SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Explainable AI In Complex Learning Systems: Trade-Offs Between Transparency And Performance

Aayushi Goel[1], Gajendra Shrimal[2], Shailendra Kumar Mishra[3], Abdul Majid[4], Srikanta Kumar Sahoo[5], Rajashri CK[6], Uma Maheswari G[7], Shajahan B[8]

[1] Assistant Professor, Symbiosis Law School, Noida, Symbiosis International (Deemed University), Pune, India. Email: aayushi.goel@symlaw.edu.in, ORCID: 0000-0002-0561-8059
[2] Assistant Professor, Department of Computer Science & Application, Vivekananda Global University, Jaipur, India. Email: gajendra.shrimal@vgu.ac.in, ORCID: 0000-0002-3812-950X
[3] Associate Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Parul Institute of Technology, Parul University, Vadodara, Gujarat, India. Email: shailendra.mishra40268@paruluniversity.ac.in, ORCID: 0000-0003-0192-198X
[4] Professor, Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India. Email: abdul.majid@presidencyuniversity.in, ORCID: 0000-0002-3769-0231
[5] Associate Professor, Centre for Cyber Security, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India. Email: srikantasahoo@soa.ac.in, ORCID: 0000-0002-3844-8308
[6] Assistant Professor, Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, India. Email: rajashrick@maher.ac.in
[7] Assistant Professor, Department of Mathematics, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, India. Email: umamahes@maher.ac.in
[8] Professor, Department of CS & IT, JAIN (Deemed-to-be University), Bengaluru, Karnataka, India. Email: shajahan.b@jainuniversity.ac.in, ORCID: 0000-0001-7199-6595

Abstract
In high-stakes fields like human resource management (HRM), Explainable Artificial Intelligence (XAI) has emerged as a crucial strategy for resolving the opacity of intricate machine learning (ML) systems. This research examines the trade-off between predictive performance and model transparency in complex learning environments. Using the Explainable HR Attrition Dataset, which includes demographic, organizational, performance, compensation, engagement, and temporal attributes, the research models employee turnover behavior with high dimensionality. Research empirically compares traditional interpretable models with an advanced ML approach, the Improved White Shark Optimized Cascaded Random Forest (IWSO-CRF). Post-hoc explainability resources like the Shapley Additive Explanations (SHAP) model are incorporated to balance interpretability and accuracy, offering both local (employee-level) and global (organizational-level) insights into model decisions. To guarantee data consistency, missing values are handled using the proper imputation techniques. Principal Component Analysis (PCA) is also used for feature extraction, which minimizes multicollinearity and reduces dimensionality while maintaining important information. Results indicate that the proposed approach enhances interpretability without significantly sacrificing predictive performance, although challenges related to explanation consistency and fidelity remain. The IWSO-CRF model achieves superior results, with 94.8% accuracy, 0.94 F1-score, 0.95 precision, and 0.95 recall, outperforming baseline interpretable models. The findings also reveal that transparency-by-design models, such as linear approaches, often fail to capture nonlinear and high-dimensional patterns in HR data. Therefore, the research combining inductive ML techniques with deductive analytical methods to develop robust, interpretable, and high-performing AI systems.

Keywords: Explainable Artificial Intelligence (XAI), Complex Learning Systems, Transparency, Trade-off, Machine Learning, Human Resource Analytics

Introduction

1. Background

Explainable Artificial Intelligence (XAI) is an emerging field of AI that focuses on making complex ML models understandable and interpretable for humans. The capabilities of AI systems have been impressive in different applications due to the widespread application of complex techniques such as Deep learning (DL) [1, 2]. In particular, these types of algorithms act as "black boxes" and do not disclose their logic of decision-making. This issue has now become very acute in those areas where it is essential that people should be aware of what reasons there are to make forecasts [3, 4]. Transparency of complex machine learning systems means that it is made clear how input data is converted into output data. This is often achieved by simplifying the model, or by integrating the explanation techniques of feature importance, attention mechanisms, or surrogate models [5, 6]. The more complex high-performance models, in turn, tend to be more complex, including a large number of layers and nonlinear transformations that conserve detailed patterns of the data. As a result, the enhancement of transparency sometimes may mean the decline in the accuracy or performance, which indicates a trade-off between interpretability and performance [7, 8]. This balance is very important in practice, and in the field of medical diagnosis, finance, and robotics, a false prediction may cause severe outcomes [9]. There is a need to provide reliable predictions and explanations at the same time. It is up to it to come up with designs that remain highly predictive and therefore, provide useful and accessible explanations at the same time [10, 11]. The emerging trends of XAI have attempted to address this requirement and have developed solutions that are a hybrid to this quandary. They are post-hoc explanation methods, models that are inherently interpretable and methods of approximation, which are able to explain complex models without essentially degrading their performance. The perception and control of trade-offs between transparency and performance is, therefore, one of the keys to the responsible and efficient introduction of AI systems into complicated settings [12, 13].

