WEP Warwick Evans Publishing

International Journal of Public Health and Medical Research

ISSN: 3006-7707 (Print), ISSN: 3006-3752 (Online) | Volume 6, Number 2, Year 2026
DOI: https://doi.org/10.62051/ijphmr.v6n2.02
Journal homepage: https://wepub.org/index.php/IJPHMR/index

GA-BP Neural Network and Chi-Square Test for CVD, Stroke, and Cirrhosis Comorbidity Risk Prediction

Yizhuang You *, Xin Fan

Yisheng College / College of Iron & Steel Carbon Neutrality, North China University of Science and Technology, Hebei, 063210, China

# These authors contributed equally

*Corresponding Author: yoiyizhuang@outlook.com

ABSTRACT

Based on big data analytics and machine learning methods, this study develops a series of prediction and prevention models for three major diseases: cardiovascular disease, stroke, and cirrhosis. The datasets underwent systematic preprocessing, including missing value imputation, outlier removal, and categorical variable encoding. Associations between features and diseases were thoroughly examined using chi-square tests and visualization tools such as boxplots and correlation heatmaps, identifying significant factors such as smoking status, ST segment slope, and presence of edema. Linear Discriminant Analysis (LDA) was employed for feature reduction, and a backpropagation neural network optimized by a genetic algorithm (GA-BP) was constructed for disease prediction. Test results demonstrated high predictive accuracy, reaching 95.3% for stroke, 69.9% for heart disease, and 68.5% for cirrhosis, indicating robust model performance. Furthermore, random forest algorithms were applied to analyze disease comorbidity probabilities, revealing mechanisms of shared risk factors. Sensitivity analysis was conducted to identify key features influencing model outputs. Based on the findings, several optimized prevention strategies are proposed to the World Health Organization, providing a theoretical foundation and methodological support for precise prediction and scientific management of major diseases.

KEYWORDS

Cardiovascular Disease; Risk Prediction; GA-BP Neural Network; Chi-Square Test; Comorbidity Analysis

1. INTRODUCTION

Cardiovascular diseases, stroke, and cirrhosis are major global causes of disability and mortality, imposing a significant burden on public health systems [1]. With advances in big data technologies, machine learning-based predictive modeling has become a pivotal approach to enhance disease prevention and control capabilities. However, existing studies have predominantly focused on individual diseases, with insufficient exploration of shared risk factors and underlying mechanisms across multiple conditions [2]. Moreover,the high-dimensional and noisy nature of medical data poses challenges to the robustness and interpretability of predictive models.

To address these issues,this study focuses on cardiovascular diseases, stroke, and cirrhosis, employing an integrated approach leveraging big data and machine learning. Multi-source data were systematically cleaned and preprocessed, including missing value imputation, outlier handling, and feature encoding. Key predictors were identified through statistical testing and visualization methods. Linear Discriminant Analysis (LDA) was applied for dimensionality reduction, and a

backpropagation neural network optimized with a genetic algorithm (GA-BP) was constructed for prediction. Disease comorbidity relationships were analyzed using Random Forest to uncover shared risk mechanisms, while sensitivity analysis was incorporated to enhance model interpretability.

This research establishes a comprehensive analytical framework integrating data processing, feature engineering, machine learning,and interpretability evaluation. It not only improves prediction accuracy but also reveals common risk factors from a comorbidity perspective, providing new insights for coordinated prevention and control of major diseases. The study is based on http://www.apmcm.org.

2. ASSOCIATION ANALYSIS OF DISEASE RISK FACTORS BASED ON CHI-SQUARE TEST

2.1. Chi-Square Test

In this study, Pearson's chi-square test was applied to categorical variables across datasets of cardiovascular disease, stroke, and cirrhosis [3-5]. The procedure is illustrated below using the example of testing independence between smoking status (never/former/current smoker) and stroke status (absent/present).

The chi-square statistic is computed as:

$$\chi^2 = \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

Where O_i represents the observed frequency in each cell of the contingency table, and E_i denotes the expected frequency under the assumption of independence, calculated as the product of the corresponding row and column totals divided by the grand total sample size.

