World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3090

A SOFT VOTING MECHANISM-BASED RANDOM FOREST MULTI-MODEL ENSEMBLE APPROACH FOR EMPLOYMENT STATUS PREDICTION

Can Luo[#,*], Zhen Liang[#]
School of Mathematics and Statistics, Hunan University of Science and Technology, Xiangtan 411201, Hunan, China.
[#]*Can Luo and Zhen Liang are both the first authors.*
[*]*Corresponding Author: Can Luo*

Abstract: Employment constitutes the cornerstone of people's livelihoods and the foundation of development. Presently, structural employment contradictions remain prominent in China, rendering the achievement of high-quality, full employment the foremost objective of current socio-economic development. Conventional single-model forecasting approaches exhibit limitations when handling high-dimensional, complex employment data, including insufficient generalisation capabilities and restricted capture of feature interactions. To address this, this study constructs a multi-model fusion employment status prediction model based on random forests, thereby enhancing predictive accuracy and stability. The study utilised a dataset comprising 4,980 employment samples, encompassing 57 raw features. Through data cleansing, Pearson correlation analysis, and feature extraction methods based on XGBoost and AUC cross-validation, 12 key feature variables were ultimately selected. Building upon this foundation, a soft voting ensemble strategy was employed to fuse three major prediction models—random forest, SVM, and LSTM—into a novel ensemble learning model. The predictive performance and stability of both the traditional trio of models and the new ensemble model were evaluated. Results indicate: The ensemble model achieved an accuracy of 85.49%, with recall and F1 scores of 97.34% and 93.21% respectively, outperforming each individual model. It effectively synergises the strengths of multiple models, enhancing adaptability to complex employment data and improving prediction robustness. This research provides a scientifically sound and effective ensemble learning approach for employment status prediction, offering practical value in supporting government precision policy-making and optimising employment services.
Keywords: Employment forecast; Random forest; Support vector machine; Long short-term memory; Soft voting ensemble strategy

1 INTRODUCTION

Employment constitutes the most fundamental pillar of people's livelihoods and a crucial underpinning of economic development. Whilst China's employment situation currently maintains fundamental stability, structural employment contradictions remain notably pronounced. Achieving high-quality, full employment represents both the core task of contemporary macroeconomic regulation and the fundamental guiding principle and strategic direction for employment work in the new era, influenced by multiple factors. Consequently, there is an even greater need for precise forecasting of employment conditions and the establishment of a more scientific employment forecasting system. This will provide decision-making support for the government to implement targeted employment policies, drive the transformation and upgrading of employment services, and implement the central government's strategy of ‘stabilising employment’.

Regarding data feature processing, owing to the multidimensional complexity of employment characteristics, studies such as those by Zhu Jiahui and Yu Liying indicate that the magnitude of the VIF value reflects the degree of multicollinearity among variables [1]. Scholars including Guan Deyong have analysed that the XGBoost algorithm can compute the relative feature importance of variables, thereby enabling rational feature analysis for subsequent dimensionality reduction research: XGBoost is an ensemble learning algorithm based on the gradient boosting framework [2]. Building upon GBDT, it incorporates multiple enhancements: optimising the loss function through second-order Taylor expansion, introducing regularisation terms to control model complexity, and employing parallel computation and weighted quantile algorithms to boost efficiency. It demonstrates outstanding performance in multi-class diagnostic tasks, effectively categorising complex employment factors. By specifically analysing the impact of significant influencing factors, the model's ability to extract key features and its training performance can be enhanced.

In terms of model prediction, traditional machine learning models have seen extensive application. For instance, in certain experiments, Ning Haotian et al. employed a random forest model to establish a data-driven job prediction framework [3]. This framework utilised multiple features including city, county, work experience, educational background, job type, and relevant technical skills to accurately forecast job titles and the requisite competencies for those positions. Addressing the non-linear dynamic characteristics of graduate employment data, Shen J proposed a university graduate employment trend forecasting method based on Long Short-Term Memory (LSTM) networks [4]. By considering multiple factors including student specialisation, current economic conditions, and academic

performance, this approach enables more accurate predictions of graduate employment trajectories. Tan Ying et al. utilised university student enrolment data to select variable parameters across four dimensions: individual attributes, family background, academic history, and campus performance [5]. Employing logistic regression algorithms, they constructed a graduate employment destination model, offering novel methodologies for career guidance. Li Ming et al. addressed the personalised demands of graduate employment services by constructing a dual-dimensional SVM classification model integrating human capital and social capital [6]. This model synthesises 18 features including academic performance and internship experience, providing a basis for universities to develop scientifically sound intervention strategies and expanding the feature boundaries of employment forecasting.

