ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Risk prediction model optimization of support vector machines in epidemiological data

Jinwu Liu[1,*]

[1] Supervision and Inspection Center of Shanxi Provincial Health Commission, Taiyuan, Shanxi 030045 Shanxi, China

SUMMARY: *Aiming at the problems of heterogeneous variables, more missing variables, class imbalance and complex risk boundaries in epidemiological data, this paper constructed an optimized support vector machine risk prediction model for multi-source monitoring information. In this study, demographic characteristics, physical signs, laboratory results, exposure information and dynamic change characteristics were incorporated into the unified computing framework, and the identification ability of the model for high-dimensional nonlinear risk patterns was improved through missing repair, feature screening, category reweighting, kernel parameter joint optimization and probability calibration. The experimental results based on structured epidemiological samples show that the Accuracy of the model is 89.47%, the Recall is 90.10%, the AUC is 0.931, the Brier Score is reduced to 0.098, and the average inference delay is 138 ms. The proposed method achieves a good balance between prediction accuracy, risk ranking and deployment efficiency, and can provide computational support for primary screening, key population identification and regional public health early warning.*

KEYWORDS: *support vector machine; Epidemiological data; Risk prediction; Model optimization*

1 Introduction

Under the background of the increasing burden of chronic diseases and the increasing requirements of emergency infectious disease surveillance, epidemiological data has gradually evolved from traditional sampling and registration data to multi-source heterogeneous data sets including questionnaire information, physical examination indicators, environmental exposure, follow-up records and regional surveillance logs. Such data not only carry the task of individual risk identification, but also serve for group early warning, resource allocation and public health decision-making. Previous studies have shown that artificial intelligence methods are being widely introduced into scenarios such as cardiovascular disease, pregnancy risk, nosocomial infection and outbreak risk assessment to improve the timeliness and fineness of risk prediction [1-4]. However, epidemiological data often have characteristics such as many missing values, complex variable correlation, unbalanced class distribution and significant nonlinear relationship. If traditional statistical modeling or empirical stratification methods are still relied on, the model is easily limited in coupling identification of high-dimensional variables, delineation of boundary samples and cross-scene generalization [5].

*spark312@163.com
https://doi.org/10.65102/is2026556 Published: 30 April, 2026

With the integration of computer technology and medical data platforms, the application of machine learning in epidemiological risk prediction has gradually expanded from a single classification task to more complex problems such as time event analysis, environmental exposure modeling, and joint prediction across institutions. You et al. established a 10-year cardiovascular disease risk prediction model based on prospective cohort data, showing that machine learning has a good application potential in long-term risk identification [6]. Fu et al. and Guo et al. respectively built prediction models around environmental volatile organic compounds exposure and chemical exposure factors, indicating that multivariate exposure data processed by the algorithm can effectively enhance the ability to identify disease risks [7, 8]. Al Mashrafi et al. applied machine learning to maternal risk classification, suggesting that the fusion of clinical and epidemiological information is helpful to improve the efficiency of physical examination of high-risk individuals [9]. See Table 1 for an overview of related studies. However, most of the existing models still focus on ensemble learning or deep network structure. For epidemiological data with limited sample size and high variable dimension, the increase of model complexity does not necessarily bring marginal benefits, but may amplify the risk of overfitting.

Table 1: Overview of research related to epidemiological risk prediction

Reference	Research Object	Main Method	Data Characteristics	Main Insight
[1]	Review of healthcare applications	Support vector machine-based models	Multi-scenario, multi-task data	SVM has broad applicability in healthcare prediction
[2]	Review of chronic disease prediction	Multiple machine learning algorithms	Heterogeneous health data	Risk prediction is shifting from statistical analysis to intelligent modeling
[6]	10-year cardiovascular risk prediction	Machine learning models	Prospective cohort data	Long-term risk assessment requires stable feature learning capability
[7]	Cardiovascular disease risk identification	Machine learning + SHAP	Environmental exposure data	Complex exposure variables can improve risk discrimination
[8]	Hypertension prediction	Machine learning models	NHANES exposure data	Multivariable environmental factors are suitable for nonlinear modeling
[9]	Maternal risk stratification	Multiple machine learning models	Clinical and epidemiological data	Individualized risk identification places high demands on model generalization
[11]	Short-term COVID-19 prediction	Support vector regression	Temporal epidemiological data	SVM is suitable for short- and medium-term dynamic prediction tasks
[12]	Cross-institutional survival analysis	Federated survival support vector machine	Distributed time-to-event data	SVM can be combined with privacy-preserving computing and federated learning

As a supervised learning method based on statistical learning theory, support vector machine (SVM) has strong advantages in dealing with small sample size, high dimension and nonlinear classification problems [10]. This method maps the original variables into a high-dimensional feature space through a kernel function, and obtains a more stable

classification boundary under the principle of maximum margin. Therefore, it has received continuous attention in medical health prediction, chronic disease screening and event risk identification. Shoko et al. applied support vector regression to the short-term prediction of COVID-19 and verified the feasibility of this kind of method in the modeling of epidemic transmission trend [11]. Spath et al. further proposed the privacy-preserving federated survival support vector machine for cross-institutional time event analysis, indicating that support vector machine is not only suitable for single-center data mining, but also has the technical space to combine with distributed computing framework [12]. However, there are still several shortcomings in the application of support vector machines in epidemiological scenarios. First, the preprocessing link of missing values, outliers and class imbalance is not fully considered, which makes the classification boundary easily disturbed by noise. Second, the selection of kernel function and the setting of penalty parameter often rely on experience trial and error, and lack of systematic optimization mechanism. Third, some studies pay more attention to the prediction accuracy itself, and lack of discussion on the connection between feature selection, model interpretability and public health application scenarios [13].

Based on the above problems, this paper focuses on the epidemiological risk prediction task, and tries to construct an optimized support vector machine model for multi-source crowd surveillance data. The objectives of this study are twofold. Firstly, the adaptation ability of the model to complex epidemiological data is improved by introducing data cleaning, feature screening, class re-weighting and kernel parameter optimization strategies. Secondly, the risk factor representation, classification boundary learning and prediction result output are completed under a unified computing framework, which enhances the stability and transferability of the model in public health surveillance scenarios. The research focus of this paper is not to simply increase the complexity of the algorithm, but to reconstruct the application process of support vector machine in epidemiological risk prediction from three levels: data processing chain, parameter optimization chain and prediction decision chain, so as to provide a more computable model basis for subsequent risk stratification and early warning intervention.

2 Methods

2.1 Optimization strategy of SVM risk prediction model

Epidemiological risk prediction is not a single source, regular and stable data set, but a heterogeneous data system composed of demographic variables, past medical history, laboratory test results, behavioral exposure information, regional surveillance records and follow-up outcomes. Once such data enters the actual modeling process, three prominent difficulties are often exposed. First, the dimensions of variables from different sources are quite different, and continuous variables and discrete variables coexist, so direct input into the model is easy to cause imbalance of feature dominant effect. Secondly, the proportion of high-risk samples in the total sample is low, and the traditional classifier is more likely to be biased towards the majority class, which weakens the recognition ability of the key population. Third, the relationship between epidemiological variables is not simple linear, and some boundary samples carry multiple risk characteristics at the same time. If the classification surface is unstable, the fluctuation of the model on the validation set will be significantly amplified. Based on this practical background, support vector machine (SVM) is used as the basic classifier in this paper, and five links are jointly improved around data standardization, anomaly suppression, class reweighting, feature selection and kernel parameter optimization, so as to improve the adaptation ability of the model on complex epidemiological data. Figure

1 shows the difference between the optimization idea and the traditional rule discrimination way.

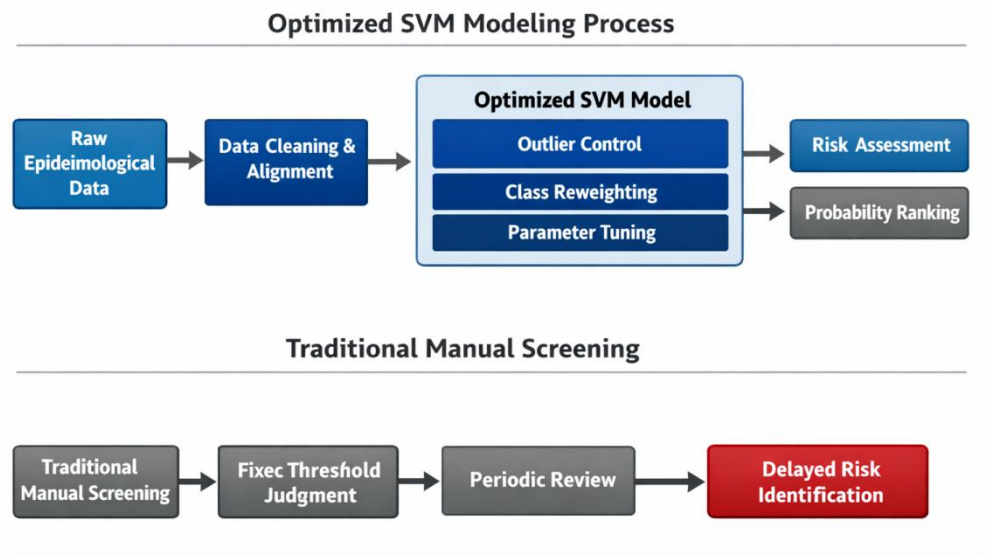


Figure 1: Comparison diagram between epidemiological risk prediction process and traditional risk discrimination methods

In Figure 1, the upper path corresponds to the computational modeling process adopted in this paper. The system first receives multi-source epidemiological surveillance data, completes cleaning, screening and feature expression after unified coding, and then generates individual risk discrimination results and ranking values by optimized support vector machine. The lower-level path corresponds to the traditional empirical identification method, which usually relies on fixed threshold stratification and post-hoc review. Although it is easy to implement, it is difficult to make full use of the nonlinear association between high-dimensional variables, and it is also difficult to identify the boundary crowd in time. The key reason why support vector machine is suitable for this task is that it does not directly pursue the lowest training error, but uses the principle of structural risk minimization to construct a classification hyperplane with better generalization ability on the basis of maximizing the sample interval. This is particularly important for epidemiological data, where applications are more concerned with the predictive stability of unknown samples rather than local fitting results on the training set.