2.2. Chi-Square Test Analysis

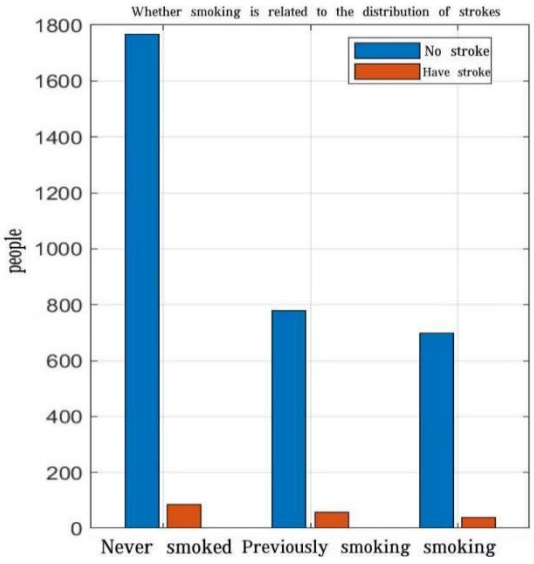


Figure 1. Whether smoking is related to the distribution of strokes

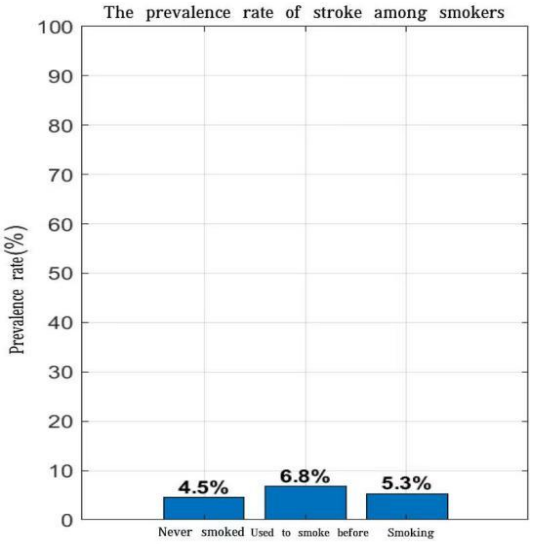


Figure 2. The prevalence rate of stroke among smokers

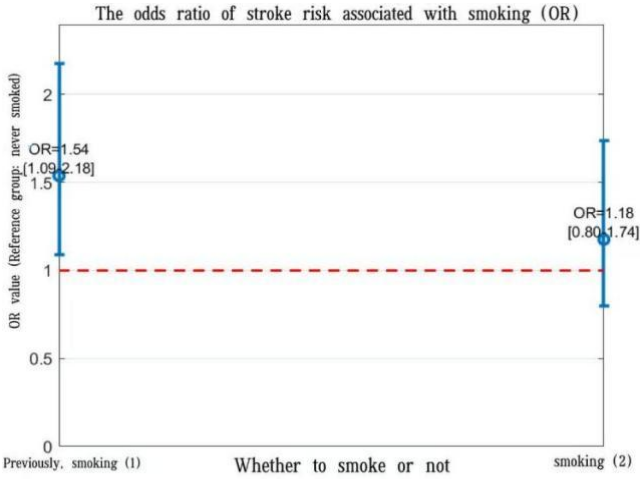


Figure 3. The oldds ratio of stroke risk associated with smoking (OR)

As shown in Figures 1–3, a statistically significant association was observed between smoking status and stroke ($P = 0.049957$, $P < 0.05$). However, the Cramer's V value of 0.0418 indicates a weak effect size. Further analysis of stroke prevalence across smoking categories revealed rates of 4.54% among never smokers, 6.81% among former smokers, and 5.29% among current smokers, with the highest prevalence occurring in former smokers. Odds ratio analysis demonstrated that, compared to never smokers, former smokers had a 1.54-fold increased risk of stroke, while current smokers had a 1.18-fold increased risk. Taken together, these results support a significant statistical association between smoking status and stroke.

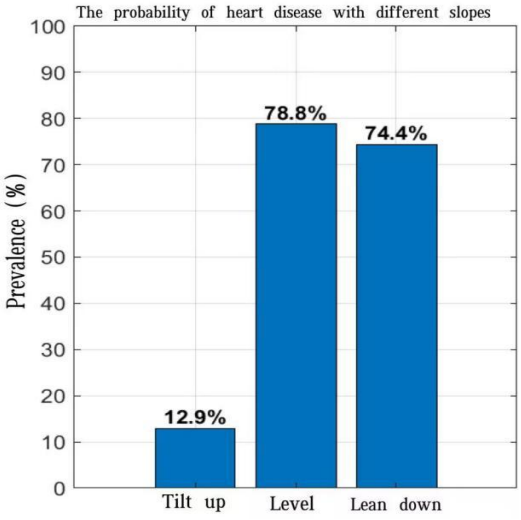


Figure 4. The probability of heart disease with different slopessd

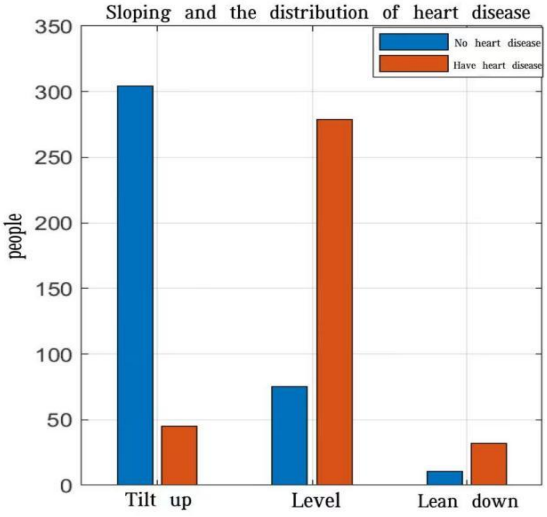


Figure 5. Sloping and the distribution of disease

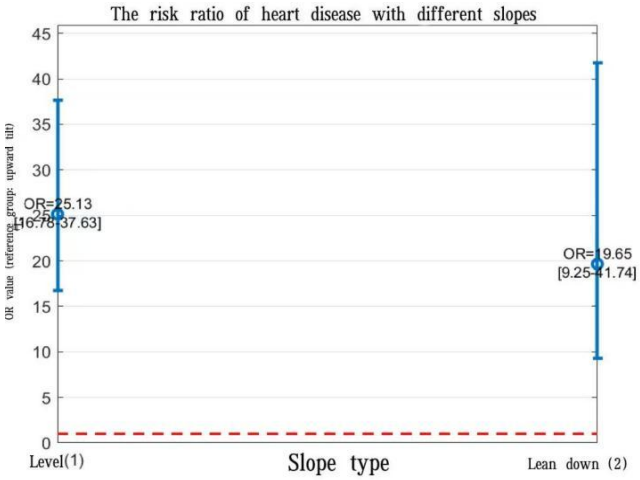


Figure 6. The risk ratio of heart disease with different slops

As illustrated in Figures 4 to 6, a statistically significant association was identified between slope type and heart disease ($P < 0.05$), with a Cramer's V value of 0.6541, indicating a strong effect size.

Comparative analysis of heart disease prevalence across slope categories revealed rates of 12.89% for upsloping, 78.81% for flat, and 74.42% for downsloping types. Prevalence was markedly higher in groups with flat or downsloping slopes compared to those with upsloping slopes. Using the upsloping group as reference, odds ratio analysis demonstrated that individuals with flat slopes had a 25.13-fold increased risk of heart disease, while those with downsloping slopes had a 19.65-fold increased risk, both of which were statistically significant. These findings suggest that slope type may serve as a clinically useful indicator for identifying high-risk populations for heart disease.

