Intelligent Assessment Method for Enterprise Employee Performance Based on Multi-Source Behavioral Data

Lijuan Yin[1], Lei Ren[2]
[1]School of Public Administration, Anhui Vocational and Technical University, Hefei 230011, Anhui, China
[2]Yiqiyin (Hangzhou) Technology Limited Company, Hangzhou 311200, Zhejiang, China

Corresponding author: Lei Ren (e-mail: r10569@sina.com)

ABSTRACT The increasing complexity of digital enterprises and intelligent industrial systems has led to the large-scale generation of heterogeneous behavioral data collected through interconnected sensing and communication infrastructures operating in complex electromagnetic information environments. In such scenarios, traditional enterprise performance evaluation remains constrained by fragmented data sources and subjective assessment, making it difficult to capture employees' actual contributions. This study proposes an intelligent performance assessment framework that integrates multi-source behavioral data by consolidating enterprise information system logs, collaboration communication records, and project management system data. A behavioral feature system is constructed across three dimensions, including task execution, knowledge contribution, and collaborative interaction. Furthermore, a hybrid model combining temporal convolutional networks with an attention mechanism is developed to adaptively learn behavioral patterns across different job categories, while a state influence coefficient is introduced to compensate for assessment deviations caused by short-term fluctuations. Validation was performed using six-month longitudinal data from R&D and sales departments of an internet enterprise, including 168 R&D engineers, 49 test engineers, and 156 sales specialists with 3.897 million behavioral records and an average of 21,500 valid records collected daily. Experimental results demonstrate that the proposed method improves assessment accuracy by 29.8% compared with a linear regression model using identical behavioral features while achieving a Gini coefficient of 0.19. By incorporating temporal decay characteristics of behavioral information and accounting for heterogeneous work patterns across specialized positions, the proposed framework enhances the reliability of intelligent decision-making and provides a scalable methodology for multi-source information fusion in data-driven management systems operating under complex communication and sensing environments.

INDEX TERMS multi-source behavioral data, performance evaluation, temporal convolutional network, information fusion, electromagnetic information environment

I. INTRODUCTION

THE deep penetration of digital technology into software development and IT operations is reshaping the underlying logic of human resource management in modern enterprises. This digital shift necessitates a transition from traditional oversight to data-driven talent optimization that aligns with the technical demands of contemporary work environments [1]. Performance evaluation, as the core link connecting employee contributions with organizational goals, directly affects the effectiveness of talent incentives and the sustainability of organizational competitiveness. Over the past decade, assessment tools have continuously evolved from balanced scorecards to Objectives and Key Results (OKR), yet a fundamental dilemma remains unresolved: the data foundation for evaluation is still fragmented, lagging, and highly dependent on manual input. This problem becomes particularly prominent in today's context of increasingly frequent cross-departmental collaboration and increasingly blurred work boundaries [2].

Traditional performance evaluation models typically use periodic assessment as the basic unit, with department managers scoring subordinates based on limited work memory and subjective impressions. Cognitive biases such as halo effect and recency effect are almost unavoidable in this process [3]. More concerning is that contributions that cannot be quantified but genuinely exist—such as helping colleagues solve technical problems, proactively taking on additional tasks at critical project junctures, and promoting cross-team knowledge sharing—are often completely ignored by formal evaluation systems [4]. These behaviors are precisely important sources of organizational social capital accumulation but become blind spots in performance

evaluation due to their difficulty in capture.

The maturity of big data technology provides new possibilities for solving this dilemma [5]. Information system logs, instant messaging records, code submission histories, project board updates, and other data accumulated during enterprise operations are becoming new windows for observing employees' actual work states. Research shows that after implementing performance management systems based on continuous behavioral monitoring, employee productivity increases by approximately 15% and on-time delivery rates improve by about 12% [6]. These data are not only voluminous but, more importantly, they naturally record the original trajectory of work processes rather than being products of post-hoc recollection.

However, data availability does not equate to evaluation effectiveness. How to extract effective signals related to performance from massive, heterogeneous, noise-dense behavioral data; how to distinguish efficient work behaviors from ineffective busyness; and how to ensure algorithmic evaluation does not replicate or even amplify biases in human decision-making—these issues constitute new technical challenges [7].

Existing research has accumulated valuable explorations in behavioral feature analysis and performance prediction. Some scholars have approached the problem from behavioral feature analysis, proposing contextualized evaluation methods that incorporate incentive and punishment requirements, introducing management intentions into evaluation model design [8]. Other work has attempted to combine cognitive computing with deep learning to dynamically adjust evaluation standards in multi-scale organizational networks [9]. Additionally, multi-source evaluation models based on adaptive neuro-fuzzy inference systems have been proven effective in integrating 360-degree feedback with objective performance indicators [10]. These studies demonstrate that hybrid intelligent algorithms exhibit superior accuracy over traditional methods when processing multidimensional employee data, with mean squared error reduced by 43.5% compared to traditional regression models and Gini coefficients controlled at the 0.18 level [9].