However, as the factors influencing employment status proliferate, the limitations of single machine learning approaches become increasingly apparent: individual models lack sufficient complexity to effectively process high-dimensional non-linear data and exhibit limited capacity to capture interactions between features. Furthermore, traditional models rely on manual feature engineering, whereas employment data often encompasses multi-source heterogeneous features, resulting in inadequate model generalisation capabilities. Consequently, in the field of employment forecasting research, ensemble models incorporating multiple models have emerged as a research focus due to their unique advantages. These models effectively address the complexity and diversity of employment data, providing robust support for analysing and predicting employment status. G R et al. integrated multiple models, including random forests, to analyse labour market data [7]. Their findings revealed that employment status is influenced by various social, economic, and regional factors, offering reference points for formulating employment-related policies. ROY M et al. fused support vector machines, neural networks, and decision trees to accurately predict job seekers' career trajectories, including promotion likelihood and career transition probability [8]. KALAISELVI B et al. employed soft voting ensemble learning, integrating model predictions via probabilistic averaging to forecast student employment status, demonstrating significantly higher accuracy than single models [9]. Huang Jianqiong et al. adopted an employment niche perspective, integrating multiple models including random forests to establish a graduate employment model [10]. This approach analysed the intrinsic mapping relationship between student employment influencing factors and career directions, achieving high classification accuracy and providing quantitative reference for university career guidance.

This study proposes a multi-model fusion employment status prediction model based on random forests. Following rigorous processing of employment data and training of three single models—random forests, SVM, and LSTM—a soft voting ensemble strategy is employed. With random forests as the core, this strategy combines two common prediction models, SVM and LSTM, through weighted fusion. This enables the ensemble model to support multi-dimensional optimisation and effectively circumvent the l LSTM. Employing a soft voting ensemble strategy with the random forest as the core, it integrates SVM and LSTM through weighted fusion. This enables the ensemble model to support multi-dimensional optimisation while effectively mitigating overfitting risks. Comparative analysis of metrics such as accuracy and precision confirms the ensemble model delivers superior, stable, and precise predictive performance across diverse datasets and real-world employment scenarios.

2 MODEL FEATURES AND DATA CONSTRUCTION

2.1 Model Principles

2.1.1 Random forest model

Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their results for prediction [11]. Its core principle involves combining multiple weak classifiers into a single strong classifier to reduce model variance and enhance generalisation capability. For classification problems, suppose there exist n decision trees T1,T2,....,Tn. For an input sample x, each decision tree yields a classification result $yi=Ti(x)$. The final classification outcome is determined through a voting mechanism, expressed as:

$$\hat{y}=argmax_{c\in C}\sum_{i=1}^{n} I(y_i=c) \tag{1}$$

Where C denotes the set of all possible categories, and $I(\cdot)$ is the indicator function, which takes the value 1 when the condition within the parentheses holds, and 0 otherwise. Among all categories, the category receiving the highest number of votes in the decision tree is selected as the final prediction result. When constructing decision trees, bootstrap sampling is employed to draw samples with replacement from the original dataset, with each sample having a probability of 1/N (where N is the sample size of the original dataset). Simultaneously, a subset of features is randomly selected for node splitting. This approach ensures that the training data and features differ across each decision tree, thereby mitigating high correlation between trees and reducing model variance.

2.1.2 SVM model

Support Vector Machines (SVM) constitute a supervised learning model primarily employed for classification and regression analysis [12]. Their core objective is to identify an optimal hyperplane within the feature space, thereby maximising the separation between samples of different categories. For linearly separable binary classification problems, consider a training dataset comprising $\{(x_1,y_1),(x_2,y_2),...,(x_N,y_N)\}$, where x_1 is the feature vector and $y_1\in\{-1,+1\}$ is the class label. The objective of SVM is to find a hyperplane defined by the following formula:

$$w^Tx+b=0 \tag{2}$$

Achieved by solving the defined optimisation problem, enabling samples of different categories to be maximally separated.