1.1 Aim of the research: This research attempts to create an explainable ML model to predict employee turnover alongside a viable balance between predictive power and model transparency. It entails a comparison between the traditional interpretable models and the optimized IWSO-CRF model, as well as incorporating the post-hoc explainability into producing meaningful local and global insights. Also, the research compares the use of PCA to improve the performance and understandability of models in HR analytics.

1.2 Research organization: This research is divided into five main sections. Section 1 presents the introduction; Section 2 discusses the literature review; Section 3 explains the methodology; Section 4 presents the results and discussion; and Section 5 presents the conclusion.

2. Previous Research

Table 1 shows the Existing research which demonstrate that combining ML with XAI improves predictive performance, while enhancing transparency across domains such as fraud detection, education, and HR analytics.

Table 1: Review of Related XAI Approaches in HR and ML Applications

Ref	Objective	Method	Key Result	Limitation
[14]	Create a transparent fraud detection system that is sensitive to privacy	Federated Learning + Explainable AI (FL-XAI)	Accurate detection with data privacy	Scalability & communication overhead
[15]	Build a multi-dimensional, explainable ML framework for automated interpreting assessment	Extreme Gradient Boosting + Data Augmentation (XGBoost-DA)	High performance (Spearman 0.87)	Small, language-specific dataset
[16]	Predict employee turnover using knowledge-graph and XAI-based framework	Graph Convolutional Network + Linear support vector machine (SVM) + XAI (GCN-LSVM-XAI)	92.5% accuracy	Complex, single dataset
[17]	Predict employee attrition and support HR decision-making using AI and XAI	Artificial Intelligence + XAI (AI-XAI)	Effective risk prediction	Data bias & high cost

3. Method

The methodology entails cleaning of the HR data, using the imputation, and the dimensional reduction using PCA. It then compares the proposed IWSO-CRF model with baseline interpretable models to evaluate performance. The process of explainability which contain SHAP is used to understand individual as well as overall model choice. Lastly, standard metrics and the quality of explanations are used to evaluate models. Figure 1 shows the proposed IWSO-CRF framework for optimizing features and improving prediction accuracy in HR analytics.

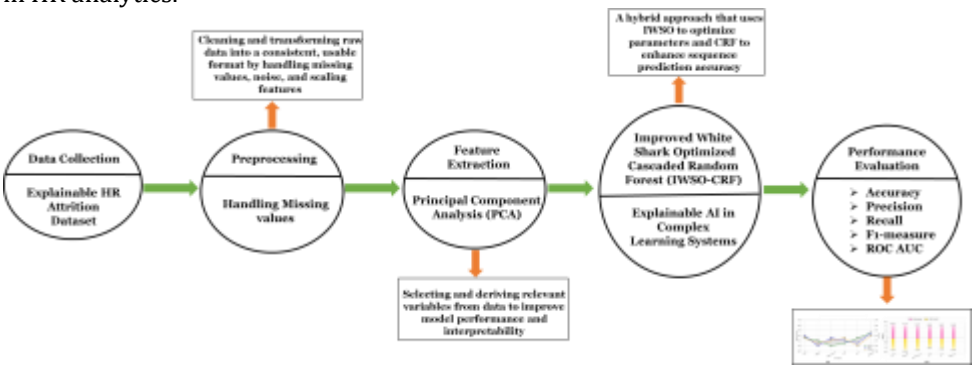


Figure 1: Proposed IWSO-CRF Model for HR Analytics

3.1 Dataset

The database contains detailed data about employees' demographic characteristics, organizational affiliations, performance, remuneration, engagement and temporal data. The database captures organization structure, job titles, departments and other data about the employment situation. This allows us to get a more realistic picture of the workplace environment. Such features as projects, performance evaluation and other behavioral factors like job satisfaction, employee engagement, and work-life balance allow us to draw conclusions about the employee's productivity. The presence of temporal variables such as tenureship and promotion time can help us understand career growth. The dependent variable indicates whether the employee left or stays in the company. The dataset was split into 80% for training and 20% for testing to evaluate model performance. Data source: https://www.kaggle.com/datasets/colabsss/explainable-hr-attrition-dataset

