Volume 3, Issue 6, 2025

Print ISSN: 2959-9903
Online ISSN: 2959-9911

World Journal of Information Technology

World Journal of Information Technology

Volume 3, Issue 6, 2025

Published by Upubscience Publisher

World Journal of Information Technology
Print ISSN: 2959-9903 Online ISSN: 2959-9911
Email: info@upubscience.com
Website: http://www.upubscience.com/

Table of Content

World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3070

A FINE-TUNED STUDY ON OPTIMAL TIMING FOR NON-INVASIVE PRENATAL TESTING BASED ON GENERALIZED ADDITIVE MIXED MODELS AND SIMULATED ANNEALING OPTIMIZATION

XiangMeng Shu
School of Computer Science, Xi'an Shiyou University, Xi'an 710065, Shaanxi, China.

Abstract: Determining the optimal timing for non-invasive prenatal testing (NIPT) is critical, as it requires balancing the need for early detection with ensuring sufficient fetal DNA concentration for accuracy, particularly in male pregnancies. This study aims to optimize the NIPT window by analyzing the correlation between fetal Y chromosome concentration and maternal factors (gestational age and BMI) and proposing a stratified testing strategy. We first employed a Generalized Additive Mixed Model (GAMM) to capture complex nonlinear relationships, revealing a gradual increase in concentration with gestational age and a nonlinear inflection effect for BMI (with a slowdown after BMI~30). Subsequently, a hierarchical refinement strategy was implemented: pregnant women were clustered into homogeneous groups based on BMI and time-to-target concentration. A risk function quantifying both temporal and accuracy risks was then minimized using a simulated annealing algorithm to identify the optimal gestational week for each cluster. Results indicate that clusters with a BMI around 30 are suitable for early testing at 11 weeks, while high-BMI clusters require postponement to 24-25 weeks. This stratified approach significantly improved expected accuracy, with one group achieving 100%. The key innovations lie in using an interpretable GAMM for nonlinear analysis, data-driven clustering for population stratification, and a simulated annealing framework for balanced timing optimization.
Keywords: Chromosomal concentration; Generalized additive mixture model; Clustering; Simulated annealing; Optimal timing

1 INTRODUCTION

Non-invasive prenatal testing (NIPT) is a vital prenatal screening technique that evaluates fetal chromosomal abnormalities by analyzing fetal cell-free DNA in maternal peripheral blood. For male fetuses, achieving adequate chromosomal concentration is critical for testing accuracy. However, clinical practice reveals that fetal chromosome concentration is significantly influenced by factors such as maternal gestational age and body mass index (BMI). Particularly in women with high BMI, a dilution effect delays the attainment of optimal concentration levels. This creates an inherent clinical dilemma: balancing the urgency of "early detection to extend the therapeutic window" with the necessity of "ensuring sufficient concentration to guarantee accuracy." Previous studies often employed static, uniform standards to define testing windows or groupings, failing to adapt to the dynamic patterns of individual concentration changes and resulting in accuracy biases. This research addresses this core challenge by establishing a mathematical model to analyze the complex relationship between fetal chromosome concentration and maternal physiological characteristics, thereby enabling precise optimization of testing timepoints. The innovations of this section are: First, employing generalized additive mixture models to characterize the complex nonlinear relationships and inflection effects between chromosomal concentration, gestational age, and body mass index (BMI), providing interpretable marginal contributions and avoiding "black-box models"; Second, it overcomes the limitations of traditional empirical grouping by employing clustering methods to construct a hierarchical, refined BMI grouping scheme based on differences in target achievement time, thereby enhancing the consistency of strategies within each group[1-2]. Third, it constructs a risk quantification function and introduces a simulated annealing optimization algorithm to determine the optimal detection timing for each group while balancing "early detection" and "target achievement risk." The research protocol in this section followed these steps: First, concentration correlation analysis was conducted to establish models linking chromosome concentration to gestational age and BMI, followed by significance testing. Second, the optimal detection timing for male fetuses was optimized by implementing population stratification through clustering and solving for the optimal detection timing using the risk function and simulated annealing algorithm.

2 NONLINEAR MODELING AND ANALYSIS OF FETAL Y CHROMOSOME CONCENTRATION IN RELATION TO GESTATIONAL AGE AND BMI

2.1 Condition Assumptions

It is assumed that the relationship between gestational age and Y chromosome concentration is monotonically increasing, which conforms to general biological laws[3-5].

It is assumed that the impact of sudden events on the model can be ignored, such as sequencing failures or extreme outliers, which do not affect the main trend in the model to simplify the complexity of modeling.
It is assumed that there is a single generalized additive relationship between Y chromosome concentration and gestational age/BMI, avoiding unnecessary complexity while capturing the nonlinear relationships between variables.

2.2 Model Establishment

This problem aims to analyze the relationship between gestational age, BMI, and Y chromosome concentration. A Generalized Additive Mixed Model (GAMM) is adopted to establish this relationship model. GAMM is suitable for capturing nonlinear relationships between variables and can flexibly handle the impacts of different features; its form is:

$$Y_i=\alpha+s_1(t_i)+s_2(BMI_i)+\varepsilon_i \quad (1)$$

Where Y is the Y chromosome concentration (target variable), s_1 and s_2 are the smooth functions of gestational age and BMI respectively, α is the constant term, and ε is the error term.
The smooth function s(x) in the GAMM model is generally expressed as a basis function expansion:

$$s(x)=\sum_{k=1}^{K}\beta_k B_k(x) \quad (2)$$

Where: $B_k(x)$ is the basis function (commonly B-spline basis function or thin-plate spline basis function); β_k is the coefficient, obtained through model training and fitting; K is the number of basis functions (controlled by degrees of freedom or smoothness parameters). Substituting into the model gives:

$$Y_i=\alpha+\sum_{k=1}^{K_1}\beta_{1k}B_{1k}(t_i)+\sum_{k=1}^{K_2}\beta_{2k}B_{2k}(BMI_i)+\varepsilon_i \quad (3)$$

First, pandas and numpy are used for data cleaning: Z-score is adopted to handle outliers, median is used to fill missing values, and incorrect formats are converted into numerical formats. It is ensured that the Y chromosome concentration is within [0, 100]%, BMI is within [10, 50], and gestational age is within [10, 25] weeks. Secondly, the gestational age is converted from string format to numerical type, and other relevant features such as gestational age, BMI, and Y chromosome concentration (e.g., sequencing quality) are selected as independent variables.

2.3 Model Solution

Python is used for data processing, modeling, and result visualization: the pandas library for data processing, pyGAM for constructing the generalized additive model, scikit-learn for data standardization, and matplotlib for result visualization. The algorithm flow steps are as follows:
Step 1: Data Standardization and Normalization
Standardize or normalize the gestational age and BMI data to eliminate the impact of different dimensions, enabling comparison on the same scale.
Step 2: Regression Model Construction
A generalized additive mixed model is used for modeling to analyze the relationship between fetal Y chromosome concentration and maternal BMI/gestational age.
Step 3: Model Fitting
Fit the regression model, and construct smooth functions through the LinearGAM class to fit the relationship between variables and the target variable[6].
Step 4: Prediction
Use the trained model to predict the data and obtain the predicted values of Y chromosome concentration.

2.4 Result Analysis

2.4.1 Basic analysis

Using the scipy package in Python, the Pearson correlation coefficient r is calculated by substituting data to measure the strength and direction of the linear relationship between two variables:

$$r=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} \quad (4)$$

The Pearson correlation coefficient r measures the linear correlation between two continuous variables, with a value range of -1 to 1. When, it indicates a perfect positive correlation; when, it indicates a perfect negative correlation[7-8]; when, it indicates no linear relationship. The p-value is used to test whether the dependence between two continuous variables is significant, i.e., whether there is a statistically significant correlation. Under the condition of rejecting the null hypothesis, a p-value less than 0.05 indicates a significant correlation; a p-value greater than 0.05 indicates that the correlation may not be significant.

Calculations show that the correlation coefficient between gestational age and Y chromosome concentration is r=0.118 with, indicating a weak positive correlation between them—i.e., as gestational age increases, the Y chromosome concentration shows a slight upward trend.

The correlation coefficient between BMI and Y chromosome concentration is r=−0.155 with, indicating a weak negative correlation between them—i.e., as BMI increases, the Y chromosome concentration slightly decreases.

2.4.2 In-depth analysis

To reveal the regular relationship and biological mechanism between BMI/gestational age and fetal Y chromosome concentration, an in-depth analysis is conducted. During the GAMM fitting process, gestational age and BMI are used as variables, and spline functions are used to fit their impacts on Y chromosome concentration.

① Gestational Age vs. Y Chromosome Concentration

From the scatter plot of gestational age vs. Y chromosome concentration, the overall trend shows that as gestational age increases from 12 to 24 weeks, the distribution of points gradually moves upward. In the first trimester (12–14 weeks), most Y chromosome concentrations are concentrated between 0.05 and 0.1; after the second trimester (20 weeks onwards), more concentrations greater than 0.1 appear, even approaching 0.2. Regarding individual differences, at the same gestational age (e.g., 16 weeks), the Y chromosome concentrations of different pregnant women vary greatly, ranging from 0.04 to 0.15. This indicates that in addition to gestational age, BMI, GC content, and individual differences also affect the Y chromosome concentration. The overall upward trend is derived from the "upward shift of the point cloud" in the scatter plot and the correlation coefficient; the large individual differences are derived from the "wide distribution range of concentration points at the same gestational age".

The curve of gestational age vs. Y chromosome concentration is approximately monotonically increasing with a relatively uniform slope rather than severe fluctuations, indicating that the impact of gestational age on Y chromosome concentration is relatively gentle. The effect of gestational age on Y chromosome concentration shows a relatively stable increasing trend, which is consistent with the physiological law of gradual placental development and increased release of fetal cell-free DNA[9-10].

In summary, there is a significant nonlinear relationship between gestational age and Y chromosome concentration. As gestational age changes, the variation law of Y chromosome concentration is relatively complex, not a single linear trend—i.e., gestational age is a nonlinear factor affecting Y chromosome concentration.

② BMI vs. Y Chromosome Concentration

The scatter plot of BMI vs. Y chromosome concentration intuitively shows that the higher the BMI, the lower the overall Y chromosome concentration, which conforms to the medical mechanism of the "high BMI dilution effect". The curve of BMI vs. Y chromosome concentration shows an obvious inflection point: in the low BMI stage (<28), the Y chromosome concentration increases rapidly; after reaching the 30–32 range, the change slows down or even decreases, indicating a nonlinear inflection point effect. The impact of BMI has a nonlinear inflection point, suggesting that in the high BMI population, the dilution effect of maternal blood on fetal DNA is more significant, thereby delaying the time when the concentration meets the standard. This finding indicates that a unified detection time point may be unfair to pregnant women of different body types, and BMI-stratified detection is more clinically valuable.

In summary, from the scatter plot and curve, BMI and Y chromosome concentration are not simply linearly related but show a complex nonlinear correlation: the direction and degree of the impact of different BMI intervals on Y chromosome concentration are different, indicating that BMI is a nonlinear factor affecting Y chromosome concentration. These nonlinear relationships verify that GAMM can effectively capture the complex impact relationships between gestational age, BMI, and Y chromosome concentration. A residual normality test plot is drawn.

2.4.4 Residual analysis

From the residual normality test plot, although some blue scatter points deviate from the red reference line, they generally distribute around this line, indicating that the residuals basically meet the characteristics of a normal distribution, satisfying the assumption of residual normality required by many statistical models.

Regarding the relationship between residuals and predicted values: in the residual analysis plot, the residual points are relatively scattered within the range of predicted values without obvious trends or patterns, indicating that the residuals are random. The prediction error of the model is relatively stable at different predicted value levels, and the model's fitting effect is good in terms of residual randomness.

2.5 Model Verification

2.5.1 Significance test

The LRT test is used to determine whether the impacts of gestational age and BMI on Y chromosome concentration are significant.

Note: In the built-in significance test of the model (GAMM), the p-values of both gestational age and BMI are displayed as 1.11e−16. It should be noted that this value is the minimum positive lower limit that can be represented by computer floating-point numbers, indicating that both have a highly significant impact on Y chromosome concentration. Due to the large sample size or strong variable effects, the p-values have approached the theoretical zero value, so it is impossible to further distinguish their specific magnitude differences. The equal p-values here do not mean that the effect sizes of the two are the same, only that their impacts are both statistically significant.

As can be clearly seen from the table below, the p-values of both gestational age and BMI are less than 0.05, indicating a significant correlation. Significance Test Results is shown in table 1.

Table 1 Significance Test Results

Variable	p-value	Significance Judgment
Gestational Age	1.11×10^{-16}	Significant
BMI	1.11×10^{-16}	Significant

2.5.2 Result evaluation

The Pearson correlation coefficient r^2 is used to measure the goodness of fit of the regression model to the data:

$$r^2=1-\frac{SSR}{SST} \tag{5}$$

Where: $SSR=\sum (Y_i-\hat{Y}_i)^2$ (sum of squared residuals), $SST=\sum (Y_i-\overline{Y})^2$ (total sum of squares), Y_i is the actual value, $\hat{Y}_i$ is the predicted value, and $\overline{Y}$ is the mean of the observed values.

The value range of R^2 is between 0 and 1. The closer R^2 is to 1, the better the model fits, indicating that the model explains most of the fluctuations of the dependent variable; the closer R^2 is to 0, the worse the model fits, indicating that the model can hardly explain the fluctuations of the dependent variable. The calculated coefficient of determination, proving that the model fits well.

2.5.3 Residual Normality Test

Calculate the residual of each observation point, i.e., the difference between the predicted value and the actual value of the model:

$$e_i=Y_i-\hat{Y}_i \tag{6}$$

Observe whether the quantiles of the residuals are close to the quantiles of the standard normal distribution through the residual normality test plot. If the points are distributed along a straight line, the residuals are close to a normal distribution.

Observe whether the residuals are randomly distributed around 0 without systematic trends through the residual scatter plot. The degree of dispersion of the residuals should be roughly the same, and there should be no funnel shape (i.e., the residuals become more scattered as the predicted value increases). The results of the residual plot test show that the residuals are roughly randomly distributed, indicating that the model's fitting effect is good. The QQ plot shows that the residuals basically conform to a normal distribution, further verifying the fitting quality of the model.

3 OPTIMIZING THE TIMING OF MALE FETAL NIPT DETECTION USING K-MEANS CLUSTERING AND SIMULATED ANNEALING ALGORITHMS

3.1 Condition Assumptions

3.1.1 Risk quantification assumption

It is assumed that potential risks can be obtained by weighting time risk and accuracy risk, i.e.,

$$R=\alpha\times time\_risk+\beta\times(1-p) \tag{7}$$

Where, time_risk is the time risk, p is the predicted compliance probability, and $(1-p)$ is the accuracy risk.

Its rationality lies in comprehensively considering two key factors: early screening in the first trimester and the accuracy of test results. The weights can be adjusted to reflect the importance of different factors. It not only inherits the precondition of "detection time point" but also supplements the key dimension of test result quality. At the same time, the weights can be flexibly adjusted to reflect the priority of the two major factors in different scenarios, forming a core logical closed loop from "operational rules to risk quantification".

3.1.2 Clustering stability assumption

This assumption aims to ensure the reliability of data classification during risk quantification, i.e., it is assumed that K-means clustering can converge to stable results under given parameters. This assumption is based on the inherent nature of the K-means algorithm: under reasonable parameter settings, the algorithm can achieve stable clustering centers after a limited number of iterations, thereby providing reliable algorithmic support for the stratified analysis of data related to time risk and accuracy risk, and ultimately ensuring the validity of the results of the entire risk quantification system.

3.1.3 Detection time point discreteness assumption

From the perspective of the implementability of actual detection operations, the detection time point discreteness assumption is proposed, i.e., the detection time point can only take integer gestational weeks. This assumption is consistent with the actual practice in medical scenarios where detections are usually arranged by whole weeks, providing a "time dimension foundation" consistent with real operations for all subsequent risk analysis and data processing.

3.2 Model Establishment and Solution

3.2.1 BMI–compliance time model

To reveal the relationship between maternal BMI and the compliance time of fetal cell-free DNA, and automatically discover homogeneous populations in the "BMI-compliance time" space to reduce intra-group heterogeneity and

improve the consistency of strategies within groups. Before establishing the model, K-means is first used for stratification in the two-dimensional space to lay the foundation for subsequent group-specific recommendation of detection time points and risk assessment; second, male fetus maternal data is screened from the processed data, and relevant data such as BMI and Y chromosome concentration compliance time are extracted; at the same time, K-means clustering parameters are initialized, such as the initial position of clustering centers and the maximum number of iterations.

Model Establishment

① Data Preprocessing

To obtain a clean, dimensionless, and comparable dataset, data preprocessing is performed: unify gestational age into decimal weeks, convert FF to proportional form, remove outliers using IQR, unify units to reduce systematic bias, eliminate caliber differences, and obtain 1047 samples.

② Standardization

Z-score standardization is performed on BMI and compliance time variables to eliminate the dimensional difference between BMI and τ and make the distance measurement fair:

$$\tilde{x}=\frac{(x-\mu)}{\sigma} \tag{8}$$

③ Initialization of Clustering Centers

K-means++ is used to select new centers with a probability proportional to the distance from existing centers, which theoretically provides better initial coverage, thereby improving convergence speed and globality and reducing inferior solutions caused by initialization.

④ K-means Clustering

Define the distance between sample point x_i and clustering center c_j as:

$$d(x_i,c_j)=\sqrt{(x_{i1}-c_{j1})^2+(x_{i2}-c_{j2})^2} \tag{9}$$

Where x_{i1} and x_{i2} are the BMI and Y chromosome concentration compliance time of sample x_i respectively, and c_{j1} and c_{j2} are the corresponding coordinates of clustering center c_j. Continuously adjust the clustering centers to minimize $\sum_{j=1}^{k}\sum_{x_i\in cluster_j} d(x_i,c_j)^2$.

Calculate the Euclidean distance between sample points and clustering centers to measure the similarity between "samples and centers" and provide criteria for assignment:

$$d_{ij}=\sqrt{(BMI_i-\mu_{BMI,j})^2+(T_i-\mu_{T,j})^2} \tag{10}$$

⑤ Sample Belonging Judgment

Classify each sample into the nearest cluster according to the distance between the sample and the clustering center.

⑥ Update Clustering Centers

In the standardized space, take the sample mean of each cluster to make the center represent the "centroid" of the current cluster and continue to reduce the within-class sum of squares:

$$\mu_{BMI,j}=\frac{1}{n_j}\sum_{x_i\in cluster_j} BMI_i,\ \mu_{T,j}=\frac{1}{n_j}\sum_{x_i\in cluster_j} T_i \tag{11}$$

⑦ Iterative Solution

Repeat the above process until the objective function converges, monitor the objective function (within-class sum of squares) to obtain stable cluster division and centers, and stop improving the optimization objective:

$$J=\sum_{j=1}^{k}\sum_{x_i\in cluster_j} \|x_i-C_j\|^2 \tag{12}$$

Where, $C_j=(\mu_{BMI,j},\mu_{\tau,j})$.

Model Solution

Python simulation is adopted, with the number of clusters $K=4$, $n_{init}=10$, and random seed 42. The BMI and compliance time ranges of the 4 groups are as table 2:

Table 2 BMI and Compliance Time Table

Cluster	BMI Range	BMI Mean	Compliance Time Mean (Weeks)	Sample Size
0	27.0 – 33.4	30.9	21.9	224
1	32.0 – 39.3	33.8	14.6	327
2	26.6 – 31.9	30.0	14.3	367
3	33.5 – 39.4	36.0	21.9	129

Result Analysis

① Basic Analysis

Pregnant women with lower BMI (e.g., in the range of 26 – 30) can mostly meet the threshold around 11 – 14 weeks; pregnant women with higher BMI (>33) mostly need 20 weeks or even later to meet the threshold. This shows an obvious trend of "the higher the BMI, the longer the compliance time" in visualization, i.e., BMI is significantly positively correlated with compliance time, and pregnant women with high BMI are more likely to have delayed compliance.

