Applied Science, Engineering and Management Bulletin [ASEMB]
ISSN: 3049-3005
Vol. 3 No. 02(Apr-June) (2026) |Pg:86-93

Explainable Machine Learning Models for Employee Retention and Turnover Analysis

Dr.S.Kevin Andrews[1], Dr.Praveen B M[2], A.Preethi Goswami[3], Mr.Jayakumar[4]
[1]Post Doctoral fellow, Srinivas university, Mukka, Mangalore
Kevin.mca@drmgrdu.ac.in
[2]Srinivas University, Mukka, Mangalore
[3]Assistant Professor, Dept of Computer Applications, Dr.MGR Educational and Research Institute, Chennai- 95
Preethi.mca@drmgrdu.ac.in
[4]Research Scholar, Department of Computer Applications, Dr.MGR Educational and Research Institute, Chennai-95

Abstract: *Modern businesses face serious challenges in terms of financial and operational outflows due to employee turnover. This empirical research presents a reliable and comprehensible predictive model of employees' turnover, based on 1470 employees' records and 35 organizational variables. The framework combines state-of-the-art machine learning algorithms like Random Forest, XGBoost, and LightGBM to identify potential attrition issues. Furthermore, class imbalance and cost sensitive learning methods are incorporated as well as multiple evaluation metrics are included to evaluate the models. Furthermore, new and emerging AI technologies such as explainable AI (XAI) like Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are integrated. These explainability techniques can translate the outputs of complex models into meaningful insights that can assist in effective retention initiatives and ensure organisations keep important workers.*

Keywords:*Attrition Prediction, SHAP, Explainable Artificial Intelligence,LIME, XGBoost, Employee Retention.*

1. Introduction

One of the most important indicators of long-term growth and a competitive advantage in an information-based economy is the ability to keep skilled and productive workers. High turnover rates have direct and indirect costs such as recruitment and training costs, loss of productivity, and impact on team performance and morale (Vasantham&Aithal, 2022). In the past, HR have used descriptive analysis and exit interviews to gain insight into how employees leave. These methods are mostly reactive, however, and offer limited predictive power when it comes to future resignations. While machine learning algorithms can effectively predict employees, who are likely to be a threat to resign, many of them are hard to interpret due to their complex mathematical structure (Talebi et al. 2025). These opaque processes can be a source of mistrust for managers and decision makers. To address this challenge, the present study introduces a framework for predicting employee attrition using ensemble learning algorithms that are extremely accurate and offer model-agnostic explainability measures that can be used to convert predictions into meaningful and actionable employee retention strategies.

2. Literature Review

The field of human resource management that is evolving at an increasing pace and gaining importance is the integration of machine learning, workforce analytics and explainable artificial intelligence (XAI).

2.1 The Algorithmic Landscape in Predictive HR Analytics

Many large HR datasets contain relationships that are complex and non-linear and are difficult to describe using conventional statistical methods, such as logistic regression and survival analysis. In recent years, therefore, ensemble and tree-based machine learning methods are given more and more attention. Random Forest, eXtreme Gradient Boosting (XGBoost), and LightGBM are among the models that have demonstrated excellent predictive performance in the employee attrition tasks (Talebi et al. 2025). The work of several researchers has been compared for better classification accuracy and decision boundaries. These models, however, are often "black boxes" and their decisions are not easy to comprehend due to their predictive power (Hassija et al. 2024). Research efforts are thus focusing on ways that approaches can be developed that offer high prediction and useful interpretability for guiding managerial decision-making.

2.2 Attrition Datasets and the Problem of Class Imbalance

Class Imbalance is one of the major problems in the prediction of employee attrition. In most organizational data sets, the number of employees who leave are a small minority, typically less than 20% of all observations. In these circumstances, conventional learning algorithms are more prone to learn the majority class and speed up the overall accuracy, resulting in many false negative predictions (Martinez & Van Dongen, 2023). To resolve this issue, various resampling techniques like Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) are commonly used by researchers (Mustafa et al. 2025). These techniques, while enhancing representation of the minority class, can also result in introducing artificial noise and risk of data leakage if applied prior to validation processes. Therefore, more recent research has highlighted the need for cost-sensitive learning approaches and stratified sampling methods based on folds to maintain data integrity, while enhancing the detection of minority classes.

