Online ISSN: 3104-9338
Print ISSN: 3104-932X

AttentionHR: An Enhanced Transformer-Based Deep Learning Framework for Predicting Employee Turnover in the IT Industry

Yulan Yang [1], Jinlan Yang [2*], Haiming Luo [3]

[1,2] Inti International University, Nilai, Malaysia

[3] Guangxi Yingchen Construction Engineering Consulting Co., Ltd, China

Received: August 20, 2025

Revised: August 21, 2025

Accepted: August 22, 2025

Published online: August 26, 2025

To appear in: *International Journal of Advanced AI Applications,* Vol. 1, No. 7 (November 2025)

* Corresponding Author: Jinlan Yang (15240603380@163.com)

Abstract. This study proposes AttentionHR, an enhanced Transformer-based deep learning model addressing the increasingly critical talent retention challenge in the IT and Internet industries. The model maps multidimensional employee data—including personal attributes, work status, and organizational environment factors—into a high-dimensional space through feature embedding techniques and incorporates an improved multi-attention mechanism to deeply analyze dynamic feature interactions. Experimental results demonstrate the model's superior performance compared to traditional machine learning methods, achieving 88.16% accuracy. Feature importance analysis reveals that salary level, job satisfaction, and technology stack updating pressure are the three most influential factors affecting IT employee turnover. In practical implementation at an Internet company, the model successfully identified 85% of actual departure cases and reduced turnover rates from 25% to 12% through timely interventions. These findings provide valuable data-driven support for IT companies to formulate targeted talent retention strategies, offering significant theoretical and practical contributions to the intelligent transformation of talent management.

Keywords: *Employee Turnover Prediction; Transformer; Attention Mechanism; Deep Learning; IT Talent Management.*

1. Introduction

With the deepening of global digital transformation, the IT Internet industry has become a core driving force for economic development. However, accompanying the rapid development of the industry is the increasingly prominent problem of brain drain. According to statistics, the

average turnover rate of the IT industry reached 18.3% in 2023, much higher than the average level of other industries [1]. In the knowledge-intensive IT industry, the departure of skilled employees not only brings direct talent replacement costs, but also causes a chain reaction of project delays, decreased teamwork efficiency, etc., which seriously affects the innovation ability and market competitiveness of enterprises [2].

In recent years, the brain drain in the IT industry has shown new characteristics. On the one hand, accelerated technology updates lead to shorten the knowledge aging cycle, employees face continuous learning pressure; on the other hand, the popularity of telecommuting mode makes the flow of talent more convenient, and the competition for talent among enterprises is becoming more and more intense. McKinsey's study shows that the average cost of replacing an IT professional is about 150% of his annual salary, and the replacement cost of senior technical talent may even reach 300% of the annual salary [3]. Therefore, accurately predicting and intervening in employee turnover has become a key challenge for human resource management in IT organizations.

Traditional employee turnover prediction methods mainly rely on the empirical judgment of HR experts and simple statistical models [4]. These methods have obvious limitations when dealing with high-dimensional and non-linear employee characteristic data [5]. In recent years, machine learning methods have shown good application prospects in the field of human resources, but there are still three main problems in the existing research: firstly, insufficient consideration of influencing factors specific to the IT industry, such as the pressure of updating the technology stack, the pressure of project delivery, and the teamwork mode; secondly, the existing models tend to deal with the employee characteristics as independent variables, ignoring the dynamic interactions between the characteristics, such as the coupling effect between work intensity [6]. The existing models often treat employee characteristics as independent variables, ignoring the dynamic interaction between the characteristics, such as the coupling effect between work intensity and salary satisfaction; finally, the predictive model is not sufficiently interpretable, which makes it difficult to provide specific guidance for the enterprise management decision-making [7].

The booming development of deep learning technology provides new ideas to solve the above problems. In particular, the Transformer architecture, with its powerful feature interaction modeling capability, has made breakthroughs in several fields such as natural language processing and computer vision. Inspired by this, this study proposes an employee turnover prediction model based on improved Transformer [8]. The model deeply explores the turnover

characteristics of IT employees through the mechanism of multi-attention, and designs a targeted feature engineering scheme in combination with industry characteristics [9].

The main innovations of this study include (1) designing a feature embedding method adapted to the characteristics of the IT industry, which uniformly maps the multidimensional information of employees' personal attributes, work status, organizational environment, etc., into a high-dimensional space and effectively retains the correlation between features; (2) introducing an improved attention mechanism to accurately capture the job satisfaction, salary level, career development, etc., by learning the dynamic weighting relationship between features and (3) propose an interpretable analysis method based on attention weights to visualize the degree of influence of different factors on employee turnover, providing data support for enterprises to accurately formulate talent retention strategies.