3.2 Data preprocessing - Handling Missing values

For the missing values have chosen an appropriate imputation strategy in order not to compromise data and results obtained by ML model. Features that had a significant amount of missing values were simply dropped. For categorical features, missing values were imputed using the mode, while for numerical variables, the median was used to reduce the impact of outliers. Missingness could contain useful information, therefore, all missingness was imputed with the unknown value. All data processing was performed prior to feature extraction.

3.3 PCA for Feature Extraction

HRM data are often high-dimensional and highly correlated, which can lead to complex models with poor interpretability. The issue becomes more difficult when considering the problem of XAI in which have to find a balance between predictive accuracy and interpretability. In the first place, the correlated features are reduced to fewer features by employing PCA, which is a technique used to minimize the dimensions of the data while retaining the maximum variance contained within the dataset.

Data Centering and Covariance Computation: The data is first standardized by subtracting the mean and then used to compute the covariance matrix:

$$C = \frac{1}{N}(X - \mu)(X - \mu)^T \quad (1)$$

In Equation (1), N is the number of observations. μ is the mean vector, X is the data matrix, and C represents the covariance matrix that captures relationships between features, and T stands for transpose. This matrix captures the variance and relationships among HR features.

Eigen Decomposition and Feature Transformation: Eigen decomposition of the covariance matrix yields the primary components:

$$V^{-1}CV = D \quad (2)$$

Where V contains eigenvectors (principal components), V^{-1} Inverse of the eigenvector matrix, used to diagonalize C,

C is the covariance matrix that contains relationships between features, and $D$is a diagonal matrix of eigenvalues that shows how much variance each component accounts for, given in Equation (2). The top components are selected to form a reduced feature space.

3.4 IWSO-CRF Model for Balancing Accuracy and Interpretability in HR Analytics

The suggested IWSO-CRF framework consists of combining IWSO with CRF to ensure an appropriate trade-off between accuracy and interpretability in HR analysis. IWSO ensures the use of maximum features and model parameters, while CRF relies on maximizing predictions using the combination of efficient trees with low correlations. The combination ensures the improvement in the prediction performance due to the reduced complexity.

Cascaded Random Forest: The challenge of XAI with respect to HRM is balancing predictive performance and interpretability. CRF improves the balance between predictive performance and interpretability by selecting a subset of highly accurate and less correlated trees, instead of aggregating all trees as in traditional approaches. In the model, bootstrap sampling and random selection of features are used to create several trees. Out-of-Bag (OOB) samples are used to evaluate each of the trees, and the trees are ranked by the accuracy of each tree:

$$\frac{1}{OOB^m}\sum_{y\in OOB^m} I(T^m(y_i)) = x_i) \quad (3)$$

In Equation (3), $T^m(y_i)$ is the prediction of the m^{th} tree, $I(\cdot)$ is the indicator function, and x_i is the true label, and The notation $y \in OOB^m$ indicates that the sample y belongs to the OOB set of the m^{th} tree, meaning it was not used during the training of that tree and is therefore used for its validation. The best N high-performing trees are then chosen and combined altogether to yield the final prediction that is stronger and without overly inter-tree. The optimal hyperparameters are selected through grid search with cross-validation. Such an aggregated method of training the model leads to lower accuracy and allows creating stable explainability, making ERF preferable among other algorithms for XAI-driven HR analytics where accuracy is important, but at the same time, the interpretability is needed. Figure 2 illustrates selecting the most accurate trees and combining them to improve prediction performance.