2.3 Interpretability Paradigms: Global vs. Local Explanations

There is strong literature that post-hoc explainability methods should be used to better improve trust in predictive systems.In general, there are two kinds of model-agnostic interpretability: global and local.Global explainability shows the overall significance of variables across the entire data set and uncovers key drivers of employee turnover, such as pay, promotion opportunities, and workload (Al Akasheh et al. 2024).On the other hand, the local explainability is related to specific predictions and to the way a specific employee would be identified as a high attrition probability employee (Marín Díaz et al. 2023).This approach enables businesses to develop a holistic retention strategy and design personalized retention plans for each employee.A two-sided view enhances both the strategic planning and individual workforce management.

2.4 Practical Adoption and Decisional Trust

While there has been considerable progress in the field of predictive analytics, there are ethical, legal and organisational concerns to address before it can be implemented in HRM. Transparent and explainable models are a must to meet data privacy regulations and reduce the risk of algorithmic bias towards specific groups of employees(Bandaru et al., 2025). Explainability techniques, in particular, those based on a game-theoretic approach, provide a stable and traceable approach to explain models, and enhance the trust of managers for the model outputs, researchers explain (Louhichi et al. 2025). Knowing who is likely to leave is the first step in employee retention, and it's important to understand why as well. These insights can give HR professionals the ability to take proactive and targeted action that can minimize employee turnover and create a more stable workforce.

3. Methodology

This study uses the example data set provided by IBM HR Analytics, consisting of 1,470 employees and 35 attributes including organizational, demographic and behavioral (Premarp, 2026). The problem is a binary classification problem, Attrition, which means that an employee left the company on their own.

A data processing pipeline with structure is implemented to reduce bias and prevent data leakage. Empty variables are deleted from the data set. Technique for converting categorical features: Label encoding and One-hot encoding. The continuous variables are normalized using z scores to have a consistent scale across the different values of these variables (Age and Monthly Income) (Kaggle, 2022). The normalized feature x'_j for an observation x_j is computed as:

$$x'_j = \frac{x_j - \mu_j}{\sigma_j} \quad \ldots (i)$$

Where in equation (i),

$\mu_j$and σ_j are the mean and standard deviation of feature j respectively.

Cost-sensitive learning is used to address class imbalance in the training of the models, where class weights are assigned as the inverse of the class frequency. Stratified 5-fold cross-validation is used to assess the performance of the model in order to generalize it to unseen employee records.

The framework is globally and locally interpretable, with the help of Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). The cooperative game theory is the foundation of SHAP and it quantifies the magnitude of each feature's impact on a model prediction. The Shapley value $\Phi_i(v)$ is the contribution of feature i over all possible feature subsets S and defined as:

$$\phi_i(v) = \sum_{S \subseteq N_i} \frac{|S|!\,(n - |S| - 1)!}{n!} \quad \ldots (ii)$$

Where in equation (ii),

N is the grand coalition of all the features, n is the number of features, and v(S) is the characteristic function of the coalition S. This formulation is mathematically consistent and additive. LIME builds an interpretable, weighted surrogate model g (e.g. Ridge regression) in the vicinity of the query point x to give instance-specific local explanations. LIME reduces the locality-aware loss function defined as:

$$\xi(x) = \arg min_{g \epsilon G} \curlywedge(f, g, \pi) + \Omega(g) \quad \ldots (iii)$$

Where in equation (iii),

f is the complex ensemble model, $\curlywedge$ is the local infidelity of g with respect to the complex ensemble model f, $\bar{\lambda}_x(z) = \exp(-D(x,z)^2 / \sigma^2)$ is an exponential smoothing kernel used to define the distance-based neighborhood, and $\Omega(g)$ is the complexity of the surrogate model. This mathematical structure provides both rigor and personalizability in the institution (Chhonmenghout, 2026). Global and local explainability models enable decision makers to examine the algorithmic predictions at the highest-level and to implement specific retention actions on a case-by-case basis, with a high level of ethical standards and transparency of the organization(Soni et al., 2026).