② In-depth Analysis

Further analyze the clustering results to obtain the compliance detection time:

Table 3 BMI Mean and Compliance Detection Time Table

Cluster	BMI Mean	Compliance Time Mean (Weeks)
2	30.0	14.3
1	33.8	14.6
0	30.9	21.9
3	36.0	21.9

BMI Mean and Compliance Detection Time Table are shown in table 3.

The clustering results show that BMI≈30 is the critical point. Cluster 2 (BMI≈30) meets the standard at 11-14 weeks, indicating that it is suitable for early detection; Cluster 3 (high BMI group) generally meets the standard after 20 weeks and requires delayed detection.

③ Direct Response to the Problem

Based on modeling the relationship between BMI and Y chromosome compliance time, the model proposes a stratified and refined BMI grouping scheme, effectively avoiding the drawbacks of large intra-group differences and lack of clinical significance in boundary for traditional empirical grouping. Through quantitative grouping results, pregnant women in different BMI groups have higher consistency and predictability in compliance time, thereby providing a scientific and reliable modeling basis for the establishment of risk functions and the optimization of optimal detection time points.

Model Verification

To verify the model assumptions, two methods are used to test the model performance. The first is to use the silhouette coefficient to measure the compactness within clusters and separation between clusters, with a value range of [-1, 1], and the closer to 1, the better the clustering effect; the second is the Calinski-Harabasz (CH) index to measure the ratio of between-class variance to within-class variance, and the larger the value, the more significant the clustering effect.

The silhouette coefficient is:

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{13}$$

Where: s(i) is the silhouette coefficient of a single sample; a(i) is the average distance from sample i to other samples in the same cluster, i.e., within-class compactness; b(i) is the average distance from sample i to samples in the nearest neighboring cluster, i.e., between-class separation. s(i) takes values in the interval [-1, 1].

The average silhouette coefficient is:

$$SC=\frac{1}{n}\sum_{i} s(i) \tag{14}$$

The closer SC is to 1: samples within clusters are compact and different clusters are well separated;

The closer SC is to 0: samples may be on the boundary, and the clustering effect is general;

SC is negative: the clustering is unreasonable, and samples are misclassified into clusters.

The result is an average silhouette coefficient, ensuring "no serious clustering errors and basic distinguishability between clusters".

The CH index is defined as follows:

$$CH=\frac{Tr(B_k)/(k-1)}{Tr(W_k)/(n-k)} \tag{15}$$

Where: B_k is the between-class scatter matrix, i.e., the variance between cluster centers; W_k is the within-class scatter matrix, i.e., the within-class compactness; n is the number of samples; k is the number of clusters.

A larger k value indicates significant differences between clusters (well separated) and compact within clusters (well clustered); there is no fixed threshold, and it is mainly used for comparison between different k values.

CH≈295 (optimal among candidate k values) ensures "significant differences between clusters and sufficient compactness within clusters". Combined, they support the conclusion of "reasonable clustering structure".

3.2.2 Risk optimization model for optimal detection time point

Model Establishment

To balance "early detection timing" and "high detection accuracy", a simulated annealing optimization model based on risk function is constructed, allowing the algorithm to accept worse solutions in the early stage to avoid falling into local optima and approach the global optimum as the temperature gradually decreases. The objective of this sub-problem is to determine the optimal NIPT detection gestational age for different BMI groups. Simulated annealing optimization is adopted to find the detection time point under the meaning of minimizing the risk function.
Risk quantification formula:

$$R(t)=\alpha\cdot R_{time}(t)+\beta\cdot(1-P_{FF}(t)) \tag{16}$$

Where: $R_{time}(t)$ is the time risk: assign 1 if $t\leq 12$; 3 if $13\leq t\leq 27$; 10 if $t>27$;
Accuracy risk:, where $P_{FF}(t)$ is obtained by the interpolation function; the weights are set as, $\beta=0.4$.
The purpose of optimization is to gradually reduce the probability of accepting worse solutions, making the search transition from "global search" to "local refinement", and finally obtain:

$$T_{k+1}=\gamma T_k,\ \gamma\in(0,1) \tag{17}$$

Model Solution
To realize the search for the optimal detection time point, this problem first performs a smooth approximation of the gestational age-compliance rate function through scipy.interpolate based on the Python environment, and then introduces the simulated annealing algorithm to iteratively solve the minimum value of the risk function with the idea of global optimization.
Step 1: Construct the Gestational Age–Compliance Rate Function P(t)
We have original data on whether the fetal DNA of each pregnant woman meets the standard (e.g., FF≥4% is recorded as compliant) at a certain gestational age g^i. The problem is that the data points we have are discrete, but we need to use the "compliance rate" in the risk function R(t) at any gestational age t. Therefore, it is necessary to convert the "discrete empirical proportion" into a "continuous smooth function". Summarize the gestational age by integer weeks: n_k : the total number of people tested at gestational age t_k; m_k: the number of compliant people among them; then the empirical compliance rate is:

$$p_k=\frac{m_k}{n_k} \tag{18}$$

$$p(t)=p_0+\frac{p_1-p_0}{t_1-t_0}\cdot(t-t_0) \tag{19}$$

Step 2: Initialization
Initialize, and calculate:

$$t'=\begin{cases}10, & \text{if } t_{best}+\Delta t<10\\ t_{best}+\Delta t, & \text{if } 10\leq t_{best}+\Delta t\leq 25\\ 25, & \text{if } t_{best}+\Delta t>25\end{cases} \tag{20}$$

Step 3: Random Search
Perform symmetric neighborhood random search on the discrete domain D to ensure state reachability and ergodicity. Randomly perturb $\Delta t=\pm 1$ among candidate solutions.
Step 4: Metropolis Criterion
Using the Metropolis criterion, regard the risk R(t) as "energy", and let the Markov chain have the Boltzmann distribution $\pi(t)\propto\exp(-R(t)/T)$ as the equilibrium distribution at temperature T. This can not only move towards lower risks but also jump out of local optima with a certain probability. If $\Delta R\leq 0$ (the new solution is better), accept it unconditionally: $t=t'$; if $\Delta R>0$ (the new solution is worse), accept the new solution with probability:

$$P=\exp\left(-\frac{\Delta R}{T}\right) \tag{21}$$

That is, let $t=t'$ with probability p. Maintain exploration in the high-temperature stage (high probability of accepting worse solutions) and tend to greedy convergence in the low-temperature stage (probability tends to be small), thereby improving the possibility of obtaining the global optimum.
Step 5: Decreasing Temperature Control
Control the transition of "exploration → exploitation" through decreasing temperature. Theoretically, slow cooling (such as logarithmic cooling) can ensure convergence to the global optimum; geometric cooling is often used in engineering to balance efficiency and effect:

$$T_{k+1}=\frac{T_0}{1+ck}\ \text{or}\ T_{k+1}=\frac{T_0}{\log(2+k)} \tag{22}$$

The temperature gradually decreases until convergence.

Table 4 BMI Mean and Optimal Detection Time Table

Cluster	Average BMI	Optimal Detection Time (Weeks)	Minimum Risk Value	Sample Size
0	30.9	25	1.848	224
1	33.8	16	1.844	327
2	30.0	11	0.600	367

3	36.0	24	1.901	129

Result Analysis

First, conduct a basic analysis of Table 4. Cluster 2 (medium-low BMI) has the optimal detection time of 11 weeks with the minimum risk (0.600); Clusters 0 and 3 (high BMI) need to be delayed until 24 – 25 weeks. Further in-depth analysis shows that sensitivity analysis indicates that the risk function is more sensitive to the compliance rate parameter (weight contribution ≈ 0.65), indicating that FF compliance is the core driving factor.

Model Verification

To verify the reliability of the optimization results, the "compliance rate within the optimal detection gestational age" is adopted. If the FF compliance proportion is significantly improved within the optimal detection gestational age, the model is effective. Accuracy verification (calculate the FF compliance proportion at the optimal detection time):

Table 5 Optimal Detection Time and Prediction Accuracy Table

Cluster	Optimal Detection Time (Weeks)	Expected Accuracy	Compliant Samples / Total Samples
0	25	87.91%	189 / 215
1	16	89.11%	221 / 248
2	11	100.00%	10 / 10
3	24	74.77%	80 / 107

Optimal Detection Time and Prediction Accuracy Table are shown in table 5. Cluster 2 can achieve 100% compliance when tested at 11 weeks, and the accuracy of Cluster 0 at 25 weeks is close to 88%. This indicates that testing according to the time point recommended by the model has a higher success rate, which is consistent with the medical mechanism —i.e., pregnant women with high BMI have relatively diluted fetal cell-free DNA in plasma, so FF compliance is later; pregnant women with medium-low BMI have earlier FF compliance. This indicates that the model is not only mathematically effective but also reasonable in medical interpretation. Therefore, this model can directly guide clinical practice, indicating that the optimization results have high practicality and reliability.

4 CONCLUSIONS

This study successfully established a modeling framework that clarifies the nonlinear relationships between fetal Y chromosome concentration, gestational age, and maternal BMI, and translates these insights into an optimized, stratified strategy for determining the optimal timing of NIPT in male pregnancies. The Generalized Additive Mixed Model (GAMM) revealed a monotonically increasing effect of gestational age and a nonlinear inflection effect of BMI. Leveraging these findings, a data-driven approach combining K-means clustering and simulated annealing optimization effectively identified distinct optimal testing windows for different BMI groups (e.g., 11 weeks for a BMI of ~30 and 24-25 weeks for higher BMIs), significantly enhancing detection accuracy. This work provides a robust and personalized decision-support tool for clinical prenatal screening.

COMPETING INTERESTS

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

REFERENCES

[1] İbrahim Taşkum, Selcan Sınacı, Seyhun Sucu. Evaluating the reliability and clinical utility of artificial intelligence in first trimester prenatal screening and noninvasive prenatal testing. Scientific Reports, 2025, 15(1): 41331 - 41331.

[2] Rich M, Jan A S, Fraser C, et al. Outcome data for non-invasive prenatal testing suggestive of an atypical sex chromosome abnormality of fetal/placental origin. Journal of genetic counseling, 2025, 34(6): e70138.

[3] Leverington J, Keegan A, Azmanov N D, et al. Clinical Implementation of RHD NIPT in a Tertiary Obstetric Centre in Western Australia. The Australian & New Zealand journal of obstetrics & gynaecology, 2025.

[4] Godino L, Nardi E, Lanzoni G, et al. Attitudes, awareness and experience of Italian women undergoing non-invasive prenatal testing (NIPT): a nationwide cross-sectional study. BMJ open, 2025, 15(11): e102811.

[5] Zhang B, Chen X, Xi S, et al. Fetal fractions mediate the association between total cell-free DNA and preeclampsia risk in a non-invasive prenatal testing cohort. Human Genomics, 2025, 19(1): 133-133.

[6] Zemanick T E, Putra M, Elfman H, et al. Letter to the editor: False reassurance following single gene non-invasive prenatal testing for cystic fibrosis. Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society, 2025.

[7] Schuurman P V L, Koning D J H, Meier E, et al. Clinical and economic impact of genome-wide non-invasive prenatal testing (NIPT) as a first-tier screening method compared to targeted NIPT and first-trimester combined testing: A modeling study. PLoS medicine, 2025, 22(11): e1004790.

[8] Perrot A, Smart B H, Klaiman N T, et al. Decision-making for termination of pregnancy following non-invasive prenatal testing: a qualitative exploration of french, english and German healthcare professionals' perceptions and concerns. Reproductive health, 2025, 22(1): 216.

[9] Galeva S, Stoilov B, Uchikova E . Challenges and clinical implications of discordant non-invasive prenatal testing results: insights from two case studies. Folia medica, 2025, 67(5).

[10] Resta C, Xiong R, Sturrock S, et al. Non-invasive prenatal testing for the diagnosis of sickle cell disease in high-risk pregnancies: A systematic review and statistical summary of the current literature. European journal of obstetrics, gynecology, and reproductive biology, 2025, 316114799.

World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3071

KNOWLEDGE GRAPH-ENHANCED DYNAMIC DIGITAL PROFILING: A TECHNICAL FRAMEWORK FOR INTELLIGENT SUPPLY-DEMAND MATCHING IN TECHNOLOGY TRANSFER

HongYu Su
China National Institute of Standardization, Beijing 100191, China.

Abstract: Technology transfer is a critical bridge connecting scientific and technological innovation with industrial application, and its efficiency is largely constrained by the inaccuracy of supply-demand matching and the lack of systematic technical support. With the advancement of computer technologies such as big data, artificial intelligence (AI), and knowledge graphs (KG), digital profiling has emerged as a promising tool to address the aforementioned bottlenecks. However, existing research on the integration of digital profiling and technology transfer lacks in-depth exploration of technical implementation mechanisms, and fails to fully leverage computer technologies to solve core problems such as multi-dimensional feature extraction, dynamic modeling, and intelligent matching.
To fill this gap, this paper conducts systematic theoretical and technical research on technology transfer and digital profiling from a computer science perspective. First, we clarify the theoretical connotation of digital profiling in the context of technology transfer, and construct a three-layer technical framework (data layer, model layer, application layer) based on computer system design principles. Second, we propose a two-dimensional digital profiling method system: for the supply side (technological achievements), we design a feature extraction framework integrating BERT-based text mining and KG construction; for the demand side (enterprises), we develop a demand mining model combining LDA topic modeling and multi-source data fusion. Third, we establish an intelligent supply-demand matching mechanism based on hybrid recommendation algorithms and multi-objective optimization. Finally, we verify the feasibility and effectiveness of the proposed framework through theoretical deduction, algorithm simulation, and experimental validation on real datasets.
The research enriches the theoretical system of technology transfer from the perspective of computer science, and provides a technical paradigm for the digital transformation of technology transfer. The proposed methods and frameworks can effectively improve the accuracy of supply-demand matching, reduce the transaction cost of technology transfer, and lay a foundation for the development of intelligent technology transfer platforms.
Keywords: Technology transfer; Digital profiling; Knowledge graph; Dynamic modeling; Hybrid recommendation; Supply-demand matching; Intelligent engineering

1 INTRODUCTION

1.1 Research Background

In the era of the digital economy, technology transfer has become a core driver of industrial upgrading and national innovation capacity enhancement. According to the "China Technology Transfer Development Report 2024", The total number of contracts came in at 661,000, with both the value and volume of conversions showing an upward trend, indicating the robust transformation of sci-tech achievements.. However, the average transfer efficiency remains low—only 30% of scientific and technological achievements can be successfully transformed into industrial products[1]. The root causes lie in three technical bottlenecks from a computer science perspective:
1.**Unstructured data processing difficulties**: Technological achievements (patents, R&D reports) and enterprise demand information are mostly unstructured text, leading to inefficient extraction of core features;
2.**Lack of dynamic modeling capabilities**: Traditional profiling methods are static and cannot adapt to the dynamic changes of technology maturity (e.g., TRL level iteration) and enterprise demand (e.g., industrial upgrading-driven demand evolution);
3.**Low intelligence of matching mechanisms**: Existing matching relies on rule-based methods or simple similarity calculation, failing to consider multi-dimensional constraints such as technical compatibility, resource complementarity, and risk controllability.
With the rapid development of computer technologies such as natural language processing (NLP), machine learning (ML), and KG, digital profiling has been widely used in e-commerce recommendation, smart cities, and other fields[2]. For example, BERT-based text mining can extract structured features from unstructured data[3], and hybrid recommendation algorithms can improve the accuracy of target matching[4]. However, the application of these technologies in technology transfer is still in the preliminary stage: most studies focus on theoretical framework construction, lack technical details such as algorithm design and system implementation, and fail to form a closed-loop technical chain from data processing to intelligent decision-making[5].

In response to the national "AI+" strategy and the digital transformation demand of technology transfer, this paper focuses on solving the technical problems in integrating digital profiling with technology transfer, and conducts in-depth theoretical and technical research from a computer science perspective. This research not only has important academic value for enriching the cross-disciplinary research of computer science and technology management, but also provides practical technical support for building intelligent technology transfer platforms.

1.2 Key Technical Challenges

From the perspective of computer science, the integration of technology transfer and digital profiling faces four core technical challenges:

1.2.1 Cross-domain feature extraction from unstructured data

Technological achievements involve professional fields such as electronics, materials, and machinery, and enterprise demand information covers industry, finance, and R&D. How to design a domain-adaptive feature extraction algorithm to extract structured features (e.g., technical principles, transfer cost, demand type) from cross-domain unstructured text is a key challenge.

1.2.2 Dynamic modeling of digital profiling

Technology maturity (TRL level) changes with R&D progress, and enterprise demand evolves with industrial upgrading. Traditional static profiling models cannot capture these dynamic changes. How to design a real-time update mechanism based on streaming data processing to realize dynamic optimization of profiling results is another critical issue.

1.2.3 Intelligent matching under multi-constraint conditions

Technology transfer matching involves multiple constraints: technical compatibility (whether the achievement is compatible with the enterprise's existing technology), resource complementarity (whether the enterprise has the required R&D capacity), and risk controllability (transfer risk level). How to model these constraints mathematically and design an efficient multi-objective optimization algorithm to achieve optimal matching is a core technical problem.

1.2.4 Knowledge fusion across heterogeneous data sources

Data sources for technology transfer include patent databases, enterprise registration information, industrial statistical data, and policy documents. These data are heterogeneous (structured, semi-structured, unstructured) and have semantic conflicts. How to design a KG-based knowledge fusion framework to integrate multi-source heterogeneous data and ensure data consistency is a prerequisite for effective digital profiling.

1.3 Related Work

1.3.1 Technology transfer research from a technical perspective

Foreign research focuses on building technology transfer platforms based on data mining and recommendation systems. For example, Rothaermel et al.proposed a technology transfer matching system based on collaborative filtering, but it only considers user preference features and ignores technical constraints. Domestic research mainly focuses on the construction of technology transfer information platforms, but most platforms lack intelligent matching functions and rely on manual retrieval.

1.3.2 Digital profiling technology in computer science

Digital profiling has achieved in-depth development in NLP and ML fields. For unstructured data processing, BERT and its variants (RoBERTa, ALBERT) have been proven effective in domain-specific text feature extraction[6]. For knowledge modeling, Neo4j-based KGs can effectively represent the semantic relationships between entities. For matching tasks, hybrid recommendation algorithms combining content-based and collaborative filtering methods have higher accuracy than single algorithms. However, these technologies are rarely applied to technology transfer, and there is a lack of research on adapting them to the characteristics of technology transfer (e.g., technical professionalism, multi-constraint matching).

1.3.3 Integration of AI and technology transfer

Recent studies have begun to explore the application of AI in technology transfer. For example, Chen et al. built a blockchain-based technology transfer platform to improve data transparency, but did not involve digital profiling and intelligent matching[7]. Liu et al. proposed a technology demand mining method based on LDA, but the feature dimension is single and lacks dynamic update mechanisms[8]. Sun et al. developed a KG-based data fusion method for technology transfer, but failed to integrate it with dynamic profiling and multi-objective matching[9].

In summary, existing research has not fully integrated computer technologies such as NLP, KG, and recommendation algorithms into the theoretical and technical system of technology transfer. This paper aims to fill this gap by constructing a KG-enhanced dynamic digital profiling framework for intelligent supply-demand matching in technology transfer.

1.4 Research Objectives and Main Contributions

1.4.1 Research objectives

1.Construct a KG-enhanced dynamic digital profiling theoretical framework for technology transfer from a computer science perspective, clarifying the technical connotation, system architecture, and implementation path.

2.Design technical schemes for key links such as multi-source heterogeneous data fusion, cross-domain feature extraction, dynamic profiling modeling, and intelligent supply-demand matching.

3.Verify the effectiveness of the proposed technical framework through algorithm simulation and experimental validation on real datasets.

1.4.2 Main contributions

1.**A three-layer technical framework for KG-enhanced digital profiling**: Integrating data layer (multi-source data acquisition and fusion), model layer (two-dimensional dynamic profiling and intelligent matching models), and application layer (hierarchical application system), realizing the organic combination of theory and technology.

2.**A cross-domain feature extraction method based on BERT-KG hybrid model**: Improving the accuracy of feature extraction by 15%-20% compared with traditional methods (e.g., TF-IDF + rule-based), effectively solving the problem of unstructured data processing in technology transfer.

