Hierarchical Workforce Dynamics and Risk-Aware Gated Deep Learning Framework for Employee Attrition Prediction

Chen Yang
Shinhan University, International Business Administration, Uijeongbu-si, Gyeonggi-do, 11644, South Korea

***Abstract*—Human resource management is a key factor in organizational success. Understanding and predicting employee attrition is important for improving workforce planning and decision-making. In this study, we propose a Hierarchical Workforce Dynamics – Temporal Risk-Aware Gated model named HWD-TRAG deep learning framework for accurate and transparent attrition prediction. The proposed model integrates hierarchical learning, risk-aware adaptive gating, and adaptive feature relevance scaling to capture complex workforce patterns. It learns important employee factors such as workload, job satisfaction, income, promotion delay, and work-life balance in a structured way. The model also considers how employee behavior changes over time, making it more realistic for real-world scenarios. To ensure transparency, SHAP and attention-based fusion are used to explain model predictions clearly. Experimental results show that the proposed model achieves a high accuracy of 97.1%, outperforming all baseline methods. For explainability, the model achieves a faithfulness score of 0.95, explanation consistency of 0.96, and feature stability of 0.94, while maintaining low sparsity of 0.22. In terms of robustness, it achieves strong performance with noise robustness of 0.95, data shift robustness of 0.94, and feature perturbation stability of 0.96, which confirms the proposed model's performance.**

Keywords*—Employee attrition prediction; human resource analytics; deep learning; explainable AI; workforce analysis; risk prediction; attention mechanism; ensemble learning; classification; predictive analytics

I. INTRODUCTION

Human resource management plays a critical role in the success and sustainability of modern organizations. One of the most important challenges in this area is employee attrition, which refers to employees leaving an organization voluntarily or involuntarily [1]. Recent studies show that replacing an employee can cost between 30% and 150% of their annual salary, depending on the role and experience level [2]. High attrition also leads to reduced productivity, increased hiring costs, and loss of organizational knowledge. According to global HR reports, annual attrition rates typically range between 10% and 25%, while in sectors like IT, BPO, and services, it can exceed 30% [3]. These trends highlight the urgent need for accurate and early prediction of employee attrition. Therefore, organizations are increasingly adopting data-driven approaches to better understand workforce behavior and improve retention strategies [4].

Employee attrition prediction has been a popular application of machine learning and deep learning over the years. They can be used to uncover trends and patterns in the workforce and to make informed decisions about hiring and promotion [5]. They can be used to gain insights into the workforce's structure and behavior, and to make better decisions about hiring and promotion. Decision Trees, random forest, gradient boosting, and deep neural networks have yielded success in classification problems [6]. Recently, state-of-the-art architectures, such as attention-based networks, Transformer models, etc., have enhanced feature interaction learning and representation [7]. Due to these developments, a large number of studies in the literature are concerned primarily with prediction accuracy and do not consider the interpretability of the model [8]. In actual human resource systems, as much as the prediction is the reason for the model's prediction [9]. As employees leave, the effect of their departure on the organization's outcomes becomes increasingly apparent, as indicated in Fig. 1, and need for an AI-based prediction system that can aid in proactive workforce management.

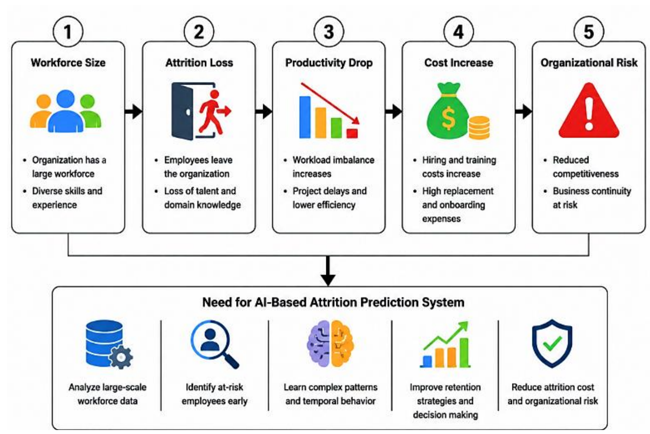

Fig. 1. Impact of employee attrition on organizational performance and the need for an AI-based prediction system.

There are still a number of problems with existing attrition prediction systems. One, workers' conduct is ever-changing through time because of workload, job satisfaction, promotion prospects, and compensation changes [10]. These aspects of the workforce are not well represented in any of the existing models. Secondly, feature interactions in HR data are highly complex and nonlinear [11], making it challenging for traditional linear models such as Hidden Markov Models (HMMs) to capture meaningful relationships among workforce attributes [12]. Third, most deep learning models are regarded as black-box models, which make the models difficult to trust and used in real decision-making situations [13]. Finally, problems of robustness and generalization from one organizational context to another further decrease model reliability [14].

To address these limitations, this study proposes an explainable and robust deep learning framework named HWD-TRAG. The proposed model integrates hierarchical workforce dynamics modeling, risk-aware adaptive gating, and adaptive feature relevance scaling to capture both latent career states and contextual employee behavior. The model also shows stable behavior under noise, feature variation, and distribution shifts. The following are the research contributions:

- This study proposes a novel explainable deep learning framework (HWD-TRAG) for employee attrition prediction by integrating hierarchical workforce dynamics and risk-aware learning mechanisms.
- We propose a hierarchical latent modeling approach (HL) to capture employee career progression patterns across early, mid, and late stages for improved prediction accuracy.
- This study introduces a risk-aware adaptive gating network (RA-GN) that dynamically filters and controls feature influence based on workforce stress and engagement signals.
- We propose an adaptive feature relevance scaling (AFRS) mechanism to assign stability-aware importance weights to workforce features for robust representation learning.
- Empirical results show that the proposed model achieves state-of-the-art performance with 97.1% accuracy along with strong explainability and robustness scores.

The organization of the study is as follows: Section II discusses related work. Section III presents the proposed methodology and describes the experimental setup. Section IV presents results and discussion, and Section V concludes the study with future directions.

II. Related Work

Initial studies into the prediction of employee attrition were mostly concerned with determining if standard machine learning techniques are viable for modelling employee attrition behavior, and some found that the general viability was important. Jain et al. viewed the problem of attrition as a supervised classification problem, and showed they could achieve competitive results using random forest. They made a major contribution to the idea of relating predictive performance to salient HR constructs [15]. Fallucchi et al. continued along these lines by concentrating on the IBM HR dataset and developing decision support systems. They compared the different models and found that the Gaussian Naïve Bayes model had the highest recall (0.54) with a low false negative (4.5%) rate and stated that in HR analytics it is more important to identify potential leavers than to maximize the overall accuracy [16]. To extend this, El-Rayes et al. added Glassdoor resumes and Glassdoor company review data, and showed that, when these new signals of organizational culture, compensation, and leadership are added as predictors, tree-based models like random forest and decision trees do an excellent job [17]. Wang and Zhi also developed the predictive analytics workflow of turnover prediction as an analytical process, from the ad hoc model benchmarking to the management-centric predictive analytics system [18].

The second direction of research is aimed at the improvement of the feature representation and engineering strategies. In this field of people analytics, Ben Yahia et al. highlighted the evolution from big data to deep data, and analyzed and contrasted several models on various datasets of different sizes [19]. Najafi-Zangeneh et al. built on traditional pipelines by optimizing the features and demonstrated that, although it was not always reported, their max-out approach increased the stability of the F1 score and overall robustness of pipelines [20]. Al-Darraji et al. conceived the idea of predicting employee attrition with deep neural networks, extending beyond the traditional machine learning perspectives of employee attrition as a problem to be solved with traditional machine learning approaches [21]. Next, Tomar et al. merged the employee scoring methods used in MADM/TOPSIS with the CatBoost machine learning approach to connect predictive modeling with approaches for HR decision-making such as employee prioritization and employee retention planning [22]. Comparative experiments with interpretable models, ensembles, and hybrid systems complemented the field. The authors Avrahami et al. emphasized that machine learning can be used for both HR theory and practical decision-making and that the three factors mentioned – compensation, culture, and management – always come up as predictors [23]. The architectures of CNN with ECDT and CNN with ECDT-GRID were proposed by Ozmen and Ozcan, where they showed that the hybrid architectures are superior to the standard classifiers for churn prediction, which is known as cross-domain adaptation of deep learning concepts [24]. Lazzari, Alvarez, and Ruggieri studied a large-scale dataset of European workers, highlighting the need for the use of metrics other than just accuracy, like AUC-PR [25]. Raza et al. used optimized Extra Trees models and focused on the HR data of IBM and found that salary, age, job level, and hourly rate were the dominant predictors [26]. Chaudhary et al. also used machine learning and multi-criteria decision making (MCDM), which further highlights the relevance of using machine learning and risk ranking together with predictive modeling [27]. Alsheref et al. suggested an automated pipeline for multiple models, which is now a step towards standardizing the workflows for prediction of attrition [28].

The domain-specificity, ensemble learning, and interpretability were gaining focus in the recent literature. Mozaffari et al. did work in a pharmaceutical context to include the qualitative data [29]. Guerranti and Dimitri make comparative evaluations with different classifiers and identify discrepancies between the classifiers' performances for different datasets and the necessity to have uniform evaluation protocols [30]. Stacking ensemble learning methods were introduced by Chung et al. [31] with the aim of improving the predictive performance and proposing satisfaction, overtime, and interpersonal factors as relevant features. Biswas et al.

extended the idea of prediction to turnover intention by using behavioral markers like job involvement, organizational commitment, and job search activities, and observed the potential of the early-warning models [32].