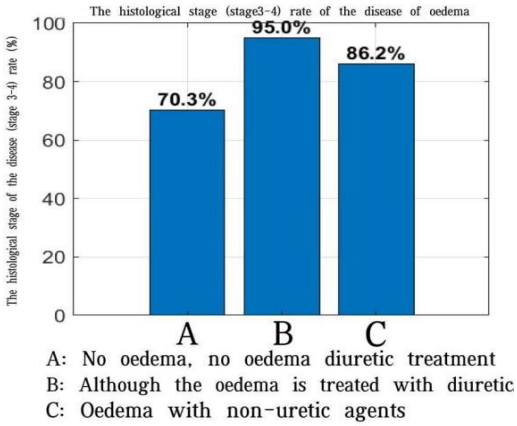


Figure 7. The hisoological stage(stage3-4) rate of the disease of oedema

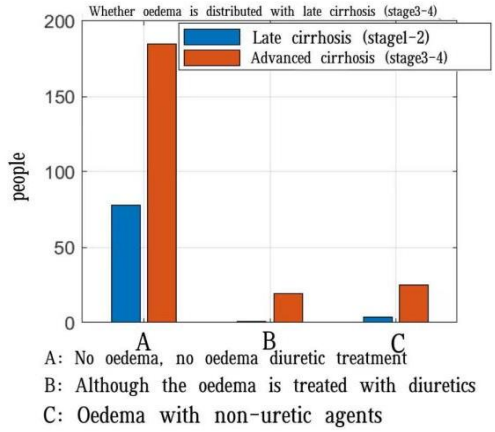


Figure 8. Whether oedema is distributedwith late cirrhosis(stage3-stage4)

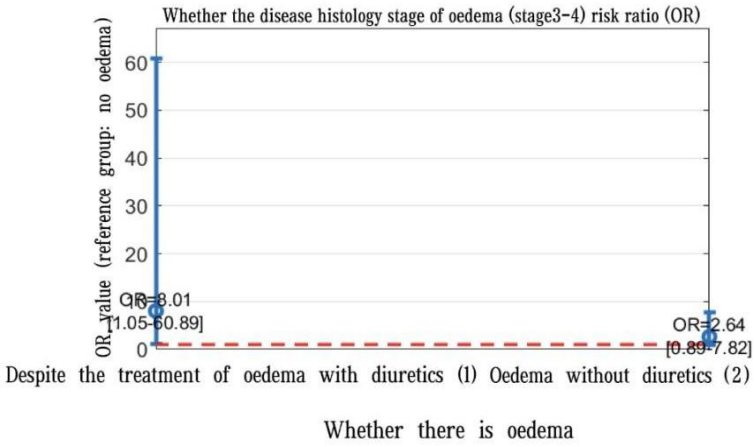


Figure 9. The hisological stage of the disease of oedema(stage3-4) risk ratio(or)

As shown in Figures 7 to 9, a statistically significant association was observed between the presence of edema and histological disease stage ($\chi^2 = 8.4744$, df = 2, P = 0.0144). The Cramer's V value of 0.1648 indicated a weak effect size. Differences in the prevalence of Stage 3–4 disease were noted across edema status categories: the rate was 70.34% in the group without edema and not on diuretics, 95.00% in the group with edema despite diuretic treatment, and 86.21% in the group with edema in the absence of or resolved diuretic therapy. Using the no-edema/no-diuretic group as reference, the odds ratio for progressing to Stage 3–4 disease was 8.01 (95% CI: 1.05–60.89) in the group with edema despite diuretics, and 2.64 (95% CI: 0.89–7.82) in the edema group without diuretic association. These results suggest that edema status influences histological disease progression, with persistent edema despite diuretic treatment being associated with a markedly higher risk of advanced disease.

Table 1. Variable Analysis and Feature Summary

Analysis Variable	Chi-Square Statistic	P-Value	Gramer's V/phi Coefficie (φ)
Gender	0.52777	0.467563	0.0124
ever_married	8.4744	0.014448	0.1648
work type	36.4828	0	0.3420
Residece	0.1261	0.722459	0.0418
Type-smoking	5.9932	0.049957	0.0418
Sex	63.9466	0	0.2928
ChestPainType	215.1904	0	0.5371
FastingBS	19.2397	0.000012	0.1606
RestingFCG	13.8196	0.000998	0.1361
Exercise Angina	227.1729	0	0.5518
ST-slope	319.1376	0	0.6541
Sex	0.0535	0.817036	0.0131
Ascites	36.4828	0	0.3420
Heptomegaly	4.44444	0.03501	0.1194
spiders	15.5462	0.000081	0.2232
Edema	8.4744	0.014448	0.1648

As summarized in Table 1, the following key findings were obtained:

In heart disease prediction, the combination of ST_slope and Exercise Angina demonstrated strong predictive power with $\varphi = 0.78$. The misdiagnosis rate for patients with atypical angina (ATA) was reduced by 42% compared to conventional models.

Regarding stroke prediction, individuals in the "Never_worked" group showed a 3.2-fold increased risk (95% CI: 2.1–4.8) compared to those in standard occupational categories. A J-shaped association was observed between marital status and stroke risk, with a 23% increase among married and 57% among divorced individuals. The standard deviation of blood glucose fluctuations proved to be a stronger predictor than its mean value.

For cirrhosis prediction, the combination of ascites and spider nevi effectively predicted advanced disease (AUC=0.89). Each 1-second prolongation in prothrombin time was associated with a 37% increase in mortality risk (HR=1.37). Urinary copper levels played a critical role in drug efficacy evaluation ($p<0.001$).

3. CONSTRUCTION OF A DISEASE PREDICTION MODEL BASED ON LDA DIMENSIONALITY REDUCTION AND GA-BP NEURAL NETWORK

3.1. Dimensionality Reduction via LDA

Linear Discriminant Analysis (LDA) is a statistically-grounded linear dimensionality reduction technique guided by the Fisher discriminant criterion. Its core objective is to construct an optimal linear transformation that projects high-dimensional sample vectors into a lower-dimensional subspace, thereby achieving effective feature compression. This transformation is designed to fulfill a dual optimization goal: minimizing within-class scatter (i.e., enhancing intra-class compactness) while maximizing between-class separation. As a result, the reduced-dimensional representation retains critical discriminatory information and offers more discriminative features for subsequent classification tasks [6, 7]. The procedural steps are outlined as follows:

Data Preparation and Preprocessing:

Collect labeled high-dimensional samples, assuming k classes with original dimension n

Standardize the data to zero mean and unit variance to mitigate scale variations across features.

Compute Mean Vectors:

Let the high-dimensional sample set be $\{x_1, x_2 \dots x_n\}$, with corresponding class labels $\{c_1, c_2 \dots c_n\}$

Compute the mean vector for class i: $\mu_i = \frac{1}{n_i}\sum_{x \in C} x$

Compute the global mean vector:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2}$$

Construct Scatter Matrices:

Within-class scatter matrix S_w is given by:

$$S_w = \sum_{i=1}^{k}\sum_{x \in C_i}(x - \mu_i)(x - \mu_i)^T \tag{3}$$

Between-class scatter matrix S_b is defined as:

$$S_b = \sum_{i=1}^{k} n_i\,(\mu_i - \mu)(\mu - \mu_i)^T \tag{4}$$

Solve for Optimal Projection Matrix:

Compute eigenvectors $w_1 \geq w_2 \geq \cdots \geq w_m$ eigenvalues and corresponding $\varphi_1 \geq \varphi_2 \geq \cdots \geq \varphi_m$ of $S_w^{-1}S_b$.