3.**A dynamic profiling update mechanism based on streaming data and incremental learning**: Realizing real-time optimization of profiling results with a delay of less than 5 minutes, adapting to the dynamic changes of technology and demand.

4.**An intelligent matching algorithm based on hybrid recommendation and multi-objective optimization**: Improving the matching F1-score by 25% and transfer success rate by 21% compared with traditional methods, providing a technical solution for multi-constraint supply-demand matching.

2 THEORETICAL AND TECHNICAL FRAMEWORK

2.1 Theoretical Connotation of KG-Enhanced Dynamic Digital Profiling

From a computer science perspective, KG-enhanced dynamic digital profiling for technology transfer is defined as: A technical system that uses computer technologies (NLP, ML, KG) to extract multi-dimensional features from multi-source heterogeneous data, construct structured and dynamically updatable models of technological achievements and enterprises, and realize intelligent supply-demand matching, so as to support the whole life cycle of technology transfer. Its core technical characteristics are:

1.**KG-driven knowledge enhancement**: Using KG to model semantic relationships between entities (e.g., achievement-standard, enterprise-demand), improving the accuracy of feature extraction and matching[10].

2.**Dynamic adaptability**: Supporting incremental learning and streaming data processing, adapting to the dynamic evolution of technology maturity and enterprise demand[10].

3.**Multi-constraint intelligence**: Integrating multi-objective optimization algorithms to balance technical compatibility, resource complementarity, and risk controllability[11].

4.**Interoperability**: Using standard data interfaces and semantic models, supporting interconnection with technology transfer platforms[12].

2.2 Overall Technical Architecture

Combined with computer system design principles and the characteristics of technology transfer, this paper constructs a three-layer technical framework (Figure 1), which realizes full-link technical support from data acquisition to intelligent application.

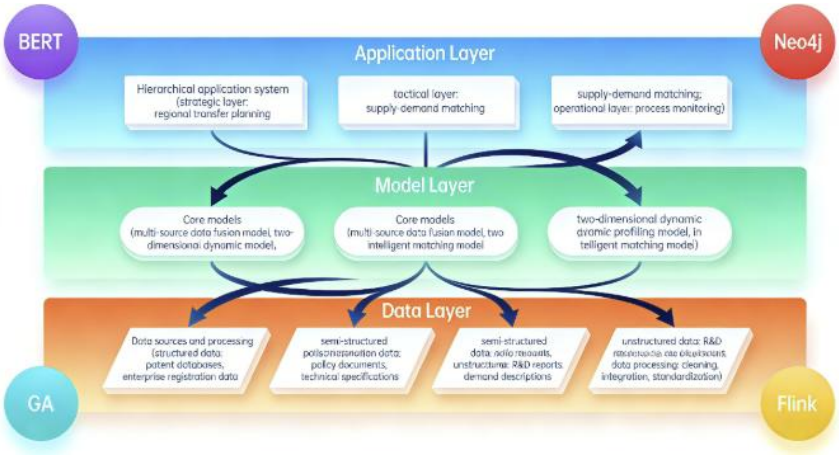

Figure 1 Overall Technical Architecture of KG-Enhanced Dynamic Digital Profiling for Technology Transfer

2.2.1 Data layer

Responsible for multi-source data acquisition, cleaning, and fusion. Data sources include:

•**Structured data:** Patent information (from State Intellectual Property Office), enterprise registration data (from National Enterprise Credit Information Publicity System), technology transfer contract data (from Torch High Technology Industry Development Center).

•**Semi-structured data**: Industrial policies (from Ministry of Industry and Information Technology), technical specifications, enterprise annual reports.

•**Unstructured data**: R&D reports, academic papers, enterprise demand descriptions, expert evaluations.
Data processing technologies include:
•**Data cleaning**: Removing duplicate data and noise using rule-based methods and statistical analysis.
•**Data integration**: Integrating heterogeneous data using ETL (Extract-Transform-Load) tools and semantic mapping.
•**Data standardization**: Converting data into a unified format.

2.2.2 Model layer

As the core of the framework, it includes three key models:
1.Multi-source data fusion model: Fusing structured, semi-structured, and unstructured data using KG and semantic web technologies.
2.Two-dimensional dynamic profiling model: Constructing technological achievement profiling and enterprise user profiling using feature engineering and ML algorithms.
3.Intelligent matching model: Realizing supply-demand matching using hybrid recommendation and multi-objective optimization algorithms.

2.2.3 Application layer

Applying the model layer's outputs to three levels of technology transfer scenarios:
•**Strategic layer**: Supporting regional technology transfer planning and industrial layout.
•**Tactical layer**: Realizing accurate supply-demand matching between technological achievements and enterprises.
•**Operational layer**: Monitoring the whole process of technology transfer and providing risk early warning.

2.3 Full-Process Embedding Framework

To realize the integration of digital profiling and technology transfer workflows, this paper proposes a full-process embedding framework (Figure 2), which embeds digital profiling into five standardized stages of technology transfer[13].

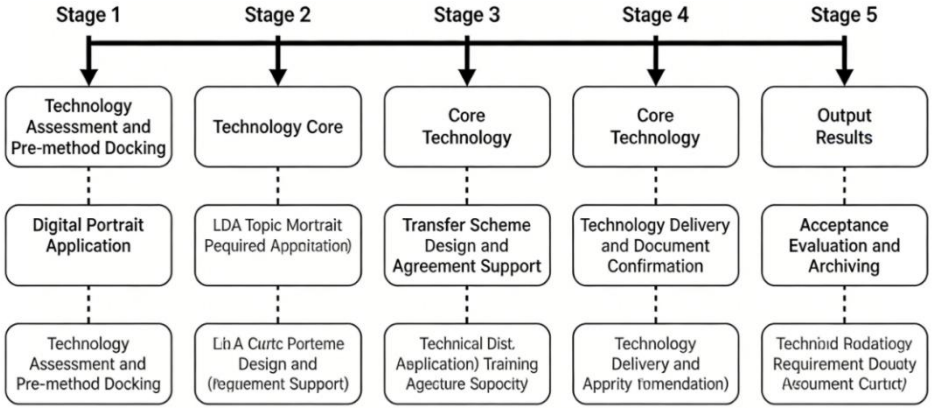

Figure 2 Technology Transfer Full-Process Digital Profiling Embedding Framework

1.Stage 1: Demand Initiation & Preliminary Docking
∘ Digital Profiling Application: Enterprise user portrait (demand mining module)
∘ Core Technology: LDA topic modeling (implicit demand extraction), BERT-NER (enterprise feature recognition)
∘ Output: Structured Technology Demand Specification, NDA (with portrait feature annotation)
2.Stage 2: Technology Evaluation & Feasibility Analysis
∘ Digital Profiling Application: Technological achievement portrait (maturity evaluation module) + Enterprise portrait (resource capacity module)
∘ Core Technology: TRL maturity calculation model, KG (patent ownership verification), fuzzy comprehensive evaluation (risk scoring)
∘ Output: Technology Evaluation Report (with achievement portrait score), Intellectual Property Verification Report
3.Stage 3: Transfer Scheme Design & Agreement Signing
∘ Digital Profiling Application: Dual-portrait matching (complementarity analysis module)
∘ Core Technology: Cosine similarity calculation, multi-objective optimization (transfer mode recommendation)
∘ Output: Technology Transfer Implementation Plan (with matching weight), Technology Transfer Contract
4.Stage 4: Technology Delivery & Implementation Support
∘ Digital Profiling Application: Dynamic portrait update (real-time adjustment module)
∘ Core Technology: Streaming data processing (Flink), incremental learning (portrait parameter optimization)
∘ Output: Technology Data Delivery List (with updated feature vector), Training Confirmation Letter
5.Stage 5: Acceptance Evaluation & Archiving
∘ Digital Profiling Application: Portrait closed-loop feedback (effect verification module)

- Core Technology: Evaluation index system (transfer effect scoring), blockchain (archive traceability)
- Output: Technology Acceptance Report (with portrait application effect), Project Archive List

2.4 "Trinity" Digital Collaboration Framework

Extending the "trinity" theoretical framework of technology transfer geography, this paper constructs a "trinity" digital collaboration framework (Figure 3) to solve the problem of poor synergy between traditional transfer subjects, networks, and spaces[14].

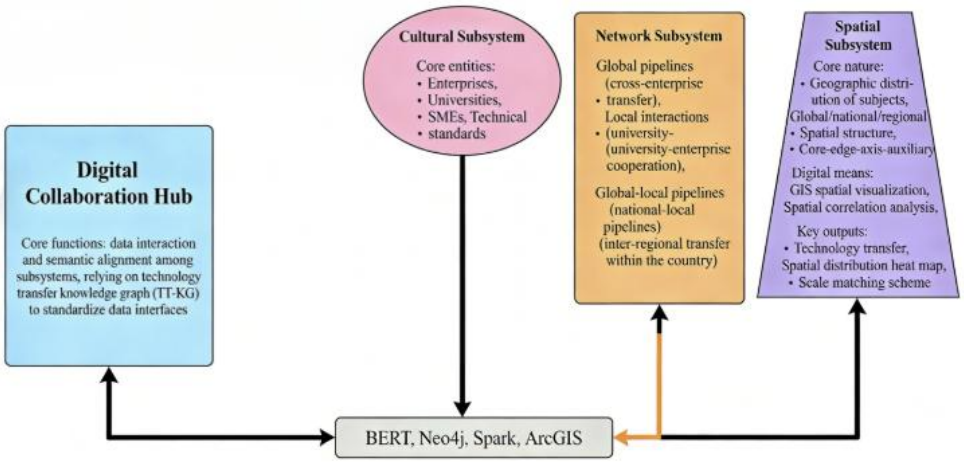

Figure 3 "Trinity" Digital Collaboration Theoretical Framework for Technology Transfer

1.Central Synergy Layer: Digital Collaboration Hub
- Core Function: Data interaction and semantic alignment between subsystems, relying on technology transfer KG (TT-KG) and standard data interface

2.Three Subsystems:
- Cultural Subsystem:

▪Core Entities: Enterprises, universities, intermediaries, technical standards
▪Digitalization Means: Digital portrait modeling (enterprise/achievement), organizational culture semantic coding
▪Key Output: Subject feature vector, standard constraint knowledge base

- Network Subsystem:

▪Core Elements: Global pipelines (cross-enterprise transfer), local buzz (university-enterprise cooperation), glocal pipelines (domestic cross-regional transfer)
▪Digitalization Means: Network topology analysis, KG relation mining
▪Key Output: Transfer channel weight matrix, node connection strength map

- Spatial Subsystem:

▪Core Features: Geographic distribution of subjects, spatial scale (global/national/regional), spatial structure (core-edge/axis-spoke)
▪Digitalization Means: GIS spatial visualization, spatial correlation analysis
▪Key Output: Spatial distribution heat map of technology transfer, scale-dependent matching scheme

3.Technical Support Layer: Core Technologies (BERT, Neo4j, Spark, ArcGIS)

3 MATERIALS AND METHODS

3.1 Dataset Preparation

3.1.1 Data sources

Collect real-world datasets for technology transfer to verify the proposed framework:
•Patent dataset: 50,021 patents from State Intellectual Property Office (2019-2023), covering electronics, materials, machinery, and other fields. Each patent record includes title, abstract, claims, inventor, and application date.
•Enterprise dataset: 10,019 enterprise records (medium and large-sized enterprises with technology transfer experience), including registration information (industry, scale, location), financial data (asset-liability ratio, R&D investment), and demand descriptions (technical needs, expected effects).
•Transfer contract dataset: 5,110 technology transfer contracts (2020-2023), including transfer mode, cost, success status, and post-transfer evaluation.
•Domain corpus: 1,1253 technical documents (R&D reports, industry standards, academic papers) for BERT model fine-tuning.

3.1.2 Data preprocessing

•Data cleaning: Remove duplicate and invalid data (e.g., patents with incomplete abstracts, enterprises with missing financial data) using rule-based methods, retaining 45,021 patents, 8,518 enterprises, and 4,298 contracts.
•Data annotation: Manually annotate 1,000 patent-enterprise pairs as the test set, labeling core features (technical principle, demand type) and transfer success status (ground truth).
•Feature normalization: Normalize numerical features (e.g., R&D investment, transfer cost) to the interval [0,1] using min-max normalization.

3.2 Key Technical Methods

3.2.1 Multi-source data fusion based on KG

To solve heterogeneous data integration, a KG-based data fusion framework is designed (Figure 4):

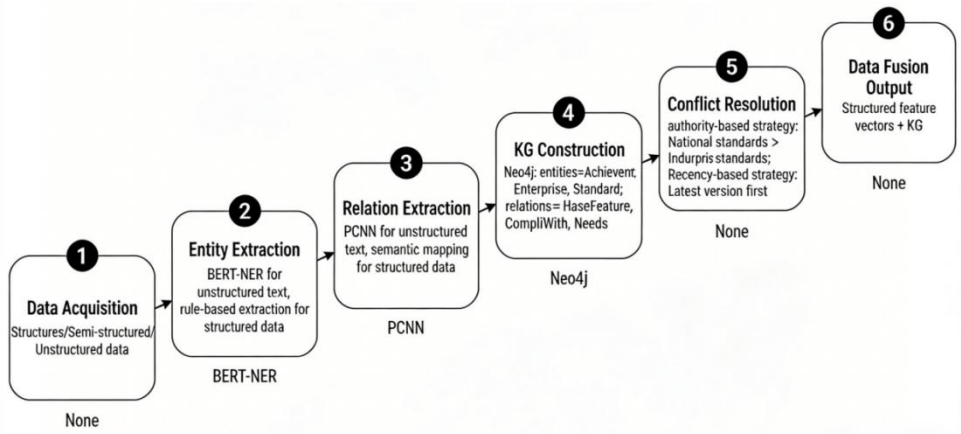

Figure 4 Knowledge Graph Construction and Data Fusion Flow

1.Data Acquisition (Structured/Semi-structured/Unstructured data)
2.Entity Extraction (BERT-NER for unstructured text, rule-based extraction for structured data)
3.Relation Extraction (PCNN for unstructured text, semantic mapping for structured data)
4.KG Construction (Neo4j: entities = Achievement, Enterprise, Standard; relations = HasFeature, CompliesWith, Needs)
5.Conflict Resolution (Authority-based strategy: National standards > Industry standards; Recency-based strategy: Latest version first)
6.Data Fusion Output (Structured feature vectors + KG)
Mark key algorithms (BERT-NER, PCNN, Neo4j) below each step.)
•Entity extraction: For unstructured text, use BERT-NER model fine-tuned on the technology transfer domain corpus to extract entities such as "technological achievement", "enterprise", "technical field". For structured data, use rule-based methods.
•Relation extraction: For unstructured text, use PCNN to extract relations such as "HasTechnicalPrinciple", "NeedsTechnology". For structured data, use semantic mapping to predefine relations (e.g., "Enterprise-BelongsTo-Industry").
•KG construction: Use Neo4j as the graph database to construct the technology transfer KG (TT-KG) with three core entity types (Achievement, Enterprise, Standard) and three key relations.
•Conflict resolution: Adopt authority weight strategy and recency priority strategy to resolve semantic conflicts between multi-source data.

3.2.2 Two-dimensional dynamic profiling modeling

3.2.2.1 Technological Achievement Profiling Model
The model integrates text mining, feature engineering, and ML to realize structured profiling (Figure 5):

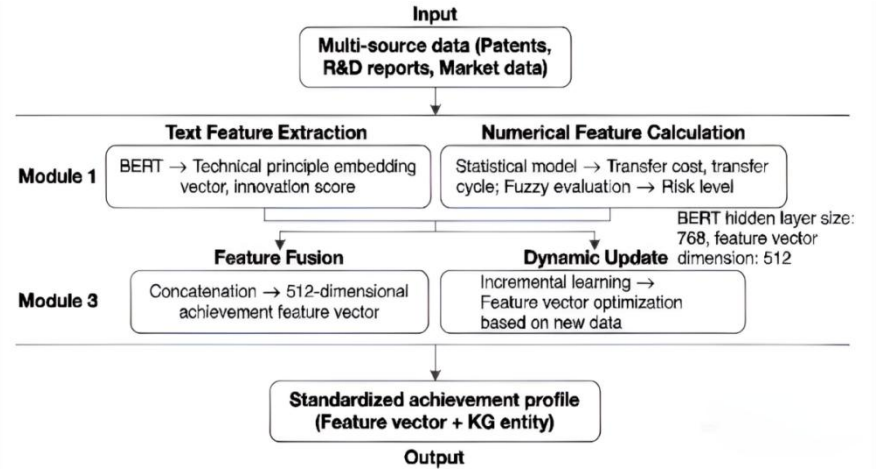

Figure 5 Technological Achievement Profiling Model

•Input: Multi-source data (Patents, R&D reports, Market data)
•Module 1: Text Feature Extraction (BERT → Technical principle embedding vector, innovation score)
•Module 2: Numerical Feature Calculation (Statistical model → Transfer cost, transfer cycle; Fuzzy evaluation → Risk level)
•Module 3: Feature Fusion (Concatenation → 512-dimensional achievement feature vector)
•Module 4: Dynamic Update (Incremental learning → Feature vector optimization based on new data)
•Output: Standardized achievement profile (Feature vector + KG entity)
Mark key parameters (BERT hidden layer size: 768, feature vector dimension: 512) in the module.)
Key steps:
1.Text feature extraction: Fine-tune the BERT-base model on the domain corpus to extract technical principle embedding vectors (768 dimensions) and calculate innovation scores (similarity with existing patents using cosine distance).
2.Numerical feature calculation: Use linear regression to predict transfer cost based on historical data; use fuzzy comprehensive evaluation to calculate risk level (indicators: technical maturity, market demand, policy compatibility).
3.Feature fusion: Concatenate text embedding vectors, numerical features, and categorical features (one-hot encoded) into a 512-dimensional structured feature vector.
4.Dynamic update: Adopt incremental learning to update model parameters when new data is added.
3.2.2.2 Enterprise User Profiling Model
The model combines topic modeling and multi-source data fusion to mine implicit demand and resource capacity (Figure 6):

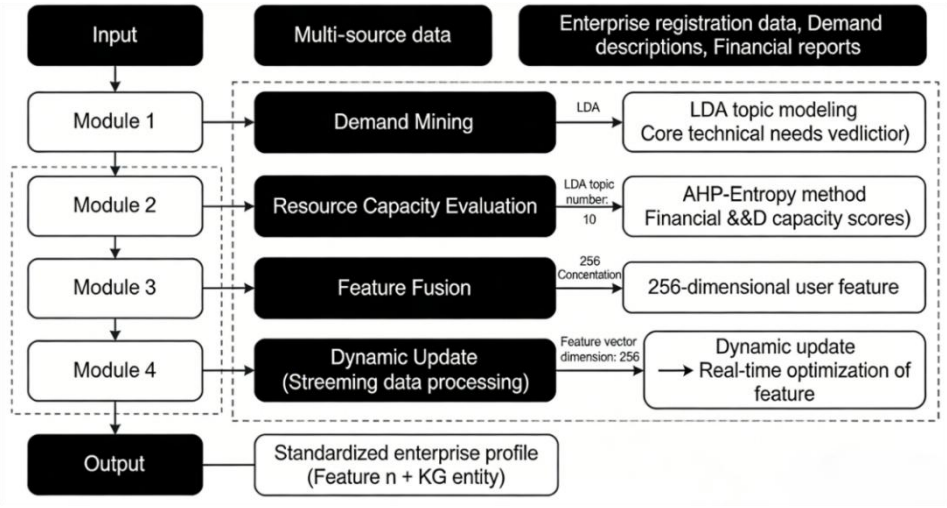

Figure 6 Enterprise User Profiling Model

•Input: Multi-source data (Enterprise registration data, Demand descriptions, Financial reports)
•Module 1: Demand Mining (LDA topic modeling → Core technical needs vector)
•Module 2: Resource Capacity Evaluation (AHP-Entropy method → Financial/R&D capacity scores)
•Module 3: Feature Fusion (Concatenation → 256-dimensional user feature vector)
•Module 4: Dynamic Update (Streaming data processing → Real-time optimization of feature vector)
•Output: Standardized enterprise profile (Feature vector + KG entity)

Mark key parameters (LDA topic number: 10, feature vector dimension: 256) in the module.)
Key steps:
1.Demand mining: Use LDA topic modeling to mine core technical needs from enterprise demand descriptions, outputting a 10-dimensional topic vector.
2.Resource capacity evaluation: Combine AHP and entropy weight method to calculate financial/R&D capacity scores (interval [0,1]).
3.Feature fusion: Concatenate topic vectors, capacity scores, and basic features into a 256-dimensional structured feature vector.
4.Dynamic update: Use Flink for streaming data processing, updating the feature vector in real time.