In performance prediction modeling, the combination of Bayesian networks and deep learning has proven effective in handling complex random relationships between evaluation indicators [7]; graph convolution-based temporal models can capture employees' dynamic evolution characteristics in collaborative networks [11]. However, existing research has not given sufficient attention to the structural characteristics of behavioral data itself—such as long-term dependencies in time series, interactive influences between different behavioral dimensions, and modulation effects of job types on behavioral patterns.

Building upon these insights, this study constructs a more explanatory intelligent performance assessment framework tailored to the intricate operational workflows of software engineering. This framework integrates high-frequency development activity data to ensure that human capital evaluations remain objectively aligned with the technical complexities of modern software and service delivery. The core idea is to view employee performance as the cumulative expression of observable behaviors over the time dimension, rather than a simple summation of static results. Specifically, the research work includes three aspects:

First, establishing a multi-source behavioral data system covering three dimensions—task execution, knowledge contribution, and collaborative interaction—and clarifying the collection specifications and quantification methods for indicators in each dimension. Second, designing an evaluation model that integrates temporal convolutional networks with attention mechanisms, enabling the model to automatically identify contribution weights of behaviors from different time periods and types.

Third, introducing a state influence coefficient to adjust model outputs to reflect external factor interference experienced by employees during specific time cycles.

The research concludes with validation using real enterprise data and discusses the model's performance in terms of fairness and interpretability, including a critical examination of its ethical limitations.

II. CONSTRUCTION OF MULTI-SOURCE BEHAVIORAL DATA SYSTEM

A. DATA SOURCES AND COLLECTION FRAMEWORK

The accuracy of performance evaluation highly depends on the quality and coverage of input data [12]. Under traditional evaluation models, data sources are typically limited to assessment records in human resource management systems and subjective scores from managers. Such data not only has low update frequency but also struggles to reflect employees' actual performance in daily work. This study attempts to break through this limitation by expanding the data collection scope to multiple information systems naturally generated during daily enterprise operations.

The collection framework connects to three main data sources:

First, project management tools, including operation logs from platforms such as Jira and TAPD that record task decomposition, work hour reporting, progress updates, and defect repairs.

Second, enterprise collaboration software, covering behavioral records from DingTalk and Enterprise WeChat regarding communication frequency, document collaboration, meeting participation, and knowledge accumulation.

Third, code and document version control systems, such as Git and Confluence, which preserve detailed histories of code submissions, review participation, and document editing.

These three data sources collectively form a three-dimensional perspective for observing employee work behaviors.

Data collection uses API interfaces for real-time synchronization, pulling incremental data from each business system every 4 hours. This high-frequency collection is

necessary to capture the granular dynamics of task switching and communication bursts. The raw 4-hour logs are subsequently aggregated into weekly behavioral vectors using a process designed to preserve behavioral trajectories, which is detailed in Section 2.3. This frequency setting is based on two considerations: on one hand, it ensures data timeliness so that evaluation results can reflect recent performance changes of employees; on the other hand, it avoids unnecessary load pressure on business systems caused by overly frequent data requests. During collection, raw data undergoes de-identification processing, removing direct identifying information such as names and employee IDs, replacing them with system-generated anonymous IDs to reduce privacy leakage risks.

B. BEHAVIORAL INDICATOR SYSTEM DESIGN

Transforming raw logs into quantifiable behavioral indicators requires systematic abstraction and design. The key to this process lies in identifying behavioral patterns that have stable associations with performance output while filtering out noise data interference. This study draws on existing literature's classification frameworks for employee behaviors and combines field research on job characteristics at the target enterprise to construct a behavioral indicator system encompassing three dimensions.

The task execution dimension focuses on employees' efficiency in advancing specific work tasks and completion quality. Quantifiable indicators include the number of tasks completed, average processing time, on-time delivery rate, and defect density. Data for this dimension primarily comes from project management tools and can relatively directly reflect employees' output levels within their designated responsibilities.

The knowledge contribution dimension measures employees' participation in knowledge sharing and technical accumulation within the organization. To avoid the pitfall of equating high volume with high quality (a common logical flaw in such metrics) [2], we employ weighted indicators: "code review comment adoption rate" (comments marked as "helpful" or leading to a code change) instead of raw comment count, and "document version net value-add" (e.g., significant additions minus deletions of outdated content) instead of total edit volume. We also include the number of technical sharing sessions delivered. Such behaviors often do not directly count toward personal performance goals but hold significant value for the organization's long-term development, and traditional evaluation systems struggle to provide fair assessment for these contributions. The collaborative interaction dimension focuses on employees' performance in team collaboration and cross-departmental communication. Quantifiable indicators include cross-team meeting participation, response time to collaborative messages, records of helping others solve problems, and mention frequency in project retrospectives. These indicators attempt to capture interpersonal interaction behaviors that promote overall team efficiency. It should be noted that different job types should have different weight configurations across these dimensions. R&D positions may need greater focus on code quality and task completion, while product operations positions may have relatively higher requirements for collaborative interaction and knowledge contribution. The model design reserves job type as a conditional input, allowing behavior feature weights to adaptively adjust according to job requirements.

