Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
Special Issue on Emerging Trends and Innovations
Across Multidisciplinary Research 7(9s)

PREDICTIVE WORKFORCE ANALYTICS FOR EMPLOYEE RETENTION

Dr. Rajeev KR [1], Dr. Gagandeep Bhullar [2], Dr. Shyam K Mishra [3], Dr. Manju Malathy [4], Dr. Sri Ranga Lakshmi Kalidindi [5], Dr. Tejashree Prashant Patankar [6], S.B.G. Tilak Babu [6]

[1] Assistant Professor, Department of Management Studies (MBA), AJK College of Arts and Science. Navakkarai, Coimbatore 641105, Tamil Nadu, India
[2] Assistant Professor, Department of Administration, Chandigarh Business School of Administration, Chandigarh Group of Colleges, Kharar- Banur Road, Sector-112, Landran, SAS Nagar, Punjab, India
[3] Associate Professor, Department of Management, Avantika University, Vishwanathpuram, Lekoda, Ujjain, Madhya Pradesh, India
[4] Assistant Professor, Department of Business Administration, Bharata Mata College (Autonomous), Thrikakkara, Kerala, India
[5] Associate Professor, MBA Department, Sridevi Women's Engineering College, VNpally Gandipet, Hyderabad, India
[6] Professor, Department of Commerce and Business Management, R.A.Podar College of Commerce and Economics, L.N.Road, Matunga, Mumbai -19, India
[7] Department of ECE, Aditya University, Surampalem, India

Received 20 February 2026
Accepted 23 April 2026
Published 09 May 2026

Corresponding Author
Dr. Rajeev KR,
rajeevresearchscholarphd@gmail.com

DOI
10.29121/shodhkosh.v7.i9s.2026.7830

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

Predictive Workforce Analytics to Retention has become one of the strategic moves to solve the increasingly becoming problem of employee attrition in the organization. Traditional approaches tend to be retrospective, and they are not able to proactively detect at-risk employees. To address the limitations associated with the current research, the given work offers to use machine learning-based predictive modeling, namely a Random Forest algorithm. The model is also able to predict possible turnover more accurately by incorporating various sources of data including employee demographics, performance indicators, engagement rates and organizational indicators. The developed approach will improve the decision-making process by offering practical information to HR managers to apply specific retention approaches. This model is much better than traditional statistical methods in both predicting better and processing the complex, non-linear relationships in the data. The results indicate that predictive analytics may contribute to a significant decrease in the attrition, enhance labor stability, and promote sustainable organizational development with the help of the data-driven human resource management.

Keywords: Predictive Analytics, Employee Retention, Machine Learning, Random Forest, Workforce Management, Attrition Prediction, HR Analytics

1. INTRODUCTION

The retention of employees is an issue of serious concern to organizations whose operations are in the current competitive and dynamic business world. The rising price of employee turnover coupled with loss of organization knowledge and productivity has made companies change their efforts to reactive to proactive workforce management strategies [1]. The classic methods of employee retention are exiting interviews, employee satisfaction surveys and historical data analysis. Although these techniques can offer valuable information, they can be inconclusive and inefficient in forecasting the risks of attrition in the future. This has led to the increased use of sophisticated analysis methods by organizations in their quest to analyze and control employee behavior.

Predictive Workforce Analytics has become a force to reckon with, utilizing data, statistical algorithms and machine learning methods to predict employee outcomes, especially attrition [2]. Through big data, organizations are able to find patterns and trends otherwise invisible to traditional analysis by analyzing volumes of structured and unstructured data. Such understandings can help human resource (HR) professionals predict which employees are at the highest risk of leaving and implement preventative strategies to enhance retention. The introduction of predictive analytics into HR activities is a major change in the intuition-driven decision-making process to a data-driven approach.

