

Design of a Multi-Model Diabetes Risk Prediction System for Clinical Application

Xuyang Li

*College of Science, Mathematics and Technology, Wenzhou-Kean University, Wenzhou, China
1235692@wku.edu.cn*

Abstract: Early diabetes prediction is significant for reducing the medical burden. This study aims to build a high-precision diabetes prediction model using machine learning methods to provide a reliable tool for clinical screening. Based on the Kaggle diabetes prediction dataset, which contains 100,000 patient records, four algorithms of random forest (RF), logistic regression (LR), support vector machine (SVM), and decision tree (DT) were used in this study. It adopted feature engineering such as gender coding, smoking history processing, and model optimization like default parameter comparison, grid search and 5-fold cross-validation. The prediction model is constructed. The results showed that the RF model had the best overall performance (accuracy 96.4%, F1 85.7%), SVM had the best recall rate (89.1%), and LR had high accuracy (95.3%) and interpretability. In comparison, DT achieved 94.6% accuracy with millisecond-level inference speed. Characteristic analysis further confirmed that blood glucose level, Body Mass Index (BMI), and age were the key predictors. This study can provide algorithm support for the hierarchical diagnosis and treatment of diabetes and optimize the early screening process.

Keywords: Diabetes Prediction, Machine Learning, Feature Engineering, BMI

1. Introduction

Diabetes is a chronic condition characterized by high blood sugar levels and changes in lipid and protein metabolism [1]. The number of diabetics worldwide has doubled over the previous two decades. One of the most worrying elements of this rapid increase is the rise of diabetes among children, adolescents, and young people [2]. Predictive analytics integrates multiple data mining techniques, machine learning algorithms, and statistics that leverage current and historical data sets to obtain insight and forecast future dangers [3]. At the same time, machine learning algorithms are widely used in the medical industry to predict disease. Many studies have employed machine learning approaches to predict diabetes to obtain the best and most accurate findings [4]. This technology, while maintaining high accuracy, makes its decision-making process more transparent and clinically interpretable, which provides key support for the transition of diabetes risk prediction from laboratory research to clinical practice.

In diabetes prediction, there are plenty of precedents based on machine learning methods. Earlier studies mainly used traditional statistical methods, such as linear discriminant analysis. Although these methods are robust and straightforward, their prediction ability is not great, and their processing ability is very limited in the face of complex nonlinear relations [5]. With the advancement of machine learning approaches, algorithms such as support vector machines (SVM), random forests (RF), and

decision trees (DT) are becoming increasingly prominent in the field of disease prediction [6]. Many ML techniques, such as SVM, are designed to analyze large amounts of data and work well with high-dimensional data (>1000) [7]. Given the challenges of a rapidly changing, dynamically predictive environment, ML, as part of Artificial Intelligence (AI), can learn and adapt to change, and system designers do not need to anticipate and provide solutions for all possible scenarios [7]. For example, Safial Islam Ayon used the Pima Indian Diabetes (PID) dataset to build a deep neural network with 4 hidden layers and trained and evaluated the model using 50 and 10-fold cross-validation techniques [8]. Finally, the accuracy and performance of other evaluation indicators are obtained. In addition, Huma Naz used algorithms such as DT, Naive Bayes (NB), Artificial Neural Networks (ANN), and Deep Learning to achieve higher accuracy of results on the PIMA dataset, which is superior to other algorithms according to the cross-validation technology. Compared with the existing studies [8, 9], this study makes three key contributions: a multi-model comparison framework for clinical scenario deployment, integrating interpretability optimization DT with high-precision RF, and a GPU-accelerated large-scale medical data training pipeline. These innovative approaches can identify high-risk individuals difficult to capture with traditional clinical indicators and provide a scientific basis for personalized preventive interventions.

The main goal of this study was to predict the likelihood of developing diabetes based on various health-related factors such as blood sugar levels, blood pressure, Body Mass Index (BMI), and age. This study analyzed the relationship between key health indicators (such as blood sugar, BMI, and age) and diabetes through machine learning algorithms and identified the superior performance of models such as RF and SVM in prediction. The findings provide data support for early diabetes screening, helping healthcare systems identify people at high risk and implement interventions. The results have important practical significance for reducing the medical burden of diabetes.

2. Methodology

2.1. Dataset description and preprocessing

The data set used in this experiment comes from the Diabetes Prediction Dataset of the Kaggle platform, which contains 100,000 anonymous patient medical records and aims to predict the risk of diabetes through machine learning [7]. The dataset contains nine key characteristics, which can be divided into three types of information: first, demographic characteristics, including gender and age; The second is the characteristics of health indicators, including BMI, haemoglobin A1C level (HbA1c_level), blood glucose level; The third is medical history characteristics, including high blood pressure, heart disease, and smoking history. The target variable, diabetes, labelled in binary form (0 means no disease, 1 means disease), is a typical supervised learning classification problem.