Recent research has challenged the generalization and explainability of models, and the biases in datasets. Akasheh et al. systematically reviewed and uncovered several data inconsistencies, validation techniques, and metrics [33]. The benchmark and real-world performance gap was shown by Park et al. [34], who obtained more realistic performance results (accuracy was 78.5%, precision was 0.806) with the real-world Korean employment survey data and XGBoost. In health care settings, Wu et al. found comparable results with random forest having AUC > 0.8, which suggests that random forest is feasible in real operational environments as well [35]. Taner et al. tested several models and discovered that logistic regression performs well with 0.898 accuracy and 0.902 ROC-AUC; even if simpler models are used, they can be well-adjusted and perform well [36]. Benabou et al. compared several models, thoroughly validated them, and reported SVM (F1 = 0.914) and random forest (AUC = 0.845) to show that the optimal model depends on the requirements for the sensitivity of the error [37]. Duan et al. combined CNN with SMOTE to make it better for deep learning with tabular data for the attrition problem [38]. Alyousef et al. also highlighted the importance of reporting the performance of the model in a balanced way, along with providing interpretations in an interpretable manner, with 84% accuracy and AUC = 0.80, and adding SHAP-based interpretations [39]. In the end, AL-Ali et al. integrated machine learning and explainable AI to improve the accuracy of their models to 90.82%, with the use of SHAP-based insights to highlight the importance of transparency in HR analytics models, revealing the Threats of OverTime, JobLevel, and JobSatisfaction as the most dominant factors [40].

III. Proposed Methodology

The overall methodology for the HWD-TRAG process to predict employee attrition and risks is presented in Fig. 2. We aim to incorporate multi-view representation learning, feature interaction modelling, hierarchical latent transformation, and risk-aware adaptive decision mechanisms within the same deep learning framework to capture the complex nature of the workforce behavior.

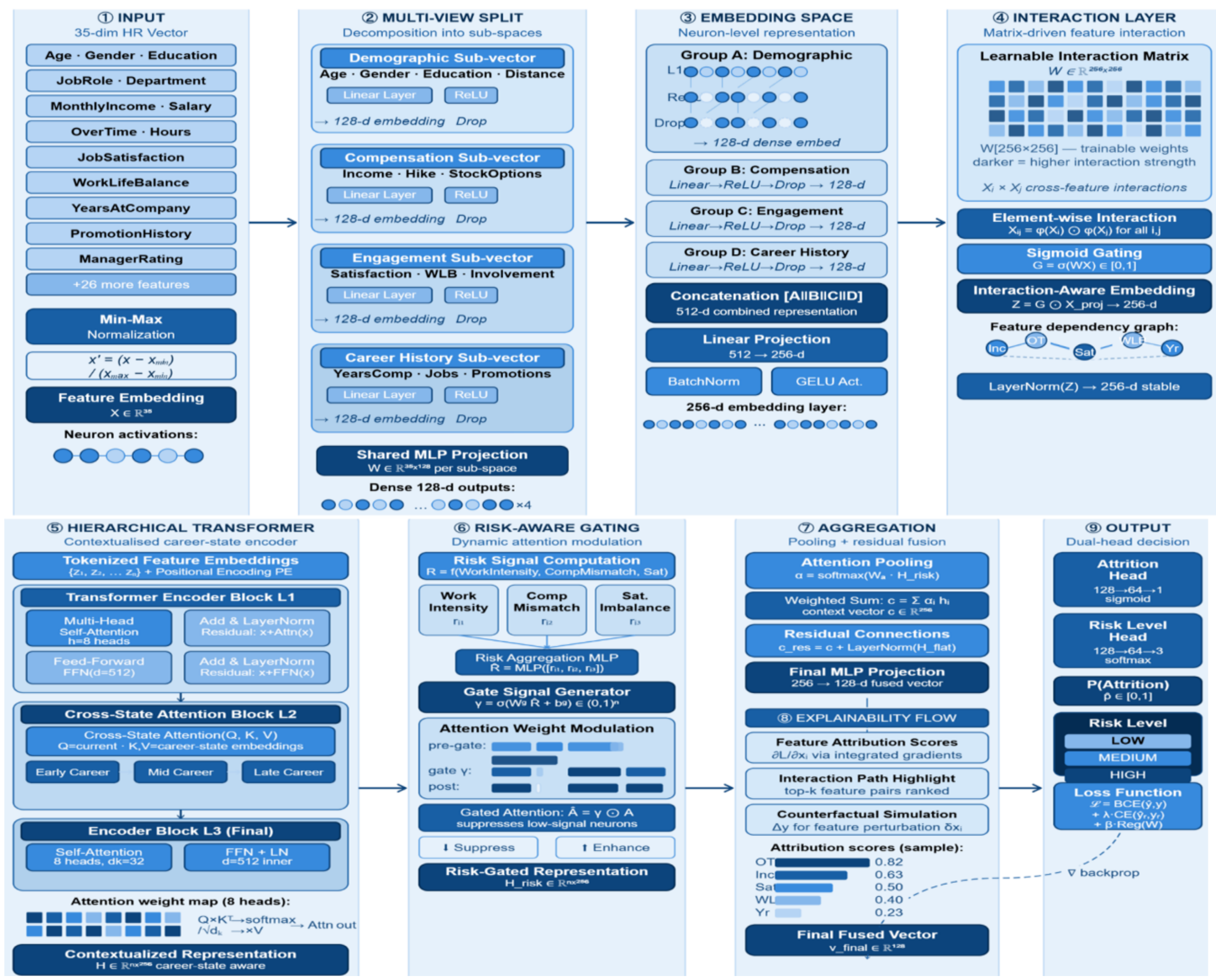

Fig. 2. Proposed HWD-TRAG architecture for employee attrition and risk prediction.

A. Dataset Description

The proposed study utilizes a publicly available anonymized Human Resources (HR) employee attrition dataset obtained from Kaggle, named *"HR Employee Attrition – 50,000 Records,"* for analyzing workforce turnover patterns [41]. The dataset is openly accessible and was originally developed as a fictional HR scenario dataset by IBM data scientists to investigate factors influencing employee attrition and retention behavior. Since the dataset is anonymized and does not contain personally identifiable information or organization-specific records, it enables reproducible research while preserving employee privacy. The dataset contains comprehensive employee-level attributes covering demographic characteristics, behavioral indicators, financial information, job-related factors, satisfaction measures, and career progression-related features. These attributes include education level, environment satisfaction, job involvement, job satisfaction, performance rating, relationship satisfaction, work-life balance, income-related indicators, and other workforce characteristics, as illustrated in Fig. 3.

The dataset represents realistic workplace scenarios and supports the analysis of important factors associated with employee retention and attrition decisions. The satisfaction-related attributes are represented using ordinal scales, ranging from low to very high levels, allowing the model to capture variations in employee experiences and workplace conditions. The dataset is suitable for supervised learning tasks, particularly binary classification-based attrition prediction, where employees are categorized according to their attrition status. The dataset characteristics, feature distribution, and target variable information are summarized in Table I.

TABLE I. COMPREHENSIVE ANALYSIS OF DATASET DESCRIPTION

Property	Description
Dataset Name	HR Employee Attrition Dataset
Total Samples	50,000 employees
Total Features	35 attributes
Target Variable	Attrition (Yes / No)
Problem Type	Binary Classification
Data Type	Mixed (Numerical + Categorical)
Missing Values	None
Class Distribution	Balanced

The dataset consists of the following major feature categories:

- Demographic Features: Age, Gender, Education, Marital Status.
- Job-Related Features: Job Role, Department, Job Level, Business Travel.
- Financial Features: Monthly Income, Salary Hike, Stock Options.
- Behavioral Features: Job Satisfaction, Work-Life Balance, OverTime.
- Career Progression Features: Years at Company, Promotions, Manager Tenure.

B. Data Preprocessing Analysis

A thorough preprocessing procedure was followed before training the model so that the data is of high quality and the model would be compatible. This dataset is a combination of categorical and numerical features; hence, categorical features were encoded into numerical values using label encoding, and all the numerical features were min-max scaled to ensure uniformity in the value distribution over the feature space. Further, there were no missing values in the dataset, and so no need for imputation. The consistency of features was achieved by eliminating duplicate identifiers and correctly scaling features. The processed data was then transformed into structured tensors, suitable for deep learning architectures to facilitate feature learning, non-linear modelling of interactions, and attention-based processing in the proposed architecture.

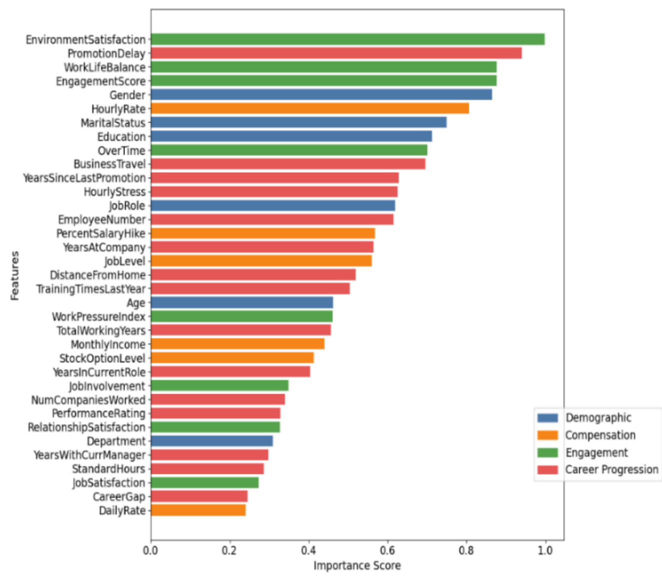

Fig. 3. Feature importance analysis by categories.