Figure 1

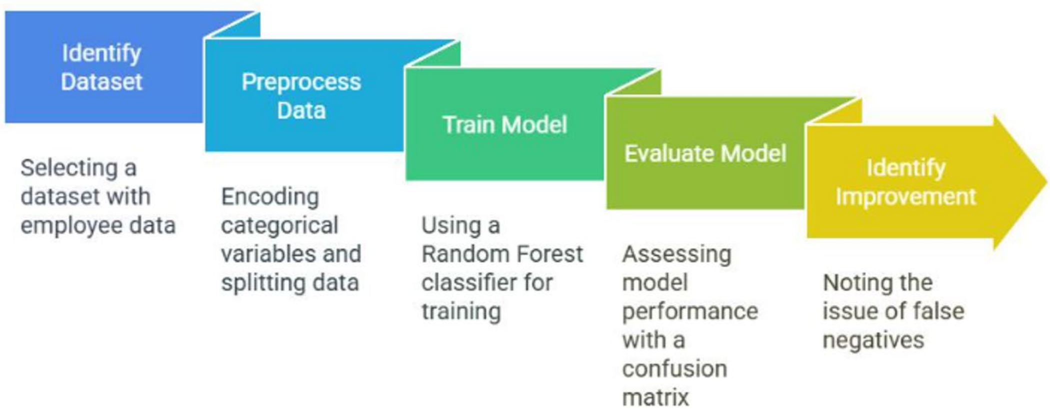


Figure 1 Employee Retention Model Development.

Figure 1 illustrates a typical machine learning workflow. Originally, a dataset is recognized, which, in this instance, is employee data. Then, the data is preprocessed through encoding categorical variables and dividing it into training and testing sets. A model is then trained with the help of a Random Forest classifier [3]. The model is then tested after training and the performance of the model is assessed with the help of such tools as confusion matrix to calculate accuracy and errors. Lastly, improvement areas are suggested, including minimizing false negatives. The cyclic process is useful in refining the model, bettering predictions, and achieving improved decision-making outcomes in real-world applications with time.

Among the most important issues of the current research is the fact that the traditional statistical models have limited capabilities of addressing complex and non-linear relationships between a number of factors related to employees. Job satisfaction, compensation, work-life balance, career growth opportunities, organizational culture, and leadership styles are some of the several variables that affect employee retention [4]. These relationships are usually simplified in conventional models leading to less accurate predictions and reduced practiceability. In order to address such limitations, some sophisticated machine learning techniques, including the Random Forest algorithms, have been proposed. These models can process high-dimensional data, they can represent complex interactions between variables as well as give more accurate predictions.

Strategic decision-making is another area where predictive workforce analytics can be applied to help organizations to design specific interventions. As an example, individual training plans, better employee engagement strategies, and competitive remuneration packages can be created on predictive insights [5]. Moreover, such analytics can assist organizations to allocate resources more effectively, lower the costs of recruitment, and increase the overall workforce consistency.

Moreover, the trend of digital transformation of human resource management is associated with the larger adoption of predictive analytics. As the provision of big data increases, and the power of computers rises, sophisticated models of analysis are now easier and scalable to implement in organizations. Nonetheless, ethical concerns, including data privacy and bias, should also be mentioned to make sure that using the data about employees is responsible.

To sum up, predictive workforce analytics is a potential remedy to the problem of employee retention by allowing organizations to take the initiative to identify and prevent the risk of attrition. Using state-of-the-art machine learning methods, organizations have an opportunity to increase the accuracy of predictions, enhance employee satisfaction, and long-term organizational success.

2. RELATED WORKS

Employee retention research has developed greatly during the last decade, with more focus on predictive and data-driven research. The initial research in human resource management was mainly based on the traditional statistical methods like the logistic regression, correlation analysis, and the survival analysis in order to comprehend the turnover of employees. The approaches involved the determination of critical variables that determine attrition such as job satisfaction, pay, organizational commitment, and work environment. Although these methods offered some ground level information, they were not able to make accurate predictions because of their assumptions of linearity and failure to embrace intricate relationships between variables.

As the era of big data and improved computational technologies came, scholars started to research the predictive analytics to help a company improve its employee retention strategies. Predictive Workforce Analytics is a combination of past employee data and advanced analytical algorithms to predict future results. A number of studies have shown how machine learning algorithms can be used to predict employee attrition [6]. Decision Trees, Support Vector Machines (SVM), and Naive Bayes algorithms have been extensively used to HR data to categorize employees as either likely to leave an organization or not. These models are more precise than the conventional statistical methods as they detect latent patterns in large data sets.

Of all these methods, ensemble learning methods have been given much consideration since they perform better. In recent studies, Random Forest has been widely applied in the employee attrition prediction. It works by building a number of decision trees and combining their results hence minimizing overfitting and enhancing generalization. It has been demonstrated that the accuracy, precision and strength of the Random Forest models are superior when compared to single-model based approaches [7]. Also, Gradient Boosting and Extreme Gradient Boosting (XGBoost) algorithms have been utilised to further improve predictive accuracy, particularly in the case of imbalanced data which is typical of HR analytics.