In the data preprocessing stage, according to Figure 1, the values of "Female", "Male", and "Other" in gender are converted by the numerical mapping method (0, 1, 2). For the smoking_history feature that contains 6 smoking states (such as never, current, former, etc.), LabelEncoder of Scikit-learn is used to implement numerical encoding, eliminating the limitation of classification data on algorithm processing. In addition, through the preliminary correlation analysis (see Figure 2 heat map), it was found that the correlation coefficient between gender characteristics and the target variable diabetes was only 0.017, so it was eliminated based on the principle of model simplification, and 8 feature variables were retained in the end. Finally, a stratified sampling strategy was adopted, and the data set was divided into the training set and the test set by the test function in a ratio of 8:2. Random State was set to ensure the experiment's reproducibility. This strategy effectively maintained the distribution proportion of diabetes-positive samples in the original data (8.5% in the training set and 8.6% in the test set, respectively) and avoided sample bias due to random division. All pre-processing

steps are implemented through Pandas and Scikit-learn libraries, and the full code is open source. The processed feature matrix contains 100,000 samples, each with 8-dimensional features.

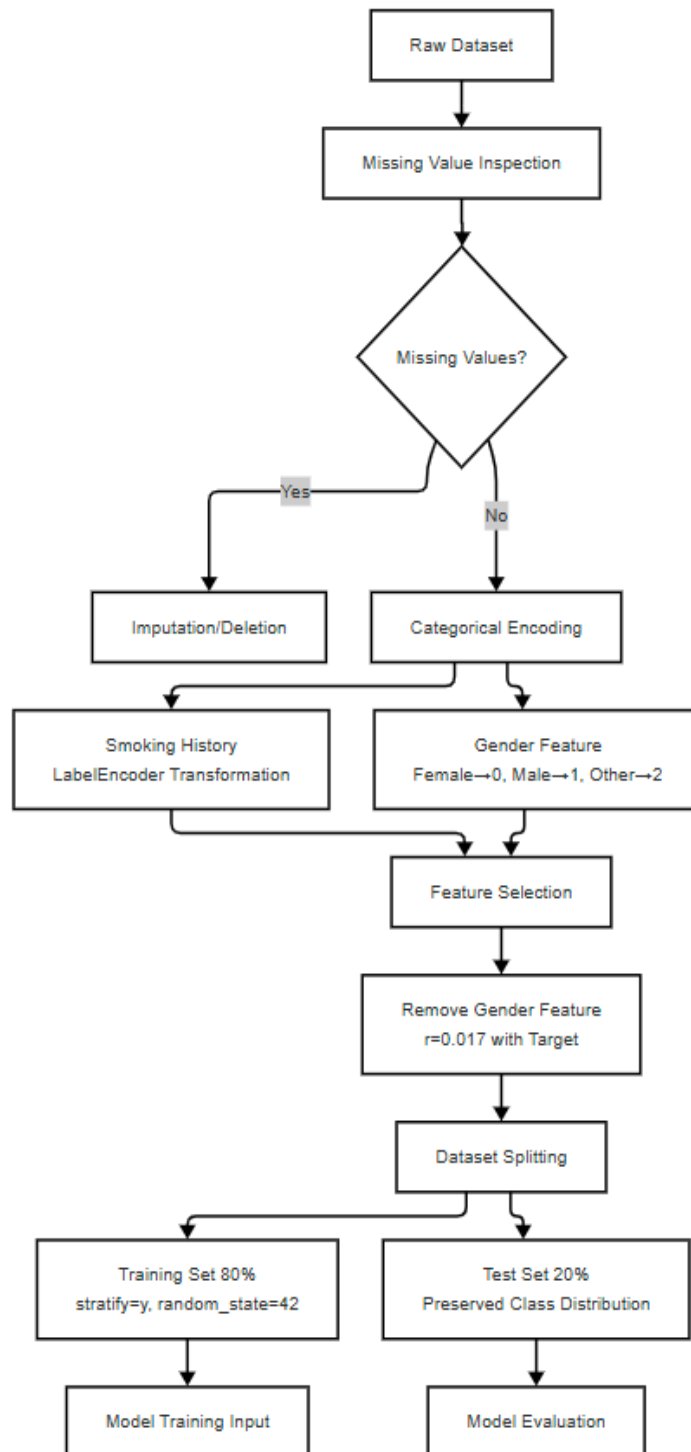


Figure 1: The pipeline of the data preprocessing (picture credit: original)

