively. This disproportion may skew the machine learning models towards the majority group. Following SMOTE (Panel B) the dataset has come to a balance, where both retained and left employees are adjusted to equal representation, 50.8% each (n=1233). This balancing is to make a fairer representation of the two classes to enhance the performance and reliability of predictive modeling.

### *F. Data Splitting*

The dataset was divided into two parts: training and testing, using an 80:20 split. This was done to make sure that the class distribution in both parts was the same as in the original dataset.

### *G. Proposed Extra Trees Classifier (ETC) Model*

ETC is an ensemble method for ML that constructs a large number of decision trees without pruning and then merges their predictions to increase their accuracy and robustness. Similar to Random Forest, ETC aggregates multiple trees but introduces greater randomness by selecting split thresholds at random rather than searching for the optimal split, which reduces variance, lowers computational cost, and mitigates overfitting. The training data is used to build a large number of bagged decision tree samples using randomly selected decision rules. Afterwards, the final class prediction is decided by means of a majority vote across all trees. Ideal for classification and regression tasks, ETC aggregates results from de-correlated trees to boost predictive ACC; this is particularly true when dealing with high-dimensional data or complex decision boundaries. By following the method for each case, and can determine the entropy of ETC, which accounts for the information gain from tree splitting. It is possible to express the computed entropy for ETC using Equation (2):

$$\text{Entropy(S)} = \sum_{i=1}^{c} -p_1 log_2(p_i) \quad (2)$$

Where:

- SSS is the dataset for which are calculating entropy.
- ccc is the number of classes in the dataset.
- pip_ipi is the proportion (probability) of samples in class iii within the dataset SSS.

### *H. Evaluation Metrics*

The proposed layout was tested using several metrics to measure its performance. The results of the categorization were displayed in a confusion matrix that highlighted the right and wrong predictions for each class. TN, FP, TP, and FN were calculated from this matrix. The following is a summary of the steps required to calculate important assessment metrics using these values: ACC, PRE, REC, and F1:

**Accuracy:** The trained model's accuracy rate relative to the total number of instances in the dataset and the input samples. This is expressed as Equation (3)-

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+Fp+TN+FN}} \quad (3)$$

**Precision:** The accuracy rate of a model's predictions (ACC) is the ratio of the total number of positive examples to the percentage of positive cases that were correctly predicted. ACC shows the accuracy with which the classifier anticipates positive classes is denoted as Equation (4)-

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \quad (4)$$

**Recall:** This measure represents the proportion of positive events that were correctly predicted relative to all instances that were expected to be positive. It can be expressed mathematically as Equation (5)-

$$Recall = \frac{\text{TP}}{TP+FN} \quad (5)$$

**F1 score:** It is a combination of the harmonic mean of PRE and REC, that is, it helps to balance REC and PRE. Its range is [0, 1]. Mathematically, it is given as Equation (6)-

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

**Receiver Operating Characteristic Curve (ROC):** The ROC plots, for a set of decision cut-off points, the ratio of accurate to incorrect positive classifications. TPR is also known as sensitivity or REC, and FPR is 1-specificity.

## IV. Results and Discussion

The experimental setup and training/testing data of the proposed model are shown in this section to demonstrate its overall and evaluation performance. Here takes a look at the outcomes and assessments of the proposed study. Analyzing and designing systems A computer system with an AMD EPYC 7B12 model, 13 GB of random-access memory (RAM), 2249.998 MHz central processing unit (CPU), and 512 KB of cache size was used for all the studies. The proposed model was evaluated using the primary key performance indicators, including ACC, PRE, REC, and F1, and it was trained on the IBM HR Employee dataset. The results are displayed in Table II. The example of the results of the classification of the proposed ETC model of employee retention prediction based on the IBM HR Employee dataset proves its outstanding performance. This model had an overall ACC of 99.1 which means that the overall predictions are very reliable. It also achieved a PRE of 98.6 and presented a strong ability to correctly identify employees who have been predicted to remain as well as the REC of 99.8, which indicates its ability to capture almost all the real cases of retention. The F1 of 99.2% also supports the fact that the

model is well-balanced and robust, and it is an extremely useful tool to predict employee retention in this dataset.