C. DATA PREPROCESSING AND FEATURE ENGINEERING

Raw behavioral logs exhibit typical characteristics of sparsity, heterogeneity, and noise density, making direct model input unlikely to yield ideal results. The data preprocessing stage needs to complete multiple transformation tasks. Crucially, this includes the aggregation of high-frequency (4-hour interval) logs into weekly vectors, which form the model's base time unit.

The aggregation process is designed to preserve behavioral trajectories, not merely produce summary statistics. For each employee per week, we calculate: (1) volumetric features (e.g., total commits, messages sent);

(2) temporal rhythm features (e.g., standard deviation of coding activity across days, proportion of activity outside core hours); and (3) trend features (e.g., change in commit volume from the previous week). This ensures that the weekly vector encodes information about the distribution and dynamics of behaviors, not just their total count.

First is the standardization and alignment of behavioral sequences. Different systems record timestamps in varying formats; some use Unix timestamps while others employ standard date-time formats. Preprocessing scripts uniformly convert all time fields to millisecond-level timestamps and reorganize them by employee ID and time sequence to form behavioral event streams for each employee.

Second is missing value handling. Missing behavioral data differs from other data types—no records during a certain period might mean the employee was on vacation, focused on offline work, or genuinely did not produce quantifiable behaviors. Simple imputation is clearly inappropriate. This study adopts an approach that preserves the original form of the time series and introduces a masking mechanism at the model input layer, enabling the model to identify which time points have valid observations.

Third is feature construction. Raw logs can be aggregated to generate higher-level features. For example, single code submission records can be aggregated into daily submission frequency and weekly code line change trends; instant messaging can be aggregated into degree centrality metrics of communication networks. These aggregated features can reflect statistical patterns of employee behaviors over longer periods. Feature standardization uses the Z-score method, but considering differences in behavioral distributions across job types, standardization parameters are calculated separately by job type. That is, code submission volume for R&D positions is compared with the mean and standard

deviation of that job group, avoiding bias caused by direct cross-position comparisons.

III. DESIGN OF INTELLIGENT PERFORMANCE ASSESSMENT MODEL

A. PROBLEM FORMALIZATION AND MODEL SELECTION LOGIC

Formalizing the performance evaluation problem as a supervised learning task: given an employee's behavioral sequence data within time window [1, T], the goal is to predict the employee's comprehensive performance score at time T+1. Formally, let $\mathcal{X} = \{x_1, x_2, \ldots, x_T\}$ be the behavioral feature sequence of an employee over T time units, where $x_t \in \mathbb{R}^d$ is the d-dimensional feature vector at the t time unit. The performance prediction target is $y \in \mathbb{R}$, representing the employee's comprehensive performance score. The model learning objective is to find a mapping function $f : x \to \mathbb{R}^{**}$, minimizing prediction error. This formalization fundamentally differs from traditional assessment methods in that the model no longer relies on concentrated year-end or quarter-end evaluations but achieves continuous performance prediction based on daily behavioral data.

Model selection needs to consider several key factors. First, behavioral data naturally has time series properties—today's performance of an employee is often related to their state in the previous period, requiring the model to capture temporal dependencies. Second, different behaviors contribute heterogeneously to performance—certain key behaviors (such as major project delivery) may have decisive impacts on performance, requiring the model to focus on these important moments. Third, the relationship between behavior and performance may vary with job type and business cycle, requiring the model to have certain adaptive capabilities.

Based on these considerations, this study selects Temporal Convolutional Network (TCN) as the base architecture, overlaying an attention mechanism on its upper layer. Compared to recurrent neural networks, TCNs have advantages of faster training speed and more controllable gradient paths. Its causal convolution structure ensures the model won't use future information to predict current performance, conforming to assessment scenario constraints. The attention mechanism is used to identify key time points in behavioral sequences that contribute significantly to performance prediction, providing some interpretability to the model.

B. TEMPORAL CONVOLUTIONAL NETWORK STRUCTURE

The input of the model is the behavior characteristic matrix $X \in \mathbb{R}^{L \times d}$ of employees in consecutive L time units, where L is the length of time window and d is the characteristic dimension. In this study, week is the basic time unit, and L is set to 8 weeks, which not only covers enough historical information, but also avoids the computational burden caused by the long sequence. We acknowledge that an 8-week window is a limitation, as it is insufficient to capture the full lifecycle of many R&D projects (e.g., major feature development often spans 3-4 months). This structural choice may introduce a recency bias that even the attention mechanism cannot fully mitigate. This issue is further addressed in the Discussion section.