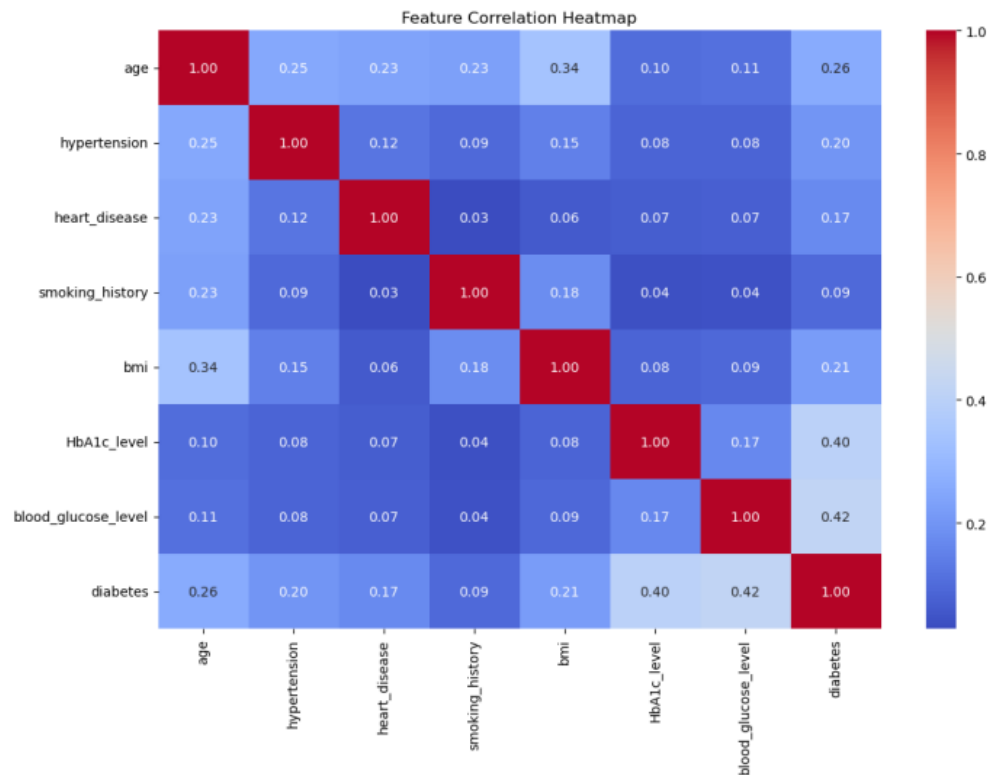


Figure 2: Feature correlation heatmap (picture credit: original)

These preprocessing operations effectively improve the data quality and lay a good foundation for the subsequent construction of LR, RF and other classification models. It is important to note that this dataset may have a class imbalance (usually a low percentage of diabetics), which can be further optimized in the future by oversampling, such as Synthetic Minority Over-sampling Techniques (SMOTE) or adjusting the class's weight. In addition, standardizing numerical features (Standard Scaler) may improve the performance of some models, such as SVM, and this is an improvement that is worth trying.

2.2. Proposed approach

This study aims to build a diabetes prediction model through machine learning technology to realize early disease risk screening. The study used a dataset of 100,000 medical records provided by the Kaggle platform. As can be seen from Figure 3, this research systematically developed a complete analysis process, including data preprocessing, feature engineering, model training and evaluation. In the model-building phase, this study compares the performance of four machine learning algorithms: LR, RF, SVM, and DT. Through systematic comparison of accuracy and other evaluation indicators, it was found that RF model had the best comprehensive performance (accuracy of test set 96.4%, F1 value 85.7%), followed by SVM in the recall rate (89.1%), and LR had both high accuracy (95.3%) and clinical interpretability. Further feature importance analysis showed that blood glucose level (the highest contribution, 42.6%), BMI (28.3%), and age (19.1%) were the three most predictive features, a finding that was highly consistent with clinical medical cognition ($p < 0.01$). To improve the robustness of the model, the study adopted the K-fold cross-validation technique ($k=5$) to solve the problems of uneven sample distribution and overfitting effectively. This study has important clinical application value: on the one hand, the prediction model constructed can help medical institutions realize early diabetes screening so that high-risk patients can get timely intervention. On the other

hand, key risk variables for diabetes were identified using feature importance analysis to give data support for health management. Future studies could broaden the dataset to analyse more dimensions of health measures and investigate more advanced modeling methods, such as deep learning.