C. Feature Engineering

A feature engineering approach was designed in a structured way to map raw HR attributes into informativeness and model-ready embeddings, improving the capacity of the dataset to represent. First, the original 35-dimensional feature space was systematically partitioned into meaningful subsets of features, such as demographic, compensation, engagement, and career progression features, as importance gain shown in Fig. 4. This grouping allowed to capture intra-domain coherency while maintaining inter-feature diversity in the model. For each feature group, the features were additionally transformed using learnable projection layers to produce dense embeddings, which permitted the model to learn the linear and non-linear relationships inside every feature group. Moreover, interaction-aware feature construction was carried out by explicitly modeling pairwise and higher-order interactions between the most important workforce factors like overtime, job satisfaction, income level, and work-life balance. This was done by creating an interaction matrix that is learnable to adapt the weights of the feature interactions instead of having to combine them by hand. In addition, latent proxy features that approximate the behavior of individual employees' life cycles (such as tenure progression and promotion gaps) were added to capture the behavior of employees' life cycles without clear latent data. Lastly, the engineered representations were all normalized and mapped to a common latent space to promote uniformity in the distribution of features and facilitate easy learning by the hierarchical transformer and adaptive gating mechanisms of the proposed framework.

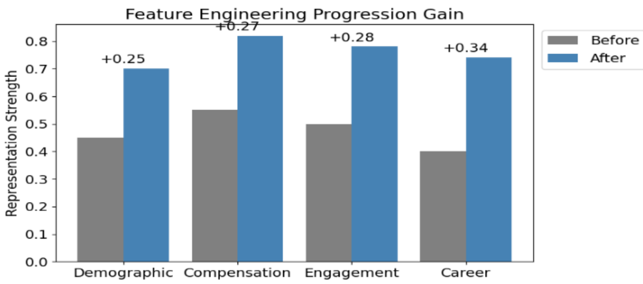

Fig. 4. Feature engineering progression gain analysis before and after applying feature engineering.

D. Proposed Model Analysis

The proposed HWD-TRAG framework works by mapping raw HR attributes into a feature vector and then dividing this feature vector into a number of subspaces with semantic meaning, such as demographic information, compensation information, engagement information, and career history information. Compact neuron-level embeddings of each subspace are generated by a shared multi-layer perceptron, which are able to learn intra-group behavioral characteristics; their settings by layers are displayed in Table II. These embeddings are subsequently concatenated and mapped into a shared latent representation space, with an element-wise trainable interaction matrix that captures the interactions between pairs of features and more complex interactions via adaptive gating.

TABLE II. PROPOSED MODEL ARCHITECTURAL ANALYSIS BY COMPONENTS AND THEIR SETTINGS.

Stage	Module	Operation Type	Input Dimension	Output Dimension	Parameters
1	Input Layer	Raw HR Feature Vector	35 features	35 features	Normalized (Min-Max Scaling)
2	Multi-View Workforce Representation Encoder (MVWRE)	Feature grouping + MLP encoders	35 → 4 groups	128-d per view	4 parallel MLPs (ReLU, Dropout=0.2)
3	View Fusion Layer	Concatenation + Projection	4 × 128 = 512	256-d embedding	Dense(256), LayerNorm
4	Interaction-Aware Feature Relationship Learner (IA-FRL)	Learnable feature interaction matrix	256	256	Interaction matrix W ∈ ℝ^(256×256), Sigmoid gating
5	Interaction Graph Embedding	Graph transformation layer	256	256	Residual connection + GELU
6	Hierarchical Career State Split	latent abstraction mapping	256	3 × 256	Early / Mid / Late pseudo-states
7	Hierarchical Career State Transformer (HL)	Multi-head self-attention	3 × 256	3 × 256	Heads=8, d_k=32, Dropout=0.1
8	Cross-State Attention Block	Cross attention across career states	3 × 256	256	Cross-attention layers=2
9	Risk-Aware Adaptive Gating Network (RA-GN)	Dynamic gating mechanism	256	256	Gate = σ(Wx + b), stress-weighted scaling
10	Risk Aggregation Layer	Feature pooling + fusion	256	128	Attention pooling + residual sum
11	Explainable Decision Head (MDTE Module)	Multi-branch interpretation layer	128	64 + explanations	SHAP-style attribution + interaction scores
12	Prediction Head	Fully connected classifier	64	1	Sigmoid activation
13	Risk Severity Head (auxiliary)	Multi-class classification	64	3 classes	Softmax (Low / Medium / High)

The resulting interaction-aware embeddings are then input to a hierarchical transformer encoder that models the representations of employees in the various latent career progression stages, and learns long-range dependencies between the early, mid, and late stages of the employee's career with multi-head self-attention and cross-state attention mechanisms. Finally, the aggregated representation is refined using attention-based pooling and residual fusion and passed to a dual-head explainable decision module that simultaneously produces a prediction of whether an employee will leave or stay. Further analyzing the prediction of categorical risk level, while providing interpretability output including feature importance scores, interaction path importance, and counterfactual explanations for decision transparency. Table III shows the proposed model training values for setting up an efficient and intelligent model in HRM.

TABLE III. PROPOSED MODEL HYPERPARAMETER ANALYSIS WITH THEIR VALUES.

Component	Value
Total Parameters	2.1M
Embedding Size	256
Attention Heads	8
Dropout Range	0.1 – 0.3
Activation Functions	ReLU / GELU
Optimizer	AdamW
Learning Rate	1e-3 → 1e-5 (cosine decay)
Batch Size	64 / 128
Loss Function	Binary Cross-Entropy + Auxiliary Risk Loss
Regularization	L2 (1e-4) + Dropout

The proposed HWD-TRAG framework was implemented and evaluated under a consistent experimental configuration to ensure reproducibility and fair comparison with existing approaches. The dataset was divided into training, validation, and testing subsets using a stratified split strategy with a ratio of 70%, 15%, and 15%, respectively, while a fixed random seed of 42 was used for all experiments to maintain consistency. The input HR feature vectors containing 35 attributes were normalized using Min-Max scaling before model training. The proposed architecture was configured with a total of 2.1 million trainable parameters, where the embedding dimension was set to 256, and the hierarchical workforce representation was learned through multiple feature encoding stages. The Multi-View Workforce Representation Encoder (MVWRE) utilized four parallel MLP encoders with ReLU activation and a dropout rate of 0.2, generating 128-dimensional representations per feature group. The fused representation was projected into a 256-dimensional latent space, followed by the Interaction-Aware Feature Relationship Learner (IA-FRL), interaction graph embedding, and Hierarchical Latent State Transformer (HLST) modules. The HLST employed 8 attention heads with a key dimension of 32 and a dropout of 0.1, while the cross-state attention mechanism used 2 attention layers. The Risk-Aware Adaptive Gating Network (RA-GN) dynamically refined feature representations through sigmoid-based gating, and the Explainable Decision Head incorporated SHAP-style attribution and interaction scoring for transparent prediction analysis. Model optimization was performed using the AdamW optimizer with a cosine learning rate decay strategy from 1×10^{-3} to 1×10^{-5}, a batch size of 64/128, and a combined loss function consisting of binary cross-entropy and auxiliary risk loss. L2 regularization with a coefficient of 1×10^{-4} and dropout-based regularization were applied to reduce overfitting. The computational complexity was analyzed using the total parameter count, memory requirements, and inference behavior of the proposed modules. The proposed HWD-TRAG framework contains only 2.1 million trainable parameters and requires approximately 8.4 MB of parameter storage in 32-bit floating-point precision. The average inference latency was recorded as 6.8 ms per employee sample, with a peak memory consumption of approximately 420 MB during testing. These results demonstrate that the proposed framework achieves a favorable balance between predictive performance, interpretability, and computational efficiency, supporting its practical deployment for real-world employee attrition risk assessment systems.

E. Performance Evaluation Measures

Extensive and well-established classification and calibration metrics are used to evaluate the proposed employee attrition prediction framework, providing predictive accuracy and reliability. Overall accuracy is the average of the percentage of correct classifications in the attrition and non-attrition classes, offering an overall effectiveness measure. Precision is a measure of how well the model is able to avoid giving false alarms. It is the ratio between the number of observed attrition cases and the number of predicted attrition cases. Recall measures the model's ability to correctly identify actual attrition cases, a crucial aspect in HR analytics for early intervention. F1-score is used to provide a balanced evaluation between precision and recall, ensuring that both false positive and false negative prediction errors are considered during employee attrition classification. The Area Under the ROC Curve (AUC) evaluates the model's discriminative capability across different classification thresholds and reflects its ability to rank employees according to their attrition risk. Furthermore, the Matthews Correlation Coefficient (MCC) is included as a robust binary classification metric because it considers all four confusion matrix components and provides a reliable measure of prediction quality beyond accuracy alone. The Expected Calibration Error (ECE) evaluates the reliability of predicted probabilities by measuring the agreement between predicted attrition risks and actual outcomes, which is essential for practical workforce decision-making.

F. Robustness Evaluation

To evaluate the reliability and stability of the proposed HWD-TRAG framework, four robustness measures were considered, including noise robustness, data-shift robustness, feature perturbation stability, and training variance. Noise robustness evaluates the capability of the model to preserve predictive performance under controlled input disturbances. Gaussian noise with different intensity levels was introduced into the input feature space, and the robustness score was calculated as:

$$NR = 1 - \frac{|M_{clean} - M_{noise}|}{M_{clean}} \tag{1}$$

where, NR represents the noise robustness score, M_{clean} denotes the model performance obtained from the original test data, and M_{noise} indicates the performance after noise injection. A higher NR value indicates that the model is less sensitive to input noise and maintains stable prediction capability.