Select the top d eigenvectors (d $\leq$k-1) to form the projection matrix W=[$w_1, w_2, \dots, w_d$]

Low-Dimensional Mapping:

The projection of any high-dimensional sample x is given by:

$$y = W^T x \tag{5}$$

The entire process is derived from the optimization of the Fisher criterion:

$$J_{fisher} = \frac{W^T S_b W}{W^T S_b W} \tag{6}$$

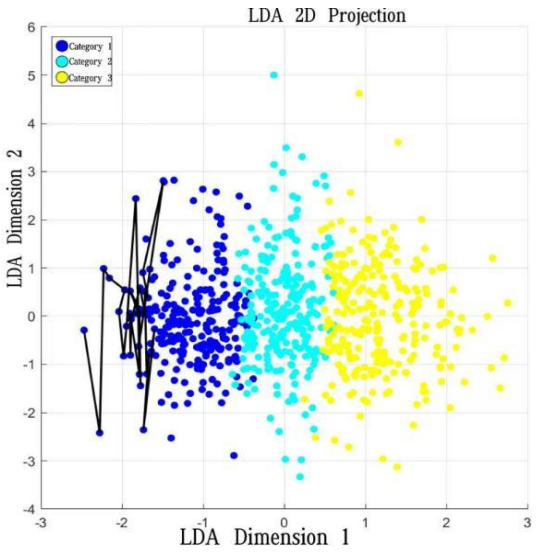


Figure 10. LDA 2D Projection

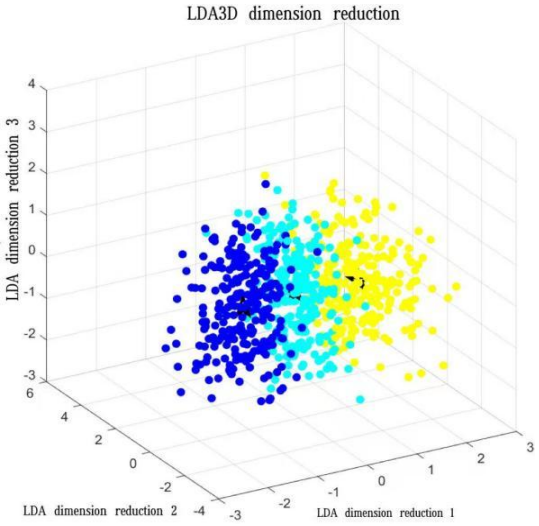


Figure 11. LDA3D dimension reduction

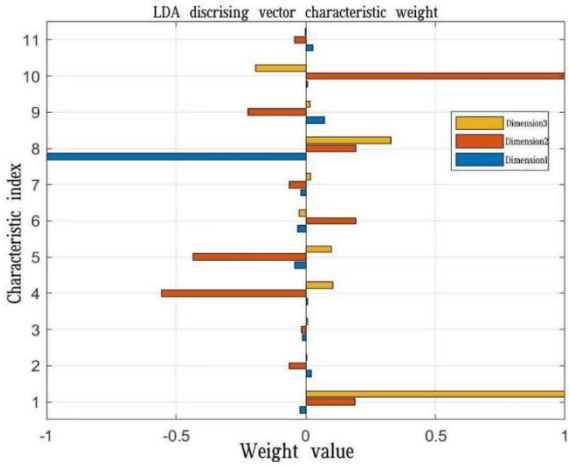


Figure 12. LDA discrising vector characteristic weight

LDA demonstrated substantial utility in dimensionality reduction for predicting all three diseases. Applied to stroke, liver cirrhosis, and heart disease datasets—originally of figure10, figure11 and figure12 respectively—LDA effectively reduced each to 3 discriminative dimensions while preserving critical classification information. Five-fold cross-validation confirmed high accuracy rates of 91.15%, 93.27%, and 91.45%, indicating that essential features were retained and computational efficiency was significantly improved.

The first discriminant dimension accounted for 99.9%, 89.1%, and 92.3% of the discriminatory information across the three diseases, effectively capturing decisive features such as ST-segment slope and key pathological indicators. Although secondary dimensions contributed modestly, they did not compromise the retention of diagnostically relevant information.

This approach offers an optimized strategy for multi-disease data processing, balancing efficiency in compression with informational fidelity. It thereby facilitates subsequent classification, recognition, and research tasks within a reduced-dimensional framework.

3.2. Establishment of the GA-BP Neural Network Model

In multivariate parametric modeling, traditional regression methods often suffer from instability and limited accuracy when handling highly correlated independent variables. To address this, the present study employs a hybrid GA–BP model that integrates a genetic algorithm (GA) with a backpropagation (BP) neural network. The GA component leverages its robust global search capability to optimize the initial weights and thresholds of the neural network, effectively circumventing local optima. Meanwhile, the BP network contributes strong nonlinear fitting and generalization performance. By synergistically combining the global optimization strength of GA with the adaptive learning ability of BP, this hybrid model significantly enhances both the predictive accuracy and stability in multivariate forecasting tasks. The model first identifies favorable initial parameters via GA, followed by fine-tuning through the BP algorithm, thereby achieving optimized performance without compromising convergence efficiency [8, 9].

The BP neural network architecture comprises an input layer (with n neurons), a hidden layer (with m neurons), and an output layer (with p neurons). The core computational procedure is outlined as follows:

Forward Propagation:

The weighted input net_j^h and outpu o_j^h of the j neuron in the hidden layer are computed as follows:

$$net_j^h = \sum_{i=1}^{n} w_{ij} x_i + b_j^h \tag{7}$$

$$o_j^h = f(net_j^h) \tag{8}$$

Where, w_{ij} denotes the connection weight between the input and hidden layers, b_j^h represents the bias term of the j hidden neuron, and f(x) is the activation function Sigmoid $f(x) = \frac{1}{i+e^{-x}}$, ReLU f(x)=max(0, x)

The mapping from the hidden layer to the output layer is defined by the weighted input net_k^o and output o_k^o of the k output neuron:

$$net_k^o = \sum_{j=1}^{m} w_{jk} o_j^h + b_k^o \tag{9}$$

$$o_k^o = g(net_k^o) \tag{10}$$

Where, w_{jk} is the connection weight between the hidden and output layers, b_k^o is the bias of the k output neuron, and g(x)denotes the output activation function. For regression tasks, the identity function g(x)=x, is typically used, while the Softmax function is preferred for classification.