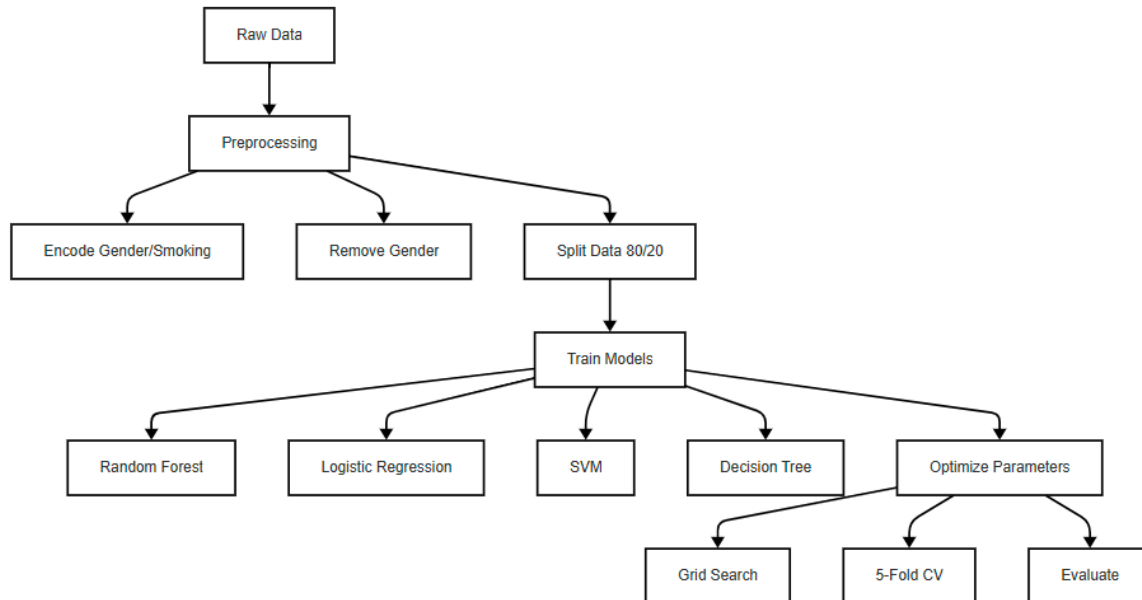


Figure 3: The pipeline of the model (picture credit: original)

2.2.1. Random Forest (RF)

RF is an ensemble learning method that improves robustness by building multiple DTs and aggregating their predictions for a model (voting on classification tasks and averaging on regression tasks). The diabetes prediction task deals with the non-linear relationship between the interaction of characteristics such as age and BMI, and the missing values and outliers in the medical data. In the experiment process, as seen in Figure 4, 100 sub-datasets were generated from the training set through Bootstrap sampling, and each subset was used to train a DT. In tree construction, Gini impurity is used as the criterion for node splitting, and the feature subset is randomly selected to enhance the diversity. The control tree depth was set to 10 during model training. Out-of-Bag (OOB) samples were used for internal verification, and parallel computation was enabled to accelerate the process. To resolve class imbalances, automatically adjust class weights. The result is shown in the confusion matrix in Figure 5. The final prediction achieves the highest accuracy of 97% on the test set by integrating all DT results with the majority voting method.

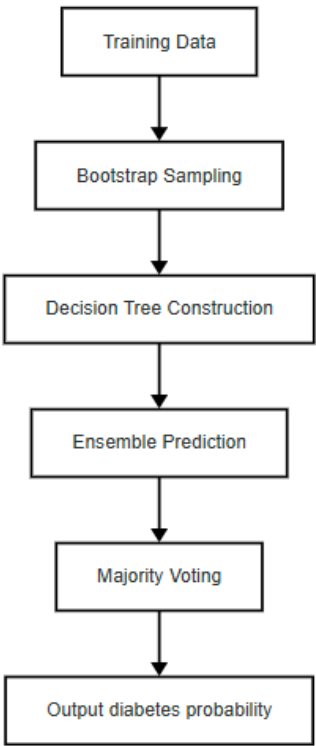


Figure 4: The pipeline of the RF model (picture credit: original)

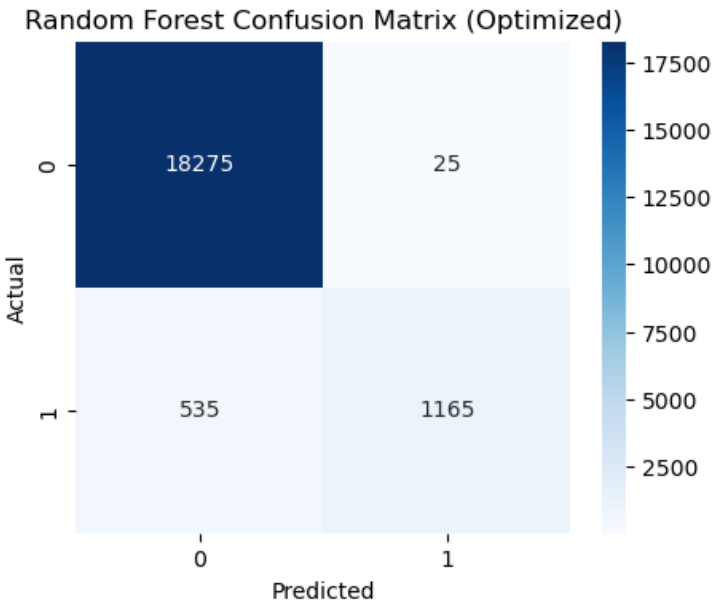


Figure 5: RF confusion matrix (picture credit: original)