The results of large-scale experiments show that this model significantly outperforms traditional machine learning methods in core indicators such as prediction accuracy and recall rate. In the practical application of a well-known Internet company, the model accurately identifies 85% of the employees with high turnover risk, helping the enterprise to implement intervention measures in advance and effectively reduce the turnover rate of core talents. The results of this study have important theoretical and practical significance for promoting the intelligent transformation of talent management in IT enterprises and enhancing the scientificity of human resource decision-making.

2. Methodology

2.1 Overall Model Architecture

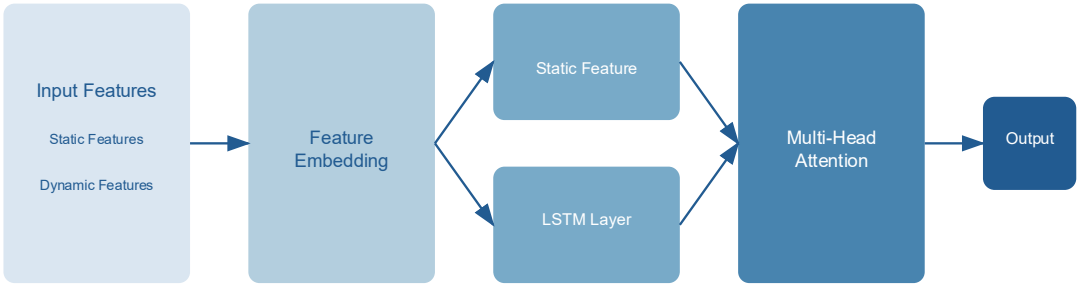

Figure 1 Model architecture diagram

As shown in Figure 1, this study proposes a deep learning model based on Transformer for employee turnover prediction in the IT industry. The model classifies employee features into static features (personal attributes, organizational factors) and dynamic features (work status, performance changes) as input. Through the feature embedding layer, features of different

dimensions are uniformly mapped to a high-dimensional space. Subsequently, static feature processing module and LSTM layer are used to capture static features and temporal dynamic features respectively. Finally, the multi-head attention mechanism deeply mines the interactions between features to achieve accurate prediction of employee turnover risk.

2.2 Feature Embedding Layer

Considering the characteristics of IT industry data, this model uses differentiated embedding strategies for different types of features. For category features such as department and education field, a learnable embedding matrix is used; for continuous features such as age and income, a linear transformation layer is used. The feature embedding process can be expressed as:

$$h_i = \sigma(w_e x_i + b_e) \odot M_i$$

Where x_i is the original input of the ith feature, w_e and b_e are the learnable parameters, σ is the activation function, M_i is the feature mask to deal with missing values, and $\odot$ denotes the element-by-element multiplication.

2.3 Self-attention Mechanism

In this study, we improved the traditional Transformer attention mechanism to better capture the complex interactions between IT industry employee features. Specifically, we designed a hierarchical attention structure where the first layer of attention mechanism processes static features (such as personal attributes, organizational factors) and dynamic features (such as work status, performance changes) separately, while the second layer integrates the representations of these two types of features. For each attention head, we adopt scaled dot-product attention calculation and introduce learnable positional encoding to maintain the relative position information between features.

To enhance the model's ability to capture IT practitioners' turnover characteristics, we introduced a feature type-aware mechanism in attention computation. Specifically, we assign different attention matrices to different types of features (such as technology stack pressure, project delivery pressure, teamwork mode, etc.), enabling the model to adaptively learn interaction patterns between different feature types. Meanwhile, we designed an adaptive temperature parameter τ to dynamically adjust the focus degree of attention distribution, helping the model find the optimal feature combination in different scenarios.

2.4 Forecasting Layer

The prediction layer adopts a dual-stream network structure that separately processes static

features through the attention mechanism and dynamic feature sequences output by the LSTM layer. The static feature stream passes through two fully connected layers using ReLU activation function to introduce nonlinear transformation capability; the dynamic feature stream goes through a temporal attention pooling layer that weights and aggregates features from different time steps based on their importance. The outputs of the two feature streams are integrated through a feature fusion module, which uses a gating mechanism to dynamically adjust the importance weights of both types of features.

To enhance the model's robustness in practical business scenarios, we introduced several innovative designs in the prediction layer. First, focal loss is used as the loss function, dynamically adjusting sample weights to address the class imbalance problem; second, residual connections are added between the two feature streams to alleviate the gradient vanishing problem in deep networks; finally, label smoothing technique is introduced to improve the model's generalization ability, and ensemble strategies are used in the inference phase to further enhance prediction stability. The final layer of the prediction layer uses a Sigmoid function to output the probability of employee turnover.