2.1.3 LSTM model

The fundamental architecture of LSTM (Long Short-Term Memory Network) centres upon the memory cell, equipped with three key gating mechanisms: the forget gate, input gate, and output gate [13]. These collaborate to regulate the flow and retention of information. The forget gate employs the σ activation function to determine the retention ratio of the previous memory cell state C_{t-1} based on the prior hidden state h_{t-1} and current input x_t, thereby enabling selective forgetting of historical information. The input gate utilises a combination of σ and tanh activation functions to filter and process new information generated from the current input x_t and h_{t-1}, precisely controlling the amount stored in the memory cell C_t. The memory cell C_t undergoes dynamic updating through the superposition of outputs from the forget gate and the input gate, thereby carrying the sequence's long-term dependency information. The output gate generates the current hidden state ht based on h_{t-1} processed by σ, x_t, and C_t transformed by $tanh$, simultaneously outputting useful information and providing input for the next time step's computation. This coordinated operation of gated mechanisms enables LSTMs to effectively circumvent the vanishing gradient problem, excel at processing long-sequence data, achieve long-term memory and precise regulation of critical information, and demonstrate formidable performance and adaptability in analytical and predictive modelling. The structural principles of LSTMs are illustrated in Figure 1.

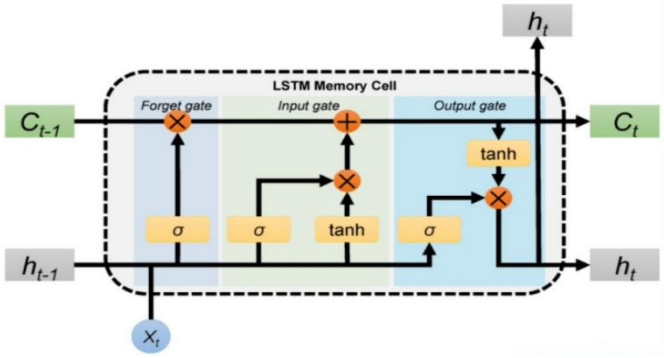

Figure 1 Structural Schematic Diagram of LSTM

2.2 Construction of the Dataset

This study constructed a raw dataset based on the 2024 Statistical Yearbook compiled by the National Bureau of Statistics and Yichang Municipal Bureau of Statistics [14]. This dataset comprises 4,980 employment samples, encompassing 57 employment characteristic variables including age, gender, ethnicity, major, graduating institution, macroeconomic indicators, and employment policies. Each sample includes an employment status indicator (1 denotes employed, 0 denotes unemployed). To reduce feature redundancy and facilitate subsequent model training and analysis, the raw data underwent preprocessing. First, irrelevant variables such as names, birthdates, and personnel IDs were excluded. Next, variables with missing values exceeding the threshold were removed, while others with low missing rates were imputed appropriately. Finally, feature recoding or fusion-based reconstruction was applied to achieve scientific dimensionality reduction. Through a series of data processing steps—irrelevant variable elimination, missing and outlier value handling, feature fusion and reconstruction, and Min-Max standardisation—a final dataset comprising 25 variables was retained for use as feature variables.

2.3 Extraction of Eigenvalues

2.3.1 Pearson correlation coefficient analysis

The Pearson correlation coefficient is a statistical measure of the linear relationship between two continuous variables. Its core logic lies in quantifying the strength of linear association between variables through the ratio of their covariance to their standard deviation. The specific formula is:

$$r=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}=\frac{\sum_{i=1}^{n}(X_I-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_I-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_I-\overline{Y})^2}} \tag{3}$$

Where r=1 denotes perfect positive linear correlation, r=−1 denotes perfect negative linear correlation, and r=0 denotes no linear correlation.