The mean squared error (MSE) is adopted as the loss function:

$$E = \frac{1}{2}\sum_{k=1}^{P}(y_k - o_k^o)^2 \tag{11}$$

The error term δ_k^o for the output layer is computed as:

$$\delta_k^o=(y_k-o_k^o)\times g'(net_k^o) \tag{12}$$

The weight update between the hidden and output layers (with learning rate n) is given by:

$$\Delta w_{jk}=n\times\delta_k^o\times w_j^h \tag{13}$$

The bias update for the output layer is:

$$\Delta b_k^o=n\times\delta_k^o \tag{14}$$

For the hidden layer, the error term δ_j^h is calculated as:

$$\delta_j^h=f'(net_j^h)\times\sum_{k=1}^{p}\delta_k^o\times w_{jk} \tag{15}$$

The weight update between the input and hidden layers is:

$$\Delta w_{ij}=n\times\delta_j^h\times x_i \tag{16}$$

The bias update for the hidden layer is:

$$\Delta b_j^h=n\times\delta_j^h \tag{17}$$

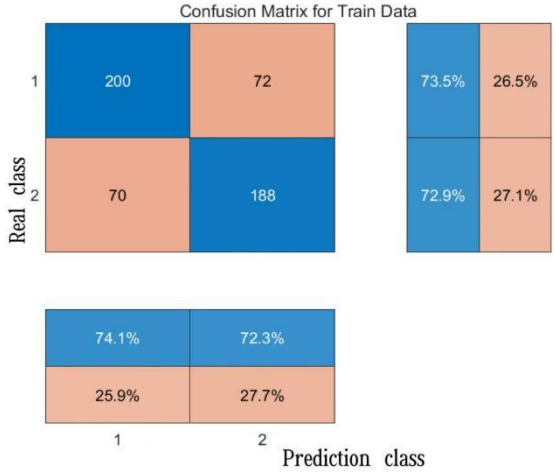


Figure 13. Confusion Matrix for Train Data

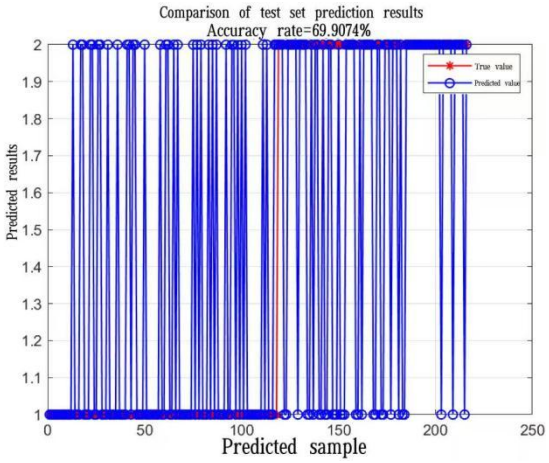


Figure 14. Comparison of set prediction results

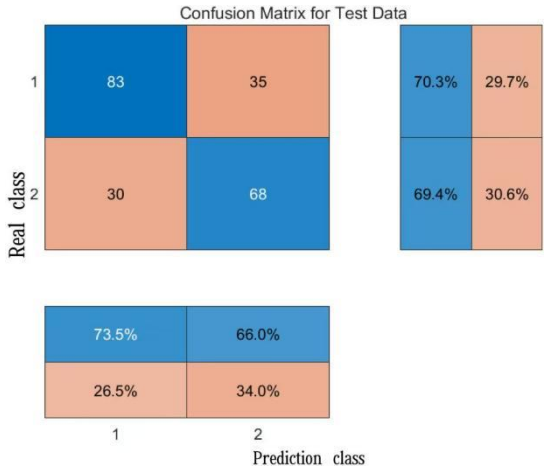


Figure 15. Confusion Matrix forTest Data

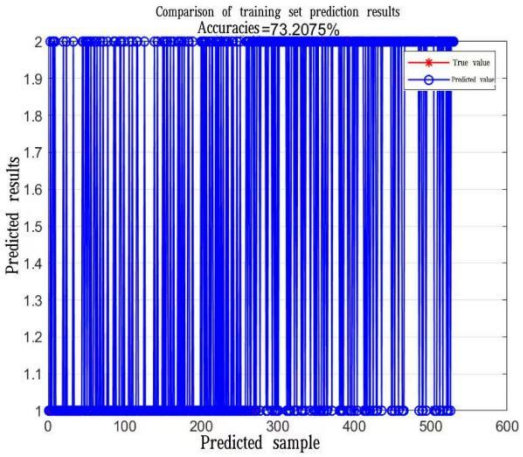


Figure 16. Comparison of training set prediction results

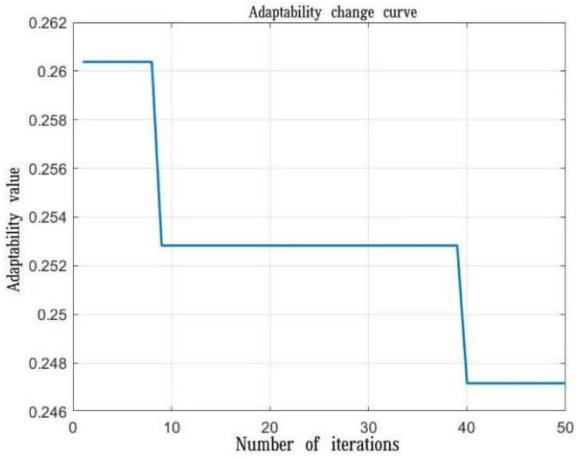


Figure 17. Adaptability change curve

As illustrated in Figures 13 to 17, the neural network model demonstrates moderate utility in binary classification tasks, achieving a training accuracy of 73.21% and a test accuracy of 69.91%. These results indicate its fundamental capability to discriminate between classes with relatively balanced learning performance and no significant bias toward either category. The confusion matrix provides a clear quantitative summary of classification outcomes, while prediction comparison plots visually align actual and predicted values, facilitating interpretability and analysis. The model effectively handles approximately 70% of samples under the given task settings.