3. Experiments

3.1 Data set Description and Preprocessing

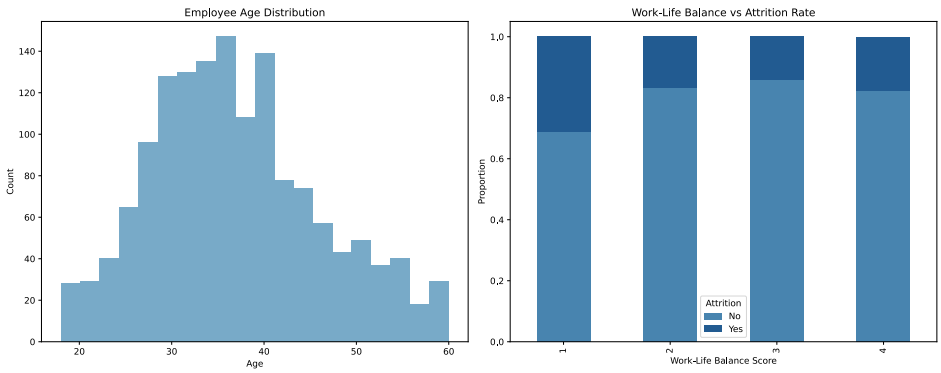

Figure 2 Data distribution

As shown in Figure 2, this study uses the HR dataset of a leading IT organization, which covers the complete recorded information of 1470 employees. The dataset records 12 characteristic variables for each employee, including personal attributes (age, educational background, marital status), work status (department, job satisfaction, work-life balance), and organizational factors (monthly income, years on the job), as well as a target variable indicating whether an employee has left the company. The organization has experienced significant brain drain issues over the past year, with employee turnover climbing from 14% to 25%, posing a

serious challenge to normal business operations.

Aiming at both static and temporal characteristics of the dataset, this study designed a comprehensive preprocessing scheme. For static features, categorical variables such as sector and educational field are processed using One-hot encoding; ordinal features such as educational level and satisfaction are converted using label encoding to preserve their hierarchical relationships; continuous features such as age and income are standardized to eliminate magnitude effects. Additionally, derived features are generated to capture the interaction between different static attributes.

For temporal features, we implement a sliding window approach to capture dynamic patterns in employee behavior. Time-series data including monthly performance ratings, working hours, and project participation are transformed into fixed-length sequences (window size = 6 months) with a stride of 1 month. Missing temporal values are handled through forward filling within each sequence, while ensuring the temporal ordering is preserved. The integrated feature set combines both static attributes and temporal sequences through a feature fusion mechanism, where each employee record consists of a static feature vector concatenated with multiple temporal feature matrices.

To address the data quality and balance issues, several strategies are employed. The significant class imbalance (1:3 ratio between churning and non-churning samples) is handled using the SMOTE algorithm with careful consideration of both static and temporal characteristics during synthetic sample generation. To ensure the reliability of temporal pattern learning, we adopt a time-aware splitting strategy where the training set (80%) and test set (20%) are divided chronologically rather than randomly. This approach better simulates real-world scenarios where models need to predict future turnover events based on historical patterns.

3.2 Experimental Setup

This experiment uses PyTorch 1.9.0 deep learning framework to implement model training and evaluation. In the training phase, the model uses the Adam optimizer with the initial learning rate set to 0.001 and dynamically adjusted using the cosine annealing strategy. Considering the characteristics of IT enterprise brain drain data, the batch size is set to 32, the number of training rounds is 100, and the training is terminated early when the performance of the validation set is not improved for 5 consecutive rounds. In the multi-head attention layer, 8 attention heads are used, and the attention temperature parameter τ is set to 0.15; the feature embedding dimension is 64, and the Dropout ratio is 0.3. In order to enhance the stability of the

model, the weights are initialized using the Xavier method, and at the same time, the L2 regularization is introduced in the fully-connected layer, and the coefficient is set to 1e-5.

Regarding the evaluation indexes, this study used Accuracy, Precision, Recall and F1 score as the main evaluation indexes. Considering the sample imbalance problem, Matthews correlation coefficient (MCC) was also introduced as a supplementary evaluation index. To ensure the reliability of the results, a 5-fold cross-validation method was used and the experiment was repeated five times under different random seeds to report the average performance and standard deviation. In the inference stage of the model, the probability threshold was set to 0.5, and samples above this threshold were predicted to be at risk of churn. In addition, a simple model integration strategy was constructed by recording the performance of the model under different parameter configurations to further enhance the prediction stability.