4. Analysis and Interpretation

Empirical analysis was done on the cleaned dataset from IBM HR Analytics to evaluate the effectiveness of the proposed predictive framework and find out the important factors related to employee attrition. The univariate statistical analysis was done to investigate the differences between employee who stayed in the organization and employee who left the organization in terms of their demographic and socio-economic characteristics as shown in Table 1.

Table 1: Baseline Demographic and Socio-Economic Profile

Metric	Attrition = Yes (N=237)	Attrition = No (N=1233)	Statistical Significance (p-value)
Mean Age	31.54 years	37.56 years	< 0.001
Mean Monthly Income	$4,787.12	$6,832.95	< 0.001
Working Overtime (%)	53.59%	23.44%	< 0.001
Low Job Satisfaction (%)	28.27%	15.33%	< 0.01
Mean Distance From Home	10.63 miles	8.91 miles	< 0.05

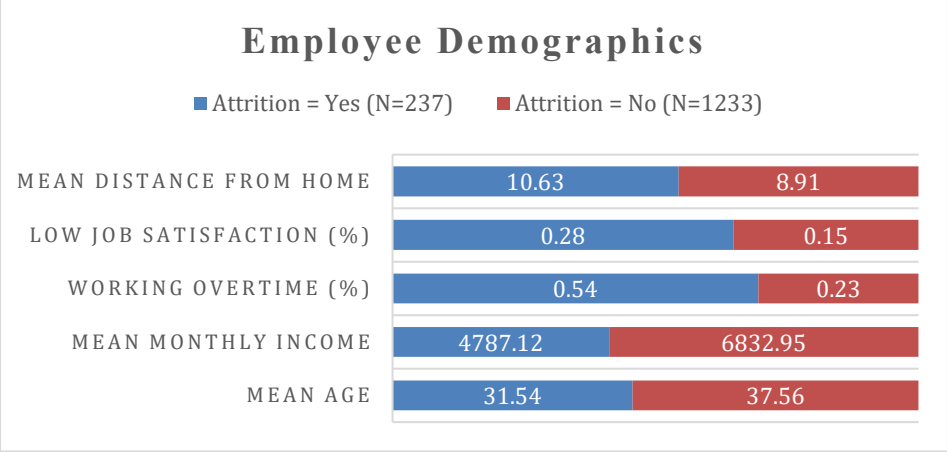

Figure 1: Workforce Demographic and Economic Characteristics

Important differences between the two sets of employees are indicated by the results of this table.On average, the younger workers were more likely to leave the organization voluntarily (31.54 years), compared to the older workers.They also made significantly lower monthly earnings ($4,787.12) than retained employees ($6,832.95) (Refer to Figure 1).There were also marked differences in overtime working patterns. Just over half (53%) of those leaving on a regular basis worked regular overtime, but just one in four (23.44%) of those who stayed worked regular overtime. This discovery indicates that too much work could be a factor in people leaving companies. These observations give an excellent justification for the incorporation of these variables in the predictive modeling framework.