TABLE II. CLASSIFICATION RESULTS OF THE PROPOSED MODEL FOR EMPLOYEE RETENTION PREDICTION USING THE IBM HR EMPLOYEE DATASET

| Matrix | Extra Trees Classifier (ETC) Model |
|---|---|
| Accuracy | 99.1 |
| Precision | 98.6 |
| Recall | 99.8 |
| F1-score | 99.2 |

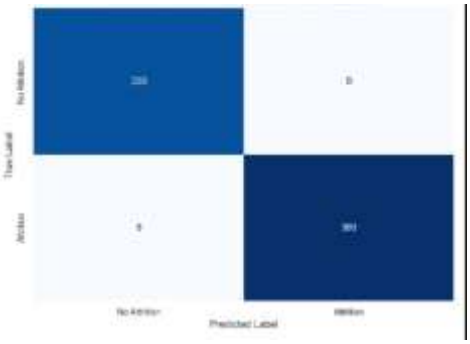

Fig. 5. Confusion Matrix for the ETC Model

The accuracy of the proposed model in predicting employee turnover, as measured by the IBM HR Employee dataset and the confusion matrix, is illustrated in Fig. 5. The original forecasts included 338 employees who remained and 386 employees who departed, with 338 falling into the "No Attrition" category. The number of employees misclassified in every category was only 8 which means that there were few errors in prediction. This high ACC of correct classification indicates the high ACC, PRE and REC of the model and shows how effective the model is in differentiating between retained and leaving employees.

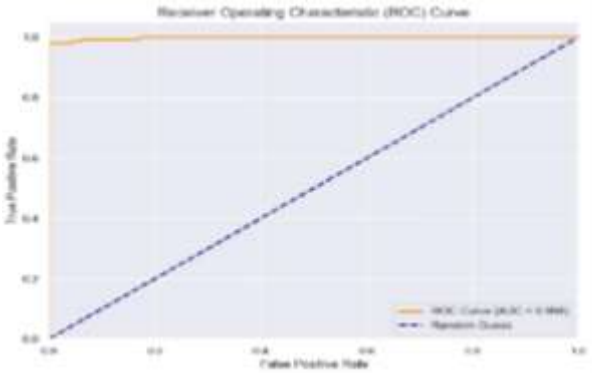

Fig. 6. ROC Curve for ETC Model

Fig. 6 shows the ROC curve, a common tool for evaluating classification model performance. A solid orange line shows the model's performance by plotting the TPR (y-axis) against the FPR (x-axis) at different threshold settings. "Random Guess," the dotted blue line that serves as a baseline, is a model that does not have any predictive power. With an AUC of 0.998, the model is virtually perfect. A testament to the model's exceptional predictive power is its near-perfect ACC in discriminating between positive and negative classes across all possible classification levels.

### *A. Comparative Analysis*

The suggested ETC model's utility was evaluated by comparing its ACC to other existing models; the outcomes of this comparison are shown in Table III. The performance of several ML and DL models for retention prediction is clearly seen when examining the IBM HR Employee dataset. Logistic Regression (LR) achieved a modest ACC of 57.34%, with lower PRE (56.68%), REC (65.92%), and F1 (63.44%), indicating limited predictive capability. Support Vector Machine (SVM) performed considerably better, achieving 87% across all metrics, demonstrating balanced and reliable predictions. The Deep Neural Network (DNN) showed strong ACC at 86.7% but lagged in PRE (73.5%), REC (66.5%), and F1 (68.9%), suggesting some challenges in correctly identifying retained or churned employees. Based on its impressive performance in ACC (99.1%), PRE (98.6%), REC (99.8%), and F1 (99.2%), the Extra Trees Classifier (ETC) emerged as the top model in this dataset for predicting employee retention.