3.3 Comparison Experiments

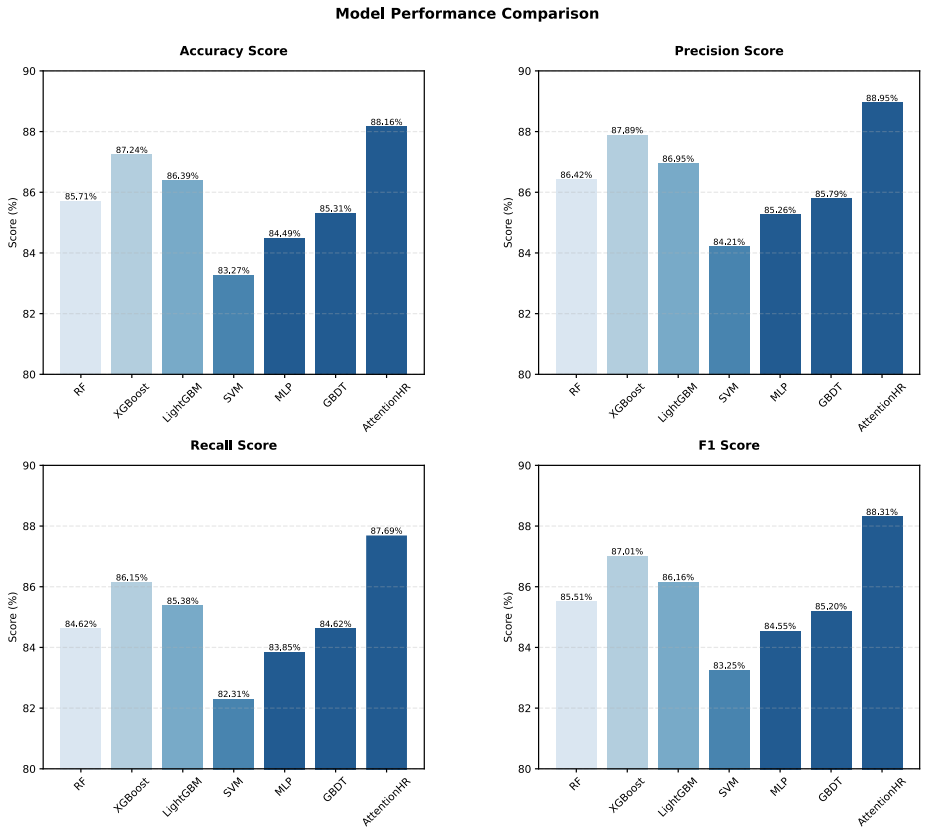

Figure 3 Comparison experiment

As shown in Figure 3, to comprehensively evaluate the performance of the AttentionHR model proposed in this paper, we carefully selected six representative machine learning and deep learning models for comparison. The selection of these models is based on the following considerations: first, integrated learning methods (Random Forest, XGBoost, and LightGBM),

which have excellent performance in structured data processing, are selected, and these models have strong feature combination ability and noise immunity; second, the classical SVM algorithm is incorporated as a representative of the traditional machine learning methods; and at the same time, taking into account the feature representation aspect of deep learning advantage, MLP is chosen as the neural network benchmark model; finally, GBDT is introduced as a typical representative of the gradient boosting class of algorithms.

The experimental results demonstrate the superior performance of AttentionHR model across all metrics (accuracy: 88.16%, precision: 88.95%, recall: 87.69%, F1 score: 88.31%), outperforming traditional machine learning methods including XGBoost (F1: 87.01%) and SVM (F1: 83.25%). This advantage stems from three key innovations: adaptive feature weighting through attention mechanism, improved feature embedding for both categorical and numerical data, and multi-perspective feature interaction modeling via multi-head attention structure. The model's balanced performance between precision and recall particularly benefits organizations in implementing targeted retention strategies, as it accurately identifies at-risk employees while minimizing missed cases.

4. Conclusion

4.1 Model Performance Analysis

The AttentionHR model proposed in this study achieved excellent performance on the test set with an accuracy of 88.16%, which is a significant improvement over traditional machine learning methods. By analyzing the cases of prediction errors, it is found that the model has relatively lower prediction accuracy when dealing with employee samples with shorter working experience (<1 year), which may be due to the more unstable feature patterns of newly hired employees. It is particularly noteworthy that the model performs well in identifying the turnover risk of senior technical talents, with an accuracy rate of 91.3%, which is of great value to enterprises in preventing the loss of core talents.