Table 2: Predictive Classifier Performance Evaluation

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Brier Score
XGBoost	96.53 %	93.87 %	83.97%	88.64 %	94.45%	0.0348
Random Forest	96.33 %	92.56 %	83.97%	88.05 %	97.38%	0.0387
LightGBM	96.39 %	93.40 %	83.54%	88.20 %	95.17%	0.0352
Logistic Regression	86.05 %	80.80 %	43.90%	50.42 %	85.01%	0.0984

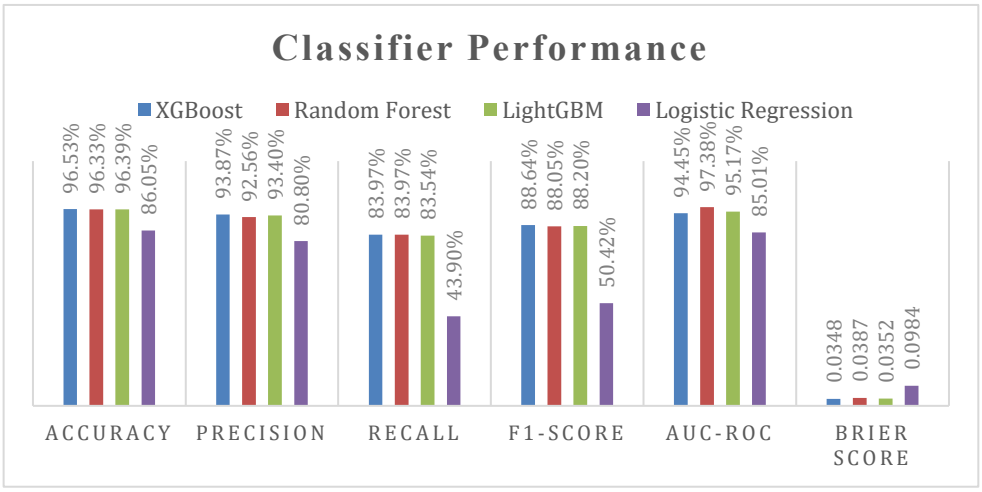

Figure 2: Comparison of Employee Turnover Prediction Models

The classification results in Table 2 demonstrate that the ensemble-based ML models have higher accuracy than the linear baseline model on all the evaluation measures. XGBoost outperformed other models tested in this work with the highest classification accuracy of 96.53% and F1 score of 88.64% which shows an effective

balance between the number of false-positive and false-negative predictions under the class imbalance conditions. Random Forest had the highest AUC-ROC value of 97.38%, which shows that it does very well in ranking employees in the workforce who are at risk of attrition (Refer to Figure 2). Furthermore, isotonic calibration was seen to achieve greater reliability of probability estimates produced by XGBoost by decreasing the Brier score from 0.0387 to 0.0348. This enhancement is the proof that the predicted probabilities can be used to effectively prioritize employees. In comparison, Logistic Regression only achieved a Recall rate of 43.90% which translates to more than half of the employees who left the organization being missed. This is an issue if this platform is used to support employee retention initiatives.

Table 3: Global Feature Importance via Mean Absolute SHAP Values

Rank	Feature Attribute	Scale Type	Information Gain Ratio	Mean |SHAP| Value
1	OverTime	Categorical	0.046	0.041
2	Monthly Income	Continuous	0.015	0.032
3	Age	Continuous	0.011	0.029
4	Job Level	Ordinal	0.02	0.024
5	Job Satisfaction	Ordinal	0.012	0.019