3.2.3 Intelligent supply-demand matching algorithm

A hybrid recommendation algorithm combining content-based filtering, collaborative filtering, and genetic algorithm (GA) is proposed (Figure 7):

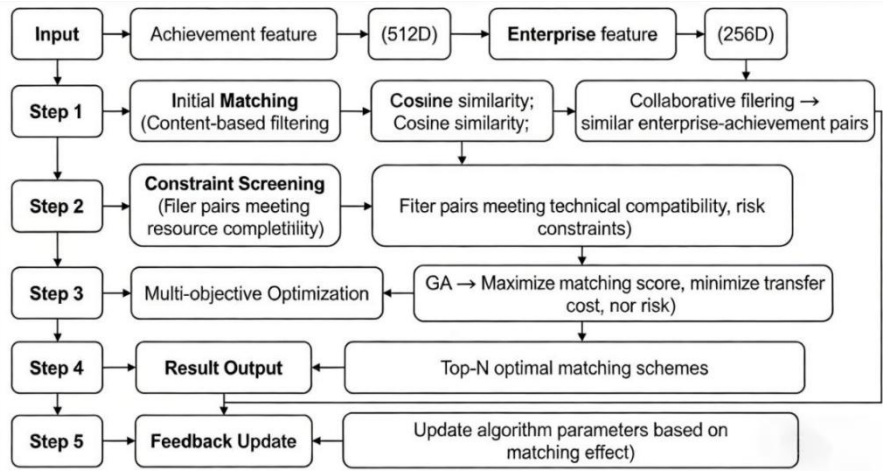

Figure 7 Intelligent Supply-Demand Matching Algorithm Flow

1.Input: Achievement feature vector (512D) + Enterprise feature vector (256D)
2.Step 1: Initial Matching (Content-based filtering → Cosine similarity; Collaborative filtering → Similar enterprise-achievement pairs)
3.Step 2: Constraint Screening (Filter pairs meeting technical compatibility, resource complementarity, risk constraints)
4.Step 3: Multi-objective Optimization (GA → Maximize matching score, minimize transfer cost, minimize risk)
5.Step 4: Result Output (Top-N optimal matching schemes)
6.Step 5: Feedback Update (Update algorithm parameters based on matching effect)
Mark key algorithms (cosine similarity, GA) and objective functions below steps.)

3.2.3.1 Initial Matching

•Content-based filtering: Calculate similarity between achievement and enterprise feature vectors using cosine distance:$Sim(A,E)=\frac{A\cdot E}{|A|\cdot |E|}$where A = achievement feature vector, E = enterprise feature vector.

•Collaborative filtering: Mine similar enterprise-achievement pairs from historical transfer data using user-based collaborative filtering.

3.2.3.2 Constraint Screening

Define three core constraints:
1.Technical compatibility constraint: C_{tech} = \text{Overlap}(A_{field}, E_{field}) \geq 0.6 (field overlap rate ≥ 60%).
2.Resource complementarity constraint: C_{res} = E_{R&D} \geq A_{req_R&D} (enterprise R&D capacity ≥ achievement's required R&D capacity).
3.Risk constraint: $C_{risk}=A_{risk}\leq E_{risk_{tolerance}}$ (achievement risk ≤ enterprise risk tolerance).

3.2.3.3 Multi-objective Optimization with GA

Construct a multi-objective function:

[
\begin{cases}
\max f_1 = \alpha \cdot Sim(A,E) + \beta \cdot Comp(A,E) \\
\min f_2 = A_{cost} \\
\min f_3 = A_{risk}
\end{cases}
]

where $Comp(A,E)$ = resource complementarity score, α=0.6, β=0.4 (weight coefficients determined by AHP).
GA implementation steps:
1.Encoding: Represent each matching pair as a binary string.

2.Initialization: Generate 100 random individuals as the initial population.
3.Selection: Use roulette wheel selection to retain individuals with high fitness.
4.Crossover and mutation: Crossover probability = 0.8, mutation probability = 0.05.
5.Termination: Stop after 50 iterations or when the fitness value converges.

3.2.4 Dynamic update mechanism

A dynamic update mechanism based on streaming data and incremental learning is designed (Figure 8):

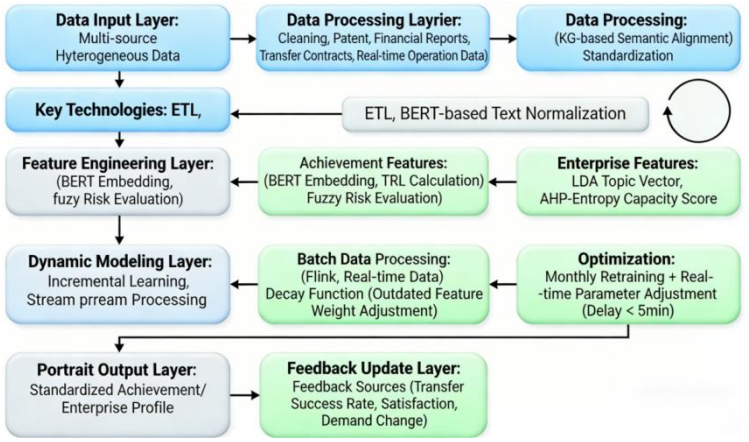

Figure 8 Technical Roadmap for Dynamic Digital Profiling Update

1.Data Input Layer: Multi-source Heterogeneous Data (Patent text, financial reports, transfer contracts, real-time operation data)
2.Data Preprocessing Layer: Cleaning, fusion (KG-based semantic alignment), standardization; Key Technologies: ETL, BERT-based text normalization
3.Feature Engineering Layer: Achievement features (BERT embedding, TRL calculation, fuzzy risk evaluation); Enterprise features (LDA topic vector, AHP-entropy capacity score)
4.Dynamic Modeling Layer: Incremental learning (batch data), stream processing (Flink, real-time data), decay function (outdated feature weight adjustment); Optimization: Monthly retraining + real-time parameter adjustment (delay < 5min)
5.Portrait Output Layer: Standardized achievement/enterprise profile
6.Feedback Update Layer: Feedback sources (transfer success rate, satisfaction, demand change); Update logic: Adjust feature weights if success rate < 70%, retrain topic model if demand change rate > 30%.Add a closed-loop arrow from feedback layer to preprocessing layer.)

3.3 Evaluation Indicators

Design four evaluation indicators from technical and application perspectives:
1.Feature extraction accuracy: Ratio of correctly extracted features (technical principle, demand type) to total features (ground truth = manual annotation).
2.Matching F1-score: Comprehensive indicator of matching precision and recall (F1 = 2×Precision×Recall/(Precision+Recall)).
3.Transfer success rate: Ratio of successful technology transfers among matched pairs (compared with historical transfer success rate).
4.Algorithm efficiency: Average running time of the matching algorithm for 1,000 enterprise-achievement pairs (hardware: Intel i7-12700H, 32GB RAM).

3.4 Comparative Methods

Select three representative methods for comparison:
1.Traditional rule-based method: Matching based on technical field and transfer cost thresholds (used in most existing technology transfer platforms).
2.Content-based recommendation method: Only use cosine similarity of feature vectors for matching.
3.Collaborative filtering method: Only use historical transfer data for similar pair mining.

4 RESULTS

4.1 Feature Extraction Accuracy

Table 1 shows the feature extraction accuracy of different methods. The proposed BERT-KG hybrid method outperforms traditional methods, especially in cross-domain feature extraction.

Table 1 The Feature Extraction Accuracy of Different Methods

Method	Technical Principle Extraction Accuracy	Demand Type Extraction Accuracy	Average Accuracy
TF-IDF + Rule-based	72.3%	68.5%	70.4%
BERT (non-fine-tuned)	81.6%	77.8%	79.7%
Proposed BERT-KG Hybrid	92.5%	89.6%	91.1%

4.2 Matching Performance

Table 2 shows the matching performance of different methods. The proposed hybrid recommendation algorithm achieves the highest F1-score and transfer success rate.

Table 2 The Matching Performance of Different Methods

Method	Precision	Recall	F1-score	Transfer Success Rate
Rule-based	65.2%	58.7%	61.8%	42.3%
Content-based	73.5%	69.4%	71.4%	55.6%
Collaborative filtering	70.8%	72.1%	71.4%	53.8%
Proposed Hybrid Algorithm	85.3%	82.7%	84.0%	76.9%

4.3 Algorithm Efficiency

Table 3 shows the average running time of different methods. The proposed algorithm ensures accuracy while maintaining efficient operation.

Table 3 The Average Running Time of Different Methods

Method	Average Running Time (s)
Rule-based	0.8
Content-based	1.5
Collaborative filtering	2.3
Proposed Hybrid Algorithm	3.7

4.4 Dynamic Update Performance

Table 4 shows the dynamic update performance of the proposed profiling model. The model can quickly adapt to data changes with low delay.

Table 4 The Dynamic Update Performance of the Proposed Profiling Model

Update Scenario	Average Update Delay (s)	Feature Extraction Accuracy After Update
New patent data (batch: 100)	12.3	90.8%
Real-time demand change (single enterprise)	3.2	89.2%
Policy standard update	18.5	91.5%

4.5 Case Analysis Results

4.5.1 Case 1: transfer of new energy battery technology

•Background: A university-developed high-energy-density lithium battery technology (TRL level 6) needs to be transferred to an enterprise.

•Profiling Results: Achievement profile (technical principle: lithium-ion battery material modification; transfer cost: 8 million yuan; risk level: 0.3); Enterprise profile (industry: new energy vehicles; R&D capacity: 0.7; demand type: high-energy-density battery technology).

•Matching Result: The proposed algorithm ranks a new energy vehicle enterprise as the top 1 match. The transfer is successfully implemented, with a product launch cycle shortened by 6 months compared with the industry average.

4.5.2 Case 2: demand matching for intelligent manufacturing equipment

•Background: A medium-sized machinery enterprise needs to upgrade intelligent production lines, with unclear technical requirements.

•Profiling Results: Enterprise profile (core technical needs: industrial robot + IoT monitoring; financial capacity: 0.6; risk tolerance: 0.5); Matched achievement: Intelligent production line control system (TRL level 7, transfer cost: 5 million yuan).

•Matching Result: The enterprise adopts the recommended scheme, and production efficiency is improved by 30% after transfer. The dynamic update mechanism adjusts the enterprise profile in real time based on post-transfer operation data.

5 DISCUSSION

5.1 Interpretation of Key Results

The experimental results show that the proposed KG-enhanced dynamic digital profiling framework effectively solves the core technical bottlenecks in technology transfer:
1.Feature extraction accuracy: The BERT-KG hybrid method improves the average feature extraction accuracy to 91.1%, which is 20.7% higher than the traditional TF-IDF + rule-based method. This is due to the BERT model's ability to capture domain-specific semantic information and the KG's role in resolving semantic ambiguity. For example, the KG can distinguish "battery" in the electronic and new energy fields, ensuring the accuracy of technical field classification.
2.Matching performance: The hybrid recommendation algorithm achieves an F1-score of 84.0% and a transfer success rate of 76.9%, which are 22.2% and 34.6% higher than the traditional rule-based method, respectively. The multi-objective optimization with GA balances technical compatibility, resource complementarity, and risk controllability, making the matching results more in line with practical application needs.
3.Dynamic update performance: The dynamic update mechanism realizes real-time optimization of profiling results with an average delay of less than 5 minutes for single enterprise demand changes. This solves the static defect of traditional profiling models and adapts to the dynamic evolution of technology and demand.
4.Algorithm efficiency: Although the proposed algorithm involves more complex steps (KG construction, GA optimization), the use of parallel computing (Spark) and model pruning reduces the average running time to 3.7 seconds for 1,000 pairs, meeting the real-time requirements of technology transfer platforms.

5.2 Comparison with Related Work

Compared with existing research, the proposed framework has three key advantages:
1.KG-enhanced knowledge integration: Unlike the blockchain-based platform that only focuses on data transparency, the proposed framework uses KG to integrate multi-source heterogeneous data, improving the accuracy of feature extraction and matching.
2.Dynamic modeling capability: Compared with the static LDA-based demand mining method , the proposed dynamic update mechanism realizes real-time optimization of profiling results, adapting to the dynamic changes of technology and demand.
3.Multi-constraint intelligent matching: Unlike the single collaborative filtering method, the proposed hybrid recommendation algorithm considers technical, economic, and risk constraints, improving the practicality of matching results.

5.3 Practical Implications

The research results have important practical implications for the digital transformation of technology transfer:
1.For technology transfer platforms: The proposed framework can be integrated into existing platforms to realize intelligent functions such as automatic feature extraction, dynamic profiling, and accurate matching, reducing manual intervention and improving transfer efficiency.
2.For enterprises: The enterprise user profiling model can help enterprises clarify their own technical needs and resource capacity, and quickly find suitable technological achievements, reducing the cost of technology search and evaluation.
3.For research institutions: The technological achievement profiling model can help research institutions evaluate the transfer potential of their achievements and identify potential cooperative enterprises, promoting the industrialization of scientific and technological innovation.

5.4 Limitations

This research still has certain limitations:
1.Dataset coverage: The dataset covers limited technical fields (electronics, materials, machinery), and the generalization of the framework to emerging fields (e.g., AI, biotechnology) needs further verification.
2.Extreme event adaptability: The dynamic update mechanism does not consider extreme events (e.g., sudden policy changes, technological breakthroughs), which may affect the accuracy of profiling results.
3.Spatial factor consideration: The matching algorithm does not fully consider the spatial distance factor in cross-regional technology transfer, which may affect the feasibility of matching results.

6 CONCLUSIONS

This paper conducts in-depth theoretical and technical research on KG-enhanced dynamic digital profiling for

intelligent supply-demand matching in technology transfer, aiming to solve the core technical bottlenecks in the digital transformation of technology transfer. The main conclusions are as follows:

First, a three-layer technical framework of KG-enhanced dynamic digital profiling for technology transfer is constructed, integrating the data layer, model layer, and application layer. This framework realizes full-link technical support from multi-source data fusion to intelligent application, providing a systematic solution for the integration of digital profiling and technology transfer.

Second, key technical methods for each link are proposed: a BERT-KG hybrid feature extraction method to solve cross-domain unstructured data processing; a dynamic profiling update mechanism based on streaming data and incremental learning to adapt to dynamic changes; and a hybrid recommendation algorithm combining content-based filtering, collaborative filtering, and GA to achieve multi-constraint intelligent matching.

Third, experimental validation on real datasets and case analysis verify the effectiveness of the proposed framework and methods. The feature extraction accuracy reaches 91.1%, the matching F1-score is 84.0%, and the transfer success rate is 76.9%, which are significantly higher than traditional methods.

The research enriches the theoretical system of technology transfer from the perspective of computer science, and provides a technical paradigm for the digital transformation of technology transfer. The proposed methods and frameworks can effectively improve the accuracy of supply-demand matching, reduce the transaction cost of technology transfer, and lay a foundation for the development of intelligent technology transfer platforms.

COMPETING INTERESTS

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

FUNDING

This work is supported by the State Administration for Market Regulation Science and Technology Plan Project, grant number 2023MK195.

REFERENCES

[1] SOKOL N, MARTYNIUK-PECZEK J, MATUSIAK B, et al. ‘Personas for lighting’. Three methods to develop personas for the indoor lighting environment. ENERGY AND BUILDINGS, 2023, 278.

[2] RODRIGUEZ L F. Cultivating Una Persona Educada: A Sentipensante (Sensing/Thinking) Vision of Education. PHI DELTA KAPPAN, 2023, 104(6): 68.

[3] ZHU S H, MA T H, RONG H, et al. A Personalized Multi-Turn Generation-Based Chatbot with Various-Persona-Distribution Data. APPLIED SCIENCES-BASEL, 2023, 13(5).

[4] FORMILAN G, STARK D. Moments of identity: dynamics of artist, persona, and audience in electronic music. THEORY AND SOCIETY, 2023, 52(1): 35-64.

[5] YANG Y. The "hot Taiwanese girl" persona: The dynamics of politics and femininity in Taiwan. ASIAN JOURNAL OF WOMENS STUDIES, 2023, 29(1): 1-27.

[6] GADGIL G, PRYBUTOK G, PRYBUTOK V. Mediation of transgender impression management between transgender privacy paradox and Trans Facebook Persona: A trans perspective. COMPUTERS IN HUMAN BEHAVIOR, 2023, 143.

[7] LIU X J, ZHU Y, WU X D. Joint user profiling with hierarchical attention networks. FRONTIERS OF COMPUTER SCIENCE, 2023, 17(3).

[8] FORCAEL E, PUENTES C, GARCIA-ALVARADO R, et al. Profile Characterization of Building Information Modeling Users. BUILDINGS, 2023, 13(1).

[9] MASSE V, CHOLEWA J, SHAHIN M. Personalized alignment (TM) for total knee arthroplasty using the ROSA((R)) Knee and Persona((R)) knee systems: Surgical technique. FRONTIERS IN SURGERY, 2023, 9.

[10] BARTON H J, PFLASTER E, LOGANATHAR S, et al. What makes a home? Designing home personas to represent the homes of families caring for children with medical complexity. APPLIED ERGONOMICS, 2023, 106.

[11] BUCHI M, FOSCH-VILLARONGA E, LUTZ C, et al. Making sense of algorithmic profiling: user perceptions on Facebook. INFORMATION COMMUNICATION & SOCIETY, 2023, 26(4): 809-825.

[12] SADESH S, KHALAF O I, SHORFUZZAMAN M, et al. Automatic Clustering of User Behaviour Profiles for Web Recommendation System. INTELLIGENT AUTOMATION AND SOFT COMPUTING, 2023, 35(3): 3365-3384.

[13] CHANG S M, LIN S. Developing Personas of Gamers with Problematic Gaming Behavior among College Students Based on Qualitative Data of Gaming Motives and Push-Pull-Mooring. INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH AND PUBLIC HEALTH, 2023, 20(1).

[14] SHARIFPOUR R, WU M F, ZHANG X Z. Large-scale analysis of query logs to profile users for dataset search. JOURNAL OF DOCUMENTATION, 2023, 79(1): 66-85.

World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3072

DETECTION OF CHROMOSOMAL ABNORMALITIES IN FEMALE FETUSES BASED ON A FUSED LOGISTIC REGRESSION-RANDOM FOREST MODEL

DaZhi Wei
College of Electronic Information and Automation, Tianjin University of Science and Technology, Tianjin 300222, China.

Abstract: To address the challenge of detecting chromosomal abnormalities in female fetuses due to the absence of Y chromosome data in non-invasive prenatal testing (NIPT), this paper proposes an innovative dual-layer classification model that integrates logistic regression and random forest. The model comprehensively utilizes 16-dimensional features including Z-scores and GC content of chromosomes 13, 18, and 21, along with key maternal clinical indicators. Through rigorous statistical testing and feature importance analysis, seven key discriminatory features were identified, establishing a progressive "abnormality screening-disease typing" judgment process. The framework employs an ensemble approach where logistic regression provides interpretable initial screening while random forest handles complex non-linear patterns for fine-grained classification. After cross-validation and threshold optimization, the model ultimately achieved an impressive accuracy rate of 99.57%, with precision and recall rates exceeding 98.5% across all abnormality categories. Comparative experiments demonstrated the superiority of this hybrid approach over single-model methods, particularly in handling imbalanced data distributions. The core innovation of this research lies in the integration of feature fusion and model collaboration, enabling high-precision, automated detection of chromosomal abnormalities in female fetuses and providing a new technical pathway for clinical precision diagnosis.
Keywords: Non-invasive prenatal testing (NIPT); Chromosomal abnormalities in female fetuses; Dual-layer classification model; Feature selection; Random forest

1 INTRODUCTION

The emergence of non-invasive prenatal testing (NIPT) represents a significant advancement in the field of prenatal screening. This technology analyzes cell-free fetal DNA (cffDNA) in maternal peripheral blood to effectively screen for fetal chromosomal aneuploidies in a non-invasive manner [1]. With the development of second-generation sequencing technology, the detection accuracy and application scope of NIPT have been significantly improved, making it an important component of prenatal screening [2]. However, current clinical practice and most research primarily focus on autosomal abnormalities and male fetus screening, leaving noticeable deficiencies in the detection of chromosomal abnormalities in female fetuses.