TABLE III. COMPARISON OF DIFFERENT MACHINE LEARNING AND DEEP LEARNING MODELS FOR EMPLOYEE RETENTION PREDICTION ON IBM HR EMPLOYEE DATASET

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| LR[34] | 57.34 | 56.68 | 65.92 | 63.44 |
| SVM[35] | 87 | 87 | 87 | 87 |
| DNN[36] | 86.7 | 73.5 | 66.5 | 68.9 |
| ETC | 99.1 | 98.6 | 99.8 | 99.2 |

The strengths of the Extra Trees Classifier (ETC) model presented include the fact that it has been applied in predicting employee retention with high ACC of 99.1%. Its higher performance guarantees good predictions that are accurate and not prone to false positives or false negatives. The model can be robust to manage complex patterns and interactions within the IBM HR Employee dataset and hence gives a thorough understanding of employee retention patterns. Besides, the ETC model can be an effective and scalable tool that organizations can use to proactively control workforce retention because it is efficient in handling large data volumes.

## V. CONCLUSION AND FUTURE STUDY

The satisfaction and trust of the stakeholders determine the success of an organization. Employees, as one of the most significant assets an organization has, are a vital part of the uplifting of an organization. A company that has a greater retention ratio be a successful one as far as its goals are concerned. A company's operations and bottom line can take a hit when a competent employee leaves because of how expensive it is to find a suitable replacement. According to the experimental outcomes, the comparative analysis of various predictive employee retention models proves that the suggested ETC is far more effective than the traditional and DL models. Although the ACC of the LR was low (57.34%), SVM (87%) and DNN (86.7%) had moderate ACC, the ETC model had an excellent ACC of 99.1%. It implies that the ETC is an effective reflection of the intricate patterns and connections of the HR data and, therefore, is a highly efficient and reliable instrument to forecast employee retention and turnover. The superior performance of ETC can be highlighted by its ability to promote strategic HR decision-making by means of predictive analytics. More granular information that may be of interest in further studies on the impact of employee happiness includes employee survey reports and data on individual benefit plans.

## REFERENCES

[1] P. Badri, A. Nerella, R. Murugesan, and K. Sundravadivelu, “Deep Learning–Based Multivariate Models for Bankruptcy and Litigation Risk Prediction,” *Adv. Consum. Res.*, vol. 2, no. 4, pp. 4442–4450,

 

2025.

[2] D. G. Allen, P. C. Bryant, and J. M. Vardaman, "Retaining Talent: Replacing Misconceptions With Evidence-Based Strategies.," *Acad. Manag. Perspect.*, vol. 24, no. 2, pp. 48–64, May 2010, doi: 10.5465/AMP.2010.51827775.

[3] H. Kali, "The Future of HR Cybersecurity: AI-Enabled Anomaly Detection in Workday Security," *Int. J. Recent Technol. Sci. Manag.*, vol. 8, no. 6, pp. 80–88, 2023.

[4] O. Ajibade and K. Ayinla, "Investigating the effect of training on employees' commitment: An empirical study of a discount house in Nigeria," *Megatrend Rev.*, vol. 11, no. 3, pp. 7–18, 2014, doi: 10.5937/MegRev1403007A.

[5] S. Tatavarthi and S. Tarakampet, "Architecting Resilient HR Automation Systems: Lessons from Enterprise-Scale Deployments," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 14, no. 01, pp. 1723–1729, January, 2026, doi: 10.22214/ijraset.2026.77214.

[6] S. Dodda, H. Volikatla, and J. R. Vummadi, "Exploring the Role of AI-Enhanced Chatbots in Automating Recruitment Processes in Human Capital Management Systems," *Int. J. Emerg. Trends Comput. Sci. Inf. Technol.*, vol. 6, no. 3, pp. 28–36, 2025.