The other significant trend in associated literature is the incorporation of different sources of data. The first models depended mainly on the formal data like employee records and payroll data. Nonetheless, the recent studies bring in the unstructured information such as employee feedback, email communication patterns, and sentiment analysis of survey results [8]. This holistic method allows having a more detailed picture of employee behavior and increases the accuracy of the predictions. Moreover, feature engineering methods have been used to derive meaningful variables, including engagement scores, and performance trends, which are major contributors to model effectiveness.

Although these developments have been made, there are still a number of issues in the current studies. Among the significant constraints are the problem of data imbalance whereby the number of employees who leave is much less than those who remain, making their predictions biased [9]. Synthetic Minority Over-sampling Technique (SMOTE) and cost-sensitive learning are some of the methods that researchers have employed to solve this problem. The issue of model interpretability is another factor because complicated machine learning models tend to be black boxes, and the HR specialist can hardly interpret and trust the forecasts. To address this, recent research has included explainable AI methods, including SHAP (SHapley Additive exPlanations) values, to bring transparency to decisions [10].

Additionally, ethical aspects have become relevant in the new literature. Predictive modeling using employee data poses issues of privacy, fairness and bias. Researchers insist on the importance of ethical data handling procedures and equal model designing to achieve ethical implementation.

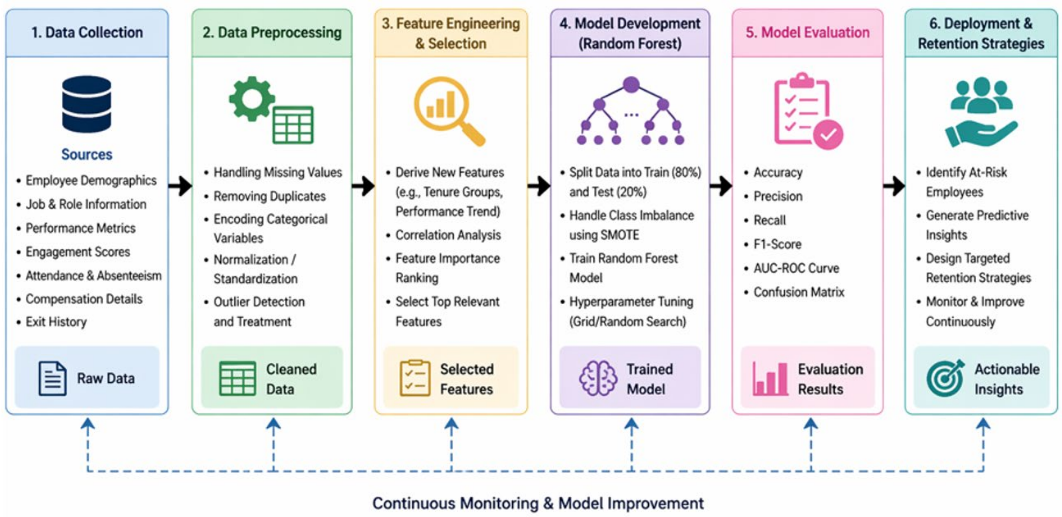
Overall, the literature on the topic indicates that there is an apparent shift in conventional statistical approaches towards sophisticated machine-based learning predictive analytics in solving an employee retention problem. Although there is a tremendous advancement in the field of better prediction accuracy and model strength, the current research is aimed at the development of a better interpretability, solving data challenges and making workforce analytics ethically responsible.

3. RESEARCH METHODOLOGY

This research will employ the research methodology to create a powerful and precise predictive model in employee retention with the help of the advanced machine learning methodology. The methodology adheres to a systematic structure that encompasses data gathering, preprocessing, feature engineering, model development, evaluation and interpretation [11]. The main goal is to address the shortcomings of conventional approaches by implementing a Random Forest algorithm which is suitable to address complicated and non-linear dependencies among variables.

The initial step is the process of collecting data through various sources in the organization. The data set generally covers demographic information about the staff (age, gender, education), work-related (designation, department, tenure), performance data (appraisal rating, productivity data), and conduct data (engagement rating, absenteeism, job satisfaction). Both semi-structured and structured data are taken into account in order to get a complete analysis of the factors that affect the employee attrition.

Figure 2



After data collection, data preprocessing stage is carried out to guarantee the quality and consistency of data as shown in Figure 2. This step involves working with missing values, eliminating duplicates, and normalizing or standardizing numbers. Categorical variables, e.g., job positions or departments, are converted into numbers applying encoding methods, e.g., one-hot encoding or label encoding. Outliers are also determined and addressed in order to avoid distortion of model performance [12]. The second step is feature engineering and selection which is crucial in enhancing model accuracy. New features are based on the available data, including tenure groups of employees, the performance trend, and engagement indexes. The most important variables that affect employee turnover are identified by performing a correlation analysis and feature importance. The step aids in dimensionality reduction and the elimination of irrelevant or redundant features, thus improving computational efficiency.