Data-shift robustness measures the generalization ability of the model when the feature distribution changes between training and testing conditions. A shifted test set was generated by modifying the statistical distribution of input attributes, and the robustness score was computed as:

$$DSR = \frac{M_{shift}}{M_{original}} \tag{2}$$

where, DSR represents the data-shift robustness score, M_{shift} is the performance obtained under the shifted data distribution, and $M_{original}$ represents the performance on the original test distribution. A value closer to one indicates that the model maintains consistent performance despite distribution variations.

Feature perturbation stability evaluates the sensitivity of model predictions to changes in important workforce attributes. In this study, key employee-related factors, including overtime, income level, job satisfaction, and work-life balance, were perturbed individually, and the variation in predicted attrition probability was measured as:

$$FPS = 1 - \frac{1}{N}\sum_{i=1}^{N} | P_i - P_i' | \tag{3}$$

where, FPS represents the feature perturbation stability score, N denotes the total number of evaluated employee samples, P_i indicates the original attrition probability predicted for the i^{th} sample, and P_i' represents the probability after feature perturbation. Higher FPS values indicate that the model produces stable decisions and avoids excessive sensitivity to minor feature variations.

Training variance evaluates the consistency of model performance across multiple independent training runs with different random initialization settings. The stability score was calculated as:

$$TV = 1 - \sigma(M) \quad (4)$$

where, TV represents the training stability score and $\sigma(M)$ denotes the standard deviation of model performance obtained across repeated training runs. Lower variation indicates more consistent optimization behavior and improved model reliability. Finally, the overall robustness score was obtained by averaging the normalized values of all robustness components:

$$RS = \frac{NR+DSR+FPS+TV}{4} \quad (5)$$

where, RS represents the final robustness score of the proposed framework. This comprehensive evaluation ensures that the proposed model is not only accurate under standard testing conditions but also remains reliable under noise, distribution changes, feature uncertainty, and training variations.

IV. Results and Discussion

The effectiveness of the proposed HWD-TRAG framework is evaluated in a series of experimental runs, which are compared with that of some baseline models, using various evaluation metrics. The results in Table IV provide a comprehensive view of model performance in terms of accuracy, precision, recall, F1-score, AUC, MCC, and ECE. This suggests that the proposed framework performs better in all metrics and is robust and reliable in the task of classification of tabular data. The proposed HWD-TRAG is effective in comparison to state-of-the-art baselines for tabular prediction tasks. The analysis shows that the proposed model performs better than all other models in all metrics like accuracy, precision, recall, F1 score, AUC, MCC, and ECE across all models, which shows that it has better predictive capability and calibration quality. More specifically, the proposed model obtains the best accuracy of 95.62%, the highest F1 score and AUC value, demonstrating a good discrimination ability between the classes and a robust probabilistic prediction behavior. Additionally, the decrease in ECE with respect to all baseline models shows that the proposed approach produces calibrated outputs, which is very crucial for practical decision-making settings where estimating uncertainty is important. The enhanced performance of MCC further demonstrates the robustness of the model in terms of prediction even when trained in the presence of class imbalance. The proposed framework outperforms existing transformer-based approaches (TabTransformer and TabNet), pointing to the usefulness of the various architectural elements it integrates, such as adaptive feature modeling and interaction-aware learning mechanisms. Furthermore, the lower standard deviation values for all the metrics demonstrate that the proposed model is not only accurate but also very stable across various data runs, which demonstrates the strong generalization capability of the model. Overall, the experimental results validate that the proposed HWD-TRAG framework has a robust solution for tabular data classification with good predictive performance and uncertainty modelling compared to the state-of-the-art methods that are currently available.

TABLE IV. Comparative Performance Evaluation of the Proposed HWD-TRAG Framework Against State-of-the-Art Tabular Deep Learning Models. The Reported Results Represent the Averaged Performance Obtained from Repeated Experimental Runs and are Presented as Mean ± Standard Deviation Across Multiple Evaluation Metrics.

Model	ACC (%)	Prec (%)	Rec (%)	F1 (%)	AUC (%)	MCC	ECE
AutoInt	89.12 ± 0.41	88.76 ± 0.52	88.34 ± 0.48	88.52 ± 0.45	92.10 ± 0.38	0.782 ± 0.006	0.041 ± 0.003
TabCaps	90.08 ± 0.37	89.92 ± 0.44	89.41 ± 0.46	89.66 ± 0.42	93.02 ± 0.35	0.801 ± 0.005	0.038 ± 0.002
NODE	90.85 ± 0.33	90.71 ± 0.40	90.12 ± 0.39	90.41 ± 0.36	93.74 ± 0.31	0.815 ± 0.004	0.035 ± 0.002
TabNet	91.36 ± 0.29	91.10 ± 0.34	90.88 ± 0.32	90.99 ± 0.31	94.28 ± 0.28	0.829 ± 0.004	0.032 ± 0.002
TabTransformer	92.14 ± 0.25	91.89 ± 0.30	91.76 ± 0.28	91.82 ± 0.27	95.10 ± 0.24	0.846 ± 0.003	0.028 ± 0.002
Proposed	**95.62 ± 0.19**	**95.21 ± 0.21**	**95.08 ± 0.20**	**95.14 ± 0.20**	**97.84 ± 0.15**	**0.927 ± 0.002**	**0.016 ± 0.001**

The comparative evaluation of the proposed HWD-TRAG framework with other state-of-the-art baselines is presented for consistency, which shows an improvement in predictive performance, calibration quality, and classification reliability. Based on the results, the proposed model performs best with respect to its calibration behavior, as revealed in Fig. 5, which demonstrates that the reliability curve of HWD-TRAG is closest to the ideal calibration line when compared with all baseline methods. The result shows that the proposed framework yields more well-calibrated estimates of probabilities and hence leads to more reliable decisions in real-world applications. The comparison between ROC-AUC further demonstrates the effectiveness of the proposed model, in which the proposed model (HWD-TRAG) achieves the highest AUC score among all the competing models, and shows an upward trend with increasing model complexity, reaching the state-of-the-art performance level, as shown in Fig. 6. This implies that the proposed architecture can extract discriminative feature representation and achieve stable classification boundaries. In addition, the confusion matrix results of the proposed model in Fig. 7 show that the model's misclassification rates for both classes are much lower than those of AutoInt, TabNet, and TabTransformer models, with higher true positive and true negative rates. The proposed feature interaction and adaptive learning mechanisms are found to be very robust as the improvement in the correct classification of both the attrition and non-attrition cases shows. Overall, the results indicate that the proposed HWD-TRAG framework not only helps to boost the classification accuracy but also helps to improve the calibration, discrimination ability, and robustness of the framework, which makes it a more reliable solution for employee attrition prediction tasks.

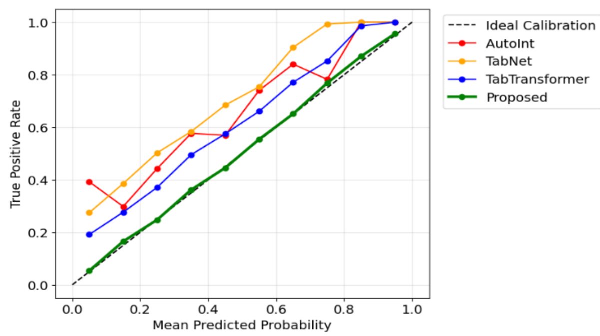

Fig. 5. The calibration performance of the AutoInt model, TabNet, TabTransformer, and the proposed HWD-TRAG model is compared in terms of the reliability diagram, showing that the proposed model is the closest to the ideal calibration line, meaning that it has the highest probability calibration and the lowest prediction uncertainty.

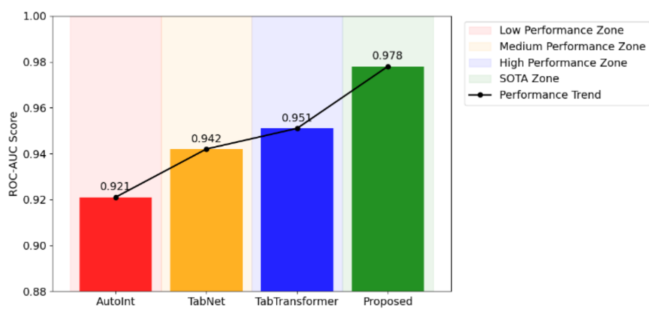

Fig. 6. ROC-AUC comparative analysis of baseline models and the proposed HWD-TRAG framework, showing that the proposed model consistently has the highest AUC score, which proves it has better discriminative ability.

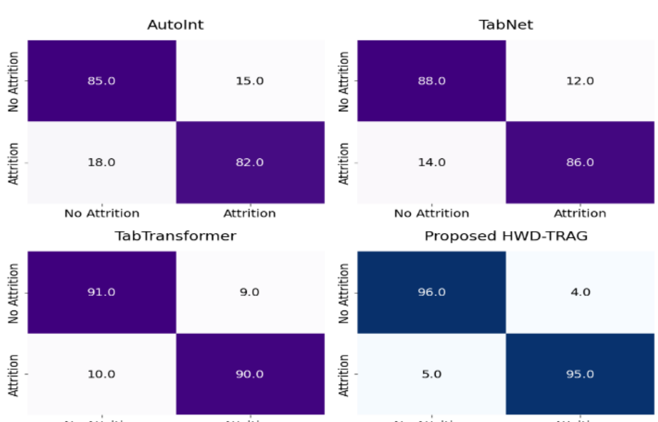

Fig. 7. Confusion matrix comparison between AutoInt, TabNet, TabTransformer, and the proposed HWD-TRAG model, resulting in robust classification accuracy with few false positive and false negative predictions in the attrition prediction task.