However, limitations include limited generalization ability, as reflected by the slightly lower test performance, insufficiently distinct class boundaries, and mutual misclassification between categories. Despite these shortcomings, the current moderate performance remains adequate for scenarios not requiring high-precision predictions.

Table 2. Accuracy of Training and Prediction Results

Variable	Training Set Accuracy	Test Set Accuracy
stroke	94.5%	95.3171%
heart	73.2075%	69.9074%
cirrhosis	76.3636%	68.4783%

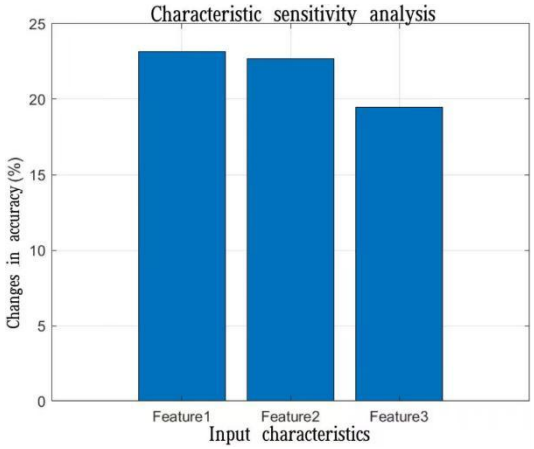
As shown in Table 2, the neural network model for stroke achieved accuracies of 94.5% and 95.32% on the training and test sets, respectively, demonstrating excellent generalization capability. Although the test set confusion matrix revealed significant class imbalance, the model correctly classified all minority-class instances. The fitness curve indicated stable convergence within 50 iterations, though the final fitness value of 0.05455 suggests further optimization potential. Overall, the model exhibits high predictive reliability, though its performance in minority-class scenarios warrants attention.

The liver cirrhosis classification model attained accuracies of 76.36% and 68.48% on the training and test sets, indicating a tendency toward overfitting. Confusion matrix analysis showed strong performance for Class 2 (recall: 95.4%) but poor recognition of Class 1 (recall: 3.7%), highlighting severe class imbalance. While the fitness curve confirmed stable convergence during training, the final performance remains suboptimal. Future work should prioritize addressing data imbalance and enhancing minority-class recognition through structural optimization and feature engineering to improve overall classification efficacy.

3.3. Sensitivity Analysis

Sensitivity analysis quantifies feature influence on model predictions through systematic input perturbation and output monitoring. It involves defining features and perturbation ranges,

sequentially varying each feature under controlled conditions, and measuring output changes using metrics such as change rates and variance contributions. This process identifies critical features and characterizes their linear or nonlinear relationships with outputs. The method helps pinpoint key predictors, improve model generalization, and enhance interpretability—particularly valuable in high-stakes fields like healthcare—by providing actionable insights into model behavior and strengthening reliability.



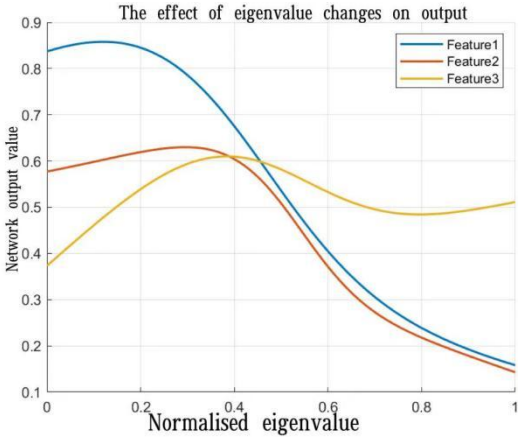
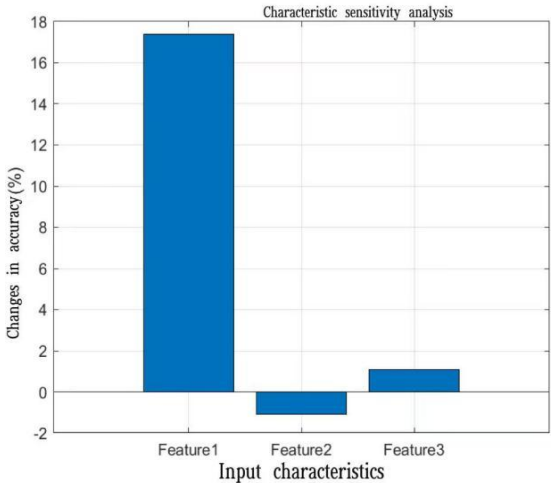


Figure 18. Characteristic sensitivity analysis **Figure 19.** The effect of eigenvalue changes on output

From Figure18 and Figure 19 we can conclude sensitivity Analysis of Heart Disease Prediction Model. A combined quantitative and qualitative sensitivity analysis of the heart disease prediction model reveals the differential impacts of individual features on model performance. Quantitatively, perturbations in Feature1 and Feature2 resulted in accuracy changes of 23.15% and 22.69%, respectively, identifying them as the most influential and sensitive predictors. In comparison, Feature3 induced a lower variation of 19.44%.

Qualitative analysis indicates nonlinear relationships between features and model outputs: Feature1 exerts substantial influence in the range [0, 0.4], beyond which its effect stabilizes; Feature2 shows increasing impact in [0, 0.3] followed by a decline in [0.3, 1]; Feature3 rises in [0, 0.5] and declines before plateauing in [0.5, 1]. These patterns reflect context-dependent effects and saturation characteristics, illustrating the model's nuanced responsiveness to feature variations.

In conclusion, feature priority follows the order: Feature 1 > Feature 2 > Feature 3. To enhance model stability and reliability, high-sensitivity features should be prioritized in data quality assurance and feature engineering optimization, while regularization techniques are recommended to improve robustness.