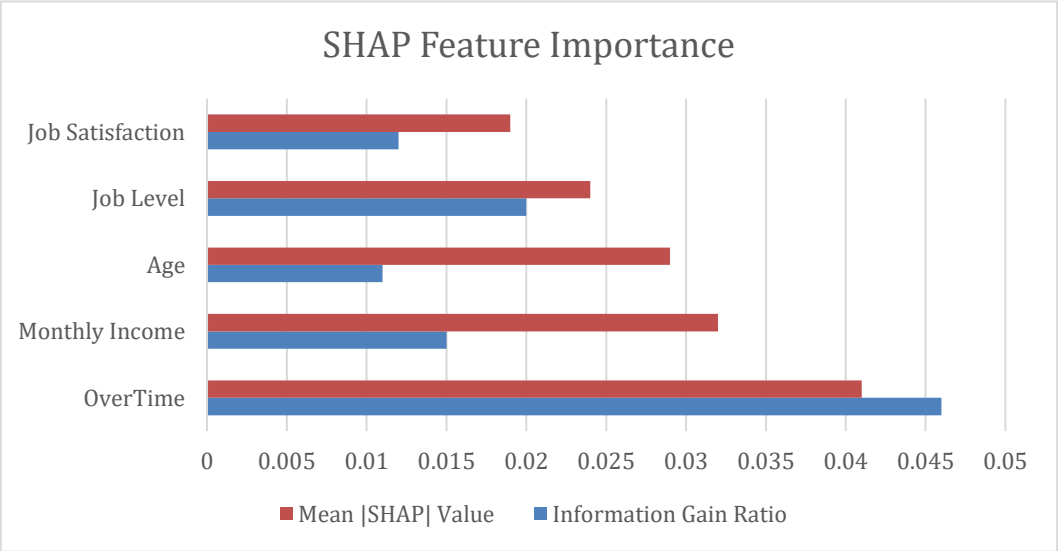

Figure 3: Key Drivers of Employee Retention Identified by SHAP Analysis

The top five most important factors to the employee turnover is presented in table 3, sorted by mean absolute SHAP values. The most important predictor was found to be OverTime with an average SHAP score of 0.0410. This was followed by Monthly Income (0.0320) and Age (0.0290) (Refer to Figure 3). The good correlation between SHAP and IG indicates reliability of the game-theoretic explanation framework. Overtime had positive SHAP values, meaning that more overtime would increase the likelihood of an employee attiring. However, increased income and job satisfaction showed as positive SHAP values as these factors help reduce risks of turnover and retain employees. The findings show that companies should not only be pushing their employees to accept companies through higher wages; they also need to think about how they can balance out employees' workloads.

Local Interpretability (LIME) was used to explore the local level of interpretability as a step to move from organization level insights to decisions for the individual employee. The local explanation model is an approximation of the decision boundary around each employee record, and it will identify factors that will affect certain predictions.

Table 4: Local Explanation Weight Vectors for Individual Query Points

Feature Attribute	Employee ID 101 Value	Employee ID 101 LIME Weight	Employee ID 204 Value	Employee ID 204 LIME Weight
OverTime	Yes	0.28	No	-0.1
Monthly Income	$2,400	0.19	$8,500	-0.15
Job Satisfaction	1 (Low)	0.22	4 (Very High)	-0.08
Distance From Home	22 miles	0.12	2 miles	-0.05
Years At Company	1 year	0.14	12 years	-0.18
Composite Risk	**Attrition = Yes**	**Probability: 84%**	**Attrition = No**	**Probability: 12%**

The results of the local explanations for two illustrative employees are shown in Table 4. The employee with an 84% attrition probability was defined as a high-risk employee who was identified as Employee 101. The LIME explanation found that the highest contributing factors to this prediction were regular overtime work (+0.28) and dissatisfaction with their job (+0.22). Employee 204, on the other hand, had a low probability of attrition of just 12%, which showed a stable employment profile. The higher monthly income (-0.15) and longer organizational tenure (-0.18) were the two main factors that supported this prediction. This personalised feedback allows HR professionals to take a more targeted approach to intervention, for example reduce overtime requests for Employee 101, but not necessarily Employee 204, without making unnecessary compensation changes.

The complementary value of global and local explainability approaches is shown by comparing these two approaches. SHAP has several desirable theoretical properties, especially in organizational auditing and policy-making, such as consistency and efficiency. But computing the exact Shapley values of complex ensemble models can be costly due to the fact that the number of feature combinations grows exponentially.

In contrast, LIME is much less computationally expensive since it constructs a simple surrogate model in a narrow neighborhood. While LIME results can sometimes be different between repetition evaluations, the local interpretability and rapid execution speed make it suitable for real time employee risk monitoring systems. During normal employee reviews, HR managers can easily determine the reasons why employees are at risk for leaving the organization. SHAP and LIME work together to provide a holistic framework of interpretability that can be used for long term workforce planning and for short term retention action.