[7] V. Agarwal, K. Mathiyazhagan, S. Malhotra, and T. Saikouk, "Analysis of challenges in sustainable human resource management due to disruptions by Industry 4.0: an emerging economy perspective," *Int. J. Manpow.*, vol. 43, no. 2, pp. 513–541, May 2022, doi: 10.1108/IJM-03-2021-0192.

[8] M. Ahmad and M. Allen, "High performance HRM and establishment performance in Pakistan: an empirical analysis," *Empl. Relations*, vol. 37, no. 5, pp. 506–524, Aug. 2015, doi: 10.1108/ER-05-2014-0044.

[9] Chetankumar Patel, "The Role of Predictive Analytics in Customer Churn Prevention Across Global Markets," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 4, no. 2, October, p. 735, Oct. 2024, doi: 10.48175/IJARSCT-19900F.

[10] N. Malali, "Augmenting Actuarial Intelligence: Defining the Future of Actuarial Work in the Age of AI Collaboration," *Int. J. Curr. Eng. Technol.*, vol. 15, no. 2, 2025.

[11] S. Bag, "Big data and predictive analysis is key to superior supply chain performance: A south african experience," *Int. J. Inf. Syst. Supply Chain Manag.*, 2017, doi: 10.4018/IJISSCM.2017040104.

[12] V. K. Singh, D. Pathak, and P. Gupta, "Integrating Artificial Intelligence and Machine Learning into Healthcare ERP Systems: A Framework for Oracle Cloud and Beyond," *ESP J. Eng. Technol. Adv.*, vol. 3, no. 2, pp. 171–178, 2023, doi: 10.56472/25832646/JETA-V3I6P114.

[13] J. W. Sajja, G. B. Komarina, and N. K. R. Choppa, "The Convergence of Financial Efficiency and Sustainability in Enterprise Cloud Management," *J. Comput. Sci. Technol. Stud.*, vol. 7, no. 4, pp. 964–992, May 2025, doi: 10.32996/jcsts.2025.7.4.110.

[14] P. Kumar Tyagi, V. Jit Singh, A. Kumar Singh, A. Saxena, P. Tyagi, and P. Mehta, "The Impact of Artificial Intelligence-Based Human Resource Management Systems on Organizational Efficiency," in *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering, UPCON 2023*, 2023. doi: 10.1109/UPCON59197.2023.10434792.

[15] C. C. Cantarelli, B. Flyvbjerg, E. J. E. Molin, and B. van Wee, "Cost Overruns in Large-scale Transportation Infrastructure Projects: Explanations and Their Theoretical Embeddedness," *Eur. J. Transp. Infrastruct. Res.*, vol. 10, no. 1, pp. 5–18, 2010, doi: 10.18757/EJTIR.2010.10.1.2864.

[16] P. W. Hom, T. W. Lee, J. D. Shaw, and J. P. Hausknecht, "One hundred years of employee turnover theory and research.," *J. Appl. Psychol.*, vol. 102, no. 3, pp. 530–545, Mar. 2017, doi: 10.1037/apl0000103.

[17] J. R. Vummadi, H. Volikatla, and S. Dodda, "Smart HR for Smart Enterprises : A Machine Learning - Based Approach to Payroll Automation and Time Optimization," vol. 6, no. 3, pp. 80–89, 2025.

[18] N. Murthy and M. Katyal, "Enhancing Employee Retention with AI - Driven Predictive Analytics," vol. 12, no. 03, pp. 48–56, 2025.

[19] A. Parupalli, "Business-Oriented Employee Performance Assessment via Machine Learning in ERP Systems," *Tijer – Int. Res. J.*, vol. 11, no. 11, 2024.

[20] V. Verma, "Optimizing Database Performance for Big Data Analytics and Business Intelligence," *Int. J. Eng. Sci. Math.*, vol. 13, no. 11, pp. 56–75, 2024.

[21] C. Tayal, S. Murumkar, and S. Biradar, "Analysing the Role of Multi-Agent AI Models for Autonomous Business Decision Systems," in *2026 IEEE 16th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA: IEEE, 2026, pp. 0058–0062, January. doi: 10.1109/CCWC67433.2026.11393746.