The temporal convolutional network consists of multiple stacked residual blocks. Each residual block contains two layers of dilated causal convolution, with dilation factors growing exponentially with layer depth, allowing the network to achieve a large receptive field with fewer layers. Let the dilation factor of the k layer be $d_k = 2^{k-1}$, and the kernel size be k_{size}, then the receptive field of each neuron at the k layer is $R_k = 1 + (k_{\text{size}} - 1) \times (2^k - 1)$. This study adopts a 4-layer network structure with kernel size set to 3 and dilation factors of 1, 2, 4, and 8 respectively. Each neuron at the final layer can perceive information from the previous 15 time steps, essentially covering all historical moments within the 8-week window.

Specifically, for input sequence X, the dilated convolution operation is defined as follows: for the l layer, the t element of the output sequence is

$$Y_t^{(l)} = \text{ReLU}\left(\sum_{i=0}^{k_{\text{size}}-1} W_{l,i} \cdot X_{t-d_l \cdot i} + b_l\right). \quad (1)$$

Where $W_{l,i} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is the i weight matrix of the convolution kernel, $b_l \in \mathbb{R}^{d_{\text{out}}}$ is the bias term, and the dimensions for input and output channels are d_{in} and d_{out} respectively. When $t - d_l \cdot i < 1$, zero padding is used to maintain constant sequence length.

Each residual block internally contains two such convolutional layers and connects the input directly to the output through skip connections:

$$H = X + \text{Conv}_2(\text{ReLU}(\text{Conv}_1(X))). \quad (2)$$

After convolution computation, feature maps are converted to fixed-length vector representations through global average pooling layers, then connected to fully connected layers to output performance prediction values. To prevent overfitting, a Dropout mechanism is introduced in each residual block with a dropout probability of 0.2. The loss function uses mean squared error.

C. ATTENTION MECHANISM AND FEATURE WEIGHT ALLOCATION

Relying solely on temporal convolutional networks makes it difficult to distinguish the importance differences of behaviors from different time periods. Behaviors closer to the assessment point should logically have higher predictive weights, but recency effects may also lead to forgetting important early contributions. The introduction of attention mechanisms aims to achieve adaptive allocation of time weights at the model level. This study uses additive attention mechanism to calculate weight coefficients for each time step. For the feature representation $h_t \in \mathbb{R}^{d_h}$ of time step

t, where d_h is the hidden layer dimension, first obtain the query vector through linear transformation:

$$q_t = \tanh(W_q h_t + b_q). \tag{3}$$

Where $W_q \in \mathbb{R}^{d_a \times d_h}$, $b_q \in \mathbb{R}^{d_a}$ are learnable parameters, and d_a is the attention dimension (set to 64 in this study). Subsequently, compute the similarity score between q_t and the learnable context vector $u_q \in \mathbb{R}^{d_a}$:

$$e_t = u_q^T q_t. \tag{4}$$

To prevent gradient vanishing and excessive weight concentration, the scores are scaled with a scaling factor of $\sqrt{d_a}$, then normalized through softmax to obtain attention weights:

$$\alpha_t = \frac{\exp(e_t/\sqrt{d_a})}{\sum_{j=1}^{L} \exp(e_j/\sqrt{d_a})}. \tag{5}$$

The final sequence representation is the weighted sum of feature vectors at each time step:

$$v = \sum_{t=1}^{L} \alpha_t h_t. \tag{6}$$

The distribution of attention weights provides intuitive explanation of model decisions—if certain weeks are assigned higher weights, it indicates these periods' behaviors have greater influence on performance prediction. In practical applications, this can provide feedback to employees and managers about which recent behaviors positively impacted assessment results and which areas still need improvement.

D. STATE INFLUENCE COEFFICIENT AND BIAS CORRECTION

Although behavioral data is rich, it cannot fully reflect all factors affecting employee performance. Situational factors such as family emergencies, personal health issues, or temporary team adjustments may cause employees' short-term behavioral performance to deviate from their true capability levels. Assessing solely based on behavioral data without distinction may result in unfair treatment of these employees. Drawing inspiration from existing patent literature, this study introduces a state influence coefficient to adjust model outputs. Specifically, a correction factor $\beta$is multiplied to the model's predicted value:

$$P_{\text{adjusted}} = P_{\text{base}} \times (1 + \beta). \tag{7}$$

The calculation of β combines the learnable part and the manual calibration part. The learnable part is defined as $\beta_{\text{learned}} = \tanh(w_s^T s)$, where $s \in \mathbb{R}^m$ is the state feature vector (including m situational variables such as employee leave days, team change index, project urgency), and w_s is the learnable weight parameter. In order to avoid excessive dependence of the model on these sparse features, β is finally taken as $\beta = \lambda \cdot \beta_{\text{learned}}$, where λ is an adjustable hyperparameter. In practical application, fine-tuning is carried out within the range of [-0.2, 0.2] according to the judgment of the supervisor, and it is reviewed and confirmed by the human resources department. A positive value of β indicates that employees have been disturbed by negative factors recently, and the predicted value needs to be revised upward; a negative value indicates that employees enjoy special favorable conditions, and their expectations are appropriately lowered during evaluation.