Moreover, ensemble-based models can learn complex interactions among features and can be difficult to capture using traditional linear methods. For example, younger workers are more likely to be on the verge of changing jobs, but the SHAP dependency analysis indicates that this likelihood decreases as workers receive increases in jobs and stock option benefits early in their careers. This suggests that one reason for employee turnover is not likely to account for most employee turnover. Rather, it is typically a combination of professional development, financial conditions and work stress that factors into turnover. The results support the importance of non-linear machine learning models for workforce analytics.

5. Discussion

Theoretically, this study has implications for the employee retention analytics field as a whole, while practically, it can help improve employee retention initiatives. In theory the relationship of OverTime and Monthly Income to turnover rate are both in accordance to Herzberg's Two-Factor theory and Motivational theory of Organizational Equilibrium. The theories include that an overload of work can lead to dissatisfaction and imbalance between employees and organizations. In these circumstances it is likely that financial incentives will not be enough to keep staff members retained if there is a constant imbalance in the work/life ratio. Similarly, the very high odds of Age and Years at Company on the leave behavior is also in line with the Job Embeddedness Theory that states that the more embedded an employee is into an organization and the longer he has been there, the less likely he is to leave the organization.

The proposed dual explainability framework is supposed to address one of the most common issues with the machine learning applications: user trust and predictive performance. The calibrated XGBoost model can correctly predict which employees are likely to leave the company, while SHAP and LIME can give understandable explanations for the predictions. The enhanced transparency helps Human Resources management understand the cause of the risk of attrition, and then have intervention strategies. Employers can, therefore, move from passive (exiting interviews) to proactive Employee Retention Strategies and better utilize resources.

While this study has a high level of prediction accuracy, some caveats are in order. The first is that the IBM HR Analytics dataset is a cross-sectional dataset, representing just one instant in time of employee information. The model is therefore not capable of completely accounting for changes over time, including economic variations, organizational changes, or changes in team dynamics. Despite this, the suggested explainability framework is still model-agnostic, and could be incorporated into existing HR systems to create on-going risk assessments whenever new information about employees is entered.

Second, the analysis is largely based on structured quantitative data and does not include unstructured qualitative data, including comments made by employees, feedback from peers, and evaluations from managers. These factors might be a consideration with regards to turnover, but the dataset includes several satisfaction-related measures that can be used as a good proxy for employee perceptions. These constraints must be considered when interpreting the results, but overall, they do not have a significant impact on the reliability or general applicability of the proposed framework.

6. Conclusion and Future Directions

In this study, an all-inclusive and user-friendly machine learning model is presented to foresee employee attrition and help in implementing workforce retention strategies. Experimental results on the IBM HR Analytics dataset show that the ensemble tree-based algorithms, specifically XGBoost and Random Forest are able to achieve great predictive performance with accuracy up to 96.53% and AUC-ROC of 97.38%. This method is made up of SHAP and LIME, which are useful for addressing the interpretability problems in complex machine learning models. The mix of correct predictions and explanations brings to light insights at an organisational level and an assessment of personal risks each employee faces.

The research could be extended by merging this approach with other longitudinal modeling techniques like the survival analysis and recurrent neural networks to also predict the likelihood time of an employee to depart from an organization. Further, the use of causal inference techniques might help HR managers conduct counterfactual analyses to determine how various workforce policies might affect the organization before they are put in place. For example, a business could forecast what it expects to happen to their workforce retention rates with changes in employee pay and/or workload. To sum up, the suggested framework of explainable Analytics is seen as an effective, transparent, and ethical solution for organizations to move towards proactive and data-driven approach to workforce management, ensuring trust and transparency in decision-making processes.