Before the data enters the classifier, this paper performs a standardization process on continuous variables to weaken the scale bias between different dimensional variables such as age, blood pressure, blood glucose, exposure concentration and biochemical indicators. The normalization is calculated as follows:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{1}$$

where,x_{ij} represents the original value of the ith sample on the JTH feature, μ_j and σ_j represent the mean and standard deviation of the feature respectively, and z_{ij} is the standardized result. After the completion of scale unification, this paper does not directly delete all abnormal samples, but uses the quantile distance constraint to truncate the extreme values, so as to reduce the traction effect of abnormal points on classification boundaries. Its expression is:

$$x'_{ij} = \min\left(\max\left(x_{ij}, Q_{1j} - 1.5IQR_j\right), Q_{3j} + 1.5IQR_j\right) \quad (2)$$

where, Q_{1j} and Q_{3j} are the lower and upper quartiles of the JTH feature, and $IQR_j = Q_{3j} - Q_{1j}$. This method preserves the overall structure of the sample and avoids unnecessary disturbance of the kernel space mapping caused by extreme records. Considering that high risk samples are usually far less than low risk samples in epidemiological data, this paper introduces a class weight mechanism into support vector machine to impose higher cost on minority class misclassification. The class weights are defined as follows.

$$w_c = \frac{N}{K \cdot N_c} \quad (3)$$

where, N is the total number of training samples, K is the number of categories, N_c is the number of samples of the CTH category, and w_c is the penalty weight of the corresponding category. In this way, the model is no longer biased only towards overall accuracy during training, but will give more importance to recall for high-risk groups. At the same time, in order to reduce the interference of redundant variables on kernel function learning, this paper constructs a comprehensive scoring criterion combining mutual information and sparse regression coefficient to screen features. The feature importance score is written as follows.

$$S_j = \alpha \cdot MI_j + (1 - \alpha) \cdot \frac{|\beta_j|}{\sum_{m=1}^{p} |\beta_m|} \quad (4)$$

where, MI_j represents the mutual information between the JTH feature and the risk label, β_j is the regression coefficient of the corresponding feature in the sparse linear model, α is the balance coefficient, and p is the total number of features. Variables with higher overall scores were retained into the final training set, thus maintaining risk explanatory power while controlling for dimensions. In the main part of the classifier, this paper uses radial basis kernel support vector machine to build nonlinear decision boundary, and its discriminant function is as follows.

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right) \quad (5)$$

where α_i is the Lagrange multiplier, y_i is the class label, b is the bias term, and K(xi,x) is the kernel function. The form of radial basis kernel selected in this paper is as follows.

$$K(x_i, x) = \exp(-\gamma \|x_i - x\|^2) \quad (6)$$

where, γ is the kernel width parameter. The penalty parameter C and the kernel parameter γ jointly determine the degree of relaxation and the local fitting ability of the classification surface. When the two parameters are not set properly, the model is prone to underfitting or overfitting. Based on this, this paper does not use simple manual trial, but uses the comprehensive evaluation value under cross-validation as the objective function to perform an adaptive search of the parameter space. The optimization objective is defined as follows.

$$J = \lambda_1 \cdot \text{Accuracy} + \lambda_2 \cdot \text{Macro} - \text{F1} + \lambda_3 \cdot \text{AUC} \quad (7)$$

where $\lambda_1, \lambda_2, \lambda_3$ are the weight coefficients, which satisfy $\lambda_1 + \lambda_2 + \lambda_3 = 1$. This design avoids the model's excessive pursuit of a single index, and can take into account the overall classification performance, class balance performance and risk ranking ability. The overall structure of the optimized SVM is shown in Figure 2.

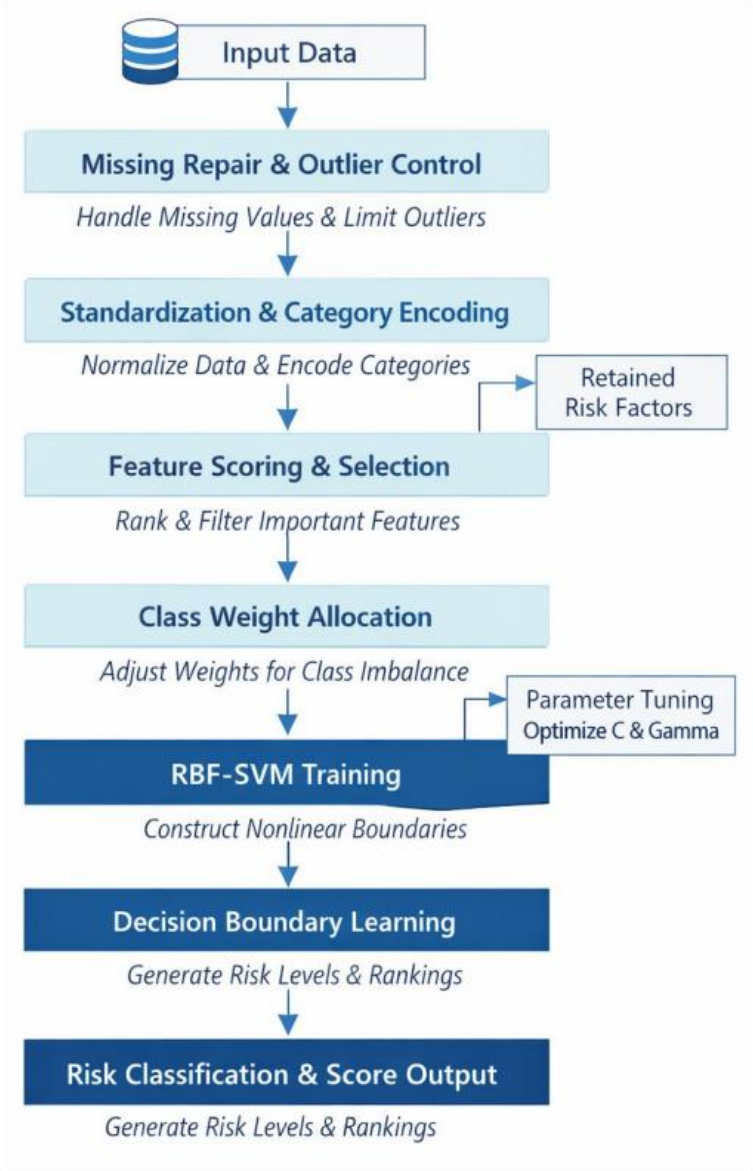


Figure 2: Structure diagram of optimized SVM risk prediction model

Figure 2 shows that the SVM is not directly placed at the end of the data as an isolated classifier, but embedded into a complete calculation link composed of "data correction, feature screening, class balance, parameter optimization, risk output". In this link, the missing repair and anomaly constraint are responsible for stabilizing the input distribution, the feature scoring module is responsible for compressing the invalid dimension, the class reweighting module is responsible for improving the ability to identify the minority class, and the parameter search module is responsible for continuously modifying the kernel space structure through cross validation. After this series of optimizations, SVM is no longer just a general classification algorithm, but a dedicated prediction model suitable for epidemiological risk identification scenarios. The value of this strategy is not limited to improving the accuracy of

a single experiment, but more importantly, it provides a model basis with better computability and transferability for subsequent screening of high-risk individuals, early warning of key groups and public health intervention.

2.2 Methods of epidemiological data collection, preprocessing and feature extraction

In order to make the above optimized support vector machine run stably in real public health surveillance scenarios, this paper constructs a complete processing flow composed of data acquisition, quality verification, time alignment, variable reconstruction and feature extraction before model training. Different from general single physical examination data, the information relied on by epidemiological risk prediction has the characteristics of dispersed sources, inconsistent collection cycles, and inconsistent record granularity. Without a unified data entry mechanism, it is difficult to obtain credible risk boundaries even if the classifier itself has high performance. Based on this consideration, the basic information of individuals, past disease history, physical examination indicators, laboratory test results, behavioral exposure factors, environmental exposure variables and follow-up outcomes are integrated into the same computing framework, and the standard sample matrix for support vector machine training is formed through structured coding. Figure 3 shows the process of epidemiological data collection and risk label generation.