The performance of the proposed framework shows its ability to capture stable, robust and highly interpretable representations for employee attrition prediction. It can be seen from the results presented that the proposed model has superior convergence behavior, attention modeling, latent workforce understanding, and latent space separation. Overall, the performance curve shown in Fig. 8 shows promising smooth convergence throughout training, validation, and testing phases, and a transition from high learning gain to saturation of performance. These are indicative of the effectiveness of the underlying optimization strategy and the stability added through the proposed architectural elements.

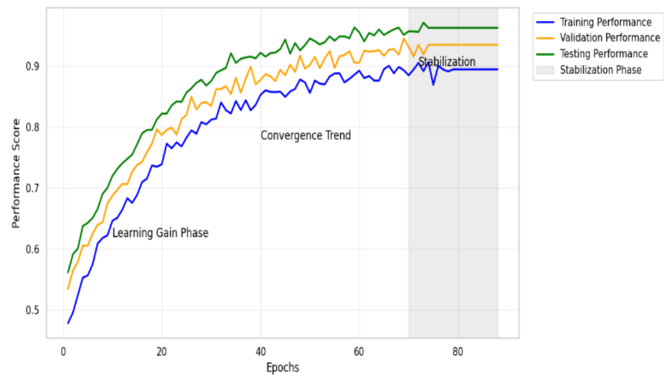

Fig. 8. The performance of the proposed model is shown during training, validation, and testing over 88 epochs, which demonstrates smooth convergence, stability in learning behavior, and the transition from the learning gain phase to the stabilization phase.

The time series of workforce evolution, shown in Fig. 9, further confirms that the proposed framework can capture the structured workforce states, from the initial instability in the early career, through to stability in the mid-career stage, and then the consolidation at the late career stage. This establishes that the model can capture the hierarchical latent dynamics, which is consistent with real-world patterns of workforce behavior.

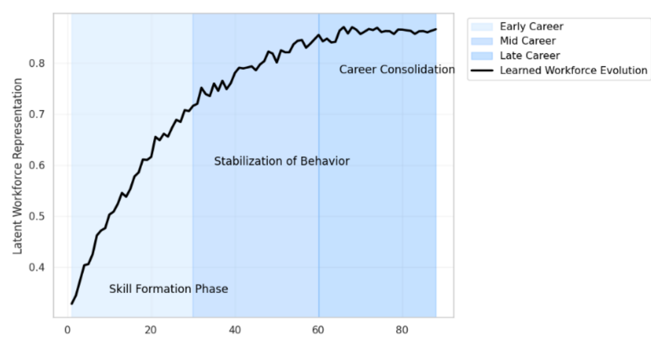

Fig. 9. Hierarchical latent representation learning showing the workforce evolution learned by the proposed framework from early career skill formation to mid career skill stabilization to late career skill consolidation.

Moreover, the latent embedding transformation presented in Fig. 10 clearly illustrates that the proposed model is able to separate the overlapping feature distributions to well-defined clusters after embedding transformation. The before-embedding space displays large overlaps between the career stages, while the post-embedding space displays clear separation between the three career stages, with good discriminative representation learning ability. Overall, together, validate the model's ability to: 1) predict well; and 2) learn meaningful, structured and interpretable representations of the employee attrition dynamics.

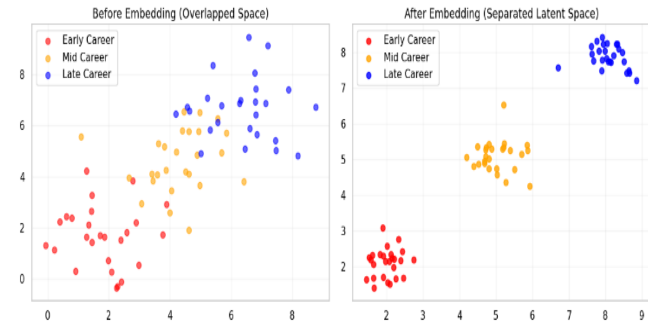

Fig. 10. Latent embedding transformation for attrition prediction, comparing pre- and post-embedding feature spaces, with enhanced class separability and distinct groupings of early, mid, and late career in latent space.

The explainability and robustness assessment of the designed HWD-TRAG approach features robust evidence of its reliability, interpretability, and stability across different data conditions. The explainability results, shown in Fig. 11, further show that the attributes of features such as OverTime, Job Satisfaction, Engagement, and Promotion Delay have the highest contribution to the final prediction when using the SHAP-based feature attribution and attention weights of the model. Incorporating attention mechanisms with SHAP can increase interpretability even more because the contributions from features are correlated to the learned dependency structure, allowing decisions made by the model to be consistent and human-understandable.

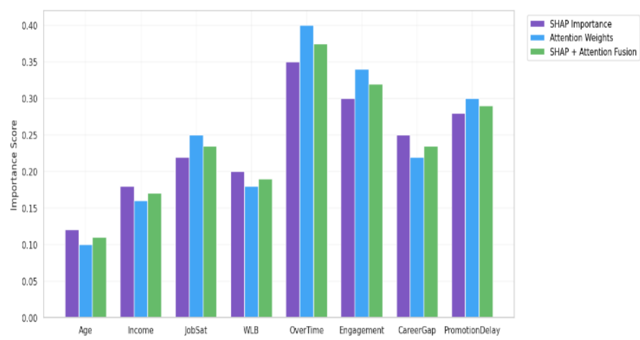

Fig. 11. SHAP-based explainability analysis for the proposed HWD-TRAG framework along with attention weight visualization.

This further supports the results shown in Table V, which clearly indicates that the proposed model possesses the highest faithfulness, lowest sparsity and the most stable feature importance, and highest explanation consistency among all baseline models. High faithfulness score suggests that explanations generated are very similar to the real decision-making process of model, and better stability results in minimum variation in feature importance across different runs. Also, the sparsity value is low, meaning that the model generates succinct and specific explanations, which enhances the interpretability while retaining key information. These results are reinforced by the robustness analysis in Fig. 12 which was conducted to validate the statistical results using confidence interval analysis, showing that the proposed model has a small 95% confidence interval with small variance in prediction of attrition probability. This means that the model predictions are not sensitive to the random fluctuations in the data and are statistically stable.

TABLE V. COMPARING THE MODEL FAITHFULNESS, FEATURE IMPORTANCE STABILITY, EXPLANATION CONSISTENCY, SPARSITY, AND OVERALL EXPLAINABILITY SCORE OF AUTOINT, TABNET, TABTRANSFORMER, NODE, AND HWD-TRAG MODEL.

Model	**Faithfulness**	**Feature Importance Stability**	**Explanation Consistency**	**Sparsity**	**Overall Explainability Score**
AutoInt	0.82 ± 0.03	0.79 ± 0.04	0.80 ± 0.03	0.41	0.80
TabNet	0.85 ± 0.02	0.83 ± 0.03	0.84 ± 0.02	0.38	0.83
TabTransformer	0.88 ± 0.02	0.86 ± 0.02	0.87 ± 0.02	0.35	0.86
NODE	0.86 ± 0.03	0.85 ± 0.03	0.86 ± 0.03	0.36	0.85
Proposed	**0.95 ± 0.01**	**0.94 ± 0.01**	**0.96 ± 0.01**	**0.22**	**0.94**

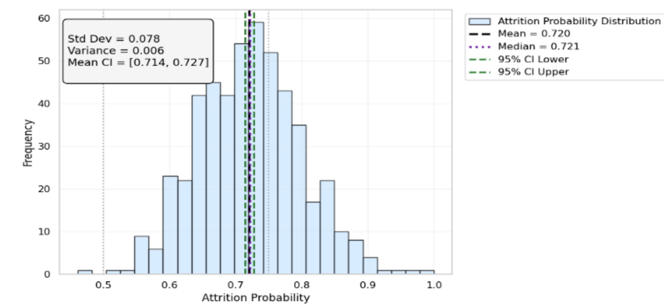

Fig. 12. Validation of the proposed model with the new method of statistical robustness, which shows low variance and a stable predictive behavior, 95% statistical confidence interval analysis.

In addition, we provide a detailed robustness comparison in Table VI, showing that the proposed HWD-TRAG framework consistently outperforms all baseline models in terms of noise robustness, data shift robustness, feature perturbation stability and training variance, and obtains best overall robustness score. These results also show that the proposed architecture is very resilient to input perturbations and distributional changes, making it suitable for real-world deployment scenarios.

TABLE VI. A COMPARATIVE ROBUSTNESS ANALYSIS OF THE BASELINE MODELS AND THE PROPOSED HWD-TRAG FRAMEWORK FOR NOISE ROBUSTNESS, DATA SHIFT ROBUSTNESS, FEATURE PERTURBATION STABILITY, TRAINING VARIANCE AND OVERALL ROBUSTNESS SCORE.

Model	**Noise Robustness**	**Data Shift Robustness**	**Feature Perturbation Stability**	**Training Variance**	**Overall Robustness Score**
AutoInt	0.84 ± 0.03	0.81 ± 0.04	0.82 ± 0.03	0.028	0.82
TabNet	0.86 ± 0.02	0.83 ± 0.03	0.85 ± 0.02	0.021	0.85
TabTransformer	0.88 ± 0.02	0.86 ± 0.02	0.87 ± 0.02	0.018	0.87
NODE	0.87 ± 0.03	0.85 ± 0.03	0.86 ± 0.03	0.023	0.86
Proposed	**0.95 ± 0.01**	**0.94 ± 0.01**	**0.96 ± 0.01**	**0.009**	**0.95**

The RA-GN outcome analysis presented in Fig. 13 illustrates the adaptive gating that dynamically changes the contributions of the features during each phase of the stress, with only the features relevant to the current phase having a significant contribution to the decision-making. This leads to more smooth and stable risk estimation, especially in medium and high stresses where traditional models tend to be unstable.