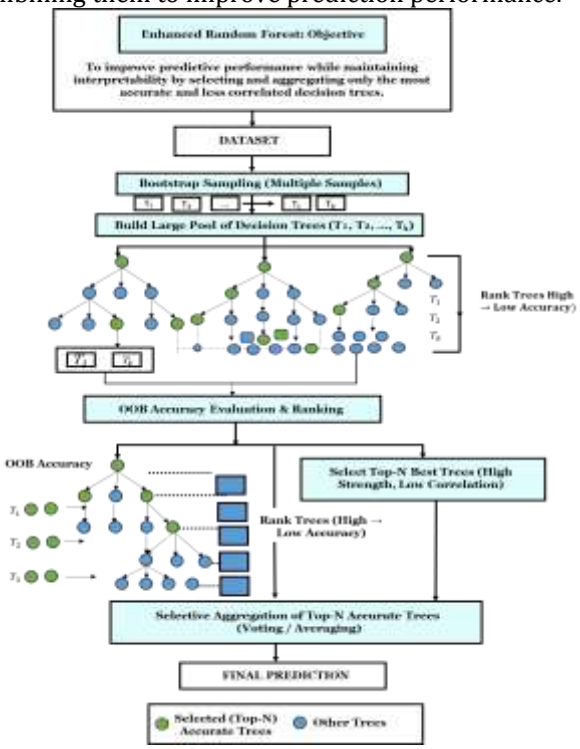


Figure 2: CRF Framework with Selective Tree Aggregation

Improved White Shark Optimization: The IWSO algorithm is a population-based meta-heuristic approach which simulates the behavior of sharks when hunting prey. It is suitable for solving complex HR problems with many dimensions, nonlinearity, and the problem of finding the right balance between accuracy and explainability of results in XAI techniques.

Velocity Update: The movement of each shark towards promising solutions is governed by:

$$u^j l + 1 = \xi[u^j l + \rho_1(\omega_{hbest_l} - \omega^j{}_l) \times d_1 + \rho_2(\omega^{u\,j\,l} best - \omega^j{}_l) \times d_2] \quad (4)$$

In Equation (4), the term $\omega^j{}_l$ denotes the current position of the j^{th} solution, and ω_{hbest_l} represents the global best position found so far by the entire population. $u^j l + 1$ represents the updated velocity of the j^{th} solution at iteration $l + 1$, while $u^j l$ is its current velocity at iteration l. The term $\omega^{u\,j\,l} best$ indicates the personal best position previously achieved by the j^{th} solution. The parameters ρ_1 and ρ_2 are control coefficients that regulate the influence of global and personal best positions, respectively, while d_1 and d_2 are random values that introduce stochastic behavior into the search process. The factor ξ is a constriction coefficient used to ensure convergence stability.

Adaptive Sensory Mechanism: The transition from exploration to exploitation is controlled by a sensory strength function:

$$T_t = |1 - f^{(-b_2 \times l / l_{max})}| \quad (5)$$

In Equation (5), l is the current iteration, T_t is an adaptive parameter controlling search behavior, f is Euler's number (~2.718), the base of exponential functions, l_{max} is the maximum number of iterations, and b_2 controls the rate of change. As iterations increase, T_t helps shift the algorithm from exploration to exploitation. The approach used by the IWSO-CRF is an implementation of optimized feature selection with selective tree aggregation. This approach makes a trade-off between performance and explainability in the application of HR analytics. The algorithm 1 provided above can be described as the use of feature reduction with the application of PCA along with the optimization of PCA features, and subsequent tree selection by WSO.

Algorithm 1: Integrated IWSO–CRF for XAI-based HRM

Input: *Dataset D*
Output: *Final Prediction Y*
1. *Preprocess data*:
   − *Handle missing values*
   2. *Apply PCA*:
   − *Compute covariance*:
   $C = (1/N)(X - \mu)(X - \mu)\text{^}T$
   − *Perform eigen decomposition*:
   $V^{-1}\,C\,V = D$
   − *Obtain reduced data X_pca*
3. *Initialize CRF population*:
   − *Generate random solutions W*
4. *Optimize using WSO*:
   Repeat until max iterations:
   − *Update velocity*:
   $v\_i\text{^}(k+1) = \xi\,[v\_i\text{^}k + \rho_1(w\_gbest - w\_i\text{^}k) + \rho_2(w\_best - w\_i\text{^}k)]$
   5. *Train Random Forest*:
   − *Build multiple trees using bootstrap samples*
6. *Evaluate trees*:
   − *Compute OOB accuracy*:
   $Acc\_m = (1/|OOB|)\,\Sigma\, I(T\_m(x\_i) = y\_i)$
7. *Select Top − N trees and aggregate*:
   − *Final prediction using voting/averaging*
Return Y

3.5 SHAP-Based Interpretability for Employee Attrition Prediction

The SHAP method is employed for explaining the predictions of the IWSO-CRF algorithm, allocating a value for each feature in terms of its contribution to the employee turnover rate. The SHAP method is able to offer both

local (at an individual employee level) and global (at a feature level) explanations. Important features, such as job satisfaction, salary, and work-life balance, are recognized as important contributors to employee turnover.