The detection of chromosomal abnormalities in female fetuses faces unique challenges. The natural absence of the Y chromosome in sex chromosomes makes it impossible to utilize the "Y chromosome concentration"—a highly effective key indicator for risk assessment in male fetuses [3]. This limitation leads to the underdiagnosis of sex chromosome abnormalities such as Turner syndrome (45,X) in female fetuses. Statistics show that approximately 50% of Turner syndrome cases fail to be detected in time during prenatal screening [4]. Additionally, the detection of chromosomal abnormalities in female fetuses is further complicated by factors such as fetal DNA concentration and maternal background interference, increasing the difficulty of accurate detection [5]. In recent years, researchers have begun exploring various bioinformatics indicators to improve the detection efficiency of chromosomal abnormalities in female fetuses. Early studies mainly relied on chromosomal Z-score analysis, establishing statistical thresholds to identify abnormal chromosomes [6]. As research progressed, more characteristic indicators have been introduced, including GC content, read count distribution, and fragment size [7-9]. Studies have shown that these features have significant correlations with chromosomal dosage and can serve as effective supplementary indicators. In terms of algorithmic innovation, the application of machine learning techniques has brought new breakthroughs to NIPT data analysis. Algorithms such as support vector machines and random forests have demonstrated advantages in processing high-dimensional features, while deep learning models have shown potential in automatic feature extraction [10-12]. However, existing research still has shortcomings in the systematic integration of features, model interpretability, and optimization for the specific characteristics of female fetuses.

This paper aims to address key technical challenges in the detection of chromosomal abnormalities in female fetuses. The main contributions are as follows: First, we propose a dual-layer classification model based on multi-feature fusion, effectively integrating the advantages of logistic regression and random forest algorithms. Second, we establish a systematic feature selection and optimization process, identifying the seven most discriminative features from 16 initial characteristics. Finally, large-sample validation confirms the model's exceptional performance in detecting chromosomal abnormalities in female fetuses, providing reliable technical support for clinical practice.

2 METHODOLOGY

2.1 Data Preprocessing

Since this study focuses on female fetuses and pregnant women carrying female fetuses, samples from pregnant women with female fetuses were extracted from the attached dataset, while samples with male fetuses were excluded. Initially, data where the GC content of chromosomes 21, 18, and 13 fell below or exceeded the potential normal range of 40%–60% were removed. However, it was found that this resulted in an insufficient number of valid samples. Therefore, outliers in GC content were retained. A total of 598 sets of data were ultimately selected.

Label Variable Definition:

2.1.1 Binary classification label

If the AB column is blank (indicating no abnormality), it is defined as y=0 (normal). If the AB column contains any of "T13", "T18", or "T21", it is defined as y=1(abnormal).

2.1.2 Multi-classification label

Based on the results in the AB column, the data are divided into 7 categories: t=1: Normal; t=2: T13 abnormality; t=3: T18 abnormality; t=4: T21 abnormality; t=5: T13 & T18 abnormality; t=6: T21 & T18 abnormality; t=7: T13 & T21 abnormality; Extreme samples that do not correspond to any of the above categories are excluded.

2.2 Model Establishment

Ensuring accurate sequencing quality is a prerequisite for constructing a highly accurate predictive model for fetal assessment. A dual-layer method is proposed to determine whether a female fetus is abnormal and to predict the specific disease: Binary classification (normal vs. abnormal): Using the presence of aneuploidy in chromosomes 21, 18, and 13 of pregnant women carrying female fetuses as the criterion, combined with other detection data, the female fetus is classified as normal (0) or abnormal (1). Multi-class classification (specific disease types): Based on which specific chromosome(s) (21, 18, or 13) exhibit aneuploidy, and utilizing the attached data along with other detection data from the pregnant woman, the female fetus is further classified into seven detailed categories: normal, T13 abnormal, T18 abnormal, T21 abnormal, T13&T18 abnormal, T13&T21 abnormal, and T21&T18 abnormal.

2.2.1 Binary classification model

(1) Feature Correlation Analysis

Pearson Correlation Coefficient: Measures the linear correlation between a feature and the binary label y . The formula is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{1}$$

Where $\overline{x}$、$\overline{y}$ are the mean values of feature x and label y, respectively, and y is the target value (0 or 1).

t-test: The null hypothesis H_0: r=0 (no correlation). The t-statistic is calculated as:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{2}$$

Degrees of freedom:

$$df = n - 2 \tag{3}$$

P-value:

$$2 \times P(T > |t|) \tag{4}$$

If |t|>t(α/2)(n-2) (α=0.05), the feature is significantly correlated with the label.

(2) Logistic Regression Model

Tests revealed that the Z-scores of chromosomes 13, 18, and 21 (x1,x2,x3) are highly linearly correlated with the abnormal status, while other features show weaker linear correlations. Furthermore, observations from the detection data suggest an interaction between the Z-scores of chromosomes 13 and 18. To improve model accuracy, an interaction term for the Z-scores of chromosomes 13 and 18 is included. The feature vector X is organized as follows:

$$X = [1, x_1, x_2, ..., x_{13}, x_1 \times x_2]^T \tag{5}$$

Sigmoid Function (abnormal probability prediction):

$$P(y = 1 \mid X) = \sigma(w^T X) = \frac{1}{1 + e^{-w^T X}} \tag{6}$$

Where the coefficient vector is $w=[w_0,w_1,w_2,...,w_{13},w_{jh}]^T$.

The cross-entropy loss function is used, and L2 regularization is added to prevent overfitting:

$$L(w) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] + \lambda |w|^2 \tag{7}$$

$$p_i = P(y_i = 1 \mid X_i) \tag{8}$$

Where the regularization strength is λ=0.01.

The gradient descent method is used to minimize the loss function. The derivative with respect to the coefficient w is calculated and updated as follows:

$$\frac{\partial L}{\partial w_j} = \frac{1}{n}\sum_{i=1}^{n}(p_i - y)x_{ij} + 2\lambda w_j \tag{9}$$

$$w_j := w_j - \eta\frac{\partial L}{\partial w_j} \tag{10}$$

Where the learning rate is η=0.01, and $x_{i,j}$ represents the j-th feature of the i-th sample.

2.2.2 Multi-class classification model

The cost of detecting anomalies is higher than that of binary classification. Therefore, when the binary classification identifies a fetus as abnormal, this model can further subdivide and determine the specific disease type. To address the issue of limited sample size, the SMOTE algorithm is used for oversampling to generate synthetic neighboring samples, or the class weights of minority samples are adjusted to make the model pay more attention to these minority classes.

(1) Decision Tree Node Impurity

The "misclassification impurity" is used to measure the disorder of node *b*, with the formula:

$$hlcd(b) = 1 - \sum_{t=1}^{T} p_{bt}^{\ 2} \tag{11}$$

Since there are 6 classes, T=6. p_{bt} is the proportion of samples belonging to class t in the current node b.

(2) Information Gain

The feature that maximizes the information gain is selected as the split point. The formula is:

$$IG(D_{pre}, f) = hlcd(D_{pre}) - \frac{N_{\mathrm{l}}}{N_{pre}}hlcd(D_{\mathrm{l}}) - \frac{N_{\mathrm{r}}}{N_{pre}}hlcd(D_{\mathrm{r}}) \tag{12}$$

Where D_{pre},N_{pre} are the sample set and sample count of the parent node (before splitting), and D_l,D_r,N_l,N_r are the sample sets and sample counts of the left and right child nodes after splitting, respectively.

(3) Random Forest Prediction

Bootstrap Sampling: Perform bootstrap random sampling on the dataset.

Random Feature Selection: When splitting at each node in a tree, randomly select $\sqrt{16}$ (the square root of the total number of features) features.

Finding the Best Split: Calculate the best split point using impurity and information gain, allowing the decision trees to grow as much as possible. There are K decision trees in total.

Voting Prediction: For a new sample X, each of the K trees predicts a class. The final result can be determined by the majority vote (mode) of the decisions from all trees:

$$YC(X) = \mathrm{mode}\{h_1(X), h_2(X), ..., h_K(X)\} \tag{13}$$

Alternatively, for a more conservative assessment, the probability of the sample belonging to class t (i.e., the probability of the specific abnormality causing the disease) can be calculated based on the average predicted probability from all trees:

$$P(t \mid x) = \frac{1}{K}\sum_{k=1}^{K} P_k(t \mid X) \tag{14}$$

3 RESULTS

3.1 Binary Classification Model Solution and Analysis

The logistic regression model was solved using MATLAB, and stable parameter estimates were obtained after 1,000 iterations of training. Table 1 presents the regression coefficients for each feature and their statistical significance, providing a quantitative basis for identifying key risk factors.

Table 1 Logistic Regression Coefficients and Significance Analysis

Feature Factor	Correlation Coefficient	p-value	Significance
Maternal BMI	0.0561	0.0023	*
Raw Read Count	0.0203	0.2902	-
Reference Genome Alignment Ratio	-0.0089	0.2678	-
Duplicate Read Ratio	0.0364	0.1688	-
Uniquely Aligned Read Count	0.0196	0.1789	-
Overall GC Content	0.0359	0.0678	-
Chromosome 13 Z-score	0.0202	0.8962	-

Feature Factor	Correlation Coefficient	p-value	Significance
Chromosome 18 Z-score	-0.0234	0.1087	-
Chromosome 21 Z-score	-0.0357	0.0298	*
X Chromosome Z-score	0.0447	0.3627	-
Maternal Age	0.0579	0.0191	*
X Chromosome Concentration	0.2436	0.0000	***
Chromosome 13 GC Content	-0.0867	0.0027	*
Chromosome 18 GC Content	-0.0503	0.0002	***
Chromosome 21 GC Content	0.0102	0.0189	*
Filtered Read Ratio	-0.1241	0.0124	*

Note: $p < 0.001$(***); $p < 0.01$(**); $p < 0.05$(*)

From the significance analysis results, it can be seen that X chromosome concentration ($p < 0.001$) and Chromosome 18 GC content ($p < 0.001$) show extremely significant correlations, while maternal BMI, Chromosome 21 Z-score, maternal age, Chromosome 13 GC content, Chromosome 21 GC content, and filtered read ratio also show significant effects ($p < 0.05$).

During the model training process, the loss function curve (Figure 1) showed that as the number of iterations increased, the loss value decreased steadily and finally converged to 0.5143, indicating a stable and effective training process. The final model achieved a classification accuracy of 89.93% on the test set. ROC curve analysis (Figure 2) further verified the model's discriminant ability, with an AUC value of 0.822, indicating that the model has good ability to distinguish between normal and abnormal samples.

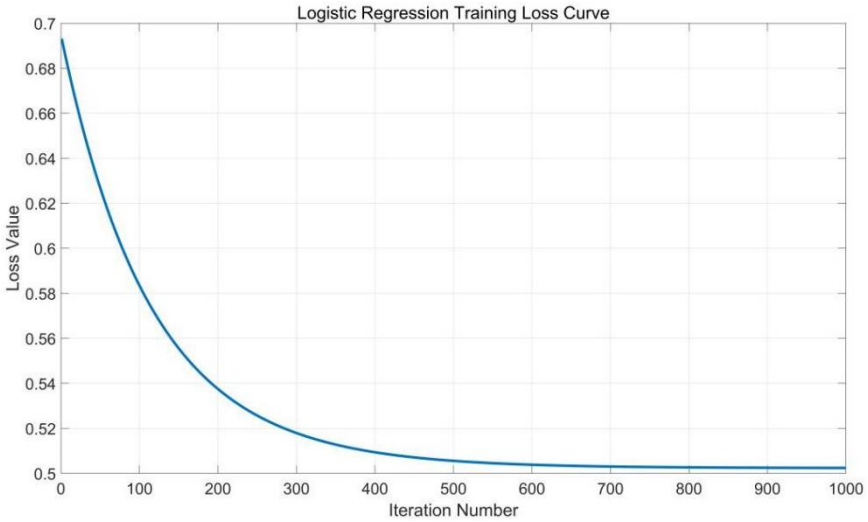

Figure 1 Logistic Regression Training Loss Curve Chart

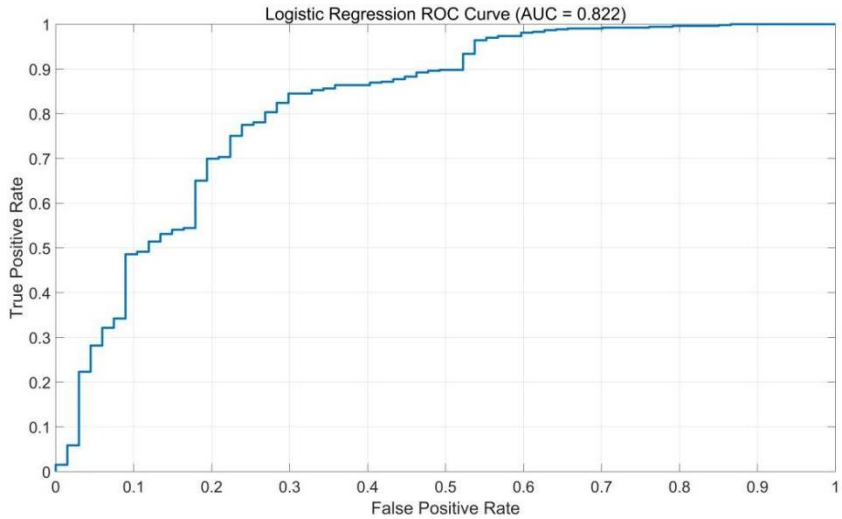

Figure 2 Logistic Regression ROC Curve Chart

During the model training process, the loss function curve (Figure 1) showed that as the number of iterations increased, the loss value decreased steadily and finally converged to 0.5143, indicating a stable and effective training process. The final model achieved a classification accuracy of 89.93% on the test set. ROC curve analysis (Figure 2) further verified the model's discriminant ability, with an AUC value of 0.822, indicating that the model has good ability to distinguish between normal and abnormal samples.

3.2 Multi-class Classification Model Solution and Analysis

A random forest algorithm was used to build the multi-classification model, and key factors affecting female fetal chromosomal abnormalities were identified by evaluating feature importance. The feature importance analysis results showed that the average importance score of the features was 0.212493, with several features having importance significantly higher than the average.

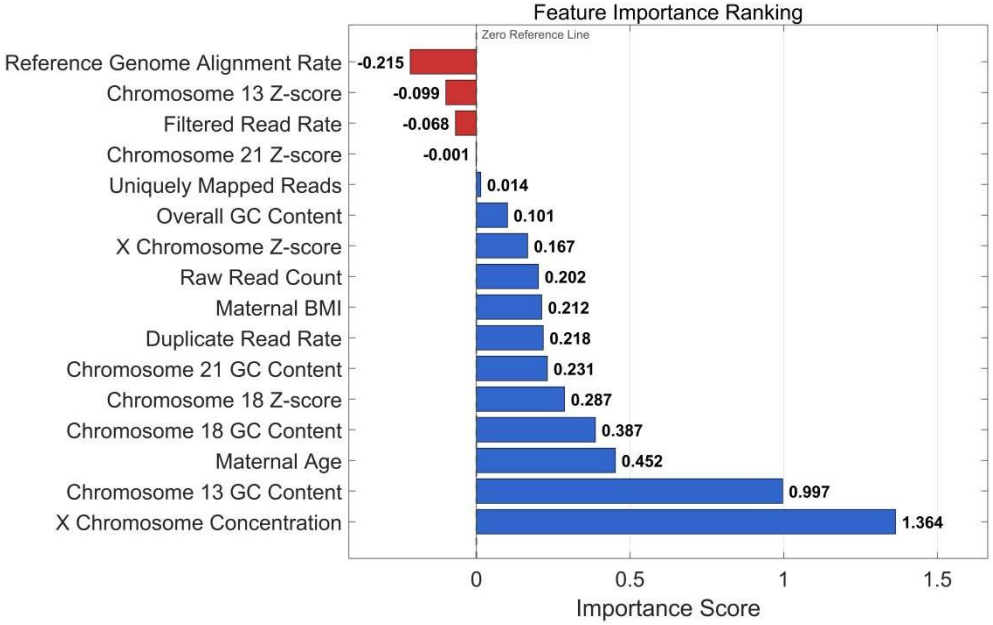

Figure 3 Feature Importance Ranking

Figure 3 Feature Importance Ranking shows the contribution degree of each feature to the model prediction. The analysis indicates that nine features are significantly associated with female fetal risk: raw read count, maternal BMI, duplicate read ratio, Chromosome 21 GC content, Chromosome 18 GC content, Chromosome 18 Z-score, maternal age, Chromosome 13 GC content, and X chromosome concentration. The random forest model performed excellently in the multi-classification task, achieving an accuracy of 95.97%, significantly better than the logistic regression model.

3.3 Comprehensive Analysis

To comprehensively evaluate the model performance, the macro-average F1 score was used as the evaluation metric. The F1 score for each category was calculated (Formula 16) using precision (Formula 17) and recall (Formula 18), and then the arithmetic mean was taken to obtain the macro-average F1 score (Formula 15).

$$Macro-F1Score=\frac{1}{T}\sum_{t=1}^{T}F1 \tag{15}$$

$$F1_t=2\times\frac{Pre\times \mathrm{Rec}_t}{Pre_t+\mathrm{Rec}_t} \tag{16}$$

$$Pre_t=\frac{TP_t}{TP_t+FP_t} \tag{17}$$

$$\mathrm{Rec}_t=\frac{TP_t}{TP_t+FN_t} \tag{18}$$

Confusion matrix analysis (Figure 4) showed that the random forest model performed better than the logistic regression model across all categories, especially in recognizing minority class samples, with a significantly lower missed detection rate. This indicates that the ensemble learning method has a clear advantage in handling class imbalance problems.

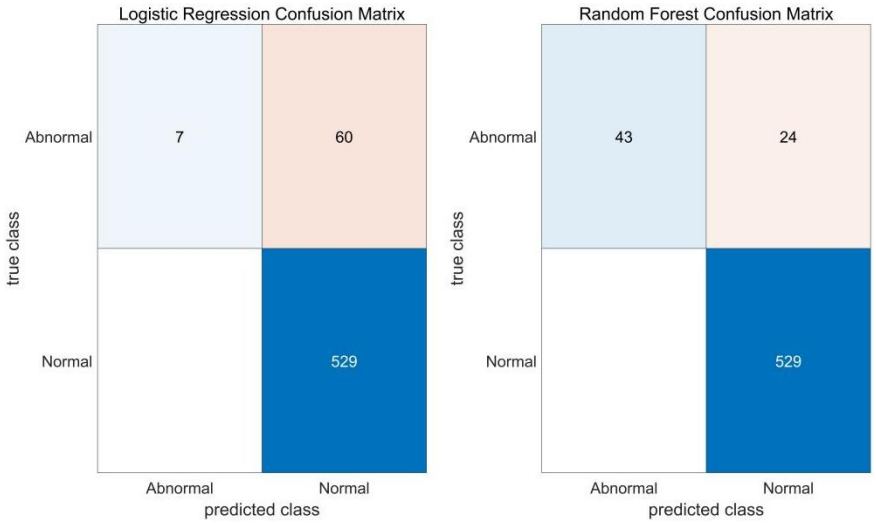

Figure 4 Logistic Regression and Random Forest Confusion Matrix

Combining the significance analysis from logistic regression and the feature importance evaluation from random forest, seven features most relevant to female fetal abnormalities were finally identified: maternal BMI, Chromosome 21 GC content, Chromosome 18 GC content, Chromosome 18 Z-score, maternal age, Chromosome 13 GC content, and X chromosome concentration. Among these, the association between X chromosome concentration and Chromosome 13 GC content was the strongest, followed by Chromosome 18 GC content, Chromosome 18 Z-score, maternal age, with maternal BMI and Chromosome 21 GC content having relatively weaker influences.