[22] C. Patel, "Integration of AI in Customer Relationship Management (CRM) for Improved Sales Outcomes," *Int. J. Emerg. Res. Eng. Technol.*, vol. 6, no. 4, pp. 137–145, 2025, doi: 10.63282/3050-922X.IJERET-V6I4P117.

[23] V. Varma, "Data Analytics for Predictive Maintenance for Business Intelligence for Operational Efficiency," *Asian J. Comput. Sci. Eng.*, vol. 7, no. 4, pp. 1–7, 2022.

[24] A. S. DeNisi and R. D. Pritchard, "Performance Appraisal, Performance Management and Improving Individual Performance: A Motivational Framework," *Manag. Organ. Rev.*, vol. 2, no. 2, pp. 253–277, Jul. 2006, doi: 10.1111/j.1740-8784.2006.00042.x.

[25] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.

[26] T. E. Ramya and B. Sanjay, "Deep Learning Approach for Prediction Employee Attrition for HR Analytics," in *2025 6th International Conference for Emerging Technology (INCET)*, 2025, pp. 1–5. doi: 10.1109/INCET64471.2025.11140331.

[27] R. Singh, Neha, R. Singh, P. Bhatnagar, D. Kaushik, and R. Chauhan, "Blockchain and Ai-Based Smart Hr Management System for Secure and Transparent Employee Records," in *2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, 2025, pp. 269–275. doi: 10.1109/ICDICI66477.2025.11135283.

[28] M. B. B. G, M. M. Hassan, B. A. Sri, P. Latha, and F. A. F. Vinola, "Prediction of Employee Attrition Classification and Retention Strategies Using Gradient Boosting with Binary Logistic Regression," in *2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS)*, 2024, pp. 1–5. doi: 10.1109/ICIICS63763.2024.10859889.

[29] A. Ismail Al-Alawi and N. Ahmed Aljawder, "The Potential for Predicting Employee Promotions Using Blended Ensemble Machine Learning Models," in *2024 International Conference on Open Innovation and Digital Transformation (OIDT)*, 2024, pp. 1–9. doi: 10.1109/OIDT59407.2024.11082724.

[30] S. Sharma and K. Sharma, "Analyzing Employee's Attrition and Turnover at Organization Using Machine learning Technique," in *2023 3rd International Conference on Intelligent Technologies (CONIT)*, IEEE, Jun. 2023, pp. 1–7. doi: 10.1109/CONIT59222.2023.10205676.

[31] B. Kaur and A. Dogra, "A Machine Learning Model for Predicting Employees Retention: An Initiative towards HR through Machine," in *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, IEEE, Nov. 2022, pp. 653–657. doi: 10.1109/PDGC56933.2022.10053249.

[32] N. Ben Yahia, J. Hlel, and R. Colomo-Palacios, "From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction," *IEEE Access*, vol. 9, pp. 60447–60458, 2021, doi: 10.1109/ACCESS.2021.3074559.

[33] G. Marvin, M. Jackson, and M. G. R. Alam, "A Machine Learning Approach for Employee Retention Prediction," in *TENSYMP 2021 - 2021 IEEE Region 10 Symposium*, 2021. doi: 10.1109/TENSYMP52854.2021.9550921.

[34] D. Sciences, "Predicting Job Satisfaction : A Machine Learning Approach to Employee Retention," no. December, 2024.

[35] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting Employee Attrition Using Machine Learning Approaches," *Appl. Sci.*, vol. 12, no. 13, 2022, doi: 10.3390/app12136424.

[36] D. Ma, M. Shu, and H. Zhang, "Feature Selection Optimization for Employee Retention Prediction: A Machine Learning Approach for Human Resource Management," *Appl. Comput. Eng.*, vol. 141, no. 1, pp. 120–130, 2025, doi: 10.54254/2755-2721/2025.21789.

 