4. Result

The results obtained from the Python algorithm can be considered one part of a trade-off between accuracy and predictability in relation to HR turnover. While simple models are easier to understand, complicated ones perform better. Key elements like work-life balance, salary, and job happiness can be found using Python explainability approaches without suffering a significant decrease in accuracy.

4.1 Exploratory Data Analysis of Features and Relationships in Employee Attrition

Figure 3 represents the correlation table of one of the main HR attributes and the individual correlations between such variables as age, income, job satisfaction, and engagement score. The majority of features are correlated with each other with low to medium integrity, meaning that there is very little multicollinearity in the data. Several related variables, e.g., experience and tenure, are correlating relatively highly. Figure 3 shows that the features are independent of each other, which contributes to the successful learning of a model and its understandability.

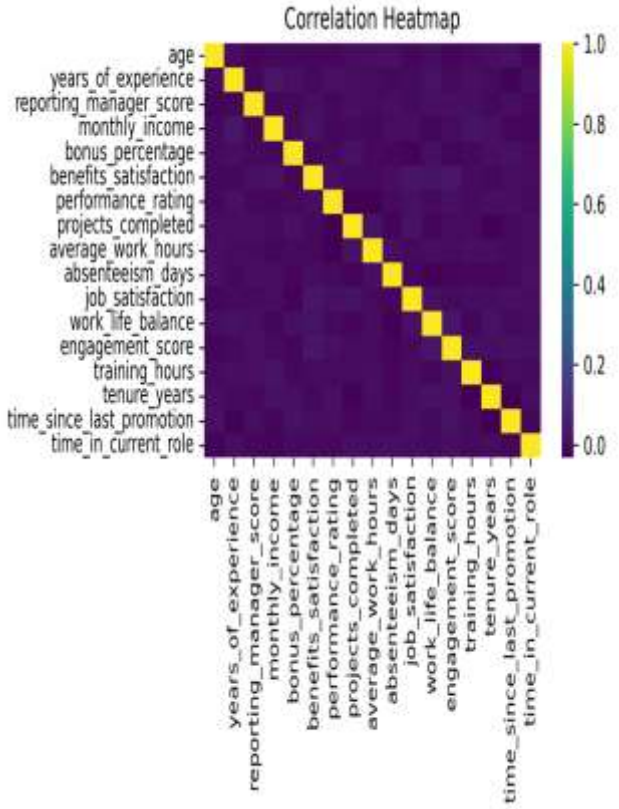


Figure 3: Feature Correlation Analysis for Employee Attrition Dataset

Figure 4 (a-c) shows the correlations between the main HR characteristics, such as age, monthly income, and job satisfaction, between the retained and the leavers. It points out observable differences in behavioral and financial characteristics of the two groups. The trends indicate that job fulfillment and aspects pertaining to engagement are significant contributors to attrition. In general, the visualization captures interactions of a complex relationship between variables, which justifies the use of sophisticated predictive models.