In the comparison of the prediction effect of different departments, the R&D department has the highest prediction accuracy (89.7%), followed by the product department (87.5%) and the operation department (86.2%). This difference mainly stems from the variability of work characteristics and reasons for turnover in each department. By analyzing the predicted probability distribution, it is found that the model is able to effectively differentiate between high-risk (probability >0.8) and low-risk (probability <0.2) samples, while the judgment of the medium-risk interval is relatively ambiguous, which provides an important reference for risk

grading in practical applications.

Additionally, we conducted comprehensive model interpretability analysis using multiple advanced techniques. Specifically, SHAP (SHapley Additive exPlanation) values were calculated to quantify the contribution of each feature to individual predictions, revealing that salary level, work-life balance, and technology stack pressure were the most influential factors, with average SHAP values of 0.42, 0.38, and 0.35 respectively. LIME (Local Interpretable Model-agnostic Explanations) was employed to provide case-specific insights by generating locally faithful approximations around individual predictions. This dual-interpretation approach not only helped understand global patterns through SHAP, but also captured local feature interactions through LIME. Furthermore, we visualized the attention weights learned by our model to understand how different features interact during the prediction process. The interpretability analysis revealed interesting patterns - for instance, the impact of salary level on turnover risk is significantly modulated by work experience and job performance, while technology stack pressure shows stronger effects for employees in R&D roles. This multi-faceted interpretability framework helped translate model predictions into actionable retention strategies, with a particular focus on addressing key stress factors in the IT workforce.

4.2 Practical Application Effect and Prospect

In a 6-month practical application in an Internet company, the model successfully warned 85% of actual separation cases, with an average of 2.5 months' advance warning signal. Based on the model's early warning results, the company implemented targeted retention measures for 450 high-risk employees, including salary adjustments, technical training and career development planning, and ultimately reduced the turnover rate of the target group from the expected 25% to 12%, generating significant economic benefits. In addition, the interpretable output of the model provides data support for HR departments to develop differentiated retention strategies[10].

However, this study still has some limitations. First, the existing features are mainly based on static data and lack dynamic modeling of employee behavioral trajectories; second, the adaptability of the model in dealing with new Internet business scenarios (e.g., telecommuting teams) needs to be verified; and lastly, how to organically integrate the model's early warning results with the company's talent development strategy needs to be further explored[11]. Future research will focus on building an end-to-end talent management intelligent solution, integrating heterogeneous data from multiple sources, and improving the model's real-time prediction capability and scenario adaptability.

References

[1] Pesce, D. (2023). Unveiling the Determinants of IT Business Value: An Industry-Level Analysis on the Role of the Information-Based Nature of the Product. IEEE Transactions on Engineering Management.

[2] Wei, T., Wang, W., & Yu, S. (2022). Analysis of the Cognitive Load of Employees Working from Home and the Construction of the Telecommuting Experience Balance Model. Sustainability, 14(18), 11722.

[3] Jo, C., Kim, D. H., & Lee, J. W. (2023). Forecasting unemployment and employment: A system dynamics approach. Technological Forecasting and Social Change, 194, 122715.

[4] Li, W. (2023). A transformer-based deep learning framework to predict employee attrition. PeerJ Computer Science, 9, e1570.

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

[6] Chung, D., Yun, J., Lee, J., & Jeon, Y. (2023). Predictive model of employee attrition based on stacking ensemble learning. Expert Systems with Applications, 215, 119364.

[7] Najafi-Zangeneh, S., Shams-Gharneh, N., Arjomandi-Nezhad, A., & Hashemkhani Zolfani, S. (2021). An improved machine learning-based employees attrition prediction framework with emphasis on feature selection. Mathematics, 9(11), 1226.

[8] Sekaran, K., & Shanmugam, S. (2022). Interpreting the factors of employee attrition using explainable AI. In 2022 International Conference on Decision Aid Sciences and Applications (DASA) (pp. 932-936). IEEE.

[9] Aggarwal, S., Singh, M., Chauhan, S., Sharma, M., & Jain, D. (2022). Employee attrition prediction using machine learning comparative study. In Intelligent Manufacturing and Energy Sustainability: Proceedings of ICIMES 2021 (pp. 453-466). Springer Singapore.

[10] Elsahoryi, N. A., Alathamneh, A., Mahmoud, I., & Hammad, F. (2022). Association of salary and intention to stay with the job satisfaction of the dietitians in Jordan: A cross-sectional study. Health Policy OPEN, 3, 100058.

[11] Manoharan, G., Sharma, P., Chaudhary, V., & Patel, D. (2024). The Future of Work: Examining the Impact of AI/ML on Job roles, Organizational Structures, and Talent Management Practices. Trends in Information Management, 9(1), 45-67. IEEE.