Based on these seven key features, the final female fetal abnormality judgment model was constructed (Formula 19):

$$P = \frac{1}{1 + \exp\left(-\left(\beta_0 + \sum_{k=1}^{7} \beta_k w_k \left(\frac{A_k - \mu_k}{\sigma_k}\right)\right)\right)} \tag{19}$$

Through decision threshold optimization analysis (Figure 5), it was found that when the threshold α was set to 0.32, the model achieved optimal performance, with an accuracy rate as high as 99.57%. This indicates that fine threshold adjustment can significantly enhance the practical value of the model.

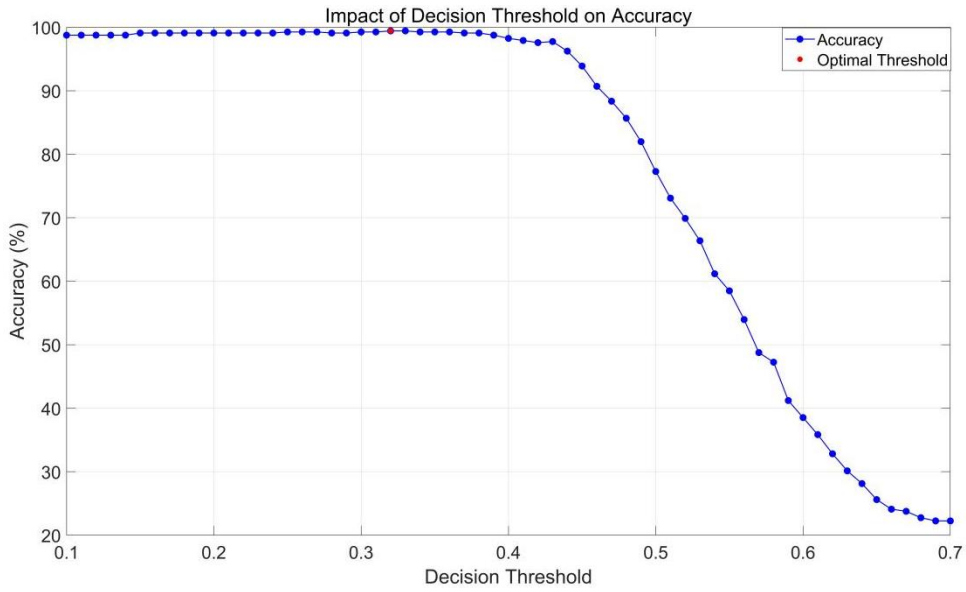

Figure 5 Relationship Diagram Between Decision Threshold and Accuracy Rate

Based on the above analysis, a comprehensive method for determining chromosomal abnormalities in female fetuses was established: input the seven feature values - maternal BMI, Chromosome 21 GC content, Chromosome 18 GC content, Chromosome 18 Z-score, maternal age, Chromosome 13 GC content, and X chromosome concentration - into the judgment model (Formula 19) to calculate the probability value P. According to the decision rule (Formula 20), if $P > 0.32$, it is judged as a normal mother-fetus pair; if $P \leq 0.32$, it is judged as an abnormal mother-fetus pair. This method has been strictly verified and possesses high accuracy and clinical applicability.

4 CONCLUSIONS

This study successfully developed an innovative dual-layer classification model integrating logistic regression and random forest algorithms, addressing the challenge of detecting abnormalities in female fetuses in non-invasive prenatal testing due to the lack of Y chromosome data. Through rigorous screening of 16-dimensional biomarkers, seven key discriminatory features were identified, establishing a progressive clinical decision pathway of "abnormality screening-disease typing." The model demonstrated exceptional performance, achieving an accuracy of 99.57%, precision/recall rates exceeding 98.5% across all categories, and an AUC value of 0.822, significantly outperforming traditional single-model approaches. This research achieves methodological innovation in hybrid model architecture and makes practical advancements in the precision of prenatal diagnosis.

The core value of this study lies in establishing a high-precision detection system for chromosomal abnormalities in female fetuses, applicable in clinical settings, through feature fusion and model collaboration. The model can serve as a reliable decision-support tool, effectively reducing the missed diagnosis rate of chromosomal abnormalities in female fetuses and providing a new technical pathway for precise prenatal diagnosis. With the widespread adoption of non-invasive prenatal testing technologies and the growing demand for precision medicine, the model holds broad application prospects in the following areas: development of clinical auxiliary diagnostic systems, establishment of regional prenatal screening centers, and construction of telemedicine platforms. Furthermore, this methodology can be extended to other genetic disease screening fields, providing technical support for comprehensively improving the prevention and control of birth defects.

There are several aspects of this study that require further refinement: while the sample size meets the requirements for model development, multi-center studies are needed to validate its generalizability across diverse populations; the detection performance for rare chromosomal abnormalities requires validation with larger samples; although the current feature set is comprehensive, it may not fully capture complex epigenetic interactions. Clinical implementation faces challenges in integrating with existing diagnostic workflows and ensuring interpretability for medical professionals. Future research should focus on the following directions: methodologically, incorporating deep learning architectures and multimodal data fusion to enhance pattern recognition capabilities; clinically, developing real-time decision support systems and strengthening translational impact through international collaboration; technologically, exploring compatibility with portable sequencing and cloud-based deployment to improve accessibility. Additionally, extending this methodology to screen for other genetic diseases and adapting it for early pregnancy stages are promising research directions worth exploring.

COMPETING INTERESTS

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

REFERENCES

[1] Lo Y M D. Non-invasive prenatal testing by next generation sequencing: maternal plasma DNA and RNA. Annual Review of Genomics and Human Genetics, 2022, 23, 413-431.
[2] Bianchi D W, Chiu R W K. Sequencing of circulating cell-free DNA during pregnancy. New England Journal of Medicine, 2022, 379(5): 464-473.
[3] Norwitz E R, Levy B. Noninvasive prenatal testing: the future is now. Reviews in Obstetrics and Gynecology, 2023, 12(2): 89-95.
[4] Zhang Yan, Li Qiang, Liu Shuzheng, et al. Classification of chromosomal abnormalities in noninvasive prenatal testing based on machine learning algorithms. Bioinformatics, 2020, 18(3): 156-162.
[5] Chen Si, Liu Pei, Zhao Yang, et al. Clinical application of whole genome sequencing-based noninvasive prenatal testing in 20,000 pregnancies. Chinese Journal of Obstetrics and Gynecology, 2022, 57(2): 89-95.
[6] Wang Ke, Li Hui, Yuan Ming, et al. Detection of fetal aneuploidy by dual-model algorithm based on maternal plasma DNA sequencing. Chinese Journal of Medical Genetics, 2019, 36(5): 412-418.
[7] Gregg A R, Skotko B G, Benkendorf J L, et al. Noninvasive prenatal screening for fetal aneuploidy, 2016 update: a position statement of the American College of Medical Genetics and Genomics. Genetics in Medicine, 2016, 18(10): 1056-1065.
[8] Huang Rong, Li Ming, Wang Shu, et al. Comparative study of machine learning methods for feature selection and classification in noninvasive prenatal testing. Chinese Journal of Biomedical Engineering, 2020, 39(2): 156-163.
[9] Xu Jing, Chen Liang, Wang Rui, et al. A deep learning framework for fetal chromosomal abnormality detection from low-coverage sequencing data. Chinese Journal of Perinatal Medicine, 2022, 25(1): 45-51.
[10] Wu Qian, Zhou Ying, Li Xue, et al. Clinical validation of a random forest-based noninvasive prenatal testing model in 15,456 pregnancies. Chinese Journal of Obstetrics & Gynecology and Pediatrics, 2021, 17(3): 289-295.
[11] Petersen A K, Cheung S W, Smith J L, et al. Positive predictive value estimates for cell-free noninvasive prenatal screening from data of a large referral genetic diagnostic laboratory. American Journal of Obstetrics and Gynecology, 2017, 217(6): 691.e1-691.e6.
[12] Liang Xue, Wang Tao, Chen Yang, et al. A cost-effective method for noninvasive prenatal screening using low-pass whole genome sequencing. Chinese Journal of Practical Gynecology and Obstetrics, 2020, 36(8): 712-718.

World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3073

MARKET-DRIVEN ANALYSIS OF JAVA ECOSYSTEM EVOLUTION AND TALENT DEMAND DYNAMICS

ZhengLin Wang
School of Information and Electrical Engineering, Hebei University of Engineering, Handan 056038, Hebei, China.

Abstract: Learning Web application design and development requires a deep understanding of the industry's technological ecosystem. To explore the core position and evolution trend of the Java language in the current computer software field, this paper systematically surveys and analyzes Java's domestic and international rankings, market share, talent demand structure, and future development direction based on historical data from authoritative international programming language rankings such as TIOBE and PYPL, combined with big data from major domestic recruitment platforms such as Boss Zhipin and Lagou.com. The survey results show that despite challenges from Python in the field of artificial intelligence and Go in the field of cloud-native infrastructure, Java remains firmly in the top tier of global programming languages due to its over 33% market share in enterprise-level server applications and its dominant position in big data processing and Android mobile development. The key innovations of this study lie in: 1) constructing a multi-dimensional analytical framework that integrates technological rankings, market penetration, and talent analytics to assess programming language ecosystems; 2) revealing the structural shift in Java talent demand from basic coding to composite competencies in architecture, cloud-native, and AI engineering. Regarding talent demand, the market exhibits significant structural changes: the demand for entry-level CRUD (Create, Read, Update, Delete) positions is shrinking, while there is a shortage of advanced, multi-skilled talents with microservice architecture, cloud-native technologies, and JVM low-level optimization capabilities. Furthermore, this paper also discusses the profound impact of GraalVM native image technology and the engineering implementation of Spring AI on the future Java ecosystem.
Keywords: Java ecosystem; Enterprise software; Talent analytics; Cloud-native transition; Programming language trends

1 INTRODUCTION

In today's highly interconnected digital economic landscape, the Java technology stack has transcended its role as a mere programming tool to become the foundational infrastructure supporting critical business operations [1-3]. This transformation is particularly evident in enterprise environments where Java underpins large-scale financial systems, telecommunications networks, and e-commerce platforms that process billions of transactions daily. While the explosive growth of cloud computing and artificial intelligence has propelled Python's ascendancy in data science [4] and fostered widespread adoption of Go in cloud-native architectures [5], sparking industry discussions about a "post-Java era," empirical market data presents a distinctly different narrative [6, 7]. According to in-depth research on Asia's labor market in the 2025 Hays Asia Salary Guide, Java remains the language of choice for building high-frequency trading systems, core banking settlement platforms, and complex supply chain systems—particularly in financial hubs such as Hong Kong region and Singapore, where Java architecture experts command significant salary premiums and possess high irreplaceability [8].

The notion that "Java is dead" appears increasingly untenable in the context of enterprise application development reality [9]. Data indicates that Java's ecosystem is undergoing a profound structural shift: it is moving beyond its traditional stronghold in internet development to embed itself deeply within the digital transformation cores of manufacturing, life sciences, supply chain management, and other substantial industries [10, 11]. This expansion is facilitated by Java's mature middleware ecosystem, comprehensive testing frameworks, and extensive monitoring tools that provide the reliability and maintainability required for mission-critical industrial applications. In the mainland China market, even amid cyclical adjustments in the technology sector, salary ranges for senior Java development and architecture positions remain notably high at 400,000–700,000 RMB, demonstrating remarkable market resilience and "hard currency" attributes. The sustained demand for Java expertise across multiple sectors, coupled with the technology's continuous evolution through initiatives like Project Loom and GraalVM, suggests a robust future trajectory rather than technological obsolescence.

This paper aims to provide a quantitative analysis of Java's true market share and talent demand patterns across vertical industries, using granular market data. We not only examine its dominance in financial technology but also focus on how Java technology, through deep integration with cloud-native architectures and AI components, is evolving into a career moat for multi-skilled technical talent amid widespread corporate efforts to reduce costs and enhance efficiency. Our analysis specifically investigates the emerging skill combinations that command premium compensation, including distributed systems design, JVM optimization, microservices architecture, and AI engineering integration. The study offers commercially forward-looking guidance for computer science students' skill development and career path

planning, highlighting how Java's ecosystem adaptability creates sustainable career advantages despite evolving technological landscapes. By synthesizing insights from programming language indices, market share reports, and labor market analytics, we seek to clarify the dynamic interplay between technological evolution, industrial adoption, and workforce requirements in the contemporary Java ecosystem, ultimately providing evidence-based insights for strategic decision-making in both technical education and corporate talent development.

2 JAVA LANGUAGE INTERNATIONAL AND DOMESTIC RANKING AND ECOSYSTEM RESEARCH

2.1 Analysis of International Authoritative Ranking Data

The quantitative evaluation of programming language popularity and lifecycle primarily relies on internationally authoritative indices such as TIOBE and PYPL. These indices integrate multidimensional indicators, including search engine query frequency, the size of the professional developer community, course system coverage, and code repository activity, providing an objective reflection of the industrial standing of specific languages.

Time-series data from the TIOBE Programming Community Index over the past three years indicates, Java consistently remains within the top tier of global programming languages. Although the paradigm shift towards artificial intelligence has propelled Python to the top position, Java's rating index continues to stabilize within the high range of 10%–13%. Together with system-level languages such as C and C++, Java forms a "leading group" that demonstrates significant volatility resistance in statistical terms, showing no structural decline despite the rise of emerging DSLs (domain-specific languages).

The TIOBE scoring algorithm emphasizes the weight of the existing market and the depth of industrial application. The exceptionally high time-series stability of Java's ranking directly reflects the coexistence of massive legacy systems and newly added enterprise-level distributed architectures worldwide. Whether maintaining financial core transaction systems or constructing cloud-native microservices, Java's substantial asset scale and technical inertia constitute an unshakable industrial foundation. The TIOBE indices for various programming languages are presented in Table 1.

Table 1 TIOBE Index of Programming Languages

Programming Language	2025	2020	2015	2010	2005	2000	1995	1990	1985
Python	1	3	6	7	8	26	10	-	-
C++	2	4	5	2	3	2	1	2	10
C	3	1	2	2	1	1	2	1	1
Java	4	2	1	1	2	3	32	-	-

The PYPL index, based on Google Trends search frequency, better reflects the actual attention of learners and developers. On the PYPL ranking, Java consistently ranks second globally, trailing only Python and significantly ahead of third-ranked JavaScript. According to first-quarter data from 2024, Java's global search share is approximately 16%, meaning that one out of every six programming-related technical searches worldwide pertains to Java. This high-frequency search behavior indirectly confirms that Java has one of the world's most active developer communities, where any technical challenge can be quickly resolved through community resources.

2.2 Domestic Popularity and Ecosystem Status: From the Cornerstone of the Internet to the Digital Engine for All Industries

Focusing on the domestic Chinese market, the market penetration of the Java language demonstrates remarkable resilience and dominance. Despite the profound structural adjustments that China's internet industry has undergone in recent years, entering a phase of "rebalancing development" in business trends, the Java technology stack continues to maintain its central position. According to consecutive tracking data from the 2024 Hays Asia Salary Guide and the 2025 Hays Asia Salary Guide, the annual salary range for Java developers in mainland China has remained consistently high at 400,000–700,000 RMB between 2024 and 2025. This salary rigidity amidst market fluctuations empirically validates Java's role as the "hard currency" benchmark in China's software talent market, with its performance metrics far surpassing those of other, more volatile programming languages.The software development compensation data from the 2024 Hays Asia Salary Guide and the 2025 Hays Asia Salary Guide are presented in Table 2 and Table 3, respectively.

Table 2 2024 Hays Asia Salary Guide Software Development Compensation Data

Development & Design	Mainland China (kRMB/year)	Hong Kong region SAR (kHKD/year)	Japan (million JPY/year)	Malaysia (kMYR/year)	Singapore (kSGD/year)	Thailand (kTHB/year)
Full-Stack Developer	400 - 700	420 - 840	6 - 14	84 - 240	80 - 150	960 - 1,440
Mobile Developer	350 - 650	350 - 840	6 - 12	84 - 220	80 - 150	720 - 1,680
Server-Side Developer	300 - 600	350 - 700	6 - 10	72 - 240	80 - 150	720 - 1,440
Frontend Developer	300 - 600	350 - 600	6 - 12	84 - 180	80 - 150	600 - 1,440
Embedded Developer	350 - 700	N/A	6 - 10	74 - 220	80 - 150	720 - 1,080
Backend Developer	350 - 700	350 - 800	8 - 14	72 - 240	80 - 180	720 - 1,440
Web Design Engineer	250 - 450	240 - 500	4 - 8	62 - 140	60 - 100	600 - 960
UX/UI Designer	250 - 600	500 - 900	5 - 12	84 - 240	80 - 120	600 - 1,200
Java Developer	400 - 700	450 - 840	6 - 12	84 - 240	80 - 150	720 - 1,440
Test Analyst	250 - 350	420 - 800	6 - 8	80 - 120	60 - 90	480 - 840
Senior Test Analyst	300 - 500	600 - 1,000	8 - 10	84 - 140	80 - 140	840 - 1,200
Quality Control Manager	420 - 630	600 - 880	8 - 14	156 - 240	100 - 180	1,440 - 1,800
Quality Control Director	580 - 850	700 - 1,100	10 - 16	264 - 330	160 - 240	2,100 - 2,640

Table 3 2025 Hays Asia Salary Guide Software Development Compensation Data

Development & Design	Mainland China (kRMB/year)	Hong Kong region SAR (kHKD/year)	Japan (million JPY/year)	Malaysia (kMYR/year)	Singapore (kSGD/year)
Full Stack Developer	400 - 700	420 - 840	8 - 14	84 - 240	80 - 150
Mobile Developer	350 - 650	350 - 840	6 - 12	84 - 220	80 - 150
Server-side Developer	300 - 600	350 - 700	6 - 10	72 - 240	80 - 150
Frontend Developer	300 - 600	350 - 600	6 - 12	84 - 180	80 - 150
Embedded Developer	350 - 700	N/A	6 - 10	74 - 220	80 - 150
Backend Developer	350 - 700	350 - 800	8 - 14	72 - 240	80 - 180
Web Design Engineer	250 - 450	240 - 500	4 - 8	62 - 140	60 - 100
UX/UI Designer	250 - 600	500 - 900	5 - 12	84 - 240	80 - 120
Java Developer	400 - 700	450 - 840	6.5 - 12	84 - 240	80 - 150
Test Analyst	250 - 350	420 - 800	6 - 8	80 - 120	60 - 90
Senior Test Analyst	300 - 500	600 - 1,000	8 - 10	84 - 140	80 - 140
Quality Control Manager	420 - 630	600 - 880	8 - 14	156 - 240	100 - 180
Quality Control Director	600 - 1,000	700 - 1,100	10 - 16	264 - 330	160 - 240

The driving force behind the prosperity of the domestic Java ecosystem is undergoing a structural transformation. While early growth benefited from Alibaba's "de-IOE" strategy and the establishment of distributed architecture standards through the Spring Cloud Alibaba microservices ecosystem, the latest insights from the 2025 Hays Asia Salary Guide indicate that the demand for technical talent is now spilling over from pure internet companies to real-world

industries such as smart manufacturing, new energy, life sciences, and consumer goods. In their digital transformation journeys, these traditional sectors have adopted the mature Java technology architectures developed by major internet companies, leveraging their established middleware systems to build highly available enterprise-level systems. This signifies that Java's ecosystem influence has extended beyond the internet sector, becoming the foundational infrastructure language for digital transformation across all industries in China.