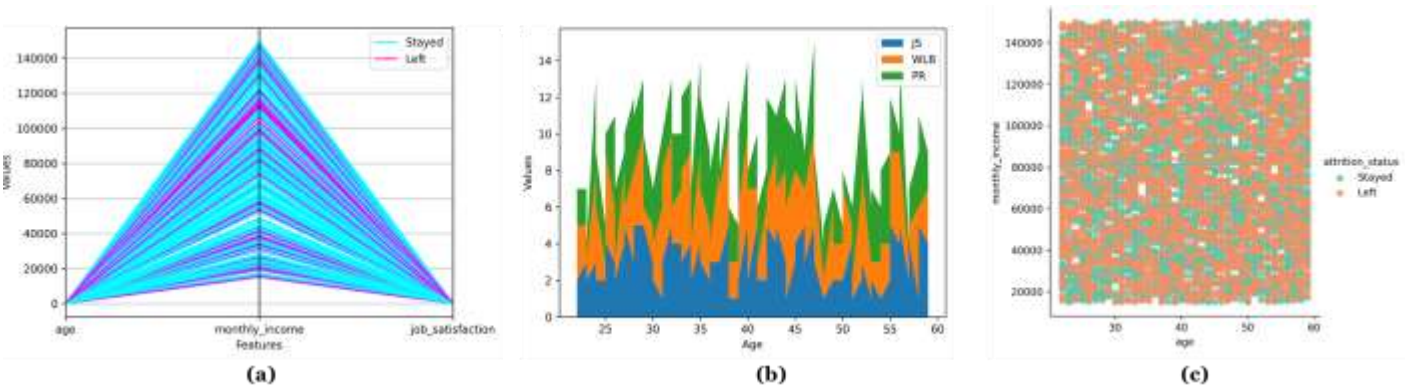


Figure 4: Employee Attribute Analysis: (a) Features by Attrition, (b) Metrics vs Age, (c) Age–Income by Attrition

4.2 SHAP-Based Feature Importance Analysis

Figure 5 shows a SHAP summary that helps understand the effect of various variables on the predictions made by the model. A dot here is indicative of the contribution made by the variable towards the making of a particular prediction, with the features arranged in order of their importance from top to bottom. Variables like tensile strength and void content have shown maximum influence on the output generated. The SHAP value of the prediction is measured along the x-axis, whereas its color value varies from blue (low) to yellow (high).

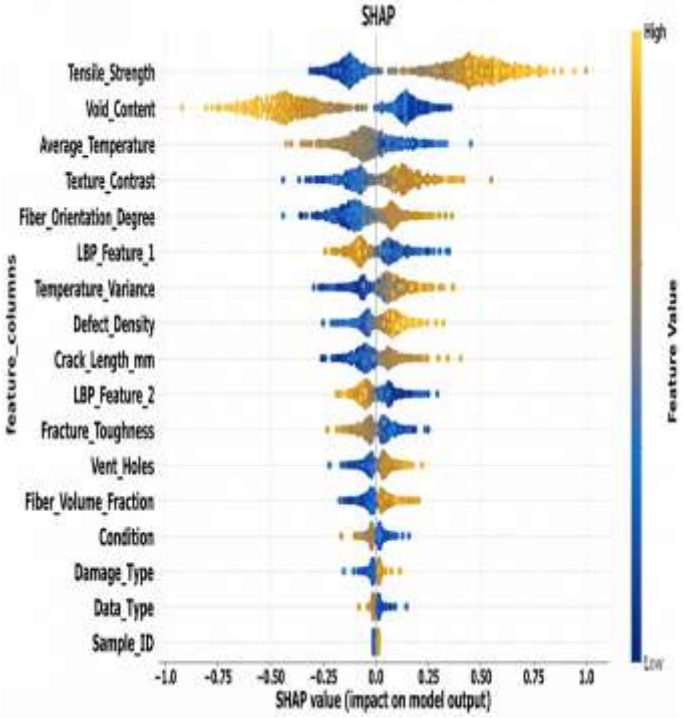


Figure 5: SHAP-Based Feature Contribution and Impact Analysis

4.3 Metrics explanation and their Comparative findings

Accuracy: Evaluates the model's overall accuracy (including attrition and non-attrition predictions).
Precision: Precisely Indicates how many employees predicted as "likely to leave" actually leave.
Recall: Calculates the number of real attrition cases that the model properly detects.
F1-Measure: The harmonic average of precision and recall, giving equal weight to false positives and false negatives.
ROC-AUC: Assesses the model's capacity to differentiate between those who leave and stay at all probability levels.

Performance evaluation compared with existing dataset [16]: Attrition is the dependent variable in this research, which is based on an IBM Watson Analytics dataset [16] of 1,500 employee datasets and 35 variables related to work, organizational, and demographic aspects. This dataset was used to train, evaluate and compare the suggested model with L-SVM. The findings of the researches are presented in Table 2 that shows that the proposed solution performs better than the baseline in all the measured aspects. This demonstrates its efficacy in HR analytics employee attrition forecasting.

Table 2: Performance Comparison on IBM Watson Analytics Dataset

Techniques	Accuracy	Precision	Recall	F1-measure
L-SVM [16]	0.925	0.92	0.93	0.92
IWSO-CRF [Proposed]	0.948	0.94	0.95	0.93

Performance evaluation compared with existing dataset [17]: The research data can be obtained by analyzing the IBM HR [17] open dataset of employee attrition, which implies understanding the domain and defining the objective, choosing the applicable features to prevent attrition, including satisfaction, performance, and workplace environment. The suggested model was trained, assessed with Existing dataset. The results demonstrate that the suggested method performs better than existing methods in terms of accuracy and ROC-AUC, demonstrating its efficacy for HR analytics in Table 3.