This design embodies the concept of combining algorithmic assessment with human judgment: letting the model handle behavioral pattern recognition in routine situations while leaving exceptional cases requiring situational understanding and value judgment to human managers. The two complement each other—fully leveraging the model's advantage in processing large-scale data while preserving human managers' situational sensitivity.

IV. EMPIRICAL RESEARCH AND RESULTS ANALYSIS

A. DATA SAMPLE AND EXPERIMENTAL SETUP

The empirical study selected an internet company's R&D center and sales division as data collection targets, with a follow-up period from September 2024 to February 2025, totaling 6 months. The sample included 217 R&D position employees (168 R&D engineers and 49 test engineers) and 156 sales position employees. The average tenure for R&D positions was 3.2 years (SD=1.8), and for sales positions was 2.7 years (SD=1.5). In terms of level distribution, P1-P3 levels (junior) accounted for 31%, P4-P5 levels (mid-level) accounted for 45%, and P6 and above (senior) accounted for 24%.

During the data collection period, a total of 3.897 million behavioral records were obtained, with an average of 21,500 valid records generated daily. Distribution across data sources was as follows: project management tool records accounted for 1.563 million (40.1%), collaboration software records for 1.782 million (45.7%), and code and document system records for 0.552 million (14.2%). In terms of data completeness, the average weekly proportion of R&D position employees with records was 94.3%, and for sales positions was 97.1%; missing data during absences or vacations was retained as null values and handled by the input layer masking mechanism.

Performance labels used the company's quarterly assessment results, which were weighted combinations of department manager scores (weight 60%), peer evaluations (weight 20%), and self-assessments (weight 20%). After standardization processing, the values ranged from [0,100]. Internal consistency reliability (Cronbach's α) was 0.87, indicating good reliability of the assessment results.

Datasets were divided by time sequence into training set (first 4 months), validation set (5th month), and test set (6th month). The training set was used for model parameter learning, the validation set for hyperparameter tuning and early stopping decisions, and the test set for final performance evaluation.

Baseline Models. To provide a scientifically rigorous comparison, we avoid the category error of comparing our deep learning model to a non-existent "KPI regression model." Instead, we compare against standard and relevant baselines: (1) Linear Regression, using the same weekly aggregated feature vectors to directly predict the performance score; (2) Elastic Net, a regularized linear model to handle feature collinearity; (3) Random Forest, a non-linear ensemble method; (4) Support Vector Machine (SVM) with an RBF kernel; and (5) a TCN Base Model (without the attention mechanism). These baselines share the same input features, ensuring a fair comparison of model architectures. Evaluation metrics used root mean square error (RMSE) and mean absolute percentage error (MAPE) to measure prediction accuracy, while calculating the Spearman rank correlation coefficient between predicted and true values to assess ranking consistency. Fairness measurement used the Gini coefficient to evaluate whether prediction error distribution is balanced across different performance level employees. However, as we critically discuss later, a balanced error distribution does not guarantee the absence of systemic bias against employees with "invisible" work styles.

B. OVERALL PERFORMANCE COMPARISON

Experimental results are shown in Table 1. The proposed TCN-Attention model achieved the lowest RMSE (6.82) and MAPE (9.3%) on the test set, reducing error by 29.8% compared to the linear regression model. Compared to the basic TCN without attention, the introduction of the attention mechanism further reduced RMSE by 8.2%, demonstrating that adaptive weighting in the time dimension indeed helps performance prediction.

To test the statistical significance of performance differences, paired t-tests were conducted on prediction errors of TCN-Attention and Random Forest on the test set: $t(372)=4.28$, $p<0.001$, indicating the error difference is statistically significant. The paired t-test result for TCN-Attention and basic TCN was $t(372)=3.16$, $p=0.002$, also reaching significant levels.

In terms of ranking consistency, TCN-Attention achieved a Spearman correlation coefficient of 0.83, meaning the performance ranking predicted by the model has strong consistency with the true ranking. This is crucial for practical applications of performance evaluation—managers care not only about absolute scores but also whether the relative positions between employees are reasonable.

C. PERFORMANCE ANALYSIS BY JOB TYPE

Considering the significant differences in work nature and behavioral patterns between R&D and sales positions, it's necessary to separately examine the model's performance on these two groups. R&D work has characteristics of long project cycles and strong outcome lag effects, while sales work is oriented toward short-term performance with more direct connections between behavior and results.