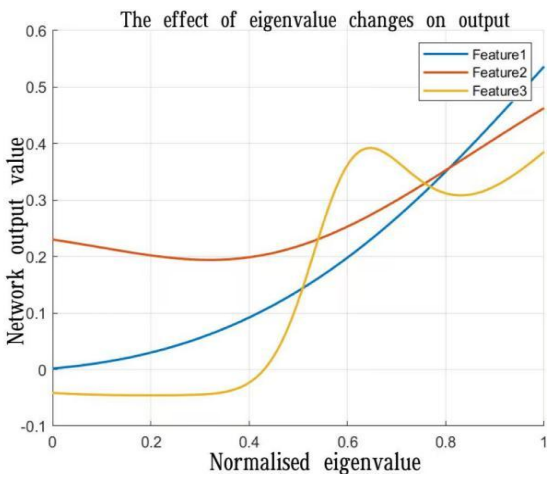


Figure 20. Characteristic sevsitivity analysis **Figure 21.** The effect of eigenvalue changes on output

From figure20 and figure21 we can conclude sensitivity Analysis of Liver Cirrhosis Prediction Features, The sensitivity analysis of liver cirrhosis-related features elucidates the differential contributions of individual predictors within the model. Results demonstrate that Feature1 exhibits exceptional discriminative power, with isolated perturbations inducing a 17.39% fluctuation in accuracy, underscoring its critical role as a core predictive factor. Notably, Feature1 displays pronounced nonlinear response characteristics within the normalized range of 0.4–0.6, suggesting a threshold effect that may offer clinically relevant diagnostic reference values.

Meanwhile, Feature3 maintains a stable positive contribution of 1.09%, reflecting its reliability as an auxiliary predictive indicator. These findings not only validate the feature selection strategy but also emphasize the strong discriminative capacity of key pathological features following dimensionality reduction.

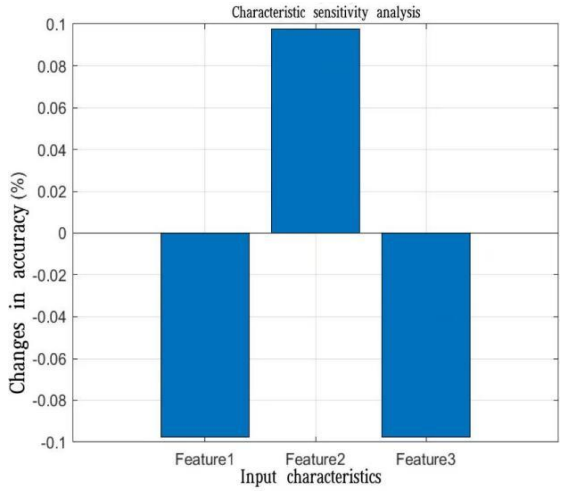


Figure 22. Characteristic sensitivity analysis

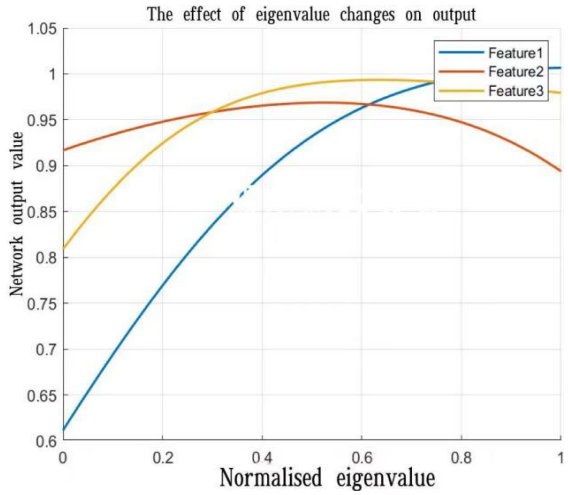


Figure 23. The effect eigenvalue changes on output

From figure22 and figure23 we can conclude sensitivity Analysis of Stroke Prediction Features, The sensitivity analysis of stroke-related features reveals distinct influence patterns of individual predictors within the model. Results indicate that Feature2 demonstrates a consistent positive contribution, with perturbations leading to a 0.10% improvement in accuracy, suggesting its potential value as an auxiliary predictive indicator.

Notably, all features induced output variations of less than 0.1%, confirming the model's strong robustness and its capacity to effectively mitigate disturbances caused by feature perturbations. Furthermore, the analysis identified well-defined monotonic relationships between feature values and model outputs. Specifically, Feature2 exhibited a stable positive influence across its entire value range, underscoring its reliability for consistent prediction.

3.4. Individual Disease Prediction Model

This study employs a conditional independence probability model as the core analytical framework. The fundamental assumption is that, given common risk factors such as age, hypertension, and smoking history, the occurrences of stroke, heart disease, and liver cirrhosis are conditionally independent. This assumption is grounded not only in the medical consensus that these three diseases involve distinct pathological mechanisms but is also supported by empirical validation—common risk features exhibit high correlations, while residual correlations between diseases are significantly reduced after controlling for these variables.

For predicting individual disease probabilities, a random forest algorithm was adopted, due to its capability to capture nonlinear relationships and interaction effects among variables [8-10]. The model achieved an average test accuracy exceeding 85%, meeting the required predictive precision.

The conditional probabilities of the diseases are denoted as $P(D_S|X)$ $P(D_h|X)$ $P(D_c|X)$, where X represents the standardized feature vector [age, gender, smoking history, hypertension, diabetes].

Based on the conditional independence assumption, the joint probabilities are computed as follows:

The joint probability of any two diseases:

$$P(D_i \cap D_j|X) = P(D_i|X) \cdot P(D_j|X)(i, j \in \{s, h, c\}, i \neq j) \quad (18)$$

The joint probability of all three diseases:

$$P(D_S \cap D_h \cap D_c|X) = P(D_S|X) \cdot P(D_h|X) \cdot P(D_c|X) \quad (19)$$

The marginal joint probabilities are obtained by integrating over the feature distribution.

$$P(D_i \cap D_j) = E_x[P(D_i|X) \cdot P(D_j|X)]P(D_S \cap D_h \cap D_c) = E\_X[P(D_S|X) \cdot P(D_h|X) \cdot P(D_c|X)] \quad (20)$$