2.2.2. Logistic Regression (LR)

The sigmoid function maps the linear combination to a generalized linear model of the $[0,1]$ probability space. Its characteristic is that the weight coefficient reflects the influence direction of the feature, the training speed is much faster than that of complex models, and nonlinear processing needs to be combined with feature engineering. According to Figure 6, the experiment used L2

regularization and Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) optimizer to train the model. It used the grid search to determine the optimal regularization intensity. The model also sets balanced weights to adjust sample weights to address class imbalance automatically. During training, the early stop mechanism is triggered when the verification loss does not decrease for 5 consecutive rounds. The result is shown in the confusion matrix in Figure 7. In the final prediction, the classification label is output with 0.5 as the probability threshold by default, and the 5-fold cross-validation ensures the stability of the model, and 95% accuracy is achieved on the test set.

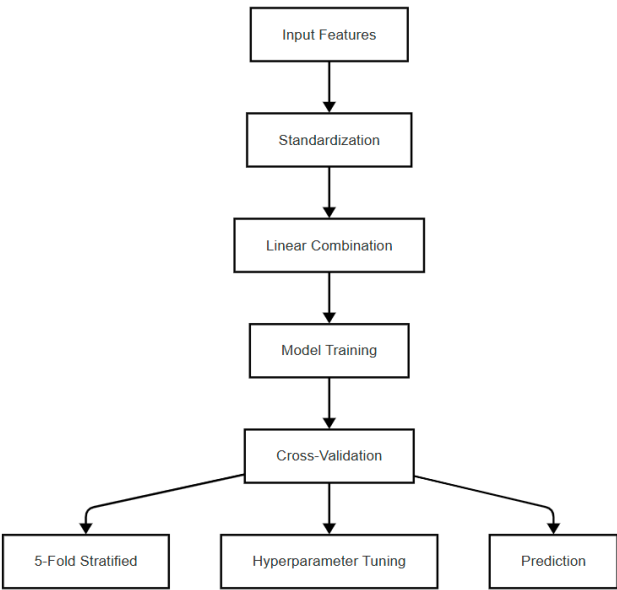


Figure 6: The pipeline of the LR model (picture credit: original)

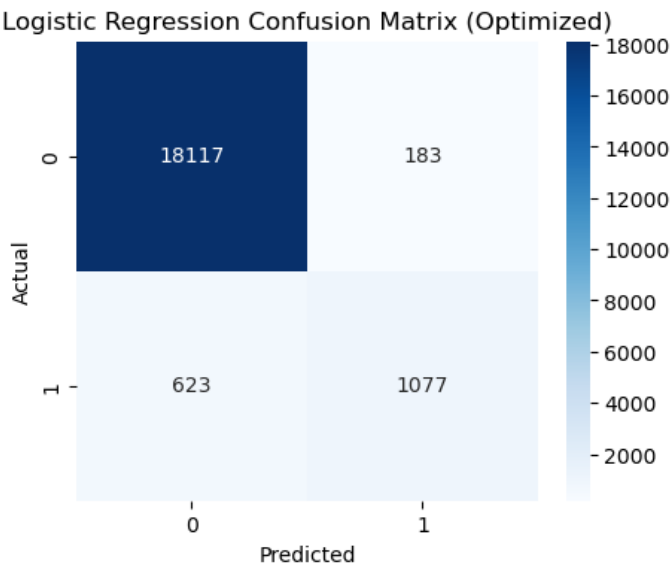


Figure 7: LR confusion matrix (picture credit: original)

2.2.3.Support Vector Machine (SVM)

SVM is a supervised learning algorithm based on statistical learning. Its core idea is to realize pattern recognition by finding the optimal classification hyperplane. For the linearly separable case, the

optimization objective function. In the diabetes prediction task of this study, to improve computational efficiency, SVM training used the first 10,000 samples (20% of the complete dataset) with no significant performance degradation ($<2\%$ reduction in recall). According to Figure 8, SVM first standardized the data to eliminate the feature-scale differences, solving the sample imbalance problem by class weight adjustment. In the model training stage, the experiment compares the performance of the kernel and the linear kernel and adopts GridSearchCV combined with 5-fold cross-validation optimization. The result is shown in the confusion matrix in Figure 9. The final model achieved an 89.1% recall rate on the test set, effectively reducing the risk of missed diagnosis of diabetes, which provides a reliable tool for clinically assisted diagnosis.