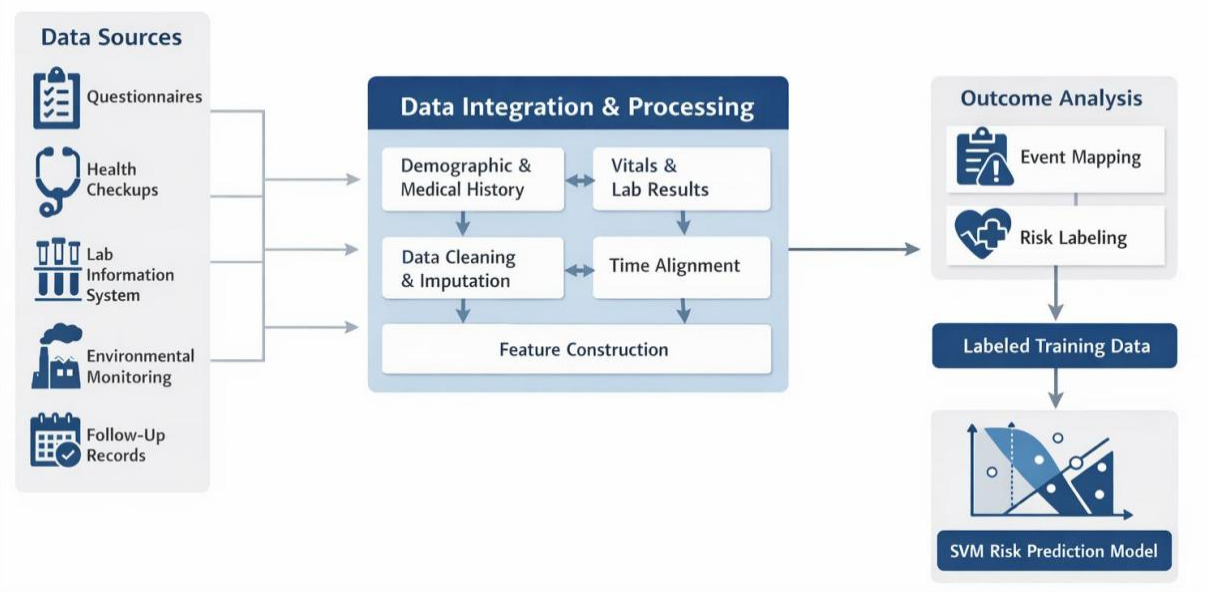


Figure 3: Flow chart of epidemiological data collection and risk outcome acquisition

In Figure 3, the sample sources are not limited to a single medical record, but cover questionnaires, physical examination systems, laboratory information systems, regional environmental monitoring platforms, and follow-up registration modules. Participants were required to complete a unified demographic registration and health questionnaire to obtain basic information such as age, gender, smoking and drinking status, past chronic disease history, family history, and physical activity level. Blood pressure, body mass index, waist circumference, resting heart rate and other physical indicators were collected, and laboratory variables such as fasting blood glucose, total cholesterol, low-density lipoprotein, creatinine and inflammatory factors were combined to form the physiological state representation. For individuals with continuous follow-up records, the system further extracts diagnostic

outcomes, hospitalization records, or risk event labels within the observation period, which are mapped into classification targets required for supervised learning. In order to avoid the impact of format differences of different data ports on subsequent calculations, this paper uses a unified data dictionary to complete field standardization, and uses a unique identity code to aggregate multi-source records to the same subject level.

After the original database is formed, the model does not directly use all the fields, but first performs missing repair and time alignment. Missing data in epidemiological data are often not completely random. If the sample containing missing records is simply deleted, it is easy to reduce the high-risk population and shift the sample distribution. Based on this, this paper adopts the neighborhood weighted imputation method to repair the missing values of continuous variables, and the calculation formula is as follows.

$$\hat{x}_{ij} = \frac{\sum_{r \in \mathcal{N}_i} \omega_{ir} x_{rj}}{\sum_{r \in \mathcal{N}_i} \omega_{ir}} \tag{8}$$

where, $\hat{x}_{ij}$ represents the imputed value of the ith sample on the JTH variable, $\mathcal{N}_i$ is the set of neighborhood samples closest to sample I, and ω_{ir} is the neighborhood weight. The weight is determined by the distance between samples and is expressed as follows.

$$\omega_{ir} = \frac{1}{d_{ir} + \varepsilon} \tag{9}$$

Here, d_{ir} denotes the distance between sample i and sample r in the space of observed variables, and ε is a minimal positive number set to avoid the denominator being zero. After this process, the missing variable is no longer replaced by rough mean, but the distribution characteristics of the same class individuals in the local structure are retained.

Because some subjects have multiple physical examinations, repeated tests or staged follow-up records, the inconsistency in the time dimension will further affect the sample expression. To this end, this paper sets a fixed observation window and aggregates the repeated measurements of the same variable within the window to form a temporal summary feature that can be used for cross-sectional classification. The aggregate result is defined as follows.

$$x_{ij}^{(w)} = \frac{1}{|T_i|} \sum_{t \in T_i} x_{ijt} \tag{10}$$

where, x_{ijt} represents the value of the JTH variable of sample i at time point t, T_i is the set of valid records of the sample in the observation window, $x_{ij}^{(w)}$ is the aggregated window feature. For the indexes with obvious dynamic change significance, this paper further constructs the change rate feature to enhance the model's ability to identify the risk evolution trend, and its formula is as follows:

$$\Delta x_{ij} = \frac{x_{ij}^{last} - x_{ij}^{first}}{t_i^{last} - t_i^{first} + 1} \tag{11}$$

where x_{ij}^{first} and x_{ij}^{last} represent the start and last recorded value of the observation window, respectively, Δx_{ij} reflects the change intensity of the variable during the observation period. For epidemiological risk prediction, this type of incremental information is often more

revealing than a single static measurement of the underlying upward trend in risk.

In the feature extraction stage, all variables are not regarded as equally effective, but reorganized according to the four subspaces of "basic attributes - physiological indicators - exposure information - dynamic changes". For multi-index variables such as environmental exposure and behavioral exposure, a single original value is difficult to completely reflect the risk level. Therefore, this paper constructs a weighted exposure intensity index:

$$E_i = \sum_{m=1}^{q} \eta_m \tilde{z}_{im} \tag{12}$$

where, $\tilde{z}_{im}$ represents the standardized result of sample i on the MTH exposure variable, η_m is the weight coefficient determined according to the stability and correlation of variables, q is the number of exposure variables, and E_i is the comprehensive exposure intensity. In this way, external risk factors that are scattered across multiple fields can be compressed into a more discriminative combination of variables. The system then concatenates the basic demographic features, window aggregation features, rate of change features, and exposure intensity features into a unified input vector:

$$h_i = [b_i \| p_i \| \Delta_i \| E_i] \tag{13}$$

Here, b_i represents the basic demographic vector, p_i represents the aggregated physiological and laboratory indicator vector, Δ_i represents the dynamic change feature vector, and the symbol $\|$ represents the vector splicing operation. This feature vector will be directly used as input to the subsequent optimized SVM. In terms of parameter setting, this paper does not regard the preprocessing process as a fixed step, but uses the validation subset to jointly adjust the size of the imputation neighborhood, the length of the time window and the exposure weight coefficient, so that the data processing chain is consistent with the classifier training chain. The purpose of this is to avoid the structural bias introduced by specifying the preprocessing parameters only empirically. The actual processing architecture of the whole data before entering the model is shown in Figure 4.

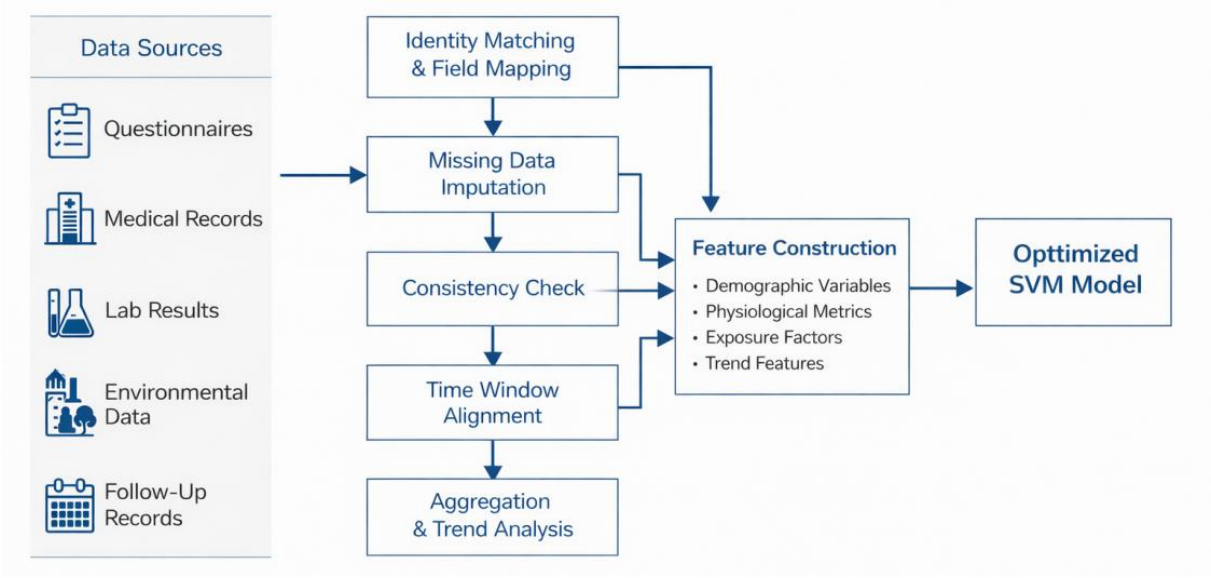


Figure 4: Architecture diagram of epidemiological data preprocessing and feature extraction

Figure 4 shows that the data processing in this paper is not simply "cleaned and fed into the model", but a structured process with a clear calculation hierarchy. Identity matching and field mapping are responsible for converging heterogeneous source data under a unified primary key, missing repair and consistency checking are responsible for ensuring input quality, time window alignment is responsible for eliminating sampling frequency differences, and feature recombination module transforms original records into compact representations suitable for support vector machine recognition. Through this process, the scattered, repeated and even conflicting information in the original epidemiological data is reorganized into a statistically comparable and machine learnable sample matrix, which provides a stable data foundation for the construction of subsequent risk prediction models.