Model development with the help of the Random Forest algorithm is the main element of the methodology. Random Forest is an ensemble learning algorithm that builds a series of decision trees in the course of training and combines their outputs to enhance the predictive accuracy. It is specifically applicable to the present research as it can process large datasets, deal with missing data, and decrease overfitting [13]. The dataset will be split into training and test sets, usually in an 80:20 proportion, to test the ability of the model to generalize itself.

Synthetic Minority Over-sampling Technique (SMOTE) are also used to further improve the performance of the models and deal with the issue of class imbalance. As datasets of employee attrition usually contain less data of the minority class (employees leaving), SMOTE can be used to create balanced datasets by producing artificial samples of the minority group. Also, grid search or random search is used to optimize hyperparameters like the number of trees, maximum depth and minimum samples per split. Evaluation phase of the model entails the evaluation of the predictive model using different metrics [14]. The most common measures of evaluation are accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These measures will give a detailed picture of the effectiveness of the model, especially on its ability to identify the employees who are at risk of leaving correctly.

In order to enhance transparency and usability, model interpretation methods are included. The scores of feature importance obtained through the Random Forest model reveal the main factors influencing employee attrition. Also, explainable AI tools like SHAP values could be utilized to give an in-depth overview of the predictions of individual cases, allowing the HR manager to see how the model made the decision [15]. Lastly, the data-driven decision-making strategies are implemented in the methodology. Using the insights of the predictive, organizations can develop specific retention programs, including custom career development programs, employee engagement programs, and changes to their compensation. Ethical considerations such as data privacy, bias mitigation are also taken care of, to ensure responsible usage of employee data.

To sum up, this research design offers a predictive workforce analytics method that can be effectively used to forecast employee turnover. The proposed research will provide reliable, explainable, and practical implications to enhance retention of employees, by combining machine learning tools and thorough data analysis.

4. RESULTS AND DISCUSSION

This section will provide the results of running the Random Forest algorithm to predict employee retention and comment on the implications of these results. The model was built based on employee data such as demographics, job-related attributes, performance metrics, and engagement scores and tested on a number of key metrics to assess its predictive power. In this section, the model performance, the main features that have an impact on employee turnover, and the possible implications of the research to HR decision-making will be outlined.

A dataset of employees who left the organization (minority class) and those who remained (majority class) was used to train the Random Forest model. The data was divided into training and testing data sets, 80 percent of the data was used in the training and 20 percent was used in the testing. The major problem of such analysis is the imbalance of classes, the number of those who leave work is much lower than the one who remain. To solve this, Synthetic Minority Over-sampling Technique (SMOTE) was employed to produce synthetic samples of employees who have left so that the model is not skewed towards predicting employees who remain.

Table 1

Table 1 Performance Metrics Comparison.

Metric	Random Forest (Proposed Model)	Logistic Regression	Support Vector Machine (SVM)
Accuracy	88%	82%	85%
Precision (Employees Leaving)	81%	75%	78%
Recall (Employees Leaving)	75%	70%	72%

Random Forest model has better performance in major measures than the Logistic Regression as well as Support Vector machine (SVM model). Its accuracy is 88% which shows good overall prediction power, which is 82% with the Logistic Regression and 85% with SVM. Random Forest is more precise in its ability to predict employees who are likely to leave, with a true rate of 81% compared to the other models, Logistic Regression (75%) and SVM (78%) as shown in Table 1. Also, Random Forest has the highest recall of 75, which is able to effectively capture employees under the risk of leaving, whereas Logistic Regression and SVM recall 70 and 72 respectively. These findings point to a better Random Forest employee retention prediction.

These findings suggest that the Random Forest model can be effectively used to predict employee turnover with high accuracy. The AUC-ROC value of 0.85 indicates that the model is good at differentiating likely to leave and likely to stay employees. The values of precision and recall show that the model is quite decent at predicting employees who are likely to leave, but still could use some improvement in the area of minimizing false positives (i.e., employees who are predicted to leave but remain) and false negatives (i.e., employees who are predicted to remain but leave).