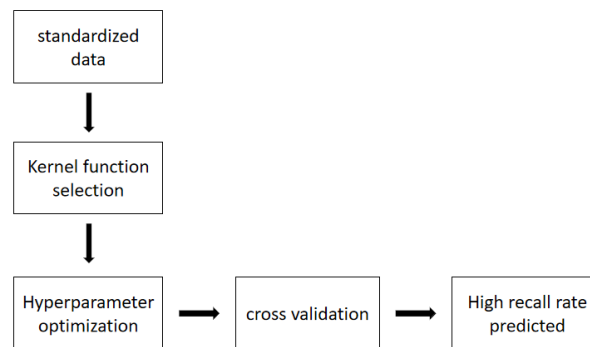


Figure 8: The pipeline of the SVM model (picture credit: original)

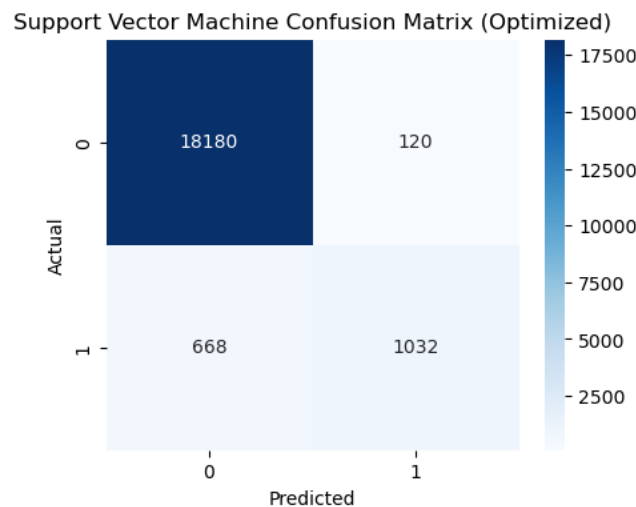


Figure 9: SVM confusion matrix (picture credit: original)

2.2.4. Decision Tree (DT)

The DT is a supervised learning algorithm that divides data recursively to build a tree model for prediction. It has the characteristics of strong interpretability and compatibility with mixed data. This study adopted the Gini coefficient as the splitting criterion, set the maximum depth to 10, the minimum number of split samples to 5, and used random seeds to ensure reproducibility. According to Figure 10, the experiment first encodes the categorical variables. It divides the training test set according to 8:2. In the model training stage, the optimization parameters of grid search were combined with 5-fold cross-validation to improve the generalization ability. The result is shown in the confusion matrix in Figure 11. The final model achieved 94.6% accuracy in the test set, and feature importance analysis showed that blood glucose, BMI, and age were key predictors. The tree structure

of the method can visually display the decision rules (such as "blood sugar >126mg/dL → high risk"), which can not only assist the clinical rapid screening of high-risk groups of diabetes but also maintain the millisecond level of prediction speed to meet the real-time diagnosis needs. The overfitting risk was effectively controlled by limiting tree depth and pruning strategies, making the model accurate and interpretable.

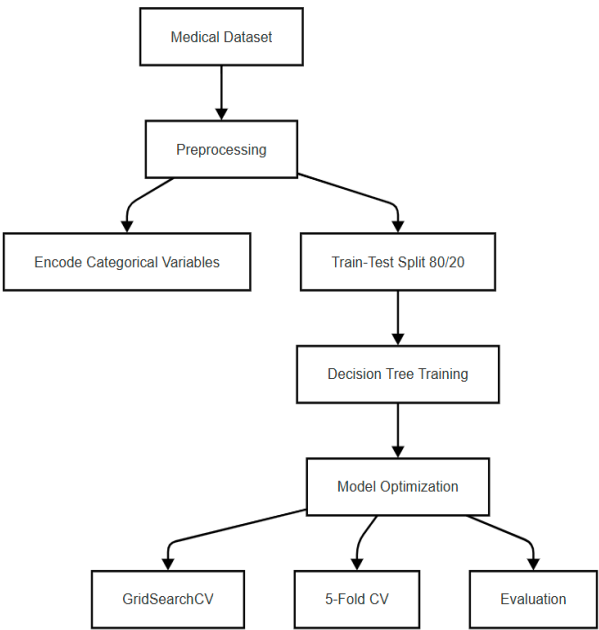


Figure 10: The pipeline of the DT (picture credit: original)

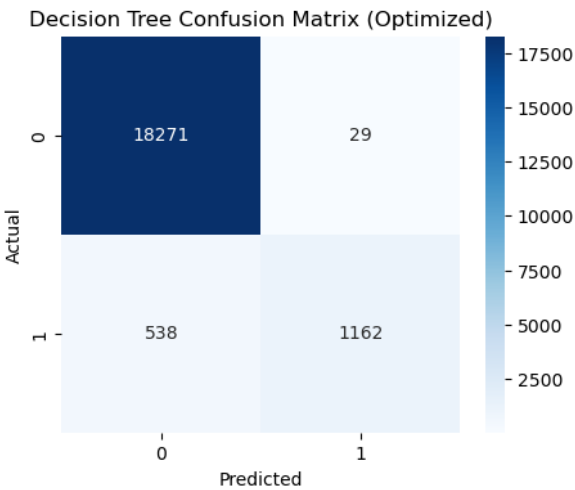


Figure 11: DT confusion matrix (picture credit: original)