2.3 Construction of epidemiological risk prediction model based on optimized support vector machine

Aiming at the coexistence of heterogeneous variables, fuzzy boundaries and unbalanced sample distribution in epidemiological risk identification, this paper further constructs an epidemiological risk prediction model based on optimized support vector machine (SVM) on the basis of the above data collection, preprocessing and feature extraction. In this model, support vector machine is not only used as the end classifier, but the structured sample representation, feature compression, class cost adjustment, probability mapping and risk stratification are integrated into the same computing framework, so that the original monitoring data can complete the stable transformation along the path of "input-discrimination-calibration-output". Compared with screening methods that rely solely on empirical thresholds, the proposed model can learn risk boundaries in high-dimensional variable space, and transform classification results into risk scores more suitable for public health applications through posterior probability calibration. Based on this, the optimized SVM epidemiological risk prediction model constructed in this paper is shown in Figure 5.

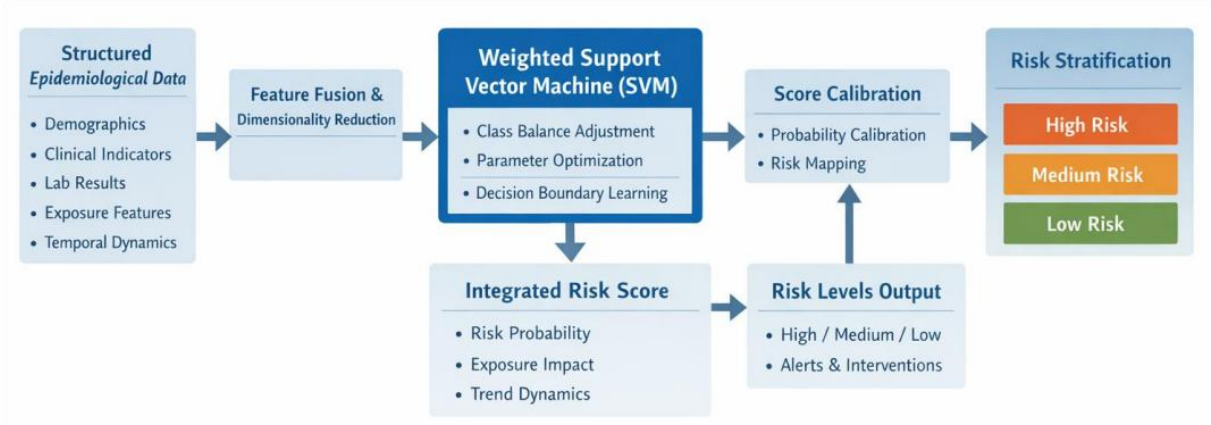


Figure 5: Structural diagram of the optimized SVM epidemiological risk prediction model

In Figure 5, the model takes the epidemiological structured feature matrix as input and enters the SVM training module after feature fusion and dimension compression. The fusion here is not just vector splicing, but mapping basic demographic variables, physical indicators, laboratory results, exposure characteristics and dynamic change characteristics into a unified risk representation space, so as to weaken the expression fragmentation between different data sources. Considering the low proportion of high-risk samples in the total samples, the class weight and parameter optimization mechanism are introduced in the model training stage to make the classification hyperplane give higher sensitivity to the minority class samples while

maintaining the overall stability. After training, the system does not directly output discrete labels, but generates continuous discrimination scores, and then converts them into risk probabilities through the probability calibration module, so as to provide a more fine-grained basis for subsequent hierarchical early warning and intervention decisions. In the support vector space, the normalized margin for sample i can be written as follows.

$$M_i = \frac{w^T \phi(h_i) + b}{\|w\|} \tag{14}$$

where h_i represents the fused feature vector of the ith sample, $\phi(\cdot)$ is the implicit feature mapping function, w is the normal vector of the classification hyperplane, b is the bias term, and M_i represents the directed interval of the sample with respect to the decision surface. This quantity not only reflects the class to which the sample belongs, but also characterizes the degree to which the sample is far from or close to the risk boundary. For epidemiological tasks, this is of practical interest because individuals near the boundary often correspond to key observations in which the clinical onset of the disease is not yet clear, but the risk factors have been clustered. In order to improve the interpretability of the model output, this paper further performs Sigmoid calibration on the discrimination scores to obtain the risk probability:

$$P_i = \frac{1}{1 + \exp(AM_i + B)} \tag{15}$$

where P_i is the high risk probability of sample i, and A and B are the calibration parameters estimated on the validation set. After this process, the interval value that originally only had classification significance was transformed into continuous risk probability, which was convenient to connect with the classification standard in public health management.

Considering that the single classification probability is still not enough to fully reflect the individual risk degree, this paper further combines the model output with the exposure intensity and dynamic change characteristics to construct a comprehensive risk score:

$$R_i = \rho_1 P_i + \rho_2 \widetilde{E}_i + \rho_3 \widetilde{\Delta}_i \tag{16}$$

where, R_i represents the individual comprehensive risk value, $\widetilde{E}_i$ is the normalized exposure intensity index, $\widetilde{\Delta}_i$ is the dynamic change characteristics after normalization, ρ_1, ρ_2, ρ_3 are the weight coefficients, and $\rho_1 + \rho_2 + \rho_3 = 1$ is satisfied. This design makes the model not only based on the static classification results, but also incorporates the current predicted probability and risk drivers into the scoring system, which is more in line with the actual process of "risk factor cumulation-state evolvation-event occurrence" in epidemiological surveillance. After obtaining the comprehensive risk value, the model performs risk stratification according to the preset threshold, and its rule is:

$$L_i = \begin{cases} \text{Low}, & R_i < \tau_1 \\ \text{Medium}, & \tau_1 \le R_i < \tau_2 \\ \text{High}, & R_i \ge \tau_2 \end{cases} \tag{17}$$

Here, L_i is the risk level of the ith sample, and τ_1 and τ_2 are the hierarchical thresholds determined based on the joint evaluation of the training set and the validation set. In this way, the output of the model was expanded from the original "high-risk or not" to the "low-medium-high" three-level risk results, which was more convenient for direct use in

grass-roots screening, key follow-up and intervention prioritization scenarios. Figure 6 shows the operation process of the optimized SVM risk prediction module.