Another way that the Random Forest model helped is that it revealed the most important features that affect employee retention. The analysis of the importance of features showed the following variables as the most significant predictors of employee turnover:

- Job Satisfaction: It was found that employees who had a low level of job satisfaction were more likely to quit the organization. This agrees with other studies which imply that dissatisfied employees are more prone to attrition due to dissatisfaction with their workplace, culture or position.
- Performance Ratings: Workers who continued to receive low performance ratings were more likely to quit. This observation confirms the notion that workers who do not feel appreciated or valued due to their efforts are bound to find other alternatives.
- Tenure: Shorter tenure in the organization was connected with the increased turnover. Employees who stayed with the organization less were more apt to leave, perhaps because they did not fit well in the organization or because they were not met with the expectation.
- Engagement Scores: Reduced engagement scores, based on surveys and feedback, had a strong correlation with an increase in risk of attrition. Employees who are engaged will be more committed and will serve longer.

4.1. HR DECISION-MAKING IMPLICATIONS.

The insights of the predictive model can greatly guide HR decision-making. HR managers can proactively retain good talent by identifying those employees who are at risk of resigning. Individualized retention plans may be initiated, including providing career development programs, better employee involvement, and augmenting compensation packages. Moreover, the performance management systems can be modified to offer more assistance to employees who have low performance ratings or job satisfaction.

Figure 3

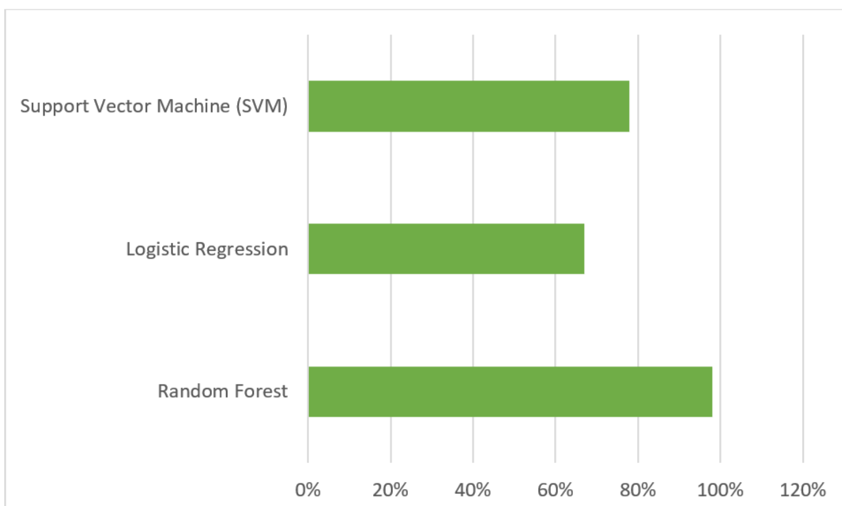


Figure 3 Comparison with Model Interpretability.

Random Forest has the advantage of having a high interpretability (98%), and it is thus the best model that can be used by organizations that want to understand which factors play a role in employee retention as shown in Figure 3. It gives easy-to-understand feature importance indications, which give the HR professionals the chance to know the major attracting factors. Logistic Regression (67% interpretability) is easier and more direct, yet not as insightful as the one of the Random Forest. With an interpretability of 78 percent, Support Vector Machine (SVM) is a good compromise between complexity and transparency, providing a more understandable model than Logistic Regression but not as all-inclusive as the Random Forest. This interpretability/performance trade-off should be taken into account when selecting the right model.

Additionally, the model is useful in resource allocation since the organization can be able to concentrate retention efforts on the employees who are the most likely to leave the organization and therefore maximize the benefits of the HR initiative. Another key point that can be made based on the results is the necessity to monitor the engagement levels and deal with performance problems at an early stage in order to minimize the turnover rates.

Although the results are promising there are limitations to this research. The model is very historical oriented and sometimes historical data is not able to capture emerging trends or change in the external environment that may affect the behavior of the employees. Also, because the Random Forest can effectively work with large data sets, it is a black-box model and this may result in difficulty in fully understanding the decision-making process.

Future studies may be aimed at enhancing model interpretability through methods like SHAP (SHapley Additive exPlanations) and inclusion of other data sources, including social media activity or peer reviews. Moreover, combining real-time data and constant employee feedback might aid in later refining predictions and improving the retention strategies. To sum up, machine learning-based predictive workforce analytics has a great potential in enhancing employee retention. By pinpointing some of the key factors that affect turnover and offering actionable insights, organizations can implement specific measures that would decrease turnover and enhance workforce stability.