As shown in Table 2, the model performs better on sales positions than R&D positions, with RMSE 15.5% lower and MAPE 16.8% lower. This difference aligns with expectations: sales behaviors have shorter transmission paths to performance results, with behavioral indicators like deal closure volume and customer visit frequency highly correlated with final performance; while R&D personnel contributions often take longer to manifest as observable outcomes, with more complex mapping relationships between behavior and performance. Notably, the Gini coefficients for both positions remain at relatively low levels, indicating the model maintains good fairness within both groups. The slightly higher Gini coefficient for R&D positions may stem from greater uncertainty in measuring R&D personnel's work outcomes themselves.

D. ANALYSIS OF ATTENTION WEIGHT DISTRIBUTION

The weight distribution output by the attention mechanism provides an interesting perspective for observing model behavior. Analysis of attention weights for all test set samples revealed that the average weights for the two most recent weeks (weeks 7-8) before the assessment point were highest, accounting for 41.3% of total weights. However, early time steps (weeks 1-4) still contributed 32.7% of the weights collectively, indicating the model doesn't simply rely on recency effects but extracts information from longer periods.

Further analysis found that for R&D positions with distinct project characteristics, when significant code submissions or critical problem fixes occurred during weeks 3-5, the attention weights for corresponding time steps would significantly increase. To quantify this phenomenon, we defined "key event weeks" as weeks when code submission volume exceeded twice the standard deviation above an individual's mean. For the 217 R&D employees, we separately calculated each person's average attention weight during key event weeks versus non-key event weeks and conducted paired t-tests. Results showed the average attention weight during key event weeks was 0.156, compared to 0.112 for non-key event weeks, with the difference being statistically significant ($t(216)=4.73$, $p<0.001$). This means the model can identify landmark events that significantly impact overall performance, even when these events occurred in earlier time windows.

This weight distribution characteristic provides valuable references for performance feedback. Employees can understand that not just recent performance, but key contributions from several months ago are also considered by the model, avoiding the common bias in year-end evaluations of "only remembering the last three months."

V. DISCUSSION

A. TIME DECAY EFFECT OF BEHAVIORAL DATA

Performance evaluation inherently faces a contradiction: more recent behaviors better reflect an employee's current state, but earlier behaviors better reflect stable traits. How

TABLE 1. Performance Comparison of Different Models

Model	RMSE	MAPE(%)	Spearman ρ	Gini coefficient
Linear Regression	9.71	13.2	0.63	0.30
Elastic Net	9.54	12.9	0.65	0.29
Support Vector Machine	8.45	11.8	0.69	0.26
Random Forest	7.81	10.9	0.74	0.23
TCN Base Model	7.43	10.2	0.78	0.21
TCN-Attention	6.82	9.3	0.83	0.19

TABLE 2. Model Performance Comparison by Job Type

Position	Sample Size	RMSE	MAPE(%)	Gini coefficient
R&D Position	217	7.28	10.1	0.20
Sales Position	156	6.15	8.4	0.17

to balance these relationships is a question that must be answered in assessment model design.

The attention mechanism in this study actually learned a data-driven time decay function. From the weight distribution, the decay is not simply exponential—when key events occur, weights locally rebound, forming multiple peaks. This indicates that the time patterns learned by the model are more flexible than preset decay functions and better align with the real processes of performance formation in actual work.

However, the fixed 8-week window is a significant limitation. For many R&D roles, a single feature development or refactoring project can span 12-16 weeks. By truncating the observation window at 8 weeks, our model structurally cannot account for contributions from the early planning and design phases of such long-term projects. This introduces a systematic recency bias. While the attention mechanism can re-weight events within the window, it cannot recover information from outside it. Future work must explore adaptive window sizing or recurrent architectures capable of handling indefinite sequence lengths.

B. ALGORITHMIC FAIRNESS AND POTENTIAL BIASES

No evaluation system can avoid fairness issues. Algorithms may appear objective but can inherit or even amplify biases implicit in training data. In this study, the Gini coefficient is controlled at 0.19, showing good error distribution balance, but this metric is far from the full meaning of fairness. The Gini coefficient measures the dispersion of prediction errors across the performance distribution. It does not measure the absence of systemic bias against specific subgroups, such as employees whose work is less "visible" to digital logging systems.

We must be vigilant about potential systematic biases in behavioral data collection itself. Some employees may have quantifiable behavioral records that are sparse due to work nature reasons (such as extensive offline communication), while others may generate more behavioral traces due to proficiency with digital tools. If the model indiscriminately equates behavior quantity with contribution level, it may systematically underestimate employees whose work styles are "invisible."