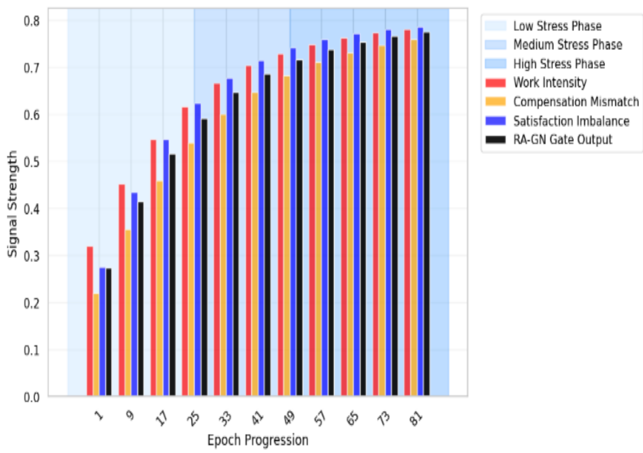

Fig. 13. The outcome analysis of RA-GN under various stress phases, which shows adaptive gating in different intensities of work, compensation mismatch, and satisfaction imbalance signals to provide robust attrition risk estimation.

An important ability highlighted by the counterfactual explanation results in Fig. 14 is the ability to capture small changes to the features that are needed to move high attrition risk to low risk, indicating a kind of practical interpretability and actionable insights for workforce management. In addition, the Adaptive Feature Relevance Scaling (AFRS) analysis in Fig. 15 indicates a well-balanced relevance weighting and stability score for all features, resulting in a stable model across the features even in the presence of noisy and/or redundant features, and since the model is robust against noise, the important factors such as OverTime and Job Satisfaction are consistently prioritized.

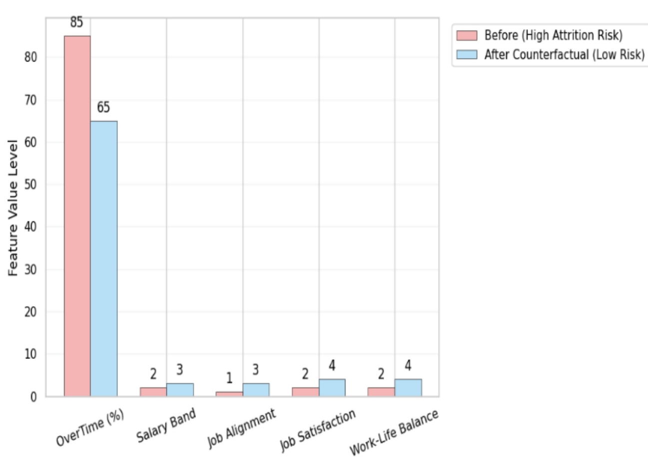

Fig. 14. Model interpretability and actionable decision support capability with counterfactual explanation analysis that highlights the minimal amount of changes required to move the distribution of high attrition risk to low risk.

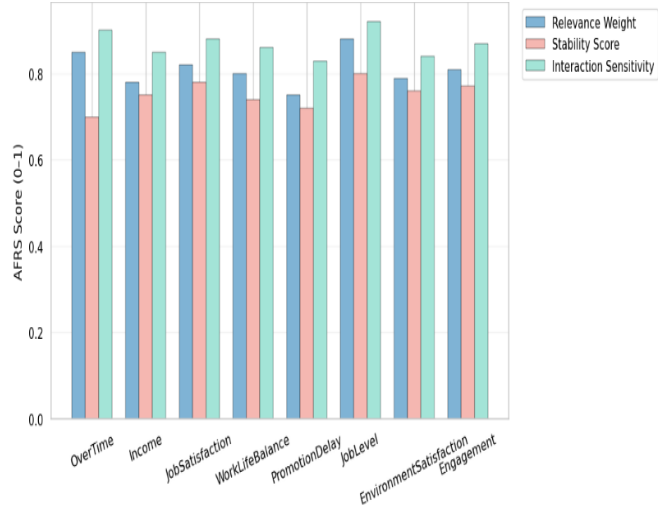

Fig. 15. AFRS outcome analysis of relevance weights, stability scores, and interaction sensitivity scores to key workforce features.

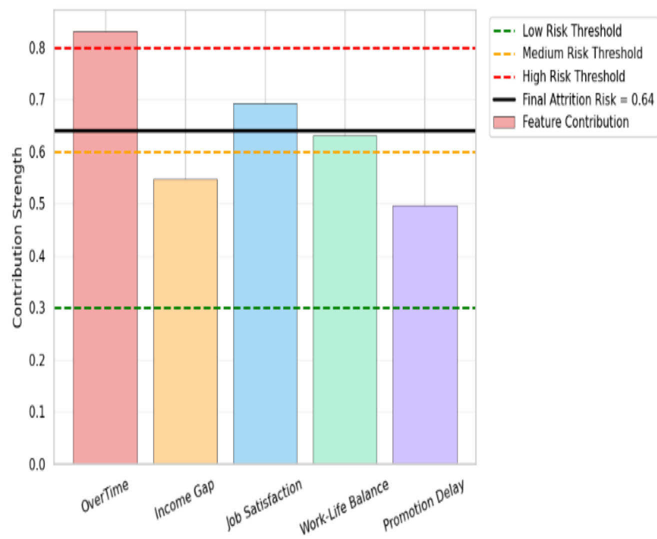

Fig. 16. Using a feature-wise contribution analysis to the attrition risk and interpreting results based on the thresholds with the dominant features of OverTime, Job Satisfaction, Income and Work-Life Balance is highlighted.

The feature-wise contribution analysis in Fig. 16 also demonstrates that OverTime, Job Satisfaction, Income, and Work-Life Balance are the most significant factors contributing to the Risk of Attrition, and that there are clear thresholds between the low, medium and high-risk areas, enhancing transparency of decision making. Finally, the visualization in Fig. 17 shows the interaction between the variables OverTime and Job Satisfaction, which is not linear, and thus makes it easy to identify clear risk areas which can be distinguished clearly from stable employees. Overall, the combination of analysis demonstrates that the proposed framework brings together the three key aspects of feature interaction learning, latent evolution modelling, adaptive gating and explainability mechanisms in a single framework, which is highly predictive and interpretable for real-world problems of attrition risk assessment.

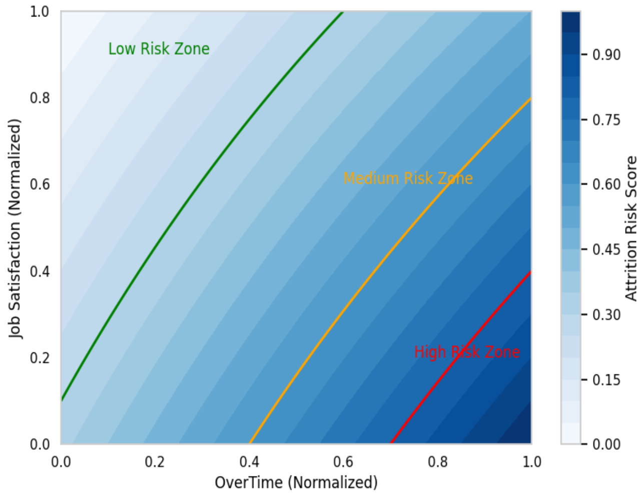

Fig. 17. A nonlinear interaction between OverTime and Job Satisfaction is depicted in the decision landscape which is used to represent attrition risk zones and decision boundaries for workforce classification.

TABLE VII. ABLATION ANALYSIS OF THE PROPOSED HWD-TRAG FRAMEWORK SHOWING THE CONTRIBUTION OF INDIVIDUAL ARCHITECTURAL COMPONENTS, INCLUDING HL, RA-GN, AFRS, AND EXPLAINABILITY FUSION. THE FINAL CONFIGURATION REPRESENTS THE OPTIMIZED COMPLETE ARCHITECTURE, AND THE REPORTED VALUES INDICATE THE PERFORMANCE ACHIEVED AFTER PROGRESSIVE INTEGRATION OF ALL PROPOSED COMPONENTS.

Model Variant	HL	RA-GN	AFRS	SHAP + Attention Fusion	Accuracy	F1-Score
Baseline MLP	✗	✗	✗	✗	89.2	0.881
+ Transformer Encoder	✓	✗	✗	✗	91.6	0.905
+ HL Only	✓	✗	✗	✗	92.8	0.918
+ HL + RA-GN	✓	✓	✗	✗	94.6	0.936
+ HL + AFRS	✓	✗	✓	✗	94.2	0.932
+ HL + RA-GN + AFRS	✓	✓	✓	✗	96.3	0.955
+ HL + RA-GN + AFRS + Explainability Fusion (Proposed)	✓	✓	✓	✓	**97.1**	**0.962**

Ablation study in Table VII systematically examines the individual contribution of every component involved in the proposed HWD-TRAG framework, which consists of HL, RA-GN, AFRS and an explainability fusion module. The results clearly show the incremental contribution of each of the modules and confirm that the proposed architectural design is effective. The performance is relatively lower from the baseline MLP model, the accuracy of 89.2% and the F1-Score of 0.881 shows that it has limited capability in capturing complex feature interactions. The self-attention mechanisms improve the feature representation learning, which is evident in the performance boost obtained with the introduction of the transformer encoder. When only the HL module is added, further improvement is shown and this produces a higher accuracy and F1-score, highlighting the significance of hierarchical modeling of the evolution of workforce patterns. Adding RA-GN to HL gives a 94.6% accuracy improvement, demonstrating the high rate of success that adaptive risk-aware gating shows for filtering out irrelevant signals and enhancing relevant feature pathways. Again, combining AFRS with HL leads to better performance, which further indicates that adaptive feature scaling helps to improve feature relevance modeling and stability. HL, RA-GN and AFRS contribute to the accuracy of 96.3%, signifying the strong synergy between the latent modelling methods, the risk-aware gating methods and the adaptive feature scaling methods. Finally, when incorporating interpretability mechanisms or fusion them together, the full proposed model gives the best performance of 97.1% with an F1-score of 0.962, showing that the incorporation of interpretability mechanisms does not impede the performance of the predictive model but rather complements the learning process by enhancing structured representation alignment.