Following standardisation of each feature variable using Pearson's correlation coefficient, the correlation analysis of the standardised variables is presented in Table 1:

Table 1 Summary of Correlation Analysis on Standardized Variables

Variable Category	Specific Variables	Correlation Features	r Range	Treatment
Demographic Attributes	Gender, Ethnicity, Political Affiliation	Weakly correlated with most variables	| r | <0.2	Retain
	Age, Elderly	Moderate positive correlation between them	0.2≤ | r | <0.4; others | r | <0.2	Retain
	Marital Status	Moderately correlated with age	0.2≤ | r | <0.4; others | r | <0.2	Retain
	Adolescent, Migration Type	No effective correlation	r ≈ 0	**Remove**
Education & Occupation	Education Level, Major, Occupation	Moderately correlated within group	0.2≤ | r | <0.4	Retain
	Graduation School, Time	Weakly correlated with most variables	| r | <0.2	Retain
Social & Economic	Hukou Type, Registered Residence, Religion	Moderately correlated within group	0.2≤ | r | <0.4	Retain
	Living Status, Living Alone, Disabled	Weak-moderate correlated within group	0< | r | <0.4	Retain
	Employment Status, Policy, Military Service	Moderately correlated for employment status	0.2≤ | r | <0.4; others | r | <0.2	Retain
	Macro Economy, CPI, Labor Market	Weakly correlated with all variables	| r | <0.2	Retain

As shown in Table 1, following standardisation, the variables exhibit overall characteristics of moderate local correlations and weak overall correlations:

(1) Demographic attributes: Age shows a moderate positive correlation with being elderly, while marital status correlates moderately with age, reflecting intrinsic links within the population structure. Conversely, 'being adolescent' and 'type of change' exhibit correlations approaching zero with all variables, rendering them analytically insignificant and thus excluded.

(2) Education and occupation category: Educational attainment, field of study code, and major exhibit moderate positive correlations with occupational type, reflecting education's guiding influence on career choices. Other variables demonstrate relatively good independence.

(3) Socioeconomic category: Household registration status and religious affiliation show moderate correlations with registered residence region. Employment status, as a core variable, exhibits moderate positive correlations with employment policies and residential status. Economic indicators exhibit overall weak correlations with other variables, presenting no risk of multicollinearity.

Overall, after excluding invalid variables, the remaining 23 variables demonstrate good independence with no severe multicollinearity issues. They may be retained for subsequent feature extraction research to ensure the reliability of the core feature variables ultimately selected.

2.3.2 Feature selection for XGBoost

XGBoost is an ensemble learning algorithm based on the gradient boosting framework [15]. Building upon GBDT, it incorporates multiple enhancements: optimising the loss function through second-order Taylor expansion, introducing regularisation terms to control model complexity, and employing parallel computing alongside weighted quantile algorithms to boost efficiency. It demonstrates outstanding performance in multi-class fault diagnosis tasks for rolling bearings.

To circumvent the potential computational complexity arising from 23 feature variables, this study employs scientific dimensionality reduction on the selected variables. First, the XGBoost algorithm calculates and ranks the importance of each feature variable. Subsequently, the final feature extraction is performed using AUC cross-validation. The results obtained from XGBoost feature selection are illustrated in Figure 2 below:

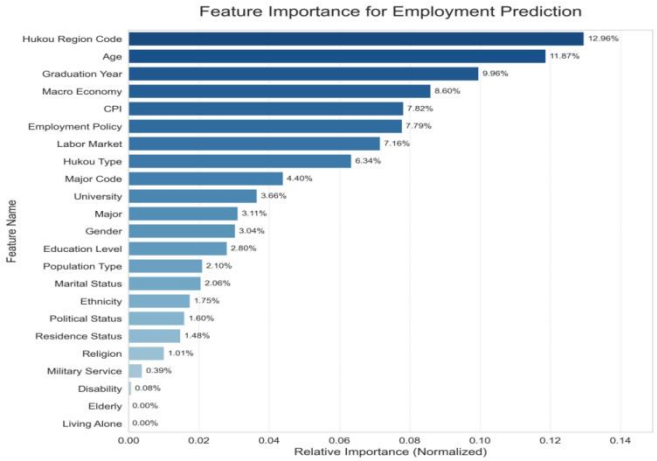

Figure 2 Feature Importance Ranking Based on XGBoost

2.3.3 Feature extraction for AUC cross-validation

(1) Cross-validation method

k-fold cross-validation partitions the dataset D into k mutually exclusive and approximately equal subsets $D_1,D_2,...,D_k$. In each iteration, one subset D_i is selected as the test set, while the union of the remaining $k-1$ subsets, denoted $D-D_i$, serves as the training set. This process is repeated k times. For each model M, during the i-th round of cross-validation, the model is trained on the training set $D-D_i$ to obtain M_i. Predictions are then made on the test set D_i, yielding the prediction result y_{pred}^{i}.