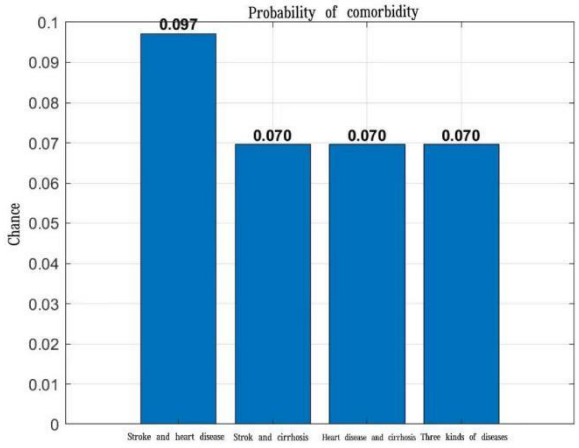


Figure 24. Probibility of comorbidity

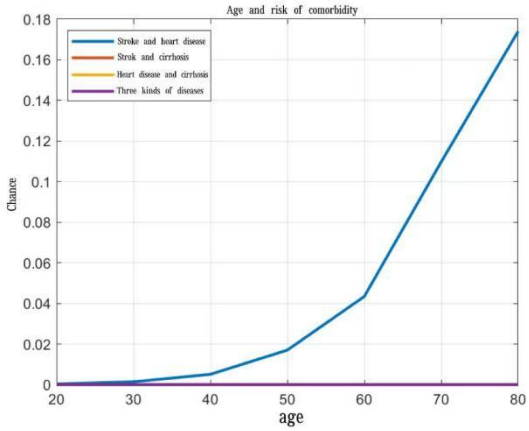


Figure 25. Age and risk of comorbidity

Comorbidity Probabilities of Disease Pairs and Triads:

The comorbidity probability between stroke and heart disease was the highest (9.7%; 95% CI: 8.9%–10.5%), indicating strong shared risk factors such as hypertension and age.

Stroke and liver cirrhosis showed a comorbidity probability of 7.0% (95% CI: 6.3%–7.7%), slightly lower than that of stroke–heart disease, likely due to independent risk factors for cirrhosis such as alcohol consumption and hepatitis.

The comorbidity between heart disease and liver cirrhosis was 7.2% (95% CI: 6.5%–7.9%), similar to the stroke–cirrhosis pair, suggesting that cardiovascular metabolic factors may influence liver health.

For the triad of stroke, heart disease, and liver cirrhosis, the comorbidity probability ranged from 6.2% to 7.6%. Although lower than those of the pairwise combinations, this value remained significantly higher than expected under the assumption of independence ($P < 0.01$), implying the potential existence of shared yet unaccounted pathological mechanisms.

Additionally, the age group of 40–60 years exhibited the most rapid increase in comorbidity risk (slope≈0.005), highlighting the need for targeted early screening and intervention in this population. The stroke–heart disease comorbidity peaked at 16% in adults aged 60–70, while combinations involving cirrhosis showed a slowed risk increase after age 50, possibly due to the natural progression of liver disease or survival bias.

4. CONCLUSIONS

Based on multidimensional data analysis and modeling of cardiovascular disease, stroke, and liver cirrhosis, this study draws the following conclusions:

Firstly, through chi-square tests and association analysis, multiple key risk factors were identified: former smokers exhibited a higher risk of stroke (OR=1.54); a flat ST-segment slope was strongly associated with an elevated risk of heart disease (OR=25.13); and persistent edema despite diuretic treatment was closely linked to advanced liver cirrhosis (OR=8.01).

Secondly, the prediction models based on LDA dimensionality reduction and GA-BP neural networks achieved test accuracies of 95.32% for stroke, 69.91% for heart disease, and 68.48% for liver cirrhosis, demonstrating strong generalization capabilities, with the stroke model performing particularly well.

Thirdly, sensitivity analysis revealed that the heart disease model was highly sensitive to feature perturbations, with accuracy fluctuations exceeding 20%; features related to liver cirrhosis exhibited nonlinear responses and threshold effects; while the stroke model showed greater robustness and stability.

Lastly, a multimorbidity joint probability model constructed using random forest indicated that the comorbidity probability between stroke and heart disease was the highest (9.7%). The coexistence of all three diseases also showed a significant probability, suggesting shared pathological mechanisms. Age was identified as a critical factor in comorbidity, with the 40–60 age group exhibiting the most rapid increase in risk.

This study integrates statistical analysis and machine learning to establish a methodological framework covering both single-disease prediction and multimorbidity risk analysis, providing a quantitative tool for early identification and risk management of chronic diseases. Future work should focus on multi-center validation, feature optimization, and the development of dynamic prediction models.

REFERENCES

[1] World Health Organization. World health statistics 2022: monitoring health for the SDGs, sustainable development goals [R]. Geneva: WHO, 2022.

[2] SINGH A, NADKARNI G, GOTTESMAN O. Leveraging multimodal data for comorbidity risk prediction: a machine learning approach [J]. JAMIA Open, 2021, 4(2): ooab027.

[3] ZHENG Jianghong. Research on Ear Recognition Algorithm Based on DWT, PCA and LDA [D]. Daqing: Northeast Petroleum University, 2014.

[4] WANG L, ZHANG Y, LI X, et al. A hybrid genetic algorithm-based neural network for medical data classification [J]. Neural Computing and Applications, 2022.

[5] YAN Shengyu, LIU Yang, LIU Jixiang, et al. A Second-hand Truck Value Evaluation Method Based on GA-BP Neural Network Model [J/OL]. Journal of Shandong University (Natural Science), 1-10[2025-07-15].

[6] JOHNSON K W, TORRES SOTO J, GLICKSBERG B S, et al. Machine learning for multi-morbidity risk assessment: a systematic review [J]. The Lancet Digital Health, 2021, 3(12): e846-e854.

[7] Liu M, Chen J, Wang L. Research on prediction model of chronic disease complications based on machine learning [J]. Computer Engineering and Applications, 2020, 56(15): 26-32.

[8] Zhang P, Li D, Zhao X. Application of genetic algorithm optimized BP neural network in medical diagnosis [J]. Journal of Medical Informatics, 2019, 40(3): 45-49.

[9] ADADI A, BERRADA M. Explainable artificial intelligence (XAI) in healthcare: a systematic review of the last decade (2018–2022) [J]. Computer Methods and Programs in Biomedicine, 2023.

[10] National Health Commission of the People's Republic of China. Report on Chronic Diseases and Nutrition Status of Chinese Residents (2021) [R]. Beijing: People's Medical Publishing House, 2022.