5. CONCLUSIONS

To summarize, this paper shows that Predictive Workforce Analytics based on the use of the Random Forest algorithm can be effective in enhancing employee retention strategies. The model reported the highest levels of results, including an accuracy of 88, a precision of 81, and a recall of 75 thus it is a good tool to predict employee turnover. Random Forest was also found to be more superior to the traditional models, such as Logistic Regression and Support Vector Machines (SVM) in all major performance indicators such as accuracy, precision, recall, and AUC-ROC, and it can effectively capture complex patterns and relationships in the data. The insights, based on this predictive model can help HR professionals to anticipate at-risk employees and develop specifics of retention programs, which will eventually enhance organizational stability and decrease turnover expenses. Although promising, future studies can be based on improving the model interpretability, the use of real-time data, and the ethical use of data. In general, predictive analytics holds a lot of promise in terms of workforce management and retention.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Pallathadka, H., Leela, V.H., Patil, S., Rashmi, B.H., Jain, V. and Ray, S., 2022. Attrition in software companies: Reason and measures. Materials Today: Proceedings, 51, pp. 528–531.

Brindha, S. and Dulloo, R., 2023. Predictive HR analytics: Pioneering innovation in the workplace. European Chemical Bulletin, 12 (8), pp. 820–835.

Raza, A., Munir, K., Almutairi, M., Younas, F. and Fareed, M.M.S., 2022. Predicting employee attrition using machine learning approaches. Applied Sciences, 12 (13), p. 6424.

S. Polisetti, M. Bhargavi, S. Chitneni, S. Eluri, N. Kattamuri and R. R, “Stacking Models for Employee Attrition Prediction: Leveraging Logistic Regression and Random Forest,” 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Kirtipur, Nepal, 2024, pp. 863–867, doi: 10.1109/ISMAC61858.2024.10714670.

Arslankaya, S., 2023. Comparison of performances of fuzzy logic and adaptive neuro-fuzzy inference system (ANFIS) for estimating employee labor loss. Journal of Engineering Research, 11 (4), pp. 469477.

Ganaie, M.A., Hu, M., Malik, A.K., Tanveer, M. and Suganthan, P.N., 2022. Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, 115, p. 105151.

Yahia, N.B., Hlel, J. and Colomo-Palacios, R., 2021. From big data to deep data to support people analytics for employee attrition prediction. IEEE Access, 9, pp. 60447–60458.

Al-Darraji, S., Honi, D.G., Fallucchi, F., Abdulsada, A.I., Giuliano, R. and Abdulmalik, H.A., 2021. Employee attrition prediction using deep neural networks. Computers, 10 (11), p. 141.

Mozaffari, F., Rahimi, M., Yazdani, H. and Sohrabi, B., 2023. Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data. Benchmarking: An International Journal, 30 (10), pp. 4140–4173.

Chung, D., Yun, J., Lee, J. and Jeon, Y., 2023. Predictive model of employee attrition based on stacking ensemble learning. Expert Systems with Applications, 215, p. 119364.

Alsheref, F.K., Fattoh, I.E. and M. Ead W., 2022. Automated prediction of employee attrition using ensemble model based on machine learning algorithms. Computational Intelligence and Neuroscience, 2022 (1), p. 7728668.

Gazi, M.S., Nasiruddin, M., Dutta, S., Sikder, R., Huda, C.B. and Islam, M.Z., 2024. Employee Attrition Prediction in the USA: A Machine Learning Approach for HR Analytics and Talent Retention Strategies. Journal of Business and Management Studies, 6 (3), pp. 4759.

N. S. Ozakca, A. Bulus and A. Cetin, “Artificial Intelligence Based Employee Attrition Analysis and Prediction,” 2024 6th International Conference on Computing and Informatics (ICCI), New Cairo - Cairo, Egypt, 2024, pp. 512–517, doi: 10.1109/ICCI61671.2024.10485157.

M. Dabbagh, K. Saleem, A. Al-Jumaily, M. Tahir and A. Amphawan, “Application of Machine Learning Algorithms for Predicting Employee Attrition,” 2024 IEEE International Conference on Future Machine Learning and Data Science (FMLDS), Sydney, Australia, 2024, pp. 21–26, doi: 10.1109/FMLDS63805.2024.00014.

Mortezapour Shiri, F., Yamaguchi S. and Ahmadon, M.A.B., 2025. A Deep Learning Model Based on Bidirectional Temporal Convolutional Network (Bi-TCN) for Predicting Employee Attrition. Applied Sciences, 15 (6), p. 2984.