(2)AUC value

The *ROC* curve plots the false positive rate (*FPR*) on the horizontal axis and the true positive rate (*TPR*) on the vertical axis. The formulas for calculating the false positive rate (*FPR*) and true positive rate (*TPR*) are respectively:

$$\mathrm{FPF} = \frac{FP}{FP + TN} \tag{4}$$

$$\mathrm{TPR}=\frac{TP}{TP+FN} \tag{5}$$

Among these, TP denotes the number of true positives, FP the number of false positives, TN the number of true negatives, and FN the number of false negatives. By altering the classification threshold, the FPR and TPR are calculated at different thresholds to plot the *ROC* curve. The AUC represents the area under the *ROC* curve, which can be approximated numerically via integration methods such as the trapezoidal rule:

$$AUC\approx\sum_{i=1}^{n-1}\frac{(TPR_{I+1}+TPR_i)}{2}(FPR_{i+1}-FPR_I) \tag{6}$$

(3)Feature selection based on cross-validation AUC

For each feature subset S, in k-fold cross-validation, compute the AUC value for each cross-validation iteration: $AUC_i(S),i=1,2,...,k$. Then compute the average AUC value:

$$\overline{AUC}(S)=\frac{1}{K}\sum_{i=1}^{k}AUC_I(S) \tag{7}$$

Select the feature subset with the maximum average AUC value as the optimal feature subset, namely:

$$S_{opt}=argmaxs\ \overline{AUC}(S) \tag{8}$$

In summary, based on the importance ranking of each feature variable concerning employment status, cross-validation was employed to calculate the AUC values for the Random Forest, SVM, and LSTM models at different feature selection thresholds. These values were then averaged. The feature selection threshold corresponding to the highest average AUC—12—was identified as the optimal number of relevant features. The selected feature variables comprise the top 12 variables in descending order of importance, as illustrated in Figure 3 below:

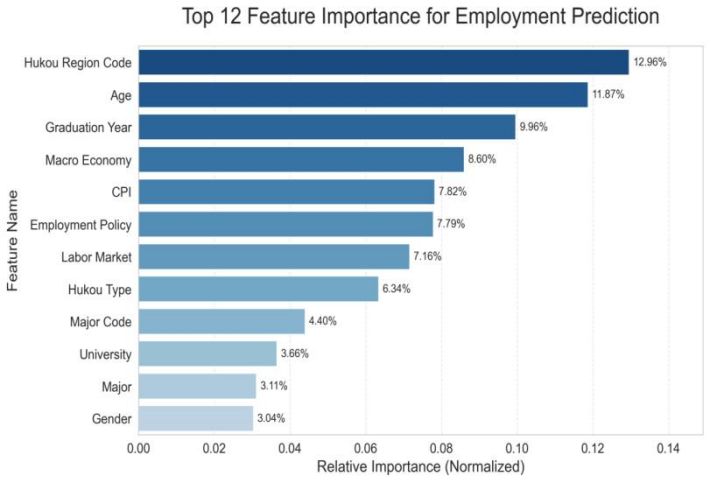

Figure 3 Selection and Ranking of the Top 12 Significant Feature Variables

3 MODEL PREDICTIONS AND RESULTS COMPARISON

3.1 Model Evaluation Metrics

Taking into account the model's characteristics, this study employs four metrics for comprehensive model evaluation: accuracy (ACC), recall, precision, and F1-score. Their respective calculation formulas are as follows:

$$\begin{cases} Accuracy=\dfrac{TP+TN}{TP+FP+TN+FN} \\ Recall=\dfrac{TP}{TP+FN} \\ Precision=\dfrac{TP}{TP+FP} \\ F1-score=\dfrac{2Precision\times Recall}{Precision+Recall} \end{cases} \tag{9}$$