Analysis of sample data found that in R&D positions, tool proficiency (measured by time from first tool use to current) correlated moderately with behavioral record quantity ($r=0.36$, $p<0.01$), suggesting some tool usage bias exists. This issue needs correction through calibration sets when deploying the model. Specifically, we propose a "non-logged work calibration set": a sample of tasks known to be high-value but which generate minimal digital trace (e.g., whiteboard design sessions, mentoring conversations). By comparing model predictions for these tasks against manager-provided value scores, a correction factor for "invisible work" can be learned and applied.

Some research attempts to achieve cross-departmental data collaborative analysis through privacy protection technologies like federated learning, enhancing model generalization capabilities while protecting individual privacy [6]. The model framework in this study can be conveniently extended to support federated training mode, with each department's data remaining local and only model gradient information exchanged, achieving cross-departmental knowledge sharing while ensuring data security.

C. ETHICAL AND PRACTICAL LIMITATIONS OF CONTINUOUS BEHAVIORAL SURVEILLANCE

The implementation of a system that monitors instant messaging response times and meeting frequency raises significant ethical concerns that must be explicitly addressed. First, the "Hawthorne effect" is highly likely: employees may alter their natural behavior simply because they know they are being monitored, potentially increasing "performative busyness" (e.g., sending more messages, attending more meetings) at the expense of productive deep work.

Second, the use of metrics like "response time to collaborative messages" penalizes employees who may have legitimate reasons for delayed responses, such as deep concentration on a complex coding task. Without careful contextualization, such metrics can discourage the very state of focused flow that is critical for R&D productivity. Third, "knowledge contribution" indicators, despite our quality-weighting efforts, may still disadvantage employees who

prefer synchronous knowledge transfer (e.g., whiteboard discussions, ad-hoc pair programming) over asynchronous, logged forms (e.g., wiki updates, code comments). To mitigate this, any practical deployment must include opt-in transparency, regular ethical audits, and employee feedback mechanisms. The goal of this research is not to advocate for unrestrained surveillance, but to explore models that, if deployed with strong governance, could provide more timely and fair feedback than traditional annual reviews.

D. MODEL INTERPRETABILITY AND MANAGEMENT ACCEPTANCE

The practical application of performance evaluation models depends not only on technical performance but also on acceptance by managers and employees. A completely unexplainable black-box model, no matter how accurate its predictions, will struggle to gain trust in actual management scenarios [13].

This study consciously reserved space for interpretability during the model design phase. Attention weights can answer the question "which time periods' behaviors are more important"; feature importance analysis can answer "which types of behaviors contribute more to performance." Combined with explanation tools like SHAP values, attribution analysis can be further refined to the level of specific behavioral indicators. In practical applications, these explanation pieces are integrated into performance feedback reports. Employees see not only their evaluation scores but also the main contributing factors—whether exceptional task completion efficiency or high weight given to collaborative contributions. This transparency helps transform performance evaluation from a process of "passively accepting judgment" to one of "understanding organizational expectations."

VI. CONCLUSION

By integrating multi-source behavioral data from software engineering environments, this study constructs a comprehensive solution—comprising data collection, indicator systems, deep learning models, and bias correction mechanisms—to optimize employee performance evaluation. This technical framework ensures that human capital assessments accurately reflect the specialized operational nuances of modern software and IT service delivery. The main contributions are reflected in three aspects:

First, establishing a behavioral feature system covering three dimensions—task execution, knowledge contribution, and collaborative interaction—providing a richer data foundation for performance evaluation. We specifically addressed the volume-versus-quality problem by introducing weighted indicators like "adoption rate of code review comments." Second, designing an assessment model that integrates temporal convolutional networks with attention mechanisms, enabling the model to adaptively learn contribution weights of behaviors from different time periods and types.

Third, introducing a state influence coefficient to achieve organic integration of algorithmic assessment and human judgment, balancing standardization and contextualization needs in evaluation to a certain extent. Empirical research results show that this method outperforms linear regression, regularized regression, and common machine learning models in both assessment accuracy and fairness. The critical discussion reveals that the Gini coefficient, while useful, is an incomplete measure of fairness, and systemic biases against "invisible work" remain a key challenge. The distribution characteristics of attention weights reveal time patterns in the performance formation process, providing more explanatory information for performance feedback. Analysis by position shows the model performs well on both R&D and sales positions with different characteristics, but prediction errors for sales positions are significantly lower than for R&D positions, suggesting positions with strong outcome lag effects remain challenging for performance evaluation.

Several limitations of this research need improvement in follow-up work. The sample comes only from two departments of one company, and the external validity of conclusions needs verification across more industries and job types. The fixed 8-week time window is insufficient for many R&D project lifecycles, a structural limitation that future work must address via adaptive or longer-range sequence models. The performance labels use the company's assessment results, which, despite good reliability, still contain subjective judgment components that may introduce indirect bias to model evaluation. The reporting of state influence coefficients depends on manager judgment, potentially introducing new forms of subjective bias; future work could explore automated methods for identifying abnormal fluctuations based on behavioral data itself. Finally, the ethical implications of continuous behavioral monitoring—including the Hawthorne effect, the penalization of deep work, and privacy concerns—require further research and robust governance frameworks before any real-world deployment. Additionally, the time resolution of behavioral data is at the weekly level, possibly missing fine-grained intra-day behavioral structures; modeling methods for higher-frequency data are worth further exploration.