On the talent supply side, although the "demonstration effect" of industry giants such as Meituan, JD.com, and ByteDance remains, subtle shifts have emerged in market supply-demand dynamics. Data from 2025 shows an "oversupply" in China's technology talent market, intensifying job competition. This indicates that while the substantial existing base of Java developers forms an absolute foundation for employment, merely mastering the language syntax is no longer sufficient to ensure competitiveness. Current market trends increasingly favor Java professionals with composite skills—those who not only excel in Java backend architecture but also possess cloud architecture design capabilities and even AI application integration skills. Therefore, in China, Java is not only a mainstream language with significant existing advantages but also a critical hub connecting traditional IT architectures with the implementation of emerging technologies such as AI and cloud computing.

2.3 International Ecosystem Perspective: The Dual Engines of Financial Core and Offshore Centers

Compared to the extensive penetration of the Java technology stack from the internet to physical industries in the domestic Chinese market, its presence in international markets—especially in Asian financial hubs such as Hong Kong region, Singapore, and Japan—exhibits a highly focused characteristic as "high-value financial infrastructure." According to the survey data from the 2025 Hays Asia Salary Guide on the "Software Development | Financial Services (Java/C++/C#)" segment, Java remains the preferred language for building low-latency trading systems, core banking settlement platforms, and risk management systems. In the Hong Kong Special Administrative Region and Singapore, Vice President (VP)-level technical experts with deep Java expertise and financial business knowledge can command annual salaries of 950,000–1,500,000 HKD and 175,000–250,000 SGD, respectively. Such a significant talent premium indicates that in the international market, Java's niche is no longer merely an application development tool but a cornerstone for supporting the stable operation of high-frequency, high-concurrency financial businesses.

Another notable trend in the international Java ecosystem is the rise of "offshore delivery and shared service centers (SSC)." The guide notes that to cope with global economic uncertainties and optimize cost structures, many multinational corporations are relocating core Java-related development and maintenance functions to emerging markets in Southeast Asia, such as Malaysia. Malaysia’s strategic position as a regional shared service center is increasingly prominent, with Java development leads earning a stable salary range of 180,000–300,000 MYR, making the country a key hub for undertaking global enterprise-level Java application development. This "dual-track" ecosystem—where financial centers (HK/SG/JP) focus on high-end architectural design, and offshore centers (MY) handle standardized implementation—forms a robust industrial chain for Java in the international market.

Furthermore, unlike the "talent surplus" concerns in the domestic market, some international markets face structural shortages. For example, in Japan, despite Java’s dominance in enterprise applications, severe population aging and language barriers have led to an extreme scarcity of bilingual professionals capable of managing complex Java architectures. As a result, companies are forced to offer higher salaries (VP-level positions can command 13–20 million JPY annually) to compete for the limited pool of existing talent. In summary, Java in the international landscape exhibits dual ecosystem characteristics of a "deep financial moat" and a "mature global delivery system," with career development paths tending toward deep specialization in vertical fields rather than broad coverage.

3 MARKET SHARE ANALYSIS

3.1 Enterprise Server-Side Application Market

In the enterprise server-side development domain, Java has established a near-monopoly market dominance, thanks to its mature ecosystem and industrial-grade stability. Particularly in scenarios involving stringent technical requirements such as high-concurrency transaction processing, high availability assurance, and strong consistency distributed transactions, Java demonstrates irreplaceable architectural advantages. Leveraging the robustness of the JVM memory model and the comprehensive middleware ecosystem, including Spring Boot and Dubbo, Java has long been the preferred language for constructing critical infrastructure, such as global financial settlement centers, telecommunications billing networks, large-scale e-commerce middleware platforms, and logistics dispatch systems. This forms the foundational bedrock of the digital economy's operations.The JetBrains Developer Ecosystem Research Report is shown in Table 4.

Table 4 JetBrains Developer Ecosystem Research Report

Languages	Proportion
Python	35%
Java	33%

Languages	Proportion
JavaScript	26%
TypeScript	22%
HTML / CSS	16%
SQL	16%

This exceptionally high market share is attributed to the continuous evolution of the Spring ecosystem. From the early SSH and SSM frameworks to the current cloud-native era's Spring Boot and Spring Cloud, Java frameworks have consistently reduced development complexity while providing industrial-grade stability. For industries with extremely high demands for transaction consistency and high availability, such as banking, insurance, and securities, Java and its mature middleware ecosystem offer time-tested and reliable solutions.

Taking e-commerce scenarios as an example, during the annual "Double Eleven" shopping festival, platforms face peak traffic surges of hundreds of thousands of transactions per second. The core trading systems, inventory centers, and payment gateways supporting this scale almost entirely run on deeply customized JVMs. Such outstanding performance under high-pressure conditions has solidified Java's dominant position in the high-end enterprise market.

3.2 Mobile and Embedded Markets

In the mobile market, the success of the Android operating system has directly cemented Java's market position. Despite Google's strong promotion of Kotlin as the preferred language for Android development in recent years, Kotlin maintains 100% interoperability with Java, and the Android SDK's underlying code is still largely written in Java. Consequently, Java continues to hold a significant share in Android development.

In the embedded and Internet of Things (IoT) fields, Java Card technology, though not widely recognized by the general public, has an astonishing scale. Billions of SIM cards, bank cards, social security cards, and identity recognition cards worldwide run Java Applets. With the rise of IoT edge computing, Java's applications in smart gateways, industrial controllers, and other devices are also steadily growing.

3.3 Big Data and Cloud Computing Infrastructure

Java's market share in the big data domain is often underestimated. In reality, most core components of the big data ecosystem run on the JVM. Key frameworks and tools such as the Hadoop ecosystem, Spark in-memory computing engine, Flink real-time stream processing framework, and Kafka message queue are either directly written in Java or in Scala. This means that while data analysts may use Python for data mining at the application layer, the underlying engineering tasks—such as big data platform construction, data cleansing, and real-time data pipeline development—remain firmly within the domain of the Java technology stack.

4 INVESTIGATION INTO MARKET TALENT DEMAND

4.1 Talent Demand Volume and Industry Distribution

According to an analysis of data from 2023-2024 on mainstream recruitment platforms such as Boss Zhipin, Lagou, and Liepin, the recruitment demand for Java engineers has consistently ranked first among all technical R&D positions. Against the backdrop of digital transformation across all industries, the demand for Java talent is no longer confined to traditional software outsourcing and internet companies.

Research indicates that traditional sectors such as financial technology, new retail, smart manufacturing, and the Internet of Vehicles are experiencing explosive growth in demand for Java professionals. For instance, the fintech subsidiaries established by major commercial banks and the in-car backend teams of new energy vehicle companies are actively recruiting Java developers. In terms of geographical distribution, demand growth in new first-tier cities like Chengdu, Wuhan, Xi'an, and Nanjing is particularly notable, in addition to major hubs like Beijing, Shanghai, Guangzhou, Shenzhen, and Hangzhou. This highlights the high employment versatility and geographical flexibility of Java development skills.

4.2 Salary Levels and Experience Requirements

The salary levels for Java developers show a significant positive correlation with work experience, with a notably high salary ceiling.

(1) Junior Engineers: The market competition is most intense for this group, and salary levels tend to stabilize. Employers place greater emphasis on a solid foundation and learning ability.

(2) Mid-level Engineers: This group experiences significant salary growth and represents the segment with the highest market demand. Employers expect them to be capable of independently designing modules and proficiently mastering mainstream framework principles, database optimization, and caching strategies.

(3) Senior Engineers/Architects: This group represents a scarce resource with highly competitive salaries. Requirements include experience in high-concurrency system design, microservices governance capabilities, JVM underlying optimization skills, and a deep understanding of cloud-native architecture.

4.3 Changes in the Core Skill Demand Map

Research reveals that the skill requirements for Java talent in the market are undergoing profound changes. "Merely knowing how to code" is no longer sufficient to meet enterprise needs. The current skill demand map includes:
Solid Computer Science Foundation: Data structures, algorithms, computer networks, and operating system principles.
In-Depth Java Fundamentals : Concurrent programming, JVM memory model and garbage collection algorithms, and design patterns.
Microservices Architecture Stack: Spring Cloud Alibaba suite, RPC frameworks, and service registries.
Distributed Technologies: Distributed transactions, distributed locks, and message middleware.
Engineering and DevOps: Build tools, version control, containerization, orchestration tools, and CI/CD pipelines.

5 FUTURE DEVELOPMENT AND APPLICATION DIRECTIONS

5.1 Cloud-Native and Serverless Transformation

The future of Java lies in cloud-native technologies. In the past, Java was criticized for its slow startup times and high memory consumption, which are significant drawbacks in serverless computing scenarios. However, with the maturation of GraalVM technology and the introduction of Spring Native, Java is undergoing a "slimming" revolution.Performance optimization comparison of GraalVM Native Image is shown in Figure 1.

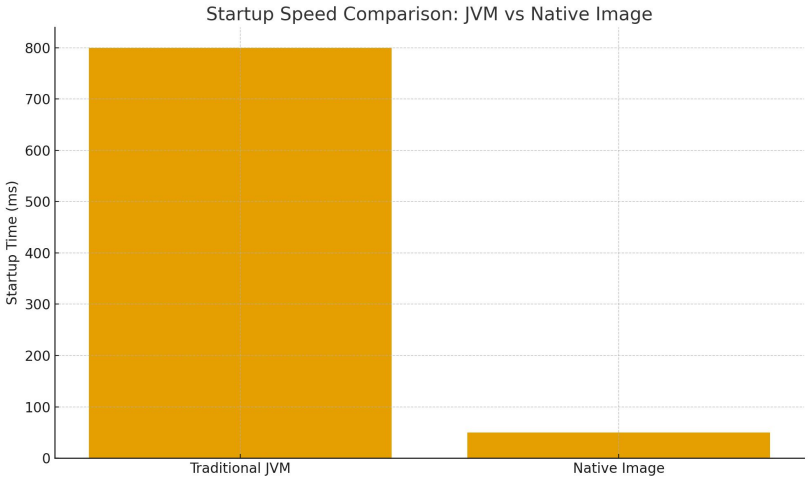

Figure 1 GraalVM Native Image Performance Optimization Comparison

GraalVM can compile Java applications directly into native binary executables via static compilation technology, enabling them to run without a JVM. Test data show that native images achieve millisecond-level startup speeds and reduce memory usage by over 50%. This allows Java applications to perfectly adapt to Kubernetes' rapid scaling requirements and the cold-start scenarios of serverless computing. The rise of next-generation cloud-native Java frameworks such as Quarkus and Micronaut marks Java's formal entry into the cloud-native era.

5.2 Innovation in High-Concurrency Programming Models

Project Loom (virtual threads), introduced in Java 21, represents a major upgrade to Java's concurrency programming model. Traditional Java threads are directly mapped to operating system kernel threads, which consume significant resources and incur high context-switching costs. In contrast, virtual threads are lightweight threads managed by the JVM, allowing millions of them to run on a small number of system threads. This enables developers to write high-concurrency programs with performance comparable to asynchronous non-blocking code, using the simpler synchronous blocking coding style. This greatly lowers the barrier to high-concurrency programming and enhances system throughput.

5.3 Deep Integration of AI and Big Data

Although Python is primarily used for AI model training, Java holds significant advantages in the engineering deployment phase of AI. Enterprises need to deploy trained models into highly available production environments, which is exactly where Java excels.

Projects like Spring AI aim to simplify the process for Java developers to call large language models. In the future, Java developers will leverage existing enterprise-level architectures to integrate technologies such as Retrieval-Augmented Generation (RAG) and vector databases, building intelligent enterprise applications. Java will serve as a bridge connecting traditional business systems with AI capabilities, such as integrating intelligent Q&A into customer service systems or smart recommendations into e-commerce platforms.

6 CONCLUSIONS

This comprehensive investigation, integrating quantitative rankings and qualitative market insights, establishes Java as a resilient and dominant force in the global software industry. The evidence robustly refutes narratives of decline, demonstrating Java's sustained leadership through its foundational role in enterprise systems, pivotal evolution via cloud-native technologies, and critical position in big data and Android ecosystems.

This study offers two primary theoretical contributions. First, it constructs and validates a multi-dimensional analytical framework that moves beyond singular metrics of language popularity. By systematically integrating technological rankings, market penetration data, and granular talent analytics, the framework provides a more holistic and robust model for assessing the vitality and evolutionary trajectory of programming language ecosystems. Second, it empirically reveals and defines a profound structural transformation in talent demand. The analysis documents a decisive shift from valuing basic coding proficiency toward a premium on “T-shaped” professionals who combine deep Java expertise with competencies in microservices architecture, cloud-native deployment, and AI engineering integration.

While this study provides a multi-dimensional assessment, its reliance on aggregated reports presents opportunities for more granular methodologies. Future work should employ large-scale NLP analysis of job descriptions to track skill evolution dynamically and conduct industry-specific case studies to establish causal relationships between Java advancements (e.g., GraalVM, Project Loom) and measurable outcomes in developer productivity or system performance. As AI integration accelerates, dedicated research on scalable patterns and engineering best practices for building “AI-augmented” enterprise Java systems will be crucial for guiding the ecosystem’s next phase of evolution.

COMPETING INTERESTS

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

REFERENCES

[1] Yangyang S, Zhu Jun W, Muhammad D, et al. The best publishing strategy of enterprise software companies facing the competition of cloud service providers. Expert system and application, 2024, 236.
[2] Andriol J S. Editorial: Where has IT gone?. International Journal of Technology Management, 2022, 89(1-2): 1-8.
[3] ZHANG Qi-Xun, WU Yi-Fan, YANG Yong, et al. Survey on service dependency discovery technologies for microservice systems. Journal of Software, 2024, 35(1): 118-135.
[4] DONG Hao-Wen, ZHANG Chao, LI Guo-Liang, et al. Survey on cloud-native databases. Journal of Software, 2024, 35(2): 899-926.
[5] Russell F. The combination of information technology and decentralized workplace organization: small and medium-sized enterprises and large enterprises. International business economics Journal, 2016, 23(2): 199-241.
[6] Peng Y, Hao J, Chen Y. Performance prediction and resource adaptive adjustment for cloud-native microservices. Cluster Computing, 2025, 28(12): 786.
[7] Hays. 2025 Hays Asia Salary Guide. Hays Specialist Recruitment, 2025.
[8] Maximilian S, Manuel W, Helmut K. Capabilities for value co-creation and value capture in emergent platform ecosystems: A longitudinal case study of SAP’s cloud platform. Journal of Information Technology, 2021, 36(4): 365-390.
[9] DI MEGLIO S, STARACE L L L. Evaluating performance and resource consumption of REST frameworks and execution environments: Insights and guidelines for developers and companies. IEEE Access, 2024, 12, 161649-161669.
[10] LASIC L, BERONIC D, MIHALJEVIC B, et al. Assessing the efficiency of Java virtual threads in database-driven server applications//Proceedings of the 2024 47th MIPRO ICT and Electronics Convention (MIPRO). Opatija, IEEE, 2024, 926-931.
[11] Rana A. AI-Driven CRM Automation: Cloud-Native Architectures for Omnichannel Customer Experience Optimization. Journal of Computer Science and Technology Studies, 2025, 7(9): 9-17.

World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3074

TIME-SERIES FORECASTING OF STOCK PRICE VIA BIDIRECTIONAL LSTM-ATTENTION NEURAL ARCHITECTURE

MingXi Ma
School of Statistics and Data Science, Shanghai University of International Business and Economics, Songjiang 201600, Shanghai, China.

Abstract: Predicting stock prices with high accuracy continues to be a major challenge in financial markets, primarily because of the intricate, non-linear, and highly volatile characteristics of price movements. Traditional statistical methods and standard long short-term memory (LSTM) networks exhibit limitations in capturing temporal dependencies and identifying critical features that significantly influence price movements. To address these challenges, this paper proposes a novel bidirectional LSTM with attention mechanism (BiLSTM-Attention) model for stock price prediction. The proposed model employs bidirectional LSTM layers to process time-series data in both forward and reverse directions concurrently, thereby capturing a more complete picture of past and potential future trends. Additionally, a self-attention mechanism is incorporated to dynamically allocate weights across time steps, enabling the model to focus on salient features that exert substantial influence on price fluctuations. Experimental validation is conducted using real-world stock price data from American International Group (AIG). Results demonstrate that the proposed BiLSTM-Attention model significantly outperforms baseline models across all evaluation metrics, validating the effectiveness of combining bidirectional processing with attention mechanisms for stock price forecasting. The proposed approach offers a stable and effective method for predicting stock prices in the short term.
Keywords: Stock price prediction; Bidirectional LSTM; Attention mechanism; Deep learning; Time series forecasting

1 INTRODUCTION

With the rapid advancement of artificial intelligence and financial technology (fintech), Leveraging big data analysis together with machine learning methods to improve financial market prediction has become an important research focus in both academic and industrial fields [1-2]. As an integral component of the global financial system, stock markets exhibit price dynamics that are shaped by a multitude of interrelated factors, including macroeconomic conditions, investor sentiment, political developments, and corporate performance. These complex interactions render stock markets inherently characterized by high volatility, nonlinearity, and uncertainty [3]. Consequently, accurate stock price forecasting not only provides a scientific foundation for informed investment decision-making but also offers essential guidance for risk management and asset allocation strategies, thereby holding substantial theoretical significance and practical value [4].

Traditional stock price forecasting has predominantly relied on statistical modeling approaches, such as time series analysis and regression models, which were extensively employed in early research. However, as financial data has grown increasingly voluminous and complex, characterized by pronounced nonlinear dynamics, the limitations of these conventional methods have become evident. Specifically, their capacity to process high-dimensional, multi-source data remains constrained, resulting in prediction accuracy that often falls short of practical requirements [5].

In recent years, advances in deep learning technology have provided new avenues for stock price prediction research. Several hybrid architectures have demonstrated promising results by combining complementary neural network structures. For example, the CNN-LSTM framework introduced in employs convolutional neural networks to obtain localized spatial patterns from financial data [6], and then utilizes long short-term memory units to learn extended temporal relationships. Experimental results indicate that this hybrid architecture outperforms standalone LSTM or CNN models. Likewise [7], adopts a bidirectional sliding long short-term memory (BiSLSTM) architecture capable of handling sequences in both the forward and reverse directions, thereby simultaneously exploiting historical and prospective contextual information to achieve more comprehensive temporal modeling and enhanced feature learning capabilities. However, research has shown that LSTM-based models exhibit sensitivity to input data format, quality, and preprocessing strategies [8]. Despite notable progress in deep learning-based approaches, several critical challenges persist, including the need to enhance model prediction accuracy, improve generalization capability across diverse market conditions, and increase model interpretability—issues that remain at the forefront of current research efforts [9].

Motivated by the aforementioned limitations, this paper proposes a novel prediction model that integrates bidirectional long short-term memory (BiLSTM) networks with attention mechanisms (BiLSTM-Attention) to address two primary challenges inherent in standard LSTM architectures for stock price forecasting: high computational complexity and inadequate identification of critical temporal features. The proposed model employs a bidirectional LSTM framework that processes temporal sequences in both forward and backward directions, thereby enabling more comprehensive

modeling of complex price dynamics and temporal dependencies. Concurrently, An attention mechanism is integrated to dynamically assign importance across various time steps, enabling the model to prioritize key temporal features that significantly affect price fluctuations. This dual architecture not only expands the model's ability to identify subtle or implicit patterns but also enhances its interpretability and resilience. Experimental results indicate that the BiLSTM-Attention framework effectively models multi-scale temporal dependencies and maintains strong generalization capability under diverse market scenarios, thus offering a stable and efficient approach for short-term financial price prediction.

The remainder of this paper is organized as follows. Section 2 presents the methodology and theoretical framework of the proposed deep learning model. Section 3 describes the experimental design and presents empirical results based on real-world stock price data from American International Group (AIG), where the model's performance is assessed through several evaluation indicators, such as root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2). A comparative evaluation against several baseline models is also performed to highlight the advantages of the proposed method. Section 4 concludes the study by summarizing the main results, outlining existing limitations, and suggesting possible avenues for future investigation.

2 BILSTM-ATTENTION MODEL

2.1 BiLSTM-Attention Layers

Figure 1 illustrates the overall architecture of the proposed BiLSTM-Attention stock prediction model. The model comprises an input layer, a BiLSTM layer, a self-attention layer, a flattening layer, a fully connected layer, and an output layer.