The meanings of each parameter in the formula are as follows:
TP (True Positive): Represents the number of instances where the actual case is positive and the classification also yields positive, i.e., the number of correctly classified positive cases.
FP (False Positive): Represents the number of instances where the actual case is negative but classified as positive, i.e., the number of incorrectly classified positive cases.
TN (True Negative): Represents the number of cases that are actually negative and correctly classified as negative.
FN (False Negative): Represents the number of cases that are actually positive but incorrectly classified as negative.
After training and testing the model using the selected feature variables during data processing, employing the aforementioned metrics for comprehensive model evaluation enables multidimensional quantification of model performance. This approach avoids the limitations of single-metric assessments, holistically reflecting the model's recognition capability, precision, and overall adaptability. It provides standardised quantitative criteria for model comparison and optimisation, objectively measuring the model's generalisation ability and practical application value.

3.2 Random Forest Model

The pre-processed dataset of 4,980 employment characteristics was divided into a training set (3,984 samples) and a test set (996 samples) at an 8:2 ratio, employing stratified sampling to ensure balanced distribution between employed and unemployed samples. The model employed bootstrap sampling to repeatedly draw samples from the training set, constructing an ensemble model comprising 200 decision trees. Predictions were generated through a voting mechanism, with evaluation metrics subsequently calculated to yield the results presented in Table 2 below:

Table 2 Evaluation Metrics Results of the Random Forest Model

Model	Accuracy	Precision	Recall	F1
Random forest model	0.8137	0.8183	0.9418	0.8757

3.3 SVM Model

The dataset partitioning scheme aligns with that of the random forest model (training set: 3,984 cases; test set: 996 cases), ensuring a consistent evaluation benchmark. The model employs a radial basis function kernel (RBF) to address the

non-linear characteristics of the employment data. Through cross-validation, the penalty coefficient C=1.0 and *gamma* parameter=0.1 were optimised to construct an optimal classification hyperplane within the high-dimensional feature space, enabling binary classification of employment status. The final results were then processed to calculate the evaluation metrics, as presented in Table 3 below:

Table 3 Evaluation Metrics Results of the SVM Model

Model	Accuracy	Precision	Recall	F1
SVM model	0.8112	0.8112	0.8712	0.8958

3.4 LSTM Model

Adhering to the 8:2 dataset partitioning standard, the training set comprises 3,984 entries and the test set 996 entries. Following temporal processing, the data is input into the model. The model design incorporates multiple hidden layers, employing the cross-entropy loss function to optimise the training process. Overfitting is mitigated through a dropout rate of 0.2, with a focus on capturing the long-term impact of temporal features—such as macroeconomic fluctuations and policy adjustments—on employment status. The final evaluation metrics are presented in Table 4 below:

Table 4 Evaluation Metrics Results of the LSTM Model

Model	Accuracy	Precision	Recall	F1
LSTM model	0.8102	0.8111	0.8679	0.8952

3.5 Optimisation Strategy Based on Random Forest Model Ensemble

3.5.1 Comparison of single model results

Through performance testing and results analysis of the three major models—Random Forest, SVM, and LSTM—each model exhibits distinct advantages and suitability characteristics in the employment status prediction task. Specific comparisons are as follows:

(1)Random forest model: Demonstrated balanced performance across metrics, achieving an accuracy of 0.8137, precision of 0.8183, recall of 0.9418, and an F1 score of 0.8757. It exhibits strong resistance to overfitting and effectively captures feature interactions, with only recall showing minor room for improvement.

(2)Support Vector Machine (SVM) model: Demonstrated balanced accuracy and recall at high levels, with outstanding capability in handling high-dimensional non-linear features. It excelled in identifying job suitability-related characteristics, though precision showed minor room for improvement.

(3)LSTM model: Excels at capturing the impact of time-series features such as macroeconomic fluctuations and policy adjustments on employment. It exhibits balanced metric values without significant weaknesses and demonstrates good adaptability to dynamic changes in the employment market.