AUTHOR CONTRIBUTIONS

Conceptualization –Lijuan Yin and Lei Ren; methodology – Lijuan Yin and Lei Ren; investigation – Lijuan Yin and Lei Ren; writing-original draft preparation – Lijuan Yin and Lei Ren. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was supported by 1. 2024 Anhui Province Higher Education Science Research Key Project (Humani-

ties and Social Sciences), Research on the impact mechanism of digital human resource management on the performance of Generation Z employees based on the mediating effect of bidirectional work experience; 2024AH052680.

2. 2024 Anhui Province Higher Education Science Research Key Project (Humanities and Social Sciences), Research on the Regulation of Online Charity from the Perspective of Social Trust: A Case Study of Anhui Province; 2024AH052673.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] S. Dang, "Optimization of Big Data-based Human Resource Performance Evaluation System," Digital Economy, no. 9, pp. 52–53, 2025, doi: 10.3969/j.issn.2096-9988.2025.09.024.

[2] W. Li, P. Yi, and L. Wang, "Contextualized Evaluation Method and Application Based on Behavioral Feature Analysis," Chinese Journal of Management Science, vol. 32, no. 10, pp. 171–180, 2024, doi: 10.16381/j.cnki.issn1003-207x.2021.1411.

[3] Y. Liu, C. Wang, and J. Zhang, "The Impact of Recency and Halo Effects in Periodic Performance Appraisals: Evidence from Field Studies," Journal of Organizational Behavior, vol. 44, no. 5, pp. 789–805, 2023.

[4] Q. Chen, "Research on Employee Performance Evaluation and Incentive Mechanism in the Big Data Era," China Human Resources and Social Security, no. 3, pp. 20–23, 2024, doi: 10.3969/j.issn.1674-9111.2024.03.010.

[5] Shanghai Huiying Information Technology Co., Ltd., "An AI-based Performance Evaluation Optimization System and Method: CN202411994364.6[P]," 2025-05-13.

[6] D. Liu, "Application Scenarios and Case Analysis of Artificial Intelligence in Enterprise Human Resource Management," Enterprise Reform and Management, no. 19, pp. 69–71, 2025, doi: 10.3969/j.issn.1007-1210.2025.19.023.

[7] Y. Qing and Y. Tang, "Big Data-based Human Resource Performance Evaluation Model Using Bayesian Network of Deep Learning," Applied Artificial Intelligence, vol. 37, no. 1, 2023, doi: 10.1080/08839514.2023.2198897.

[8] Y. Wang, Q. Cong, and H. Yan, "An Empirical Study on the Impact of Internal Marketing on Task Performance of Frontline Service Employees," Nankai Business Review, vol. 11, no. 6, pp. 28–36, 2008, doi: 10.3969/j.issn.1008-3448.2008.06.006.

[9] Z. Luo and K. Lu, "Application research and effectiveness evaluation mechanism of hybrid intelligent algorithm integrating cognitive computing and deep learning for dynamically adjusting employee performance evaluation in multi-scale organizational networks," Discover Artificial Intelligence, vol. 5, no. 1, 2025, doi: 10.1007/s44163-025-00285-x.

[10] Baiyin Yinzhu Power (Group) Co., Ltd., State Grid Gansu Electric Power Company, and State Grid Gansu Electric Power Company Baiyin Power Supply Company, "Multi-dimensional Data-based Employee Performance Evaluation Method and System: CN202511058846.5[P]," 2025-10-31.

[11] X. Wang, Y. Liu, and J. Zhang, "Integrating Multi-Source Enterprise Data for Real-Time Performance Assessment: Evidence from Manufacturing and Tech Firms," Journal of Organizational Behavior, vol. 44, no. 6, pp. 1021–1038, 2023.

[12] Y. Zhang, H. Wang, and L. Chen, "Explainable AI for Performance Management: Enhancing Trust and Acceptance in Organizational Settings," Journal of Business and Psychology, vol. 38, no. 4, pp. 887–905, 2023, doi: 10.1007/s10869-023-09921-7.

[13] Khomsun Lelavijit and Supaporn Kiattisin, "An Integrated Conceptual Model of 360-Degree Performance Appraisal and Candidate Forecasting Using Adaptive Neuro-Fuzzy Inference System," Journal of Mobile Multimedia, vol. 16, no. 4, pp. 449–476, 2020, doi: 10.13052/jmm1550-4646.1642.