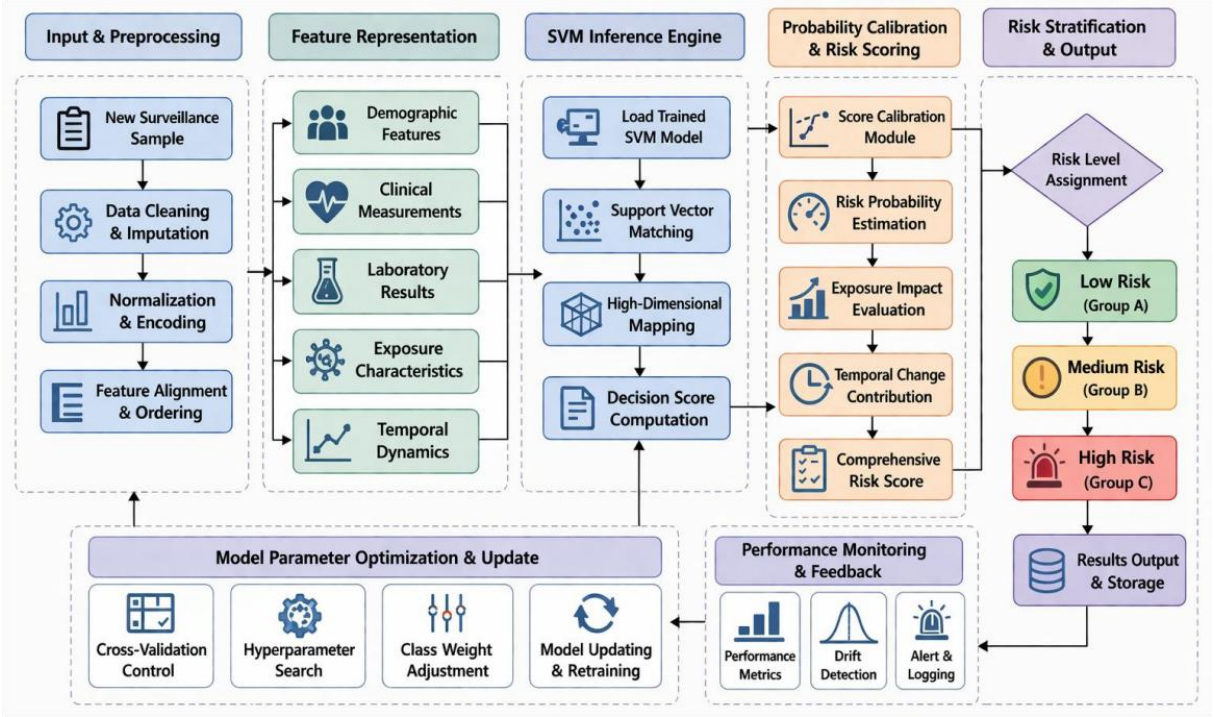


Figure 6: Operation flow chart of optimized SVM risk prediction module

Figure 6 shows that the feature vector generated by the new input sample is matched with the support vector structure formed in the training phase to calculate the discrimination margin in the high-dimensional space. The calibration function is subsequently used to map the interval values into high risk probabilities, which together with the exposure contribution and the time change contribution compose the comprehensive risk score. In this way, the output of the model is no longer limited to the abstract classification labels, but can provide the expression of risk intensity closer to the actual application requirements. For continuously arriving monitoring data, the system can also periodically update the parameter search results and threshold setting according to new samples, so that the model can maintain good adaptability in the case of regional population structure changes, exposure pattern adjustment or monitoring period extension.

3 Results

3.1 Prediction performance analysis of optimized SVM model

In order to verify the effectiveness of the optimized support vector machine model constructed in this paper in the epidemiological risk prediction task, a unified experimental platform is built and compared with the unoptimized standard support vector machine model. The experimental platform and core parameter configuration are shown in Table 2. Considering that epidemiological data has the characteristics of multiple dimensions of variables, unbalanced distribution of categories, and time update differences in some fields, the experimental stage not only focuses on the classification accuracy, but also investigates the comprehensive performance of the model in terms of recall ability, probabilistic ranking effect, resource consumption and response delay to determine whether it has the feasibility to

enter the real public health surveillance scenario.

Table 2: Experimental platform and optimized SVM model parameter configuration

Category	Parameter Name	Parameter Value
Hardware Platform	Processor	Intel Core i7-12700
	Memory	32 GB
	Storage	1 TB SSD
Software Environment	Operating System	Ubuntu 22.04
	Programming Language	Python 3.11
	Machine Learning Library	scikit-learn 1.4
Data Processing Parameters	Number of Neighbors for Missing Value Imputation	5
	Time Window Length	6 months
	Retained Feature Selection Ratio	70%
Optimized SVM Parameters	Kernel Function Type	RBF
	Penalty Coefficient Search Range	2^-3 to 2^7
	Kernel Parameter Search Range	2^-8 to 2^3
	Class Weight Strategy	Balanced
	Number of Cross-Validation Folds	10
Output Parameters	Risk Stratification Thresholds	0.35, 0.70

The experimental data came from the epidemiological risk prediction data set organized in this paper, covering community health records, physical examination records, laboratory test results, behavioral exposure information, environmental monitoring variables and 12-month follow-up outcomes. A total of 4680 valid samples were included, of which 31.24% were high-risk samples and 68.76% were medium-low risk samples. Each sample included 43 demographic variables, physical indicators, metabolic indicators, exposure indicators and dynamic change characteristics. In order to ensure the stable distribution of training and testing, this paper uses stratified sampling to conduct 10 rounds of repeated validation. In each round of experiment, the training set and the test set are divided by 8:2, and 10-fold cross validation is performed inside the training set to complete parameter optimization. The purpose of this process is to reduce the accidental fluctuations caused by a single partition as much as possible, so that the model performance evaluation is closer to its real application level.

Under the condition of gradually increasing sample size, the AUC and Macro-F1 changes of optimized SVM and standard SVM are shown in Figure 7.

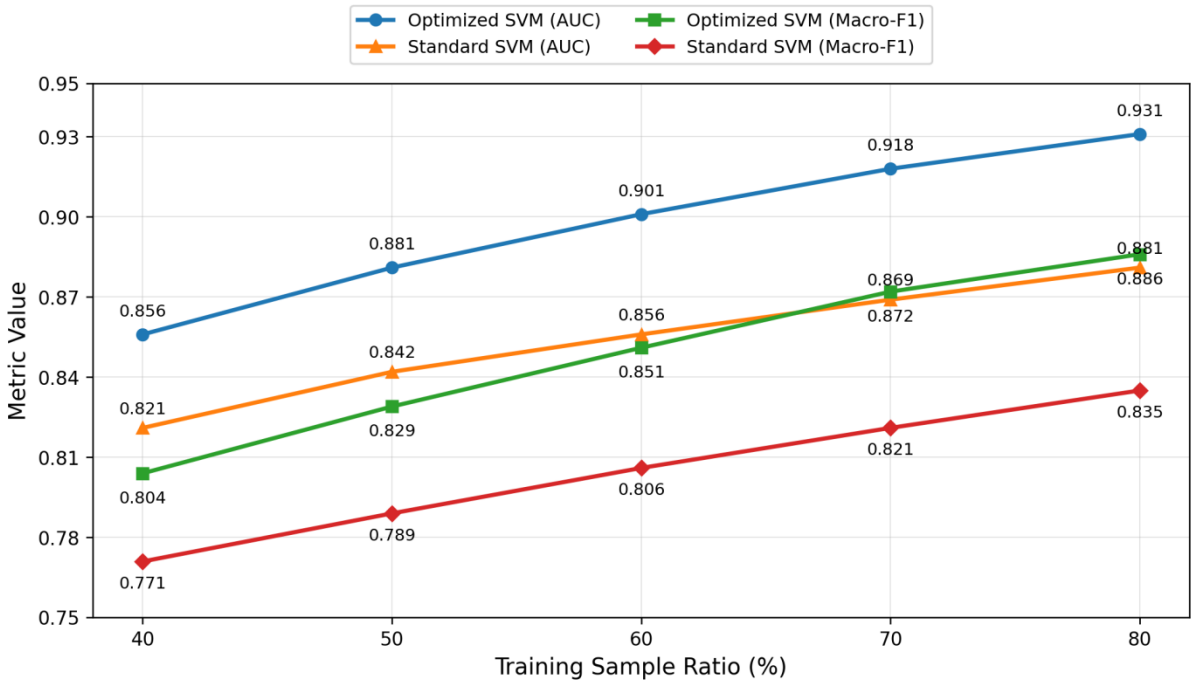


Figure 7: AUC and Macro-F1 variation of the two types of SVM models under different proportions of training samples

It can be seen from Figure 7 that as the proportion of training samples increases from 40% to 80%, the performance of both types of models shows a steady upward trend, but the optimized SVM maintains a higher level in all sample intervals. When the proportion of training samples is 80%, the AUC of the optimized model reaches 0.931, while the standard support vector machine is 0.881, and the difference is 0.050. The corresponding Macro-F1 are 0.886 and 0.835, respectively, indicating that the feature screening, category reweighting and parameter collaborative optimization strategies proposed in this paper not only improve the overall ranking ability, but also enhance the balanced recognition effect of the model for minority samples. It is worth noting that the optimized model still achieves an AUC of 0.856 under the sample proportion of 40%, which indicates that the method can form a relatively stable risk boundary under the condition of moderate sample size, and has strong adaptability to the limited sample size scenario common in epidemiology.

In order to further investigate the behavior of the model in the real risk output link, this paper counted the recall rate, precision rate and average response time under different risk thresholds, and the results are shown in Figure 8.