Overall, no single model can comprehensively accommodate the complex characteristics of employment data, which simultaneously exhibits high dimensionality, non-linearity, and time-series properties. To enhance the efficiency and precision of employment status forecasting, a multi-model ensemble learning strategy was developed: this project employs the Random Forest model—which demonstrated the most balanced performance across all metrics—as the base model. This is integrated with two complementary models, SVM and LSTM, both characterised by high recall rates. The fusion algorithm synergises the strengths of multiple models, thereby effectively improving the accuracy, stability, and generalisation capability of the employment status prediction model. This approach better addresses the complex and volatile nature of employment data.

3.5.2 Optimisation based on Random forest model ensembles

Soft Voting integration synthesises the predictive probabilities of base models while enhancing forecast reliability. This approach employs a soft voting strategy to consolidate the predictive strengths of three models—Random Forest, SVM, and LSTM—achieving highly accurate employment status predictions. The process comprises two primary steps:

Step 1: Independent Base Model Prediction: Each base model is trained independently on the input features, comprising all core features that have undergone screening. Following training, each base model outputs prediction probabilities for the sample belonging to different categories, denoted respectively as pRF, pXGB, and pLR.

Step 2: Probability-Weighted Fusion: The ensemble model performs a weighted sum of the prediction probabilities from each base model, deriving the final prediction result through the following formula:

$$\hat{y}_{ensemble}=argmax_c \sum_{M=1}^{M} w_m \cdot \hat{p}_{m,c} \tag{10}$$

Here, M denotes the number of base models, w_m represents the weight of base model m, with equal weights assigned to all base models by default. $\hat{p}_{m,c}$ denotes the predicted probability of class c by base model m.

3.5.3 Analysis of integrated optimisation results

The newly integrated model achieved high evaluation metrics upon training and testing, with an accuracy of 0.8549, precision of 0.8476, recall of 0.9734, and an F1 score of 0.9321. Its average performance surpassed that of individual models including Random Forest, SVM, and LSTM. Table 5 below presents the evaluation metrics for each model:

Table 5 Evaluation Metrics Results of Different Models

Model	Accuracy	Precision	Recall	F1
Random forest model	0.8137	0.8183	0.9418	0.8757
SVM model	0.8112	0.8112	0.8712	0.8958
LSTM model	0.8102	0.8111	0.8679	0.8952
New ensemble prediction model	0.8549	0.8476	0.9734	0.9321

Observations indicate that the newly integrated model leverages a soft voting strategy to synthesise the strengths of multiple models. It retains the random forest's resilience against overfitting and its ability to capture feature interactions, while enhancing the SVM's processing capabilities for high-dimensional sparse features such as specialisation and alma mater. Concurrently, the incorporation of LSTM improves the weighting assigned to temporal features, resulting in superior overall performance across complex data scenarios. The comparative prediction results, presented in confusion matrix format, are illustrated in Figures 4 and 5 below:

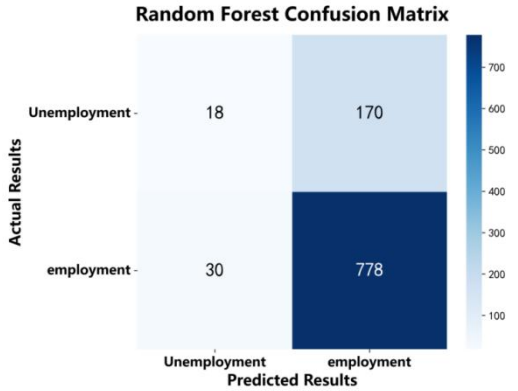

Figure 4 Prediction Performance of the Random Forest Model

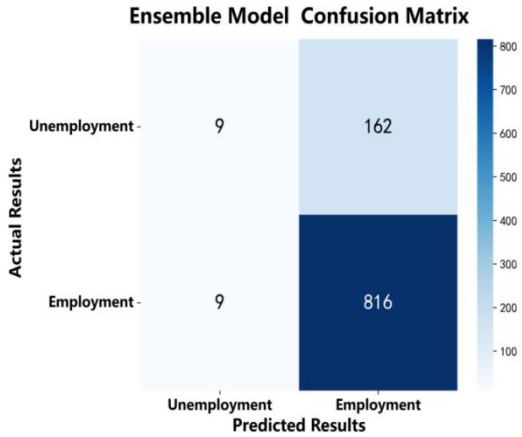

Figure 5 Prediction Performance of the New Ensemble Model

4 CONCLUSIONS

High-accuracy employment status prediction represents a significant focus in contemporary employment initiatives. Traditional single-model approaches exhibit limited generalisation capabilities within complex, multi-factor scenarios, with considerable scope for enhancing predictive precision. To accurately assess the employment choices of the workforce, this study processed and analysed 4,980 employment samples from Yichang City. An ensemble learning model integrating Random Forest, Support Vector Machines (SVM), and Long Short-Term Memory (LSTM) was constructed to forecast the influence of various characteristic factors on employment status. Comparing the results against single-model predictions, the ensemble model demonstrated superior average performance across all four key metrics—accuracy, precision, recall, and F1 score—compared to individual models. This approach better accommodates the complex and dynamic nature of employment data, enhancing both the stability and generalisation capability of the predictions.

In summary, the ensemble learning model constructed in this study demonstrates commendable performance and application potential in predicting and evaluating employment status across multiple feature variables. It holds significant reference value for scientifically formulating employment policies and optimising labour resource allocation. However, certain limitations exist: firstly, the model does not sufficiently incorporate unquantifiable dynamic factors

such as sudden industry events or regional temporary policies; secondly, predictive accuracy for specialised fields and flexible employment arrangements requires enhancement. Future research may incorporate attention mechanisms to enhance the capture of temporal features, while simultaneously expanding the sample coverage to further improve the model's adaptability and robustness in complex employment scenarios.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Cheng Qiyun, Sun Caixin, Zhang Xiaoxing, et al. Short-Term load forecasting model and method for power system based on complementation of neural network and fuzzy logic. Transactions of China Electrotechnical Society, 2004, 19(10): 53-58.
[2] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.
[3] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.
[4] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.
[5] Shi Biao, Li Yuxia, Yu Xhua, et al. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.
[6] Li M, Wang F. Prediction and countermeasures of college graduates' career destinations based on machine learning SVM algorithm: From the perspective of human capital and social capital framework. Educational Research and Experiment, 2023(5): 78-84.
[7] G R, Rajendran S, Pavul J. Exploring factors influencing female employability in Bengaluru using XGBoost. 2025 2nd International Conference on Circuits, Power and Intelligent Systems (CCPIS). Bhubaneswar: IEEE, 2025: 1-6.
[8] Roy M, Bhoi A K, Sharma K. Multimodal machine learning approaches for career prediction. 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC). Bhubaneswar: IEEE, 2022: 1-5.
[9] Kalaiselvi B, Geetha S. Ensemble voting classifier-based machine learning model for predictive modeling of campus student placements. Science and Technology: Recent Updates and Future Prospects. Bhopal: Book Publisher International, 2024: 1-10.
[10] Huang J Q, Guo W L, Li Q Y. Research on influencing factors of college graduates' employment based on random forest model. Journal of Jiangsu Normal University (Natural Science Edition), 2019, 37(04): 55-58, 74.
[11] Dong Xibin, Yu Zhiwen, Cao Wenming, et al. A survey on ensemble learning. Frontiers of Computer Science, 2020, 14(2): 241-258.
[12] Hu C, Meng F, Luo W, et al. Fault diagnosis method for wind turbine bearings based on CEEMDAN and ISSA optimized SVM. Journal of Machine Design, 2025, 42(4): 109-119.
[13] Xue Y, Huang Y S, Yang C J. Design method for contra-rotating propellers with optimal circulation based on wake based on neural network surrogate model and genetic algorithm. Journal of Ship Mechanics, 2025, 29(4): 517-527.
[14] Yichang Municipal Bureau of Statistics. Statistical Yearbook of Yichang 2024. https://tjj.hubei.gov.cn/tjsj/sjkscx/tjnj/gsztj/ycs/202504/P020250428344682431565.pdf.
[15] Li Jiangtao, An Xingqin, Li Qingyong, et al. Application of XGBoost algorithm in the optimization of pollutant concentration. Atmospheric Research, 2022, 276: 106238.