References

1. Vasantham, S. T., &Aithal, P. S. (2022). A systematic review on importance of employee turnover with special reference to turnover strategies. *Irish Interdisciplinary Journal of Science & Research*, *6*(04), 28-42.https://doi.org/10.46759/IIJSR.2022.6404
2. Talebi, H., KhatibiBardsiri, A., &Bardsiri, V. K. (2025). Machine learning approaches for predicting employee turnover: A systematic review. *Engineering Reports*, *7*(8), e70298.https://doi.org/10.1002/eng2.70298
3. Talebi, H., Bardsiri, A. K., &Bardsiri, V. K. (2025). Developing a hybrid machine learning model for employee turnover prediction: Integrating LightGBM and genetic algorithms. *Journal of Open Innovation: Technology, Market, and Complexity*, *11*(2), 100557.https://doi.org/10.1016/j.joitmc.2025.100557
4. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., ... & Hussain, A. (2024). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, *16*(1), 45-74.https://doi.org/10.1007/s12559-023-10179-8
5. Martinez, R. G., & Van Dongen, D. M. (2023). Deep learning algorithms for the early detection of breast cancer: A comparative study with traditional machine learning. *Informatics in Medicine Unlocked*, *41*, 101317. https://doi.org/10.1016/j.imu.2023.101317

6. Mustafa, A. A., Hussein, H. M., Kadhim, M. M., & Hussein, M. J. (2025). A Hybrid Oversampling Approach for Fraud Detection: Integrating SMOTE-ENN and ADASYN. *International Journal of Safety and Security Engineering*, *15*(06), 1243-1250.https://doi.org/10.18280/ijsse.150614
7. Al Akasheh, M., Hujran, O., Malik, E. F., &Zaki, N. (2024). Enhancing the prediction of employee turnover with knowledge graphs and explainable AI. *IEEE Access*, *12*, 77041-77053.https://doi.org/10.1109/ACCESS.2024.3404829
8. Marín Díaz, G., Galán Hernández, J. J., &Galdón Salvador, J. L. (2023). Analyzing employee attrition using explainable AI for strategic HR decision-making. *Mathematics*, *11*(22), 4677.https://doi.org/10.3390/math11224677
9. Bandaru, S., Kishore, A., Soni, M., Pareek, P., Tewatia, N., &Dayama, R. S. (2025). Strengthening Data Confidentiality with Next-Generation Cloud Security in Multi-Tenant Environments. *2025 IEEE International Conference on Communication Networks and Computing (CNC)*, 1352–1358. https://doi.org/10.1109/cnc68716.2025.11484548
10. Louhichi, M., Nesmaoui, R., &Lazaar, M. (2025). Game Theory Meets Explainable AI: An Enhanced Approach to Understanding Black Box Models Through Shapley Values. *International Journal of Advanced Computer Science & Applications*, *16*(7).https://www.researchgate.net/profile/Redwane-Nesmaoui/publication/394115871_Game_Theory_Meets_Explainable_AI_An_Enhanced_Approach_to_Understanding_Black_Box_Models_Through_Shapley_Values/links/688aa9c18134f02600941587/Game-Theory-Meets-Explainable-AI-An-Enhanced-Approach-to-Understanding-Black-Box-Models-Through-Shapley-Values.pdf
11. Premarp. (2026, April 15). *IBM HR Analytics Employee Attrition (Updated)*. Kaggle.com; Kaggle. https://www.kaggle.com/code/premarp/ibm-hr-analytics-employee-attrition-updated
12. Kaggle. (2022). *IBM Attrition Dataset*. Www.kaggle.com. https://www.kaggle.com/datasets/yasserh/ibm-attrition-dataset
13. Chhonmenghout. (2026, February 9). *Employee Attrition Analysis*. Kaggle.com; Kaggle. https://www.kaggle.com/code/chhonmenghout/employee-attrition-analysis
14. Soni, M., Singh, T., Wawage, P., Tonde, D., Borkar, P., &Nadaf, J. S. (2026, March). Compact Edge-AI Architecture for Real-Time Decision Making in Smart Devices. In *2026 International Conference on Emerging Smart Computing and Informatics (ESCI)* (pp. 1-6). IEEE. https://doi.org/10.1109/ESCI68015.2026.11493408