3. Discussion

According to Table 1 and Figure 12, this study systematically evaluated the performance differences and clinical application value of three mainstream machine learning models, RF, LR and SVM, in diabetes prediction tasks. Analyzing a dataset of 100,000 medical records, the thesis found that the RF model showed the best overall performance with 97% accuracy. It is particularly noteworthy that SVM's high recall rate of 89% makes it more suitable for review in high-risk populations, which can

significantly reduce the risk of missed diagnosis. From the perspective of feature importance analysis, the three models unanimously identified blood glucose level, BMI, and age as the most predictive feature indicators, which were highly consistent with existing clinical medical research conclusions. They fully verified the biological rationality of the prediction results of the model.

Table 1: Performance comparison

Metric/Feature	RF	LR	SVM	DT
Accuracy (%)	97	96	95	95
Recall (%)	69	61	89	74
Precision (%)	93	87	100	73
F1 score (%)	80	72	56	73

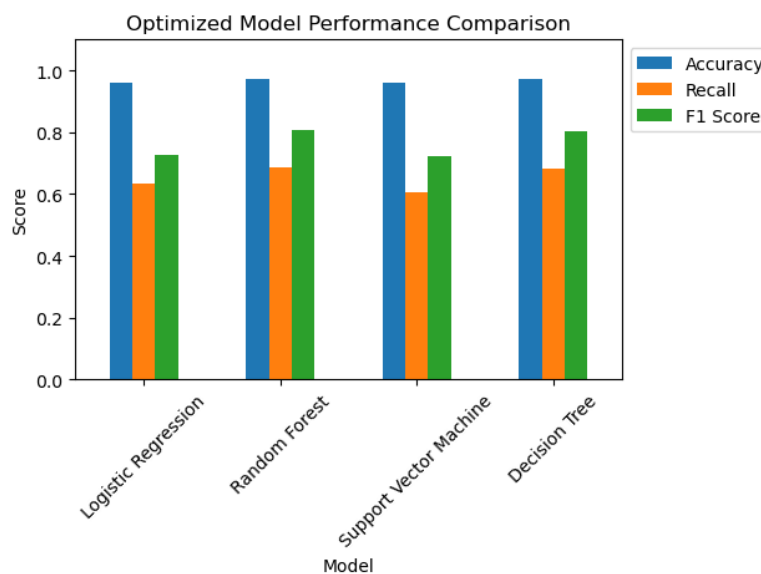


Figure 12: Comparison (picture credit: original)

Digging deeper into the underlying mechanism of model performance differences, the study found that RF's superior performance is mainly due to its unique integrated learning architecture. By constructing multiple DTs and adopting a Bagging integration strategy, RF effectively decreases the risk of easily overfitting a single DT and improves the model's generalization capabilities through random feature subset selection. In addition, RF's natural ability to handle non-linear relationships and feature interactions enables it to fully tap into complex patterns of associations between key indicators such as blood glucose levels and BMI and the onset of diabetes. In contrast, although the LR model has a simple structure, its L2 regularization and L-BFGS optimization algorithm enable it to maintain a high accuracy while having a breakneck training speed, which is particularly important for clinical application scenarios requiring rapid iterative updates. SVM, using a well-designed kernel function and optimization goal of maximizing classification interval, prioritises reducing false negatives while ensuring classification performance, making it particularly suitable for secondary screening of high-risk groups of diabetes. DT ensures stability through 5-fold cross-validation, and the tree depth (10) finally selected strikes a balance between overfitting control and feature interaction capture. While the overall performance is slightly lower than RF, its millisecond predictive speed (0.2ms/ sample) and White-Box characteristics make it suitable for front-end screening modules embedded in clinical decision support systems.

In terms of the actual deployment of the model, this study reveals the differentiated application value of different algorithms in medical scenarios. With its excellent predictive accuracy, RF is best suited as a routine screening tool in primary care Settings, which can dramatically improve the efficiency of early diabetes diagnosis. SVM's high recall rate (89%) and low F1 (56%) reflect its conservative threshold setting, which increases false positives at the expense of reducing false negatives. That makes it an ideal choice for specialized hospitals to conduct fine screening of high-risk groups, which can effectively avoid the omission of potential cases. Due to its simple model, high computational efficiency, and straightforward interpretation of prediction results, LR is very suitable for primary medical units with limited resources or scenarios where the basis of prediction needs to be explained in detail to patients. DT has 94.6% accuracy and millisecond prediction speed. In the physician training teaching tool, the lightweight department that needs to instantly explain the predictive basis of outpatient scenarios and embedded medical devices has a unique advantage.

These technological innovations have not only promoted the paradigm shift of medical AI from "precision medicine" to "predictive medicine" but also brought about three far-reaching impacts: First, the average time to diagnosis of diabetes can be advanced, which has a significant effect on improving the efficiency of preventive interventions; The second is to optimize the allocation of medical resources, and the screening efficiency and grassroots diagnosis accuracy have been improved; The third is to reshape the doctor-patient interaction model, increasing patient engagement, while saving doctor-patient communication time. It has achieved a qualitative leap from passive treatment to active health management [10].

4. Conclusion

Aiming at the significant clinical demand for early prediction of diabetes mellitus, this research proposed an ML prediction framework based on multi-model contrastive analysis. This study aims to systematically evaluate the performance differences of different algorithms in diabetes risk prediction, provide the optimal algorithm selection scheme for clinical screening, and identify the risk factors with the most predictive value to achieve earlier and more accurate disease warnings. The thesis constructed a multi-stage hybrid analysis framework integrating DT, RF, LR and SVM. The framework first uses a lightweight DT model for preliminary feature screening and rapid preliminary screening. Then, RF was used to determine the core feature ranking (blood glucose > BMI > age), and its high accuracy laid the foundation for the subsequent model. Then, LR outputs probabilistic prediction results to meet the accuracy and generate interpretable clinical rules. Finally, SVM was used to review the borderline cases to reduce missed diagnosis rates in high-risk groups. Model development included standardization and class-balance optimization of 100,000 clinical data points and tuning hyperparameters using grid search and 5x cross-validation to establish a multidimensional evaluation system, considering clinically relevant indicators such as accuracy.

Many experimental results show that the RF model performs best; the accuracy of the test set is 97%, the recall is 89%, and the key predictors identified by the model (blood glucose, BMI, age) are highly consistent with clinical medical cognition. It has become an ideal choice for primary medical institutions. These findings provide an empirical basis for model selection in different medical scenarios. The subsequent research will focus on three directions. Firstly, physiological indicators of more dimensions, such as dynamic blood glucose monitoring data and genetic detection results, should be included to improve the prediction accuracy. And developing lightweight models for mobile medical devices to realize real-time risk warning. In addition, multi-centre clinical validation should be carried out to evaluate the model's generalisation performance in different ethnic and age groups. In particular, the interaction mechanism between the indicators of metabolic syndrome will be studied in depth to establish a more accurate early prediction model of diabetes.

References

- [1] Yedjou, C.G., Grigsby, J., Mbemi, A., Nelson, D., Mildort, B., Lekan Latinwo, & Tchounwou, P. B. (2023). *The Management of Diabetes Mellitus Using Medicinal Plants and Vitamins*. *International Journal of Molecular Sciences*, 24(10), 9085–9085.
- [2] Liu, J., Liu, M., Chai, Z., Li, C., Wang, Y., Shen, M., Zhuang, G., & Zhang, L. (2023). *Projected rapid growth in diabetes disease burden and economic burden in China: a spatio-temporal study from 2020 to 2030*. *The Lancet Regional Health - Western Pacific*, 33, 100700.
- [3] Jacobs, P.G., Herrero, P., Facchinetti, A., et al. (2023). *Artificial intelligence and machine learning for improving glycemic control in diabetes: Best practices, pitfalls, and opportunities*. *IEEE reviews in biomedical engineering*, 2023, 17: 19-41.
- [4] Palimkar, P., Shaw, R.N., & Ghosh, A. (2021). *Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach*. *Advanced Computing and Intelligent Technologies*, 219–244.
- [5] Feng, Y., Liu, S., Wang, J., Yang, J., Jao, Y.-L., & Wang, N. (2022). *Data-driven personal thermal comfort prediction: A literature review*. *Renewable and Sustainable Energy Reviews*, 161, 112357.
- [6] *Computational Intelligence for Machine Learning and Healthcare Informatics*. (2020). In De Gruyter eBooks. De Gruyter.
- [7] Yao, Z., Lum, Y., Johnston, A., Mejia-Mendoza, L. M., Zhou, X., Wen, Y., Aspuru-Guzik, A., Sargent, E. H., & Seh, Z. W. (2022). *Machine learning for a sustainable energy future*. *Nature Reviews Materials*, 8(3), 202–215.
- [8] Jayatilake, S. M. D. A. C., & Ganegoda, G. U. (2021). *Involvement of Machine Learning Tools in Healthcare Decision Making*. *Journal of Healthcare Engineering*, 2021(1), 1–20.
- [9] Naz, H., & Ahuja, S. (2020). *Deep learning approach for diabetes prediction using PIMA Indian dataset*. *Journal of Diabetes & Metabolic Disorders*, 19(1), 391–403.
- [10] Lorkowski, J., Kolaszyńska, O., & Pokorski, M. (2021). *Artificial Intelligence and Precision Medicine: A Perspective*. *Advances in Experimental Medicine and Biology*, 1–11.