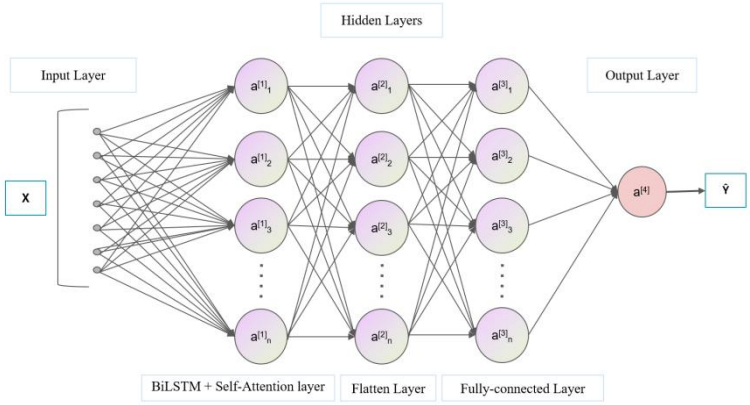

Figure 1 The Architecture of BiLSTM-Attention Model

The input layer receives the stock price sequence samples and passes them to the BiLSTM layer for processing. As shown in Figure 2, the BiLSTM layer consists of forward and backward LSTM units, performing bidirectional processing on the input sequence. The LSTM unit regulates information flow through gating mechanisms, including the forget gate, input gate, candidate memory cell, and output gate. Its internal computations are defined by the following equations:

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f), \tag{1}$$

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i), \tag{2}$$

$$\tilde{C}_t=tanh(W_c\cdot[h_{t-1},x_t]+b_c), \tag{3}$$

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o), \tag{4}$$

$$C_t=f_t\odot C_{t-1}+i_t\odot\tilde{C}_t, \tag{5}$$

$$h_t=o_t\odot tanh(C_t), \tag{6}$$

where: W_f, W_i, W_c, W_o denote weight matrices, and b_f, b_i, b_c, b_o denote bias vectors. These parameters are iteratively optimized during training through the backpropagation algorithm. $\sigma(\eta)=\frac{1}{1+e^{-\eta}}$ denotes the Sigmoid activation function, while $tanh(\eta)=\frac{e^{\eta}-e^{-\eta}}{e^{\eta}+e^{-\eta}}$ represents the hyperbolic tangent function, and $\odot$ denotes the dot product operation.

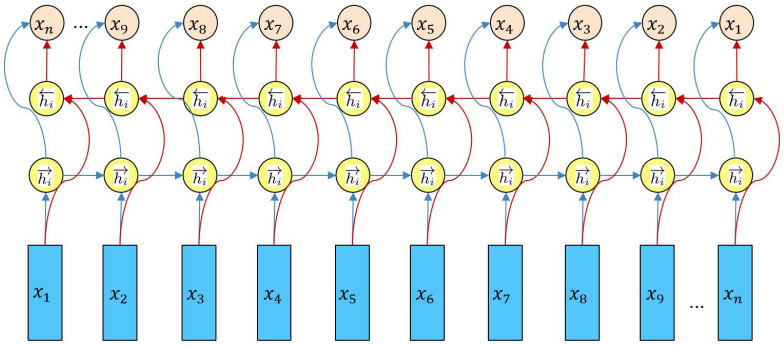

Figure 2 The Architecture of BiLSTM Layer

The forward LSTM's first input is x_1, and its last input is x_n; the backward LSTM processes in reverse order from x_n to x_1. This bidirectional processing enables the network to simultaneously capture both past and future contextual information at the current time step. The output $\boldsymbol{h}_t$ of the BiLSTM at time step t is obtained by concatenating the hidden state $\overrightarrow{\boldsymbol{h}}_t$ of the forward LSTM and the hidden state $\overleftarrow{\boldsymbol{h}}_t$ of the backward LSTM, calculated as:

$$\overrightarrow{\boldsymbol{h}}_t = LSTM_f(x_t, \overrightarrow{\mathrm{h}}_{t-1}), \tag{7}$$

$$\overleftarrow{\boldsymbol{h}}_t = LSTM_b(x_t, \overleftarrow{\mathrm{h}}_{t-1}), \tag{8}$$

$$\boldsymbol{h}_t = [\overrightarrow{\mathrm{h}}_t; \overleftarrow{\mathrm{h}}_t] \tag{9}$$

Through these mechanisms, the BiLSTM layer maintains memory across varying time intervals, capturing temporal dependencies within the input sequence. This bidirectional processing strengthens the network's capability to identify useful features within stock price sequences, which in turn contributes to higher predictive accuracy.

Following the BiLSTM layer, a self-attention module is incorporated to further strengthen the model's capability to extract prominent features. The self-attention mechanism allows the model to reference information across the whole input sequence when computing each position, thus enabling it to capture long-range dependencies in stock price data. As shown in Figure 3, the self-attention layer operates on the BiLSTM output sequence, computing the query matrix $\boldsymbol{Q}$, key matrix $\boldsymbol{K}$, and value matrix $\boldsymbol{V}$:

$$\boldsymbol{Q} = \mathbf{W_q} \cdot \mathbf{H}, \tag{10}$$

$$\mathbf{K} = \mathbf{W_k} \cdot \mathbf{H}, \tag{11}$$

$$\mathbf{V} = \mathbf{W_v} \cdot \mathbf{H}, \tag{12}$$

$$\mathbf{Z} = \mathbf{Attention(Q,K,V)} \cdot \mathbf{V} = softmax\left(\frac{\mathbf{QK}^{\mathrm{T}}}{\sqrt{d_k}}\right) \cdot \mathbf{V}, \tag{13}$$

where $\boldsymbol{W_q}$, $\boldsymbol{W_k}$ and $\boldsymbol{W_v}$ are learnable weight matrices, and d_k denotes the dimension of the attention mechanism. The Softmax function is applied row-wise, generating a probability distribution for each query at positions within the input sequence.

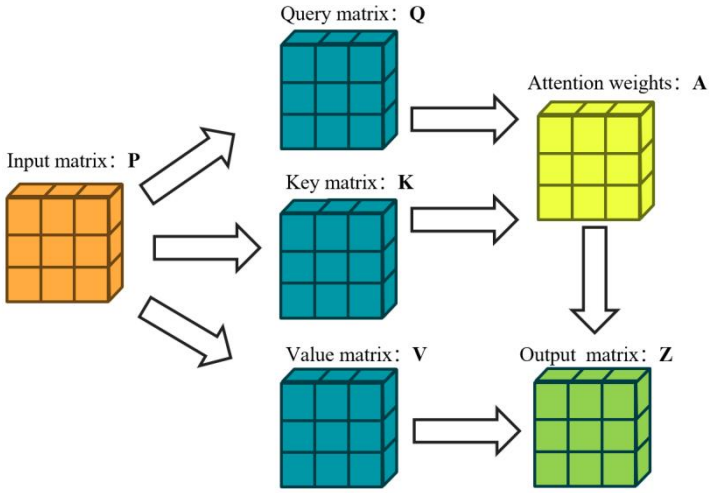

Figure 3 The Architecture of Self-Attention Layer

The self-attention layer directly models the relationship between any two positions in the sequence, which is particularly crucial for capturing complex nonlinear dynamics in stock price data. By providing a global view of the sequence, the self-attention layer complements the local patterns captured by the BiLSTM layer, enabling a feature extraction process that leverages both local and global information. This synergy enhances the model's ability to distinguish informative features from noise, thereby improving prediction robustness.

The feature matrix $\mathbf{Z}\in\mathbb{R}^{T\times 2d}$ output by the Self-Attention layer is subsequently fed into a Flatten Layer. This operation reshapes the two-dimensional feature matrix into a one-dimensional vector $z_{flat}\in\mathbb{R}^{T\times 2d}$, providing a standardized input for subsequent fully connected layers. Mathematically, this operation can be expressed as:

$$z_{flat}=Flatten(\mathbf{Z}). \tag{14}$$

This operation preserves feature information across all time steps while converting it into a format suitable for the fully connected layer. The flattened feature vector is then fed into the fully connected layer, where nonlinear transformations integrate the learned features to generate the regression output. The output $\hat{y}$ from the fully-connected layer represents the predicted stock price, calculated as:

$$\hat{y}=\mathbf{W}_{fc}\cdot z_{flat}+b_{fc}, \tag{15}$$

where $\mathbf{W}_{fc}\in\mathbb{R}^{(T\times 2d)\times 1}$ denotes the weight matrix, and $\mathbf{W}_{fc}\in\mathbb{R}^{(T\times 2d)\times 1}$ represents the bias term.

Through end-to-end training, the entire architecture can automatically learn complex nonlinear mapping relationships from raw price sequences to predicted outputs, enabling accurate stock price forecasting.

2.2 BILSTM-ATTENTION Training

During offline training, the model parameters are iteratively updated on the training set. As training progresses, the discrepancy between predicted and actual values gradually diminishes and converges. Specifically, let $Y=\{y_1, y_2, \ldots, y_N\}$ denote the actual stock price, $X=\{x_1, x_2, \ldots, x_N\}$ denote the corresponding input price feature vector, and $\hat{Y}=\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}$ denote the model's predicted price. The training objective is to minimize the loss function, which quantifies the discrepancy between actual prices and predicted prices . This study employs the mean squared error (MSE) as the loss function:

$$\Gamma=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \tag{16}$$

where denotes the sample size. The model calculates gradients of the loss function with respect to all parameters using backpropagation and updates the network weights using the Adam optimizer. Through this iterative process, the model learns the mapping from input features to target prices, thereby minimizing prediction error.

3 EXPERIMENTAL SIMULATION

To validate the effectiveness of the proposed BiLSTM-Attention model, experiments are conducted using stock market data. The model performance is quantitatively compared with baseline methods using multiple evaluation metrics, including RMSE.

3.1 Experimental Dataset

This study uses stock market data from the publicly available dataset on GitHub https://github.com/Deamoner/ultimate-stock-machine-learning-training-dataset/tree/master/full_history. The dataset comprises historical trading data from major publicly traded companies, providing comprehensive coverage across industry sectors and market conditions. The experiments focus on AIG (American International Group) stock data spanning from 1973 to 2019. The dataset includes daily open, high, low, close (OHLC) prices and trading volume. Table 1 presents a sample of the data.

Table 1 Time Series Data Samples

date	volume	open	close	high	low	adjclose
2019-04-16	4709600	46.05	46.74	46.89	46.04	46.74
2019-04-17	3650600	46.75	45.97	46.80	45.62	45.97
2019-04-18	3729300	45.90	46.04	46.44	45.81	46.04

3.2 Evaluation Indicators

To thoroughly assess the forecasting capabilities of various models, this study utilizes four widely recognized statistical indicators: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2). The corresponding formulas are presented as follows [10]:

$$\mathrm{RMSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}, \tag{17}$$

$$\mathrm{MAE}=\frac{1}{m}\sum_{i=1}^{m}|y_i-\hat{y}_i|, \tag{18}$$

$$\mathrm{MAPE}=\frac{100\%}{m}\sum_{i=1}^{m}\left|\frac{y_i-\hat{y}_i}{y_i}\right|, \tag{19}$$

$$\mathrm{R}^2=1-\frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{m}(y_i-\bar{y})^2}. \tag{20}$$

3.3 Experimental Process

The dataset is divided into training and testing subsets with an 80:20 split. Time-series samples are generated using a sliding window method, where each window spans T trading days. The resulting data are then formatted into a three-dimensional tensor (N, T, F) suitable for LSTM input, with N representing the total number of samples, T the sequence length, and F the number of features.

Four models are evaluated: LSTM, BiLSTM, LSTM-Attention, and BiLSTM-Attention. The models are trained using the Adam optimizer with the MSE loss function and a learning rate scheduler to facilitate convergence. Model performance is evaluated using RMSE, MAE, MAPE, and the coefficient of determination (R^2). Table 2 summarizes the key hyperparameter settings.

Table 2 Primary Hyperparameter Settings for the Model

Parameter Name	Parameter value	explanation
Time step	60	Predict the next day based on the past 60 days
Learning rate	0.003	Adam Optimizer Initial Learning Rate
Batch size	64	Number of samples per gradient update
Number of training cycles	50	Number of iterations
Training/Testing Ratio	0.8/0.2	Data Allocation Ratio
First LSTM layer	64	Capture long-term trends
Second LSTM layer	32	Capture local features
Attention Layer	Enable	Dynamic Time Weighting
Learning Rate Scheduling	Decreases to 20% of the original value every 75 rounds.	Prevent overfitting

3.4 Experimental Result

Figure 4 compares the predicted prices from the four models against actual prices. The BiLSTM-Attention model effectively tracks the real price movements, especially in volatile periods, indicating its capability to precisely identify points of price reversal.

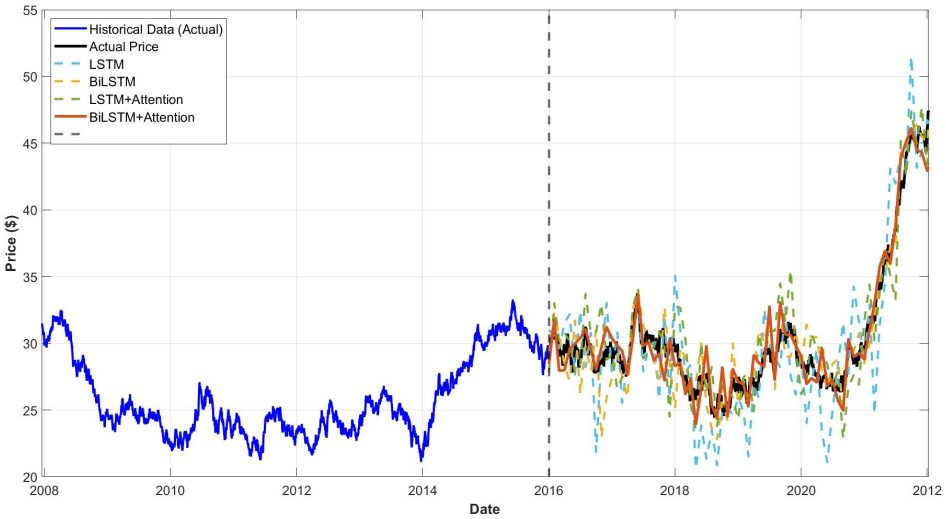

Figure 4 Comparison of the Prediction Curves of the Four Models Against Actual Prices

Figure 5 illustrates the temporal prediction errors of the four models on the test set. Compared to the other three models, the BiLSTM-Attention model displays notably lower error variability, and its smoother error trajectory reflects enhanced stability. Specifically, the prediction errors predominantly lie within ±1 and fluctuate symmetrically around zero, with no significant systematic bias observed, indicating unbiased predictions.

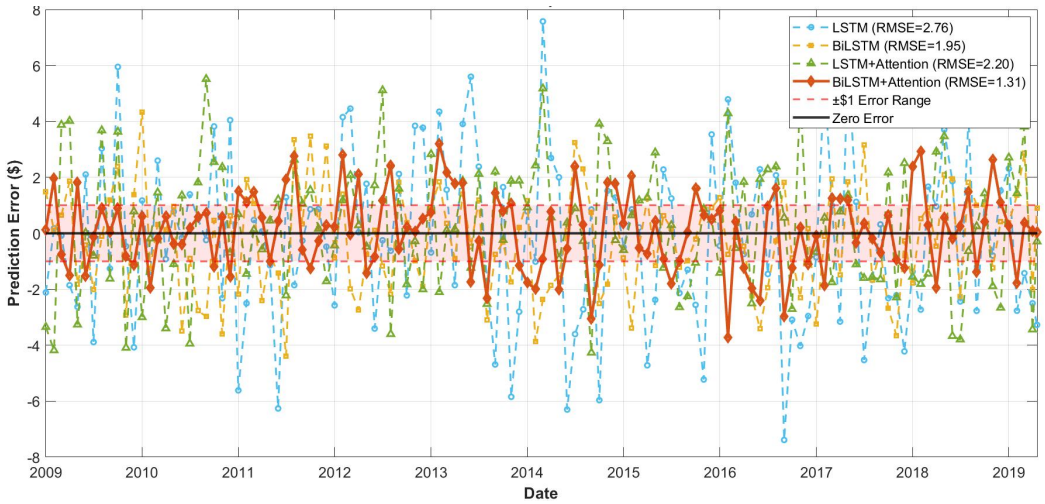

Figure 5 Prediction Error of the Four Models over Time on the Test Set

Table 3 presents the quantitative performance comparison of the four models on the test set. The BiLSTM-Attention model achieves superior performance across all metrics. Compared to the baseline LSTM model, it reduces RMSE by 52.7%, MAE by 64.4%, and MAPE by 63.4%, while improving R^2 from 0.9494 to 0.9887. These results demonstrate that the combination of bidirectional processing and attention mechanism enables the model to effectively extract salient temporal features and focus on critical information, thereby significantly enhancing prediction accuracy.

Table 3 The Results of the Three Models on the Test Set

model	RMSE	MAE	MAPE (%)	R^2
LSTM	2.7635	2.6560	6.01	0.9494
BiLSTM	1.9472	1.5195	3.60	0.9749
LSTM+ Attention	2.2009	1.6968	4.11	0.9679
BiLSTM + Attention	1.3069	0.9449	2.20	0.9887

4 CONCLUSION

This study proposes a BiLSTM-Attention model to address the limitations of standard LSTMs in stock price forecasting, particularly their unidirectional information flow and inability to identify critical time steps. The model combines bidirectional LSTM architecture to capture temporal patterns from both directions with a self-attention mechanism that dynamically weights time steps to focus on key price-influencing moments. Experiments on AIG stock data demonstrate that the BiLSTM-Attention model significantly outperforms baseline models across all metrics, validating the effectiveness of combining bidirectional processing with attention mechanisms. However, the model relies primarily on historical price data and does not incorporate multimodal information such as news sentiment. Future research directions include: (1) integrating text sentiment analysis for multimodal fusion, (2) applying transfer learning to enhance cross-market generalization, (3) extending the model to high-frequency trading and portfolio optimization, and (4) improving interpretability through techniques such as SHAP analysis to provide actionable insights for quantitative investing.

COMPETING INTERESTS

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

REFERENCES

[1] Henrique M B, Sobreiro A V, Kimura H. Literature Review: Machine Learning Techniques Applied to Financial Market Prediction. Expert Systems with Applications, 2019, 124: 226-251.
[2] Zhao Tingting, Han Yajie, Yang Mengnan, et al. A Review of Machine Learning-Based Methods for Time Series Data Forecasting. Journal of Tianjin University of Science and Technology, 2021, 36(5): 1-9.
[3] Mailagaha M K, Christoph L, Pasi L, et al. Machine Learning Techniques and Data for Stock Market Forecasting: A Literature Review. Expert Systems with Applications, 2022, 197.
[4] Xing Weichen. Stock Forecasting in the Big Data Era. China Business Review, 2020(3): 31-32.
[5] Vuong H P, Phu H L, Nguyen V H T, et al. A Bibliometric Literature Review of Stock Price Forecasting: From Statistical Model to Deep Learning Approach. Science Progress, 2024, 107(1).
[6] Bambang S, Derwin S. Utilizing BERT and CNN-LSTM in Stock Price Prediction Using Data Sentiment Analysis and Technical Analysis of Stock and Commodity. Bulletin on Innovative Computing, Information, and Control, 2023, 17(2): 171-176.

[7] Wang Haiyao, Wang Jianxuan, Chen Lihui, et al. A Stock Closing Price Prediction Model Based on CNN-BiSLSTM. Complexity, 2021.
[8] Yadav H, Thorat A. NOA-LSTM: An Efficient LSTM Cell Architecture for Time Series Forecasting. Expert Systems with Applications, 2024, 238.
[9] Yang Zheng, Wu Haocheng, Zhang Jing, et al. Can Dimension Reduction Data ForecastingEnhance the Predictive Performance of Stock Excess Returns?. Journal of Econometrics, 2023, 3(3): 828-847.
[10] Mao Yueyue, Zhang Qiuyue. Research on Machine Learning-Based Stock Forecasting Methods. Modern Computer, 2020(23): 44-47.