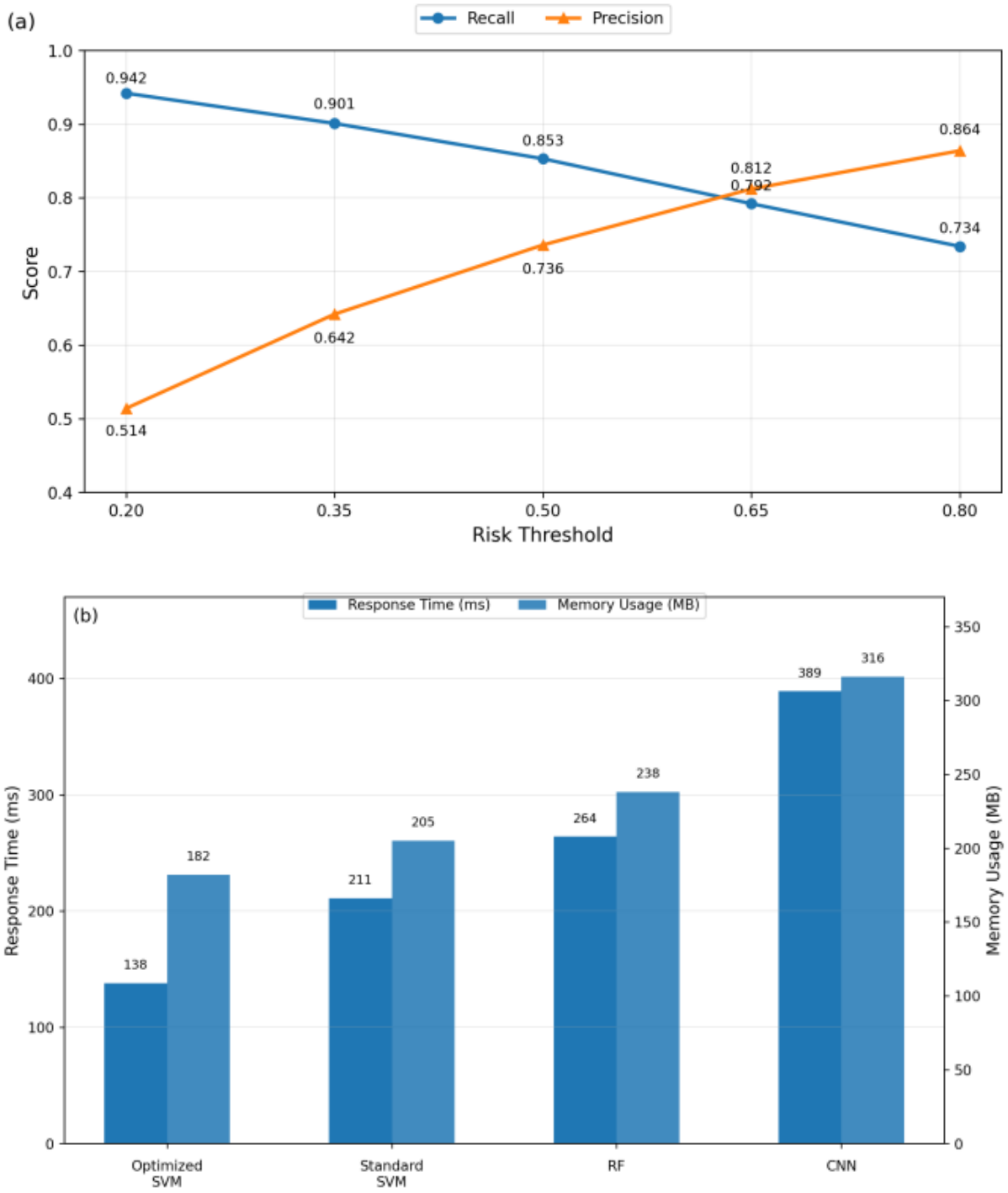


Figure 8: Comparison of recognition performance and computational efficiency of optimized SVM under different risk thresholds

Figure 8(a) shows that with the increase of risk threshold, the recall rate of the model gradually decreases, while the precision rate continues to increase, which is the typical performance of risk prediction models in screening tasks. When the threshold is set to 0.35, the recall rate is 0.901 and the precision rate is 0.642, which can control a certain degree of false positives while maintaining a high detection ability. When the threshold was increased to 0.65, the precision increased to 0.812, but the recall decreased to 0.792, which was more suitable for key screening under the condition of limited resources. Figure 8(b) reflects the

differences of different algorithms at the deployment level. The average single batch response time of the optimized SVM is 138 ms, and the memory usage is 182 MB, which are lower than the standard SVM, random forest, and convolutional neural network. Compared with the standard support vector machine, the proposed model does not increase the inference burden after adding the preprocessing and parameter optimization mechanism, but reduces the response time due to the reduction of redundant features and more stable boundary learning. The results show that the proposed model has both offline modeling ability and deployment feasibility in periodic monitoring scenarios.

In order to evaluate the classification performance of the optimized support vector machine from multiple dimensions, this paper further statistics its main indicators on the test set and compares it with the standard support vector machine. The results are shown in Table 3.

Table 3: Comparison of prediction performance between optimized SVM and standard SVM

Model	Accuracy / %	Precision / %	Recall / %	Specificity / %	Macro-F1 / %
Standard SVM	84.36	78.91	81.74	85.58	83.52
Optimized SVM	89.47	85.68	90.10	89.19	88.63

According to Table 3, the optimized SVM outperforms the standard model in the six indicators of Accuracy, Precision, Recall, Specificity, Macro-F1 and AUC. Among them, Recall increases from 81.74% to 90.10%, which means that the missed judgment rate of high-risk samples is effectively suppressed. For epidemiological risk prediction, this improvement has more practical value than simply improving the overall accuracy, because the public health system usually pays more attention to the timely identification of high-risk individuals, rather than the classification accuracy of low-risk samples. Macro-F1 is increased from 83.52% to 88.63%, which indicates that the proposed model maintains a better balanced performance under the condition of class imbalance.

Considering that epidemiological risk prediction often needs to face different types of population risk scenarios, we further test the adaptability of the optimized SVM on three types of sub-tasks, including high risk identification of chronic diseases, high risk identification related to environmental exposures, and risk identification of infection events. The results are shown in Table 4.

Table 4: Prediction effects of optimized SVM in different epidemiological risk tasks

Risk Task	Accuracy / %	Precision / %	Recall / %	Macro-F1 / %	AUC
High-Risk Identification of Chronic Diseases	90.12	87.45	91.38	89.21	0.938
High-Risk Identification Related to Environmental Exposure	88.74	84.93	89.57	87.16	0.922
Risk Identification of Infectious Events	89.56	84.66	89.35	87.88	0.927

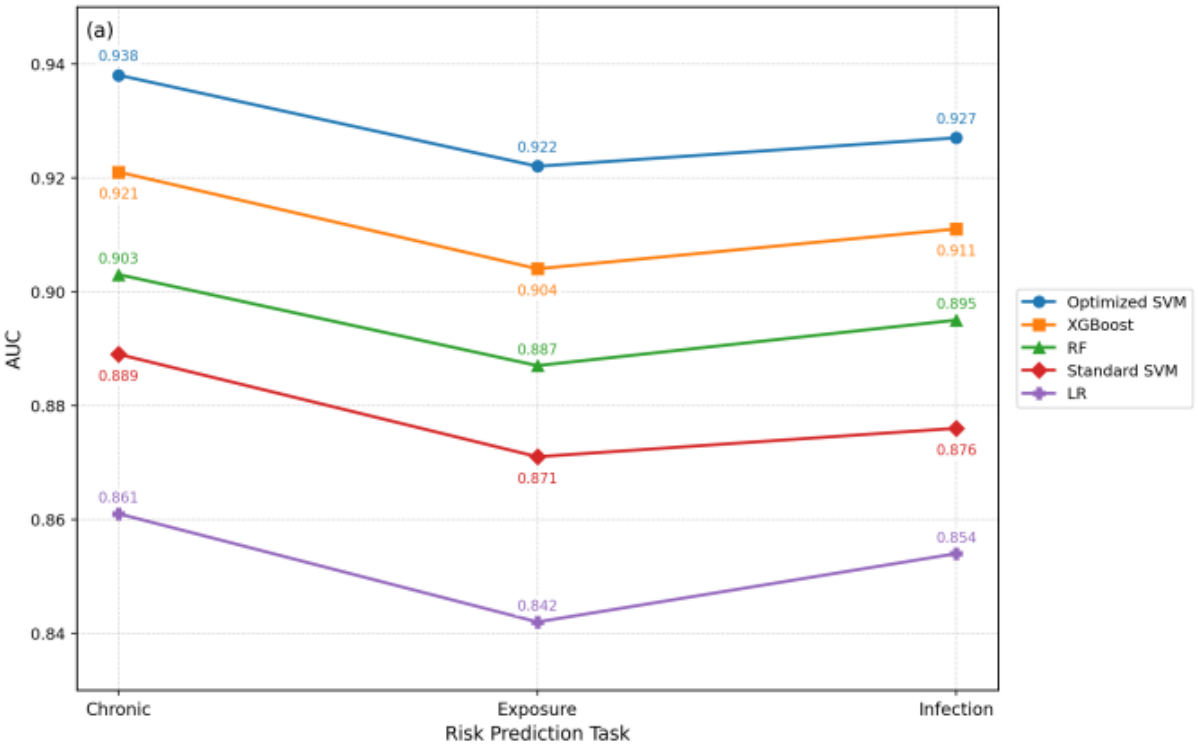
It can be seen from Table 4 that the optimized support vector machine maintains high performance in different risk tasks, and the high risk identification effect of chronic diseases is the best, with an AUC of 0.938, indicating that the model has a strong ability to characterize stable structured variables such as blood pressure, blood glucose, blood lipid, past history and behavior factors. The AUC of high risk identification related to environmental exposure was slightly lower, 0.922, because the external exposure variable in this task fluctuates more, and there is often a time delay effect between individual exposure level and outcome, which

makes the decision boundary more complex. The Recall of the infection event risk recognition task reaches 89.35%, which indicates that the model has a good response ability to short-term abnormal changes after adding the time window aggregation features and dynamic change features.