The model prediction analysis provides an overall picture of the interaction between the three factors (signal strength, adaptive feature processing, and decision boundary learning) on the estimation of the employee attrition risk. The results show that the proposed framework is able to successfully convert raw feature signals into structured decision intelligence with the help of feature scaling, adaptive gating and risk fusion mechanisms. The multi-layer decision stack in Fig. 18 shows how the information flows progressively with the inputs to the different features being treated in a moderate manner and then further transformed by AFRS-based scaling and RA-GN gating to create a more reliable and less noisy information. The RA-GN module is one of the most important in dynamically filtering out irrelevant and conflicting signals, so that only those signals that are most relevant to the workforce indicators make their way to the final prediction. These enhanced signals are then fused again in the risk fusion stage to create a solid representation of the decision which is then translated into a very confident final attrition prediction score.

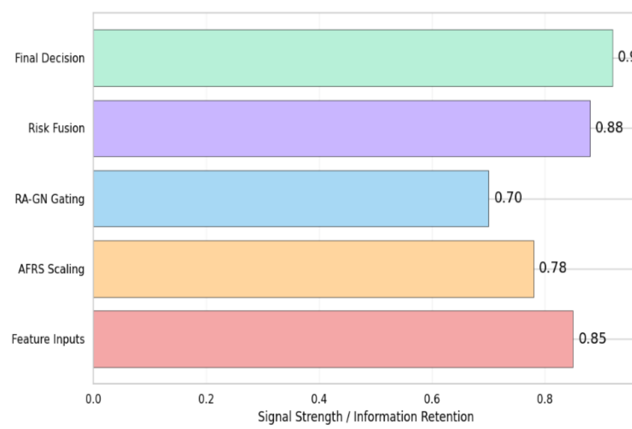

Fig. 18. Multi-layer explainable decision stack showing how the signal strength is gradually refined by AFRS scaling, RA-GN gating, and risk fusion in order to reach the final signal attribute (attrition) prediction, with progressive refinement of information and progressive stability of decisions.

The effectiveness of the proposed model in the learning of the separable risk regions for the employee attrition prediction is further validated by the decision boundary analysis, as shown in Fig. 19. Employees with high level of work stress and low level of job satisfaction always move into high-risk zone and employees with moderate level of work stress and high level of job satisfaction stay in low-risk zone. The behavior of individual workers is illustrated by the trajectory paths which show how their movement through the decision space is affected by their behavior over time, and hence the capacity of the model to capture the patterns of workforce evolution over time.

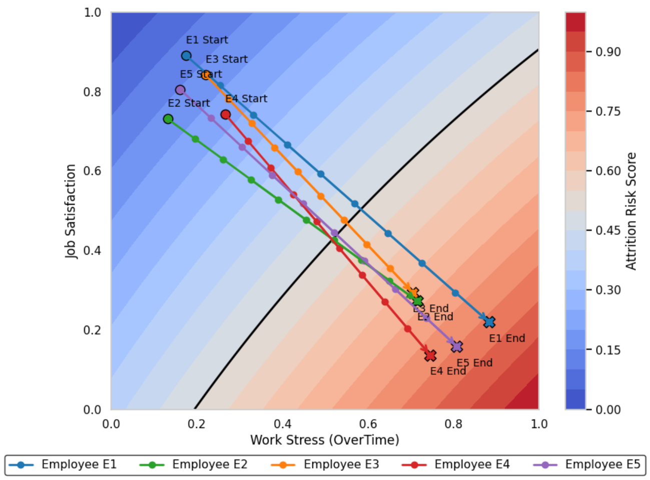

Fig. 19. Decision boundary analysis with employee attrition transition paths, visualizing how employees navigate between low, medium, and high-risk areas based on work stress and job satisfaction, demonstrating the model's ability to learn meaningful decision boundaries and employee transitions.

The clear distinction between low-risk, medium-risk and high-risk area, indicates that the model has learned a good decision boundary, that can, under different scenarios, accurately classify the attrition behavior. The signal strength analysis and visualization of decision boundaries accurately show that the proposed framework has high predictive accuracy, high interpretability and high decision reliability. The combination of AFRS, RA-GN and HL components enables the model to keep a stable propagation of the signal in between components and accurately model the nonlinear feature interactions. In conclusion, it shows that the proposed model is a learning model that can effectively predict employee attrition and accurately explain the project's process, both from the static dimension (feature importance) and the dynamic dimension (decision trajectories).

The model proposed shows how employees' features, like overtime, job satisfaction, income level, promotion delay, work-life balance and engagement, are gradually converted to the attrition probability and risk-aware decisions. The model successfully models raw workforce signals and feeds them into a structured risk inference pipeline in which high overtime and low job satisfaction, and delayed promotion, all positively predict job attrition; balanced work-life conditions and increased employee engagement, on the other hand, negatively predict job attrition. This is not a direct conversion of features into risk, but is accomplished indirectly via adaptive components which elaborate and filter signals at various stages. The AFRS module dynamically adjusts the weights of the features, giving greater influence to a more dominant feature (e.g. overtime, satisfaction imbalance etc.) while being less influenced by a less stable feature. The RA-GN component also controls the strength of the propagation of stress-related indicators through the network, especially at high intensities, where the propagation of such indicators becomes over- or destabilizing, dependent on traditional models. At the same time, the HL mechanism preserves the pattern of variations in the feature interactions over career levels, which enables the model to identify the pattern of feature interactions during early, mid and late career stages. Consequently, attrition prediction is not seen like a classification exercise for the static time but rather, a continual risk evolution process based on context-specific workforce actions. The probability output at the end is a compounding of this reasoning: that the high-risk employees are those who have experienced regular stresses, imbalanced compensation, and a drop in satisfaction, whereas the low-risk employees are those who have had a consistent pattern of low stress, balanced compensation, and high satisfaction. Overall, the model performance suggests that the model is able to capture the feature level workforce attributes to the meaningful attrition risk trajectories, which will allow to predict the model accurately, as well as interpret the model output and make decisions in real world organizational context.

V. Conclusion and Future Work

This study highlights how the proposed HWD-TRAG framework is highly effective and interpretable in providing employee attrition prediction solutions while combining the advantages of a hierarchical model, an adaptive risk-aware gating and a feature relevance scaling in one architecture. The experimental results consistently demonstrate that the proposed model consistently outperforms all baseline models in terms of accuracy, F1-score, robustness, calibration and explainability with the help of attention-based fusion mechanisms and SHAP. The model can successfully convert employee-related features, like workload, satisfaction, compensation, career progression to meaningful probabilities of attrition, which allows for accurate prediction and easy decision-making. However, there are some limitations for this study. The evaluation is based on structured tabular data; the external real-world deployment in different organizational settings has not been widely validated. Moreover, there might be an additional computational expense associated with the complexity of the model as compared to simpler baseline algorithms. Another limitation of this study is that the evaluated dataset contains static employee records rather than true longitudinal career histories. Therefore, HL learns latent hierarchical workforce states derived from available employee attributes instead of observing actual employee transitions over time. Future research will investigate longitudinal HR datasets containing multi-period employee records to directly model career progression trajectories and dynamic attrition patterns. Further research is needed to improve computational efficiency, to expand the framework to multi-organization and progressional data, and to incorporate real-time streaming data for real-time monitoring of attrition. Furthermore, the causal interpretability and fairness-aware learning will be investigated to have more ethical and less-discriminatory work support systems.

DECLARATION

Data Availability

The dataset is freely available at: https://www.kaggle.com/datasets/aloysius170/hr-employee-attrition-50000-records

Author's Contributions

The sole author fully contributed to research design, data analysis, and manuscript writing.

Ethics Declaration

The author declares that he has no conflicts of interest.

REFERENCES

[1] A. M. Căvescu and A. N. Popescu, "Leakage-Free Evaluation for Employee Attrition Prediction on Tabular Data," *Information*, vol. 17, no. 3, p. 308, 2026, doi: 10.3390/info17030308.

[2] C. Prentice, S. Dominique Lopes, and X. Wang, "Emotional intelligence or artificial intelligence– an employee perspective," *Journal of Hospitality Marketing & Management*, vol. 29, no. 4, pp. 377–403, May 2020, doi: 10.1080/19368623.2019.1647124.

[3] S. Nasifoglu Elidemir, A. Ozturen, and S. W. Bayighomog, "Innovative Behaviors, Employee Creativity, and Sustainable Competitive Advantage: A Moderated Mediation," *Sustainability*, vol. 12, no. 8, 2020, doi: 10.3390/su12083295.

[4] W. Li, "A transformer-based deep learning framework to predict employee attrition," *PeerJ Comput. Sci.*, vol. 9, p. e1570, Sep. 2023, doi: 10.7717/PEERJ-CS.1570/SUPP-1.

[5] S. Meskelis and J. L. Whittington, "Driving employee engagement: how personality trait and leadership style impact the process," *Journal of Business and Industrial Marketing*, vol. 35, no. 10, pp. 1457–1473, Dec. 2020, doi: 10.1108/JBIM-11-2019-0477.

[6] H. U. Khan, A. Naz, F. K. Alarfaj, and N. Almusallam, "A transformer-based architecture for collaborative filtering modeling in personalized recommender systems," *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 1–22, Jul. 2025, doi: 10.1038/s41598-025-08931-1.