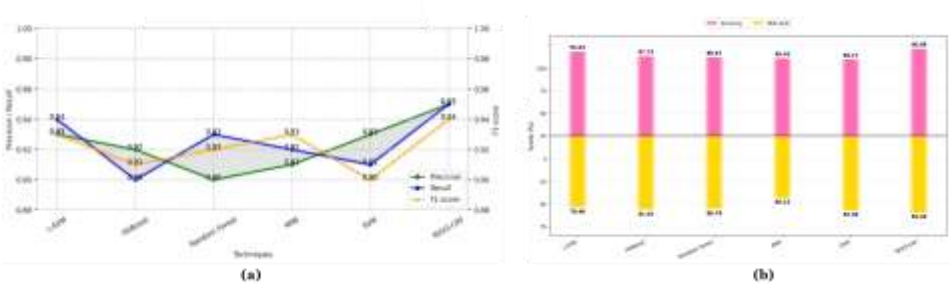
Table 3: Comparative Evaluation of Models on IBM HR Dataset

Techniques	ROC AUC	Accuracy
XGBoost [17]	79.23	85.91
RF [17]	79.20	85.81
KNN [17]	66.94	84.55
SVM [17]	80.78	83.87
IWSO-CRF [Proposed]	82.40	87.58

Performance evaluation compared with proposed dataset: The proposed research uses an Explainable HR Attrition Dataset [Proposed], which represents a full subset of employee data modeled in demographic, organizational, performance, compensation, engagement, and temporal (tenure, promotion history). Such data can be used to gain a more profound insight into employee turnover through the modeling of both personal factors and work environment. The models including L-SVM [17], XGBoost, RF, KNN, and SVM [18], were implemented and evaluated on the proposed dataset. The findings reveal the proposed model to be more effective in employee attrition prediction in HR analytics than the baseline methods shown in Table 4 and Figure 6 (a and b).

Table 4: Evaluation of ML Models on Explainable HR Attrition Dataset

Techniques	Accuracy (%)	Precision	Recall	F1-measure	ROC AUC
L-SVM	93.01	0.93	0.94	0.93	78.66
XGBoost	87.75	0.92	0.90	0.91	81.35
RF	86.67	0.90	0.93	0.92	80.78
KNN	85.42	0.91	0.92	0.93	68.15
SVM	84.17	0.93	0.91	0.90	82.04
IWSO-CRF	96.08	0.95	0.95	0.94	84.40



(a) (b)

Figure 6: Model Performance Analysis: (a) Model Comparison Across Techniques, (b) Overall Model Evaluation Across Techniques

5. Conclusion

The key issue that is explored in this research is the trade-off problem between interpretable and accurate ML models for HR analysis, as models that offer explainable decision-making fail to provide enough capacity for recognizing non-linear interactions between variables, whereas models with good predictive performance lack interpretability. In particular, such an issue could be solved using the implementation of the IWSO-CRF model along with post-hoc explainability methods and the construction of an end-to-end process chain, which includes preprocessing and PCA having the maximum predictive capacity and minimum interpretability. The suggested solution shows high predictive capacity, namely, 96.08% accuracy, 0.95 precision and recall, 0.94 F1-score, and 84.40 ROC-AUC and, therefore, proves itself applicable to predicting employee turnover. However, there are some potential difficulties associated with the implementation of the approach, such as the possibility of inaccuracy and inconsistency in explaining predictions post hoc, the decrease of feature interpretability due to PCA, low generalizability caused by the dependence on the dataset, and additional complexity because of the hybrid architecture. Further research should pay attention to the development of intrinsically interpretable models.