3.2 Comparison experiment of risk prediction model and analysis of comprehensive results

In order to further test the comprehensive advantages of the optimized support vector machine model constructed in this paper in epidemiological risk prediction, this paper sets up five sets of comparison experiments between logistic regression, random forest, XGBoost, standard support vector machine and the proposed model in the unified data division, the same feature input and consistent training environment. The purpose of the comparison is not only to determine which method is higher on a single index, but to examine the overall performance of the model under different risk tasks, different evaluation scales, and actual deployment conditions. Considering that epidemiological risk identification requires not only high classification Accuracy, but also good probability ranking ability and stable system response performance, we calculate the accuracy, Macro-F1, AUC, Brier Score, inference delay and high-risk recall at the same time in the experiment to form a more complete comparison framework.

The AUC and Macro-F1 performance of different models in the multi-task scenario are shown in Figure 9. Here, three representative tasks are selected, including high risk identification of chronic diseases, high risk identification related to environmental exposures, and risk identification of infection events, to cover the typical epidemiological application scenarios dominated by structured variables, dominated by external exposure variables, and with obvious characteristics of time series fluctuations.



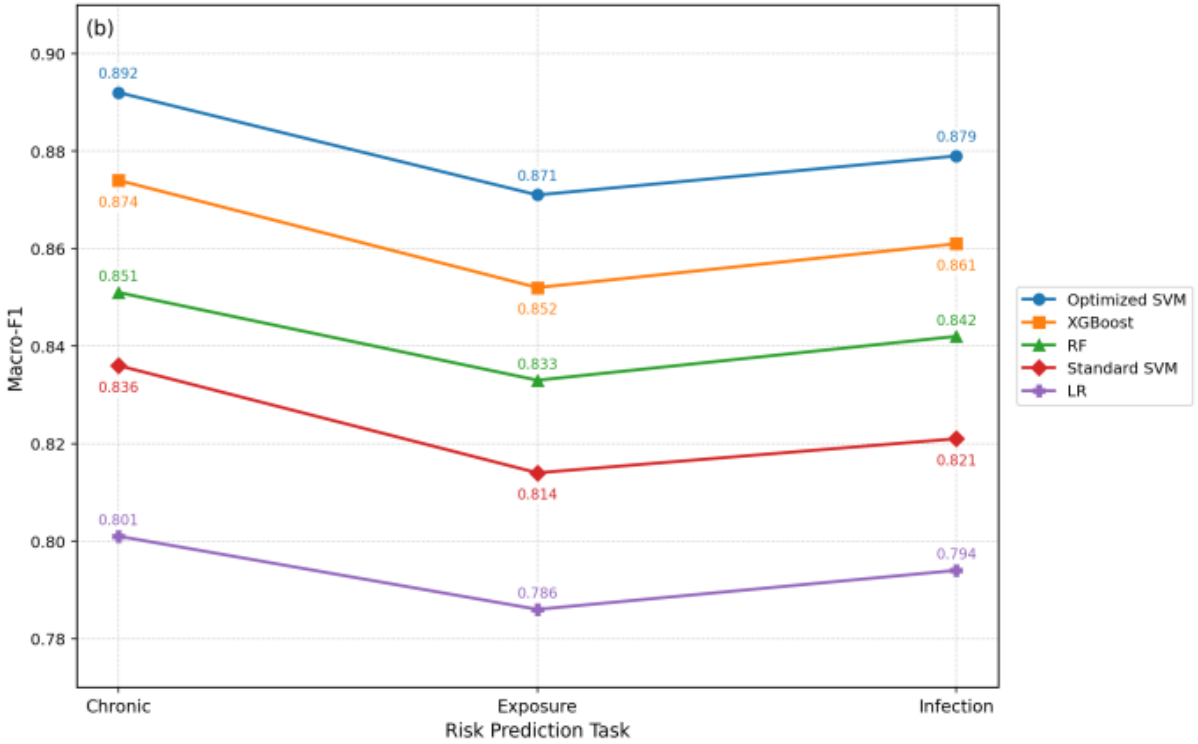


Figure 9: Comparison of AUC and Macro-F1 for different models in the multi-class epidemiological risk task

It can be seen from Figure 9 that the proposed model keeps ahead in the three types of tasks, and the AUC of the chronic disease high-risk identification task reaches 0.938, which is 0.049 higher than that of the standard support vector machine and 0.077 higher than that of the logistic regression. In the task of environmental exposure related risk identification, the Macro-F1 of the proposed model is 0.871, which is still higher than 0.852 of XGBoost and 0.833 of random forest. The results show that after the adjustment of class weight, feature selection and joint optimization of kernel parameters, the ability of SVM to describe complex nonlinear boundaries is enhanced, especially in the epidemiological scenarios with strong correlation of variables and imperfectly balanced sample distribution, its advantage is more obvious than that of linear model. Compared with the tree model, although the proposed method does not rely on a deeper ensemble structure, it performs more stable in terms of high-risk sample recall and class balance.

In addition to prediction performance, probabilistic calibration quality and computational cost should also be considered when the model enters the real monitoring system. To this end, this paper further compares the Brier Score of each model with the average inference delay, and the results are shown in Figure 10. The lower the Brier Score, the closer the model output probability is to the true risk distribution. The shorter the inference latency, the more conducive it is to deploy to periodic refresh or batch alert systems.

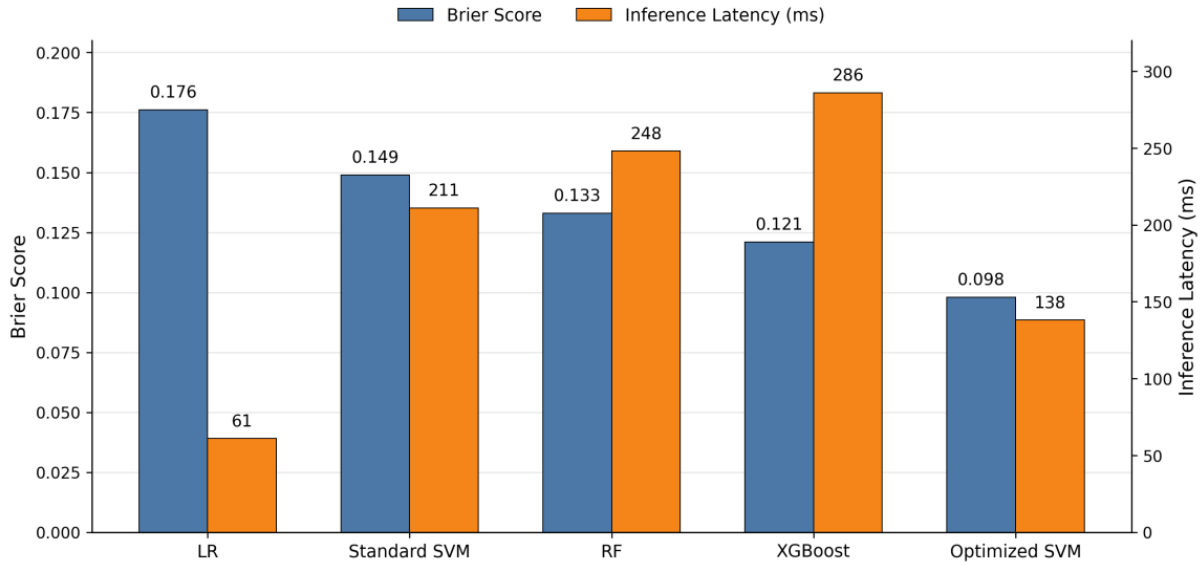


Figure 10: Comparison of probabilistic calibration error and average inference delay for different models

Figure 10 shows that the Brier Score of the proposed model is 0.098, which is the lowest among all comparison methods, indicating that its risk probability output is closer to the real outcome distribution, which is particularly critical for the hierarchical management of public health. In practical applications, decision-makers usually do not only care about "whether the risk is high", but also care about "how high the risk is and whether it needs to intervene in advance". At the same time, the average inference delay of the proposed model is 138 ms, which is significantly lower than that of random forest and XGBoost, and only higher than that of logistic regression with the simplest structure. Combining Figure 9 and Figure 10, it can be seen that although logistic regression runs fastest, it is insufficient in fitting complex boundaries. The tree model has high accuracy in some tasks, but brings higher computational burden. The standard support vector machine has a certain degree of balance, but it is not enough to capture the high-risk samples under the condition of class imbalance. In contrast, the proposed model achieves a more reasonable compromise between performance and efficiency.

To comprehensively show the overall performance of different models on key indicators, Table 5 presents the core results of the five methods.

Table 5: Comprehensive performance comparison of different risk prediction models

Model	Accuracy / %	Precision / %	Recall / %	Macro-F1 / %	AUC	Brier Score	Average Inference Latency / ms
Logistic Regression	82.94	77.36	78.42	79.37	0.852	0.176	61
Random Forest	87.18	82.64	85.31	84.20	0.895	0.133	248
XGBoost	88.46	84.95	87.28	86.29	0.912	0.121	286
Standard SVM	84.36	78.91	81.74	83.52	0.881	0.149	211
Optimized SVM	89.47	85.68	90.10	88.63	0.931	0.098	138

According to Table 5, the proposed model achieves the best results in the five core indicators of Accuracy, Recall, Macro-F1, AUC and Brier Score. Among them, the Recall reached 90.10%, indicating that the model was more effective in controlling the missed judgment of high-risk individuals. The AUC reached 0.931, indicating that its ability to sort samples with different risk levels was stronger. The Brier Score drops to 0.098, which further indicates that the model output probability has good calibration. Compared with XGBoost, the proposed model improves the Accuracy by 1.01 percentage points, improves the Recall by 2.82 percentage points, and reduces the inference delay by 148 ms. This means that optimizing SVM does not simply increase the complexity of the model in exchange for performance, but achieves more effective computational optimization in three levels: feature organization, class balance and kernel space learning.