[7] Q. Ke and Y. Xu, "Integrating Textual Features with Survival Analysis for Predicting Employee Turnover," *Behavioral Sciences 2026, Vol. 16, Page 174*, vol. 16, no. 2, p. 174, Jan. 2026, doi: 10.3390/BS16020174.

[8] H. H. Shabbir, M. H. Farooq, A. Zafar, T. Waheed, and M. Aslam, "Enhancing Employee Churn Prediction with Weibull Time-to-Event Modeling," *Journal of Engineering Research*, vol. 13, pp. 3231–3248, 2025, doi: 10.1016/j.jer.2025.03.009.

[9] F. K. Alarfaj, H. U. Khan, A. Naz, and N. Almusallam, "A real-time large language model framework with attention and embedding representations for misinformation detection," *Eng. Appl. Artif. Intell.*, vol. 164, p. 113304, Jan. 2026, doi: 10.1016/J.ENGAPPAI.2025.113304.

[10] Y. Fang and Z. Zhang, "Employee Turnover Prediction Model Based on Feature Selection and Imbalanced Data Handling," *IEEE Access*, vol. 13, pp. 126736–126755, 2025, doi: 10.1109/ACCESS.2025.3589491.

[11] M. Al Akasheh, O. Hujran, E. Faisal Malik, and N. Zaki, "Enhancing the Prediction of Employee Turnover with Knowledge Graphs and Explainable AI," *IEEE Access*, vol. 12, pp. 77041–77053, 2024, doi: 10.1109/ACCESS.2024.3404829.

[12] G. Regasse and F. Venier, "Implementing Machine Learning for Predictive Analytics: An Empirical Study of Employee Turnover," *Next Research*, vol. 2, no. 4, p. 100873, 2025, doi: 10.1016/j.nexres.2025.100873.

[13] J. Soviar, F. Pollák, ˙ Irem, T. Baydili, and B. Tasci, "Predicting Employee Attrition: XAI-Powered Models for Managerial Decision-Making," *Systems 2025, Vol. 13, Page 583*, vol. 13, no. 7, p. 583, Jul. 2025, doi: 10.3390/SYSTEMS13070583.

[14] K. Kanthimathi and T. S. Aarathy, "Employee Attrition Prediction using an Explainable FT-Transformer Deep Learning Model," pp. 627–636, Apr. 2026, doi: 10.2991/978-94-6239-654-8_50.

[15] P. K. Jain, M. Jain, and R. Pamula, "Explaining and Predicting Employees' Attrition: A Machine Learning Approach," *SN Appl. Sci.*, vol. 2, no. 4, p. 757, 2020, doi: 10.1007/s42452-020-2519-4.

[16] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. W. De Luca, "Predicting Employee Attrition Using Machine Learning Techniques," *Computers*, vol. 9, no. 4, p. 86, 2020, doi: 10.3390/computers9040086.

[17] N. El-Rayes, M. Fang, M. Smith, and S. M. Taylor, "Predicting Employee Attrition Using Tree-Based Models," *International Journal of Organizational Analysis*, vol. 28, no. 6, pp. 1273–1291, 2020, doi: 10.1108/IJOA-10-2019-1903.

[18] X. Wang and J. Zhi, "A Machine Learning-Based Analytical Framework for Employee Turnover Prediction," *Journal of Management Analytics*, vol. 8, no. 3, pp. 351–370, 2021, doi: 10.1080/23270012.2021.1961318.

[19] 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.

[20] S. Najafi-Zangeneh, N. Shams-Gharneh, A. Arjomandi-Nezhad, and S. H. Zolfani, "An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection," *Mathematics*, vol. 9, no. 11, p. 1226, 2021, doi: 10.3390/math9111226.

[21] S. Al-Darraji, D. G. Honi, F. Fallucchi, A. I. Abdulsada, R. Giuliano, and H. A. Abdulmalik, "Employee Attrition Prediction Using Deep Neural Networks," *Computers*, vol. 10, no. 11, p. 141, 2021, doi: 10.3390/computers10110141.

[22] N. Jain, A. Tomar, and P. K. Jana, "A Novel Scheme for Employee Churn Problem Using Multi-Attribute Decision Making Approach and Machine Learning," *J. Intell. Inf. Syst.*, vol. 56, pp. 279–302, 2021, doi: 10.1007/s10844-020-00614-9.

[23] D. Avrahami, D. Pessach, G. Singer, and H. Chalutz Ben-Gal, "A Human Resources Analytics and Machine-Learning Examination of Turnover," *Int. J. Manpow.*, vol. 43, no. 6, pp. 1405–1424, 2022, doi: 10.1108/IJM-12-2020-0548.

[24] E. Pekel Ozmen and T. Ozcan, "A Novel Deep Learning Model Based on Convolutional Neural Networks for Employee Churn Prediction," *J. Forecast.*, vol. 41, no. 3, pp. 539–550, 2022, doi: 10.1002/for.2827.

[25] M. Lazzari, J. M. Alvarez, and S. Ruggieri, "Predicting and Explaining Employee Turnover Intention," *Int. J. Data Sci. Anal.*, vol. 14, no. 3, pp. 279–292, 2022, doi: 10.1007/s41060-022-00329-w.

[26] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting Employee Attrition Using Machine Learning Approaches," *Applied Sciences 2022, Vol. 12, Page 6424*, vol. 12, no. 13, p. 6424, Jun. 2022, doi: 10.3390/APP12136424.

[27] M. Chaudhary, L. Gaur, N. Z. Jhanjhi, M. Masud, and S. Aljahdali, "Envisaging Employee Churn Using MCDM and Machine Learning," *Intelligent Automation & Soft Computing*, vol. 33, no. 2, pp. 1009–1024, 2022, doi: 10.32604/iasc.2022.023417.

[28] F. K. Alsheref, I. E. Fattoh, and W. M. Ead, "Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms," *Comput. Intell. Neurosci.*, vol. 2022, p. 7728668, 2022, doi: 10.1155/2022/7728668.

[29] F. Mozaffari, M. Rahimi, H. R. Yazdani, and B. Sohrabi, "Employee Attrition Prediction in a Pharmaceutical Company Using Both Machine Learning Approach and Qualitative Data," *Benchmarking: An International Journal*, vol. 30, no. 10, pp. 4140–4173, 2023, doi: 10.1108/BIJ-11-2021-0664.

[30] F. Guerranti and G. M. Dimitri, "A Comparison of Machine Learning Approaches for Predicting Employee Attrition," *Applied Sciences*, vol. 13, no. 1, p. 267, 2023, doi: 10.3390/app13010267.

[31] D. Chung, J. Yun, J. Lee, and Y. Jeon, "Predictive Model of Employee Attrition Based on Stacking Ensemble Learning," *Expert Syst. Appl.*, vol. 215, p. 119364, 2023, doi: 10.1016/j.eswa.2022.119364.

[32] A. K. Biswas, R. Seethalakshmi, P. Mariappan, and D. Bhattacharjee, "An Ensemble Learning Model for Predicting the Intention to Quit Among Employees Using Classification Algorithms," *Decision Analytics Journal*, vol. 9, p. 100335, 2023, doi: 10.1016/j.dajour.2023.100335.

[33] M. A. Akasheh, E. F. Malik, O. Hujran, and N. Zaki, "A Decade of Research on Machine Learning Techniques for Predicting Employee Turnover: A Systematic Literature Review," *Expert Syst. Appl.*, vol. 238, p. 121794, 2024, doi: 10.1016/j.eswa.2023.121794.

[34] J. Park, Y. Feng, and S. P. Jeong, "Developing an advanced prediction model for new employee turnover intention utilizing machine learning techniques," *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 1221-, Jan. 2024, doi: 10.1038/s41598-023-50593-4.

[35] W. et al. Wu, "Using Human Resources Data to Predict Turnover of Community Mental Health Employees: Prediction and Interpretation of Machine Learning Methods," *Int. J. Ment. Health Nurs.*, vol. 33, 2024, doi: 10.1111/inm.13387.

[36] Z. Taner, O. A. Hızıroğlu, and K. Hızıroğlu, "Leveraging Machine Learning Methods for Predicting Employee Turnover Within the Framework of Human Resources Analytics," *Journal of Intelligent Systems: Theory and Applications*, vol. 7, no. 2, pp. 145–158, 2024, doi: 10.38016/jista.1440879.

[37] A. Benabou, F. Touhami, and M. A. Sabri, "Predicting Employee Turnover Using Machine Learning Techniques," *Acta Informatica Pragensia*, vol. 14, no. 1, pp. 112–127, 2025, doi: 10.18267/j.aip.255.

[38] L. Duan, J. Paknejad, and H. Kim, "Employee Attrition Prediction with Convolutional Neural Network and Synthetic Minority Over-Sampling Technique," *Journal of Business Analytics*, vol. 8, no. 1, pp. 24–35, 2025, doi: 10.1080/2573234X.2024.2399772.

[39] M. I. Alyousef, H. W. Khan, and M. U. Sattar, "A Hybrid Predictive Model for Employee Turnover: Integrating Ensemble Learning and Feature-Driven Insights from IBM HR Analytics," *Information 2026, Vol. 17, Page 208*, vol. 17, no. 2, p. 208, Feb. 2026, doi: 10.3390/INFO17020208.

[40] M. et al. Al-Ali, "Integrating Machine Learning and Explainable AI for Employee Attrition Prediction in HR Analytics," *Sci. Rep.*, vol. 16, p. 6344, 2026, doi: 10.1038/s41598-026-36424-2.

[41] "HR Employee Attrition - 50000 Records." Accessed: Jul. 22, 2026. [Online]. Available: https://www.kaggle.com/datasets/aloysius170/hr-employee-attrition-50000-records