References

1. Love, P.E., Fang, W., Matthews, J., Porter, S., Luo, H. and Ding, L., 2023. Explainable artificial intelligence (XAI): Precepts, models, and opportunities for research in construction. Advanced Engineering Informatics, 57, p.102024.https://doi.org/10.1016/j.aei.2023.102024
2. Mathew, D.E., Ebem, D.U., Ikegwu, A.C., Ukeoma, P.E. and Dibiaezue, N.F., 2025. Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of AI models for human. Neural Processing Letters, 57(1), p.16.https://doi.org/10.1007/s11063-025-11732-2
3. Khan, N., Nauman, M., Almadhor, A.S., Akhtar, N., Alghuried, A. and Alhudhaif, A., 2024. Guaranteeing correctness in black-box machine learning: A fusion of explainable AI and formal methods for healthcare decision-making. IEEE Access, 12, pp.90299-90316.https://doi.org/10.1109/ACCESS.2024.3420415
4. Papadakis, T., Christou, I.T., Ipektsidis, C., Soldatos, J. and Amicone, A., 2024. Explainable and transparent artificial intelligence for public policymaking. Data & Policy, 6, p.e10.https://doi.org/10.1017/dap.2024.3
5. Ouared, A., May, M., Piau-Toffolon, C. and Dugué, N., 2026. Automating transparent learner profiling through explainable AI. Automated Software Engineering, 33(1), p.33.https://doi.org/10.1007/s10515-025-00574-w
6. Meas, M., Machlev, R., Kose, A., Tepljakov, A., Loo, L., Levron, Y., Petlenkov, E. and Belikov, J., 2022. Explainability and transparency of classifiers for air-handling unit faults using explainable artificial intelligence (XAI). Sensors, 22(17), p.6338.https://doi.org/10.3390/s22176338
7. Choubin, B., Jaafari, A., Henareh, J., Karimi, O. and Hosseini, F.S., 2025. Explainable artificial intelligence (XAI) for interpreting predictive models and key variables in flood susceptibility. Results in Engineering, 27, p.105976.https://doi.org/10.1016/j.rineng.2025.105976
8. Karahan, S.N., Güllü, M., Karhan, D., Çimen, S., Osmanca, M.S. and Barışçı, N., 2025. Realistic Performance Assessment of Machine Learning Algorithms for 6G Network Slicing: A Dual-Methodology Approach with Explainable AI Integration. Electronics, 14(19), p.3841.https://doi.org/10.3390/electronics14193841
9. Ortigossa, E.S., Gonçalves, T. and Nonato, L.G., 2024. Explainable artificial intelligence (xai)—from theory to methods and applications. IEEE Access, 12, pp.80799-80846.https://doi.org/10.1109/ACCESS.2024.3409843
10. Hoffman, R.R., Mueller, S.T., Klein, G., Jalaeian, M. and Tate, C., 2023. Explainable AI: roles and stakeholders, desirements and challenges. Frontiers in Computer Science, 5, p.1117848.https://doi.org/10.3389/fcomp.2023.1117848
11. Nannini, L., Alonso-Moral, J.M., Catalá, A., Lama, M. and Barro, S., 2024. Operationalizing explainable artificial intelligence in the european union regulatory ecosystem. IEEE Intelligent Systems, 39(4), pp.37-48.https://doi.org/10.1109/MIS.2024.3383155
12. Nazar, M., Unar, S., Ahmed, A., Su'ud, M.M., Alam, M.M. and Rahmat, A., 2026. Requirements Driven Explainable Artificial Intelligence Framework for Secure and Transparent Clinical Decision Support Systems. IEEE Access, 14, 29132 - 29144.https://doi.org/10.1109/ACCESS.2026.3664500

13. Swamy, V., Frej, J. and Käser, T., 2025. The future of human-centric eXplainable Artificial Intelligence (XAI) is not post-hoc explanations. Journal of Artificial Intelligence Research, 84.https://doi.org/10.1613/jair.1.17970
14. Awosika, T., Shukla, R.M. and Pranggono, B., 2024. Transparency and privacy: the role of explainable ai and federated learning in financial fraud detection. IEEE access, 12, pp.64551-64560.https://doi.org/10.1109/ACCESS.2024.3394528
15. Jiang, Z. and Zhang, Z., 2025. From black box to transparency: Enhancing automated interpreting assessment with explainable AI in college classrooms. Research Methods in Applied Linguistics, 4(3), p.100237.https://doi.org/10.1016/j.rmal.2025.100237
16. Al Akasheh, M., Hujran, O., Malik, E.F. and Zaki, N., 2024. Enhancing the prediction of employee turnover with knowledge graphs and explainable AI. IEEE Access, 12, pp.77041-77053.https://doi.org/10.1109/ACCESS.2024.3404829
17. Marín Díaz, G., Galán Hernández, J.J. and Galdón Salvador, J.L., 2023. Analyzing employee attrition using explainable AI for strategic HR decision-making. Mathematics, 11(22), p.4677.https://doi.org/10.3390/math11224677