2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

MECFE: A Novel Consensus Feature Engineering Approach for Enhanced Diabetes Risk Prediction

Jijun Tong[1], Lisi Ye[2], Congcong Yang[2], Yang Chen[2], Yuqiang Shen[2], Qingli Zhou[3], and Shudong Xia[3(✉)]

[1] School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China
[2] School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China
[3] The Fourth Affiliated Hospital of School of Medicine, and International School of Medicine, International Institutes of Medicine, Zhejiang University, Yiwu 322000, China
Shystone@zju.edu.cn

Abstract. Diabetes mellitus has emerged as a global health crisis, with its prevalence rising sharply and placing significant strain on healthcare systems. Early and accurate prediction of diabetes risk is crucial for effective prevention and management. While machine learning and deep learning techniques have made advancements in diabetes prediction, feature engineering remains an underemphasized area and faces several challenges, particularly in terms of interpretability, depth, and stability. To address these challenges, this study proposes a novel Medical Enhancement Consensus Feature Engineering (MECFE) approach that integrates structured medical knowledge with data-driven insights. MECFE consists of two key components: Medical-Data Collaborative Feature Construction (MD-CFC), which incorporates clinical knowledge and ClinicalBERT for enriched feature generation; and Heterogeneous Model Consensus Feature Selection (HM-CFS), which employs a three-layer weighted fusion strategy across multiple models to improve feature stability and clinical relevance. Experimental results show significant improvements in the performance metrics of all models, with Bayesian-optimized LightGBM achieving the best results: R^2 increasing by 0.108, RMSE decreasing by 23.66%, MSE reducing by 41.75%, and MAPE dropping by 16.42%, demonstrating the effectiveness of the MECFE. Additionally, feature importance analysis in LightGBM is employed to further enhance the model's interpretability, providing deeper insights into the factors influencing diabetes risk.

Keywords: Diabetes Prediction, Feature Engineering, Heterogeneous Model, LightGBM, ClinicalBERT, Interpretability.

1 Introduction

Diabetes mellitus, a chronic metabolic disorder marked by hyperglycemia, has become a pressing global health concern. According to the International Diabetes Federation [1,2], the number of diabetic patients is expected to reach 783 million by 2045, with

China experiencing a 56% increase over the past decade alone [3,4]. Diabetes is a major driver of complications such as cardiovascular and renal diseases, significantly burdening healthcare systems worldwide [5]. Consequently, improving early diagnosis and predictive accuracy is critical for effective disease prevention and healthcare resource management.

Recent advancements in machine learning and deep learning have shown promise in diabetes risk prediction. While conventional models such as logistic regression are valued for their interpretability, they often fail to capture complex nonlinear patterns inherent in clinical data [6]. In contrast, advanced machine learning algorithms—such as Random Forest (RF) [7], XGBoost [8], and Support Vector Machines (SVM) [9]—can partially address nonlinearity, whereas deep learning architectures, including Deep Neural Networks (DNNs) [10] and Long Short-Term Memory networks (LSTMs) [11], are particularly effective in handling large-scale, high-dimensional datasets. However, most studies focus on model optimization, often overlooking the foundational role of feature engineering in improving prediction quality and clinical applicability.

Feature engineering—comprising preprocessing, feature construction, and feature selection—remains underutilized and insufficiently explored in medical predictive modeling. Clinical datasets are typically high-dimensional, heterogeneous, and noisy, and inadequately engineered features may mask clinically relevant biomarkers, thereby degrading both model performance and interpretability [12]. Current feature engineering methodologies face three primary challenges: limited interpretability, insufficient feature information depth, and instability of selected feature subsets. For instance, techniques like Principal Component Analysis or correlation-based selection [13,14] often lack pathological relevance, while inconsistency in selection criteria across models reduces feature stability [15].

To address these limitations, this study proposes a novel MECFE approach, composed of two synergistic modules: MD-CFC and HM-CFS. MD-CFC integrates structured medical knowledge and ClinicalBERT [16] with data-driven construction to enrich feature representation. HM-CFS employs six model-based scoring methods fused via a three-layer strategy—including a self-attention mechanism and medical refinement layer—to select more stable and clinically meaningful features. These features are validated across multiple algorithms, with a Bayesian-optimized LightGBM model [17] ultimately achieving superior performance. To enhance clinical insight, this study further leverage LightGBM's feature importance and regression tree visualization to explain model decisions and highlight influential risk factors.

The main contributions of this study are as follows:

1. Innovative Feature Engineering Approach: This study proposes MECFE, which fuses medical knowledge and data-driven learning through MD-CFC and HM-CFS for enriched and interpretable feature design.

2. Improved Predictive Performance: Extensive experiments demonstrate that MECFE improves predictive accuracy across various algorithms, with LightGBM achieving the best results.

3. Enhanced Model Interpretability: Feature importance analysis offers transparent insights into diabetes risk, aiding clinical decision-making.

2 Materials and Methods

2.1 Dataset

The dataset employed in this study is sourced from the Tianchi Precision Medicine Competition, a publicly available clinical dataset comprising 6,768 samples. The feature space consists of 39 biomedical indicators, categorized into nine distinct groups: four liver function indicators, four serum protein indicators, four lipid-related indicators, three renal function markers, five hepatitis B-related indicators, eight hematological cell counts, four platelet-associated indicators, five immune cell proportions, and one demographic attribute. The target variable is the Blood Glucose (BG), which serves as a clinical proxy for diabetes risk.

2.2 LightGBM

LightGBM is adopted as the core prediction model in MECFE for its efficiency and scalability in high-dimensional biomedical data. It is based on Gradient Boosted Decision Trees (GBDT), and its objective function combines prediction loss with a regularization term:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \tag{1}$$

At each iteration, gradients and Hessians are computed as:

$$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}, h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} \tag{2}$$

A greedy algorithm is then used to build decision trees by maximizing information gain. Bayesian Optimization with Expected Improvement is employed to fine-tune hyperparameters, ensuring optimal model performance.

2.3 ClinicalBERT

ClinicalBERT, a Transformer-based model pretrained on large-scale medical texts. It employs the standard self-attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{3}$$

and multi-head attention:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{4}$$

In this study, ClinicalBERT contributes in two ways: (1) during feature construction, it interprets clinical narratives to extract medically relevant features, mitigating the limitations of manual or rule-based approaches; and (2) during model training, attention weights from the fine-tuned ClinicalBERT—capturing the interactions between BG and 38 biomedical indicators—are leveraged as adaptive importance scores to inform feature selection.

3 Overview

MECFE encompasses a multi-stage pipeline of data preprocessing, feature engineering, model optimization, and interpretability analysis, as illustrated in Fig. 1. The pipeline begins with systematic data preprocessing to ensure data quality and consistency for subsequent modeling tasks. The MECFE consists of two core components: 1) MD-CFC, which integrates exploratory data analysis with domain-informed feature construction by combining structured clinical knowledge and semantic representations extracted via ClinicalBERT; 2) HM-CFS, which leverages multiple learning algorithms to evaluate feature importance from diverse perspectives. These features are then refined through a three-layer process: a feature evaluation layer, a weighted fusion layer, and a medically-informed refinement layer. A self-attention mechanism is introduced to enhance the interpretability and relevance of selected features. The resulting feature subset is passed into a LightGBM model, with its hyperparameters optimized via Bayesian methods. Finally, interpretability analysis is conducted to provide insight into the model's decision process, ensuring clinical relevance and transparency.

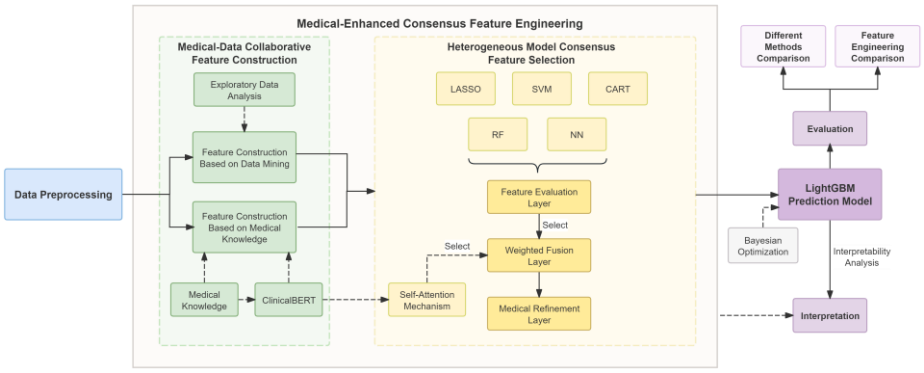


Fig. 1. Overview of Prediction Process.

4 Medical-Enhanced Consensus Feature Engineering

4.1 Medical-Data Collaborative Feature Construction

Data Preprocessing. To ensure modeling quality, raw data undergo systematic preprocessing. Missing values are imputed with the mean, and outliers are normalized via Z-score transformation. The dataset is split into 5,414 training and 1,354 test samples using an 8:2 ratio. To address class imbalance, oversampling is applied to create a balanced 1:1 distribution between diabetic and non-diabetic samples in the training set.

Feature Construction Based on Data Mining. This study conducts exploratory data analysis (EDA) using Pearson correlation and scatterplot matrices to identify redundancy and relevance among features.

Fig. 2 shows that certain indicators exhibit strong mutual correlations. For example, red blood cell count correlates highly with hemoglobin (r = 0.80) and hematocrit (r = 0.87), indicating strong collinearity due to their common role in reflecting oxygen-carrying capacity. Platelet count and plateletcrit exhibit a very high correlation (r=0.92), highlighting a linear relationship between platelet count and total platelet volume. Globulin is also strongly correlated with total protein (r=0.76), reflecting physiological relationships in plasma composition. A strong negative correlation(r=0.95) is observed between neutrophils and lymphocytes, suggesting an immune dynamic balance. Importantly, over 70% of the features exhibit weak correlations (|r|<0.5), suggesting low redundancy across most features. This indicates a strong potential for feature engineering to enhance the representational capacity of the feature space and improve model generalizability.

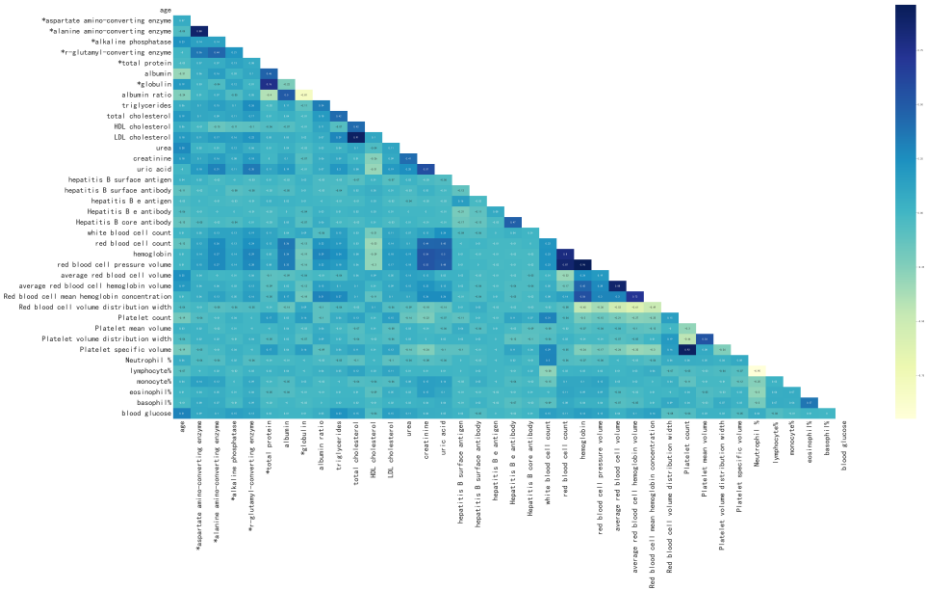


Fig. 2. Pearson Correlation Coefficient Matrix.

The scatterplot matrix is used to analyze the linear relationship between each feature and the target variable BG. As shown in Fig. 3, the overall linear correlation between most features and BG levels is weak. Certain features, such as globulin, triglycerides, and hepatitis B core antibody, show no clear correlation with BG levels. In contrast, features like platelet count, red blood cell indicators, and age exhibit a more noticeable relationship with BG. These findings suggest that subsequent feature engineering should focus on strengthening the relevant features while eliminating weakly associated ones, in order to optimize the feature set and enhance model performance.

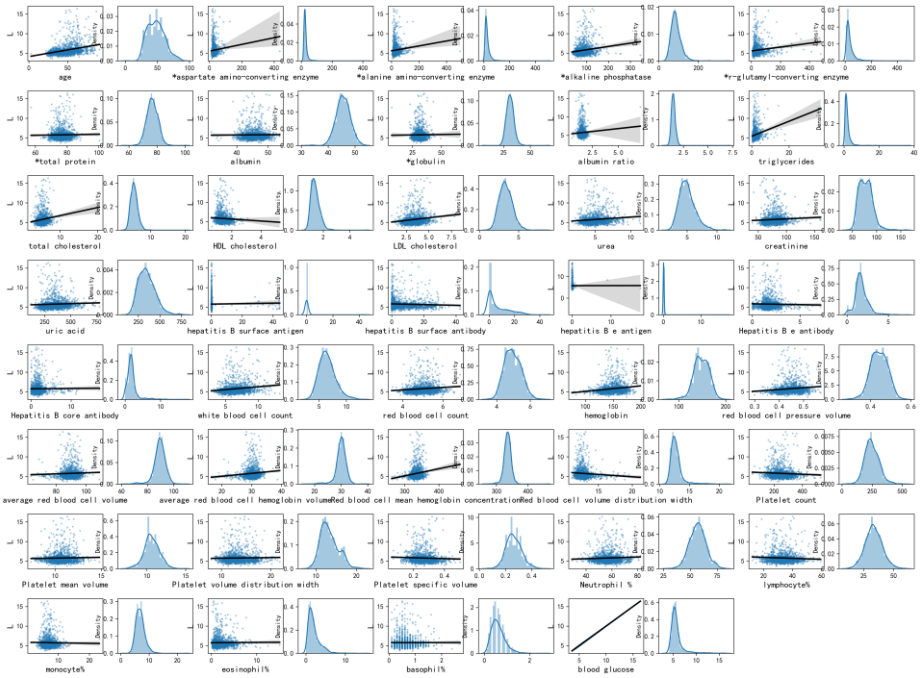


Fig. 3. Linear Correlation Scatter Matrix.

Based on these observations, this study applies arithmetic operations—such as addition, subtraction, multiplication, and division—as well as advanced transformations like feature crossing and polynomial expansion. These methods generate 51 composite features that aim to capture nonlinear interactions and latent physiological relationships within the data.

Feature Construction Based on Medical Knowledge. To integrate domain knowledge, this study utilizes ClinicalBERT to extract semantic embedding from medical text T:

$$E = \text{ClinicalBERT}(T), \quad X_{\text{gen}} = h(E) \tag{5}$$

These embeddings encode contextualized medical semantics, which are then transformed into candidate features via a nonlinear mapping function $h(\cdot)$. To ensure clinical validity, a domain knowledge filter M—constructed based on structured ontologies and rule-based constraints—is applied. This filter selects feature pairs (x_i, x_j) with confirmed clinical associations. The final refined feature set is computed as:

$$X_{\text{final}} = \{x_i \times x_j \mid (x_i, x_j) \in M\} \tag{6}$$

This dual construction strategy combines data-driven derivation with domain-guided validation. It yields 37 clinically meaningful composite features that improve both the

expressiveness and interpretability of the model input space. As a result, the proposed features enhance the performance of diabetes risk prediction and provide a more reliable basis for clinical decision-making.

4.2 Heterogeneous Model Consensus Feature Selection

To ensure stable, interpretable and clinically relevant feature selection, this study introduces the HM-CFS mechanism that integrates statistical, machine learning, and large language model–driven perspectives.

Heterogeneous Model Feature Importance Estimation Approach. This section introduces the feature importance estimation methods based on a set of heterogeneous models, including least absolute shrinkage and selection operator (LASSO) [18], SVM, classification and regression tree (CART) [19], RF, neural network (NN) [20].

LASSO : Performs linear regression with L1 regularization to enforce sparsity.

$$\min_{w} \frac{1}{2N}\sum_{i=1}^{N} (y_i - w^{\top}x_i)^2 + \lambda \parallel w \parallel_1 \tag{7}$$

Features with zero coefficients are excluded, making LASSO particularly effective for high-dimensional data where interpretability and parsimony are critical.

SVM: Incorporates L1 regularization into its margin-based objective:

$$\min_{w,b,\xi} \frac{1}{2} \parallel w \parallel_2^2 + C\sum_{i=1}^{N} \xi_i + \lambda \parallel w \parallel_1 \tag{8}$$

Only features that influence the optimal separating hyperplane retain non-zero weights. Kernelized versions further capture complex, nonlinear dependencies relevant to biomedical prediction tasks.

CART: Selects features based on their ability to reduce mean squared error (MSE) at each split:

$$\Delta\mathrm{MSE} = \mathrm{MSE}_t - \frac{|D_L|}{|D_t|}\mathrm{MSE}_L - \frac{|D_R|}{|D_t|}\mathrm{MSE}_R \tag{9}$$

This recursive partitioning yields interpretable decision rules and naturally handles mixed data types.

RF: Computes feature importance by measuring the increase in out-of-bag (OOB) error when a feature is randomly permuted:

$$I_j = \frac{1}{T}\sum_{t=1}^{T} \left(\mathrm{MSE}_{\mathrm{OOB},t} - \mathrm{MSE}_{\mathrm{OOB},t}^{(j)}\right) \tag{10}$$

This ensemble-based approach stabilizes selection across resampled subsets and captures nonlinear interactions missed by individual trees.

NN: Estimates feature relevance based on the input-layer weights, typically using the weight matrix $W^{(1)}$ from the input to the first hidden layer:

$$I_j = \sum_{i=1}^{H} |W_{ij}^{(1)}| \tag{11}$$

Larger cumulative weights indicate greater influence on the output. The capacity of NNs to model high-order dependencies makes them valuable in uncovering latent interactions in complex clinical datasets.

Hierarchical Multi-Perspective Weighted Fusion Mechanism. To consolidate the diverse feature importance scores derived from heterogeneous models, this study proposes a Hierarchical Multi-Perspective Weighted Fusion Mechanism (HMP-WFM), as illustrated in Fig. 4**.** This three-layer framework integrates model-derived importance, consistency-based weighting, attention-guided medical relevance, and external clinical validation, yielding a robust and interpretable feature subset for downstream modeling.

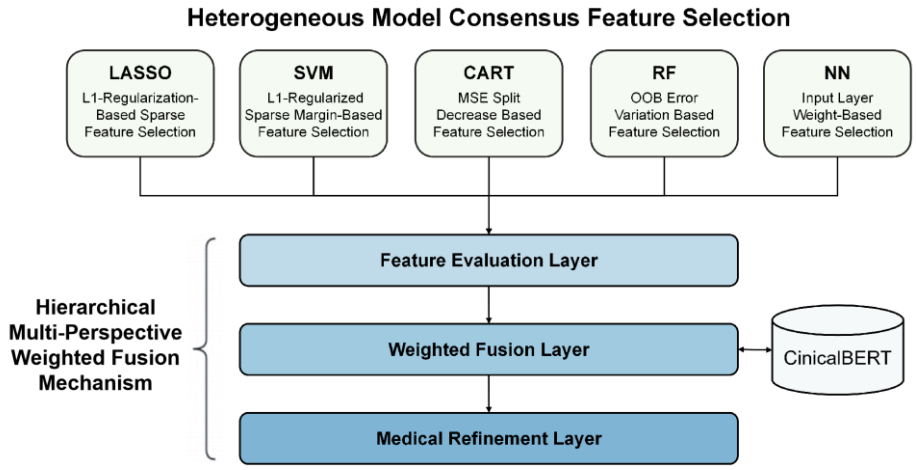


Fig. 4. Heterogeneous Model Consensus Feature Selection Process.

Feature Evaluation Layer. This study first computes a raw importance matrix $W_{\text{raw}} \in \mathbb{R}^{M\times 5}$, where each entry corresponds to the score of feature x_j as determined by one of five models.

Weighted Fusion Layer. To account for model agreement and medical relevance, this study designs two complementary weighting schemes:

1) Model Consistency is assessed by computing the average rank and variance of each feature across models. Features with high mean rank and low variance are emphasized using a stability-enhancing weight:

$$W_j^{\text{stability}} \propto \exp(-\text{Var}_j) \cdot \text{RankScore}_j \tag{12}$$

2) Clinical Relevance is derived using ClinicalBERT, which estimates the attention score α_j between each feature and BG. These scores are normalized and enhanced based on a confidence threshold:

$$W_j^{\text{BERT}} = \frac{\exp(\alpha_j)}{\sum_{k=1}^{M} \exp(\alpha_k)}, \quad W_j^{\text{BERT-adjusted}} = W_j^{\text{BERT}} \cdot (1 + \eta \cdot \Delta\alpha_j) \tag{13}$$

Medical Refinement Layer. Features verified through biomedical ontologies or literature receive an additive adjustment:

$$W_j^{\text{medical}} = W_j \cdot (1 + \rho_j) \tag{14}$$

where ρ_j reflects the strength of external clinical evidence.

Final Feature Importance Calculation. The final importance score of feature x_j is computed by weighted summation across all layers:

$$W_j^{\text{final}} = \lambda \cdot W_j^{\text{raw}} + \mu \cdot W_j^{\text{stability}} + \nu \cdot W_j^{\text{BERT-adjusted}} + \tau \cdot W_j^{\text{medical}} \tag{15}$$

The weight λ, μ, ν, τ are hyperparameters tuned via grid search.

Features with $W_j^{\text{final}} > \theta$ (threshold determined via cross-validation) form the selected subset:

$$X_{\text{selected}} = \{x_j \mid W_j^{\text{final}} > \theta\} \tag{16}$$

By fusing quantitative evaluation, model consensus, language-model insight, and domain knowledge, HMP-WFM enables interpretable, stable, and clinically grounded feature selection—laying a strong foundation for accurate diabetes risk modeling.

5 Experiments

5.1 Settings

To enable fine-grained diabetes risk modeling, this study adopts BG as a continuous predictive target, rather than a binary label. As a richer signal, BG better captures nonlinear relationships with features and supports more informative risk assessment.

Two comparative experiments are designed: (1) evaluating performance before and after MECFE-based feature enhancement under the same model, and (2) comparing multiple mainstream models post-enhancement to identify the optimal architecture. Models include LR, SVM, DT, RF, NN, XGBoost, CatBoost, and LightGBM. All models are trained using five-fold cross-validation on an 8:2 train-test split, with input features standardized.

Model performance is evaluated using the coefficient of determination (R^2), mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). Hyperparameters are tuned individually: LightGBM uses Bayesian optimization; XGBoost and CatBoost apply grid search with cross-validation; SVM and

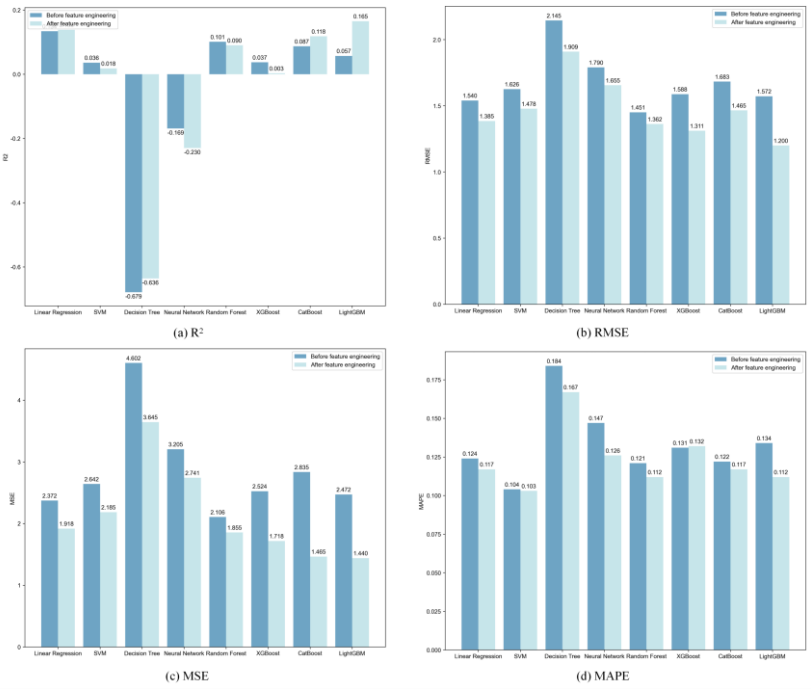
NN undergo targeted adjustments to kernels, regularization, and architecture. Experiments are executed on Ubuntu 22.04 with an NVIDIA RTX 3090 GPU and CUDA 12.3, using frameworks such as scikit-learn, XGBoost, LightGBM, and CatBoost.

To enhance interpretability, the final model is analyzed via feature importance rankings, offering insights into key predictive factors and further validating MECFE's ability to select clinically relevant features.

5.2 Results

The dataset is preprocessed through outlier removal, missing value imputation, and standard normalization to ensure consistency across numerical features. Beyond the original variables, 51 features are derived via data-driven methods and 37 are constructed based on medical domain knowledge, resulting in 126 candidate features. The HM-CFS module is employed to quantify feature importance and select the top 40 informative features. Finally, a Bayesian-optimized LightGBM model is trained on the refined feature set to achieve improved predictive performance.

Feature Engineering Effectiveness Evaluation. To evaluate the effectiveness of MECFE, this section compares model performance under scenarios with and without the proposed feature engineering approach.



(a) R^2 (b) RMSE (c) MSE (d) MAPE

Fig. 5. Comparative Assessment Before and After Feature Engineering.

As shown in Fig. 5(a), the R^2 values of most models increase after applying MECFE, with notable improvements observed in DT, CatBoost, and LightGBM, whose R^2 values rise by 0.043, 0.031, and 0.108, respectively. It indicates that the feature enhancement process substantially improves model fitting. Meanwhile, Fig. 5(b) and Fig. 5(c) show consistent reductions in RMSE and MSE. Specifically, the RMSE of LightGBM decreases by 23.67%, while its MSE drops by 41.75%. On average, the RMSE across all models reduces by 12.24%, and MSE by 40.67%, demonstrating that MECFE effectively lowers prediction errors. Fig. 5(d) further reveals improvements in MAPE, with LightGBM achieving a reduction of 16.42%, while SVM shows the smallest change with a 0.96% decrease. The average MAPE reduction across models is 11.85%. The results highlight MECFE's effectiveness in reducing prediction errors and mitigating model bias.

Model Comparison Evaluation. In this section, a comprehensive evaluation of various models' performance is conducted.

Table 1. Model Prediction Results.

Algorithm Name	R^2	RMSE	MSE	MAPE
LR	0.138	1.385	1.918	0.117
SVM	0.018	1.478	2.185	0.103
DT	-0.636	1.909	3.645	0.167
NN	-0.230	1.655	2.741	0.126
RF	0.090	1.362	1.855	0.112
XGBoost	0.108	1.239	1.536	0.117
CatBoost	0.118	1.465	1.465	0.117
LightGBM	0.165	1.200	1.440	0.112

Table 1 shows that LightGBM achieves the highest R^2 score of 0.165, indicating the strongest fitting capability. It also records the lowest RMSE and MSE values, at 1.200 and 1.440 respectively, reflecting minimal prediction error. CatBoost and XGBoost also demonstrate competitive performance, particularly in error control. In contrast, Decision Tree and Neural Network models perform poorly, with negative R^2 values and relatively high RMSE and MSE, suggesting weak predictive capability. In terms of MAPE, SVM achieves the best performance, followed closely by LightGBM and Random Forest. Overall, LightGBM offers the most balanced and reliable performance, making it the optimal choice for downstream analysis of diabetes-related risk factors.

Interpretability Analysis. To better understand and analyze the intermediate process of model prediction and identify the key features in diabetes risk prediction, an in-depth analysis of the feature importance plot of the LightGBM model is conducted.

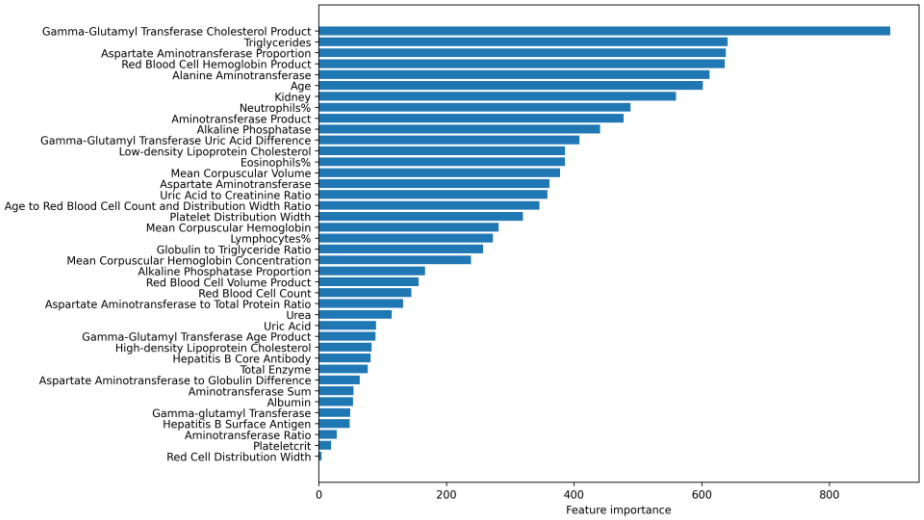


Fig. 6. LightGBM Feature Importance Results.

Fig. 6 presents the feature importance distribution derived from the LightGBM model, highlighting the relative contribution of each variable to the final prediction. Notably, the Gamma-Glutamyl Transferase Cholesterol Product exhibits the highest importance, underscoring its critical role in reflecting liver metabolic dysfunction. Other top-ranking features such as Triglycerides, Aspartate Aminotransferase Proportion, and Red Blood Cell Hemoglobin Product further reinforce the clinical relevance of liver enzymes and lipid metabolism in diabetes risk prediction. Several high-impact features are derived through MECFE's enhancement strategy, including Aminotransferase Product and Gamma-Glutamyl Transferase Uric Acid Difference, demonstrating the approach's ability to capture meaningful interactions among biological markers. Conversely, features like Red Cell Distribution Width and Plateletcrit show minimal influence. Overall, the highly ranked features align closely with those selected by the MECFE framework, validating its effectiveness in enhancing model feature representation.

6 Conclusion

This study proposes a novel MECFE approach to address the challenges in early diabetes prediction. By integrating both data-driven methods and medical prior knowledge in the MD-CFC component, and employing a multi-model consensus approach for feature selection in the HM-CFS component, MECFE enhances diabetes risk prediction performance, feature stability, and interpretability. Experimental results demonstrate consistent gains across multiple metrics (R^2, RMSE, MSE, MAPE), while feature importance visualization further enhances model explainability. Nevertheless, as validation is currently limited to public datasets, future work should evaluate MECFE in real-world clinical settings to assess its practical applicability.

Acknowledgments. This study is supported by the Key Research and Development Program of Zhejiang Province (2025C01135).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Heald, A.H., Stedman, M., Davies, M., et al.: Estimating life years lost to diabetes: Outcomes from analysis of National Diabetes Audit and Office of National Statistics data. Cardiovascular Endocrinology & Metabolism 9(4), 183–185 (2020)
2. International Diabetes Federation.: IDF diabetes atlas, 10th ed. Belgium: International Diabetes Federation. Retrieved from https://diabetesatlas.org/atlas/tenth-edition/ (2021)
3. Lin, X., Xu, Y., Pan, X., et al.: Global, regional, and national burden and trend of diabetes in 195 countries and territories: An analysis from 1990 to 2025. Scientific Reports 10(1), 1–11 (2020)
4. Ong, K.L., Stafford, L.K., McLaughlin, S.A., et al.: Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: A systematic analysis for the Global Burden of Disease Study 2021. The Lancet 402(10397), 203–234 (2023)
5. Abel, E.D., Gloyn, A.L., Evans-Molina, C., et al.: Diabetes mellitus—Progress and opportunities in the evolving epidemic. Cell 187(15), 3789–3820 (2024)
6. Uddin, S., Khan, A., Hossain, M.E., et al.: Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making 19(1), 1–16 (2019)
7. VijiyaKumar, K., Lavanya, B., Nirmala, I., et al.: Random forest algorithm for the prediction of diabetes. In: 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), pp. 1–5. IEEE (2019)
8. Prabha, A., Yadav, J., Rani, A., et al.: Design of intelligent diabetes mellitus detection system using hybrid feature selection-based XGBoost classifier. Computers in Biology and Medicine 136, 104664 (2021)
9. Yu, W., Liu, T., Valdez, R., et al.: Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making 10, 1–7 (2010)
10. Zhou, H., Myrzashova, R., Zheng, R.: Diabetes prediction model based on an enhanced deep neural network. EURASIP Journal on Wireless Communications and Networking 2020, 1–13 (2020)
11. Alex, S.A., Jhanjhi, N.Z., Humayun, M., et al.: Deep LSTM model for diabetes prediction with class balancing by SMOTE. Electronics 11(17), 2737 (2022)
12. Ali, M.S., Islam, M.K., Das, A.A., et al.: A novel approach for best parameters selection and feature engineering to analyze and detect diabetes: Machine learning insights. BioMed Research International 2023(1), 8583210 (2023)
13. Kakoly, I.J., Hoque, M.R., Hasan, N.: Data-driven diabetes risk factor prediction using machine learning algorithms with feature selection technique. Sustainability 15(6), 4930 (2023)
14. Khanam, J.J., Foo, S.Y.: A comparison of machine learning algorithms for diabetes prediction. ICT Express 7(4), 432–439 (2021)
15. Barbieri, M.C., Grisci, B.I., Dorn, M.: Analysis and comparison of feature selection methods towards performance and stability. Expert Systems with Applications, 123667 (2024)

16. Alsentzer, E., Murphy, J.R., Boag, W., et al.: Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323 (2019)
17. Ke, G., Meng, Q., Finley, T., et al.: LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30 (2017)
18. Zou, H.: The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101(476), 1418–1429 (2006)
19. Breiman, L., Friedman, J., Olshen, R.A., et al.: Classification and Regression Trees. Routledge (2017)
20. Gurney, K.: An Introduction to Neural Networks. CRC Press (2018)