4 Discussion

The optimized support vector machine epidemiological risk prediction model constructed in this paper is superior to the comparison methods in terms of prediction accuracy, risk recall, probability calibration and computational efficiency. The experimental results show that the Accuracy of the model reaches 89.47%, the Recall reaches 90.10%, the AUC reaches 0.931, the Brier Score is reduced to 0.098, and the average inference delay is 138 ms, indicating that the model can not only identify high-risk individuals more accurately. It also has the computational feasibility to enter the actual monitoring system. Compared with the standard support vector machine, logistic regression and tree model, the proposed method has a more significant improvement in the detection rate of high-risk samples and class balance performance, which is particularly critical for epidemiological scenarios, because public health screening pays more attention to miss detection control, rather than simply pursuing the overall classification accuracy.

The model performance improvement mainly comes from the synergy of three aspects. First, feature selection and variable reorganization enhance the discrimination of the input space, so that demographic features, physiological indicators, exposure information and dynamic change features can form a clearer risk boundary in the unified representation, and reduce the perturbation of redundant variables on the classification plane. Secondly, the class reweighting mechanism improves the learning bias under the condition of sample distribution imbalance, which significantly enhances the sensitivity of the model on the minority class samples, thus improving the recall ability of high-risk individuals. Thirdly, the joint optimization of kernel parameters and probability calibration strategy improve the stability of model output, so that the support vector machine no longer stays in discrete classification, but can output continuous scores closer to the real risk distribution, which is an important reason for the obvious decline of Brier Score. Combined with the above experimental results, it can be seen that the improvement of the optimized support vector machine compared with the standard model in AUC and Macro-F1 is not a local fluctuation, but is directly related to the improvement of feature organization, boundary learning and output mapping. However, this model still has some limitations. First, although the current experimental data cover multi-source epidemiological variables, the samples are still mainly from structured monitoring records, and the adaptability to cross-regional and cross-institutional data distribution drift still needs to be further verified. Second, risk label construction relies on existing follow-up outcomes, and model boundaries may be affected in scenarios with short observation Windows or delayed outcomes. Third, although support vector machine has advantages in medium scale samples and high dimensional features, its training efficiency and online update ability still need to be optimized when the monitoring data continues to expand

to larger scale. Further research can be carried out in the directions of federated learning, incremental update and transfer modeling to improve the generalization ability and deployment flexibility of the model in complex public health environments.

5 Conclusion

In order to improve the accuracy of risk identification in epidemiological data and the stability of the model, this paper constructs an optimized risk prediction model for multi-source monitoring data around the support vector machine method, and completes the system design from the aspects of data acquisition, preprocessing, feature extraction, kernel parameter optimization, class reweighting and risk stratification output. In this study, demographic variables, physical signs, laboratory results, exposure information and dynamic change characteristics are integrated into the computational framework, so that the support vector machine is no longer a single classification tool, but can undertake continuous risk scoring and grading early warning tasks. Experimental results show that the Accuracy of the proposed model reaches 89.47%, the Recall reaches 90.10%, the AUC reaches 0.931, the Brier Score is reduced to 0.098, and the average inference delay is 138 ms. The proposed model is superior to the standard support vector machine and other comparison models in terms of prediction performance and computational efficiency. This shows that the proposed optimization strategy can effectively improve the modeling difficulties caused by the high-dimensional, nonlinear and class imbalance problems in epidemiological data. In general, the proposed method can provide a relatively stable computational support for primary screening, key population identification and regional public health surveillance. However, the model is still mainly validated based on structured monitoring data, and its adaptability to cross-regional sample distribution drift and long-term follow-up scenarios still needs to be further tested. Future research can combine federated learning, incremental update and transfer modeling mechanisms to improve the generalization ability of the model in a multi-institution environment. At the same time, longer-term longitudinal data can be introduced to further improve the dynamic risk prediction system for real public health applications.

Acknowledgements

This work was supported by Seismic Analysis and Optimization of Multi story and High rise Steel Structures -Taking the Standardized Factory Building Project of Minning Textile Industrial Park as an Example.(2022 General Project of Scientific Research in Higher Education Institutions in Ningxia)

Jinwu Liu was Born in 1982 in Shuozhou City, Shanxi Province, China. He graduated from Shanxi Medical University. He works at the Supervision and Inspection Center of Shanxi Provincial Health Commission. He is an intermediate nutritionist whose research areas include health management, hygiene supervision and inspection, medical and health supervision, supervision of drinking water hygiene, and supervision of infectious disease prevention and control. E-mail: spark312@163.com

References

[1] Guido R, Ferrisi S, Lofaro D, et al. An overview on the advancements of support vector machine models in healthcare applications: a review[J]. Information, 2024, 15(4): 235.

[2] Islam R, Sultana A, Islam M R. A comprehensive review for chronic disease prediction using machine learning algorithms[J]. Journal of Electrical Systems and Information Technology, 2024, 11(1): 27.

[3] Teshale A B, Htun H L, Vered M, et al. A systematic review of artificial intelligence models for time-to-event outcome applied in cardiovascular disease risk prediction[J]. Journal of medical systems, 2024, 48(1): 68.

[4] Cai Y Q, Gong D X, Tang L Y, et al. Pitfalls in developing machine learning models for predicting cardiovascular diseases: challenge and solutions[J]. Journal of Medical Internet Research, 2024, 26: e47645.

[5] Späth J, Sewald Z, Probul N, et al. Privacy-preserving federated survival support vector machines for cross-institutional time-to-event analysis: Algorithm development and validation[J]. JMIR AI, 2024, 3(1): e47652.

[6] You J, Guo Y, Kang J J, et al. Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study[J]. Stroke and vascular neurology, 2023, 8(6).

[7] Fu Q, Wu Y, Zhu M, et al. Identifying cardiovascular disease risk in the US population using environmental volatile organic compounds exposure: A machine learning predictive model based on the SHAP methodology[J]. Ecotoxicology and environmental safety, 2024, 286: 117210.

[8] Guo K, Ni W, Du L, et al. Environmental chemical exposures and a machine learning-based model for predicting hypertension in NHANES 2003–2016[J]. BMC Cardiovascular Disorders, 2024, 24(1): 544.

[9] Al Mashrafi S S, Tafakori L, Abdollahian M. Predicting maternal risk level using machine learning models[J]. BMC Pregnancy and Childbirth, 2024, 24(1): 820.

[10] Tang Y, Liu Y, Du Z, et al. Prediction of coronary artery lesions in children with Kawasaki syndrome based on machine learning[J]. BMC pediatrics, 2024, 24(1): 158.

[11] Wu Y, Xiang C, Wang Z, et al. Interpretable prediction models for disability in older adults with hypertension: the Chinese Longitudinal Healthy Longevity and Happy Family Study[J]. Psychogeriatrics, 2024, 24(3): 645-654.

[12] Briggs E, de Kamps M, Hamilton W, et al. Machine learning for risk prediction of oesophago-gastric cancer in primary care: comparison with existing risk-assessment tools[J]. Cancers, 2022, 14(20): 5023.

[13] Shoko C, Sigauke C. Short-term forecasting of COVID-19 using support vector regression: An application using Zimbabwean data[J]. American Journal of Infection Control, 2023, 51(10): 1095-1107.

[14] Zhang T, Rabhi F, Chen X, et al. A machine learning-based universal outbreak risk prediction tool[J]. Computers in biology and medicine, 2024, 169: 107876.

[15] Wang J, Wang G, Wang Y, et al. Development and evaluation of a model for predicting

the risk of healthcare-associated infections in patients admitted to intensive care units[J]. Frontiers in Public Health, 2024, 12: 1444176.

[16] Chen Y, Zhang Y, Nie S, et al. Risk assessment and prediction of nosocomial infections based on surveillance data using machine learning methods[J]. BMC Public Health, 2024, 24(1): 1780.

[17] Allen A, Iqbal Z, Green-Saxena A, et al. Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus[J]. BMJ open diabetes research & care, 2022, 10(1).

[18] Wang Y, Yao H X, Liu Z Y, et al. Design of machine learning algorithms and internal validation of a kidney risk prediction model for type 2 diabetes mellitus[J]. International Journal of General Medicine, 2024: 2299-2309.

[19] Wang H, Tucker W J, Jonnagaddala J, et al. Using machine learning to predict cardiovascular risk using self-reported questionnaires: Findings from the 45 and Up Study[J]. International Journal of Cardiology, 2023, 386: 149-156.

[20] Dalal S, Goel P, Onyema E M, et al. Application of machine learning for cardiovascular disease risk prediction[J]. Computational Intelligence and Neuroscience, 2023, 2023(1): 9418666.

[21] Chinnasamy P, Kumar S A, Navya V, et al. Machine learning based cardiovascular disease prediction[J]. Materials Today: Proceedings, 2022, 64: 459-463.

[22]