WEP Warwick Evans Publishing

Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 14
TPCEE 2026

Data-Driven Statistical Analysis and Predictive Modeling of Diabetes Risks Based on Physiological Indicators

Yunyu Xie *

College of Sciences, Shanghai University, Shanghai, China

* Corresponding Author Email: tom_tse0655@shu.edu.cn

Abstract. Nowadays, T2DM is a serious issue across the world that needs earlier detection. Physiological monitoring is commonly carried out, but the traditional clinical model is unable to deal with the complexity and non-linearity of the data. Currently, a serious research problem of how to build a predictive system with high computational accuracy and also be interpretable enough for clinical application has yet to be solved. This paper will conduct research using the PIDD to examine the problem of T2DM risk assessment with a data-driven approach. A custom MLP with an 8-32-16-1 architecture is used to learn the various associations between physiological indicators and diabetes. Dropout regularisation was added to prevent overfitting, and simultaneously, weight visualisation methods were employed to improve the interpretability and transparency of the model's internal decision-making process. According to the statistical analysis, the most significant contributors are glucose concentration, BMI, and age in order of impact. Based on the above, the optimal MLP has high precision and is relatively interpretable in terms of variable importance. The above method can be utilised to construct a general model of modern medical information systems and support the implementation of evidence-based medicine.

Keywords: Type 2 Diabetes Mellitus; Multi-Layer Perceptron; Risk Assessment; Model Interpretability; Medical Informatics.

1. Introduction

The pioneering research on Type 2 Diabetes Mellitus (T2DM) risk assessment and predictive models has generally focused on identifying the biological and genetic factors that underlie glucose metabolism, such as the role of the PBX1 gene [1]. To address the above problem in time, prompt intervention and suitable graded management for T2DM need to be carried out. As a result, the construction of an early warning system for this widespread disease is urgently required. With the increase in the number of diabetic people with diabetes, many scientists have begun studying all kinds of physical changes that take place in diabetic patients and quantifying them. The 11th edition of the IDF Diabetes Atlas also shows that approximately 11.1% of adults in the world are living with diabetes today, and by 2050, this is expected to reach as high as 853 million [2].

Recently, a systematic review has shown that machine learning models can be built based on these readily available metabolic markers so as to conduct early non-invasive diagnosis of various diseases and uncover many high-dimensional correlations that traditional linear models may fail to recognise [3]. Meta-analyses of clinical data have also confirmed that, in terms of all other races, fasting plasma glucose and BMI are still the most reliable indicators for predicting the risk of T2DM, thus providing a standard foundation for large-scale predictive informatics [4]. Therefore, early computational work used special methods to handle the noise in the biological data, such as fuzzy support vector machines and ensemble learning like XGBoost [5, 6]. Although the models have shown some improvements in prediction stability, they are still not capable of generating deep and hierarchical feature representations autonomously to capture the full complexity of metabolic syndrome.

Lately, some scholars have introduced deep neural networks so as to improve the accuracy of diabetes decision-making systems by automatically learning knotty features from raw physiological data [7]. A Multi-Layer Perceptron (MLP) is not a simple linear model but it has hidden layers and non-linear activation functions to learn from non-linear relationships among the data. Advanced configurations,

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

such as dropout layers for regularisation and the Adam optimiser, have shown good generalisation performance in diabetes risk classification within PIDD [8].

At the same time, given that deep learning is a black box model, it has difficulty meeting the interpretability requirements for clinical application. For the time being, the two paths of simple but interpretable statistics and high-precision but opaque modelling are mutually exclusive commonly. Therefore, the practical application of high-level models in medicine is limited, and at the same time, understanding the weight of each physiological predictor is as crucial as the prediction itself.

In response to the above problems, this paper utilises the PIDD to address the growing global epidemic of T2DM and achieve accurate, understandable early-warning functions. To address the shortcomings of previous studies, a transparent 8-32-16-1 MLP architecture has been constructed in this paper to identify non-linear diabetic risks in the PIDD, as traditional clinical assessments are based on static thresholds and fail to figure out the complicated non-linear relationships among physiological indicators. Dropout regularisation and weight visualisation have been applied in this study to help the high-precision model become more interpretable for clinical causality analysis. In combination with deep predictive modelling, this work will also aim to enhance the interpretability of the structure for clinical decision-making.

2. Method

2.1. Data Source and Explanation

The empirical foundation of the paper is the Pima Indians Diabetes Dataset (PIDD), which was first compiled by the National Institute of Diabetes and Digestive and Kidney Diseases, and is now maintained by the UCI Machine Learning Repository [9]. The raw dataset contains 768 clinical observations, and each observation is composed of 8 physiological indicators and a binary outcome for female patients of Pima Indian ancestry.

A high-precision pipeline has been established to ensure the accuracy and reliability of the analysis results. First of all, biological zero values were identified in five essential features, namely glucose, blood pressure, skin thickness, insulin, and BMI, which are physiologically implausible in living subjects. These abnormal items were treated as missing values. Then, several cleanups were performed subsequently. Redundant records were eliminated, and samples with missing indicators exceeding a threshold of more than three null values were discarded to maintain the stability of the feature space. The refined dataset has been reduced to 761 high-fidelity examples.

2.2. Selection of Features and Explanation

The four general indicators of glucose, BMI, age, and insulin were selected as the input feature space after preprocessing for predictive modelling. They have been picked out as they are clinically relevant and serve as basic metabolic and demographic indices of T2DM risk.

The glucose concentration of this group's average and range were 120.9±32.0 mg/dL and 44 to 199 mg/dL, respectively. The mean BMI of the physical profile was 32.0±7.9 kg/m^2, and the range was from 18.2 to 67.1 kg/m^2. Therefore, most subjects are corpulent. Age and endocrinology information were added to the previous list, including an average of 33.2±11.8 years across the 21-to-81 age range, and insulin levels had a large spread, ranging from 14 μU/mL to 846 μU/mL, with a mean of 79.8±115.2 μU/mL.

2.3. Research Methodology

The methodology uses a hierarchical analytical structure to connect the relationship between statistical correlation and precise predictive modeling, starting from a simple base to a complicated multi-layer perceptron.

2.3.1. Foundational Statistical Baselines

According to traditional statistical results, older people are generally more likely to have high blood sugar. However, as a first-order approximation, linear regression is inherently incapable of addressing the multidimensional variation in T2DM. A white-box Decision Tree (DT) classifier was used to obtain the initial indicators. The DT recognises glucose and BMI as the basic reasons, but due to a shallow architecture of depth 4, it only obtains coarse, axis-aligned decision boundaries. The aforesaid baseline models only show a general trend of the data and cannot reveal the internal structure or how the physiological indicators work together to cause this change.

2.3.2. The Architecture of the Multi-layer Perceptron

Given the deficiencies of the prior linear and shallow baselines, a custom MLP high-precision analytical engine has been designed in this study. An MLP with an 8-32-16-1 hierarchical structure has been set up to solve the problem of non-linearity and other deficiencies in older models. While linear regression is just a single-variable model, a 32-neuron first hidden layer is used in this case to map the 8 input features to a high-dimensional space. A 16-neuron abstraction layer is utilised to further compress these signals for stable diagnostic prediction. Using the ReLU activation function and multiple layers of depth, the MLP can learn the complex, non-linear relationship between BMI and insulin sensitivity in a given region and achieve a higher level of prediction granularity than the rigid if-then rules of a DT or the fixed slope of a linear model.

2.3.3. Training Mechanisms and Regulation Approaches

Multiple regularisation layers are added in the training pipeline to ensure good generalisation and avoid learning stochastic noise. A dropout rate of 0.2 is used between hidden layers, and some neurons are randomly disabled during training to help the network learn more robust, stable features. Additionally, L2 regularisation with a weight decay of 10^{-4} is added to the cost function to reduce the magnitude of weights and thus make the decision surface smoother. Binary cross-entropy with logits-loss is used to calculate the prediction error, and then the Adam optimiser is employed with a learning rate of 0.001 for optimisation. The above arrangement ensures a steady convergence over 20 epochs and is thus more stable than a fast but unstable method.

2.3.4. Evaluation and Structural Understandability

Multiple performance assessments are carried out to find a good threshold for the diagnostic boundary based on three indicators, namely accuracy, sensitivity, and AUC. To reduce the black-box nature of the architecture, weight heatmap analysis of the first 32×8 weight matrix is performed. The aforesaid operation helps to determine the importance of different input features reflected in hidden units and aligns the results with real clinical practice.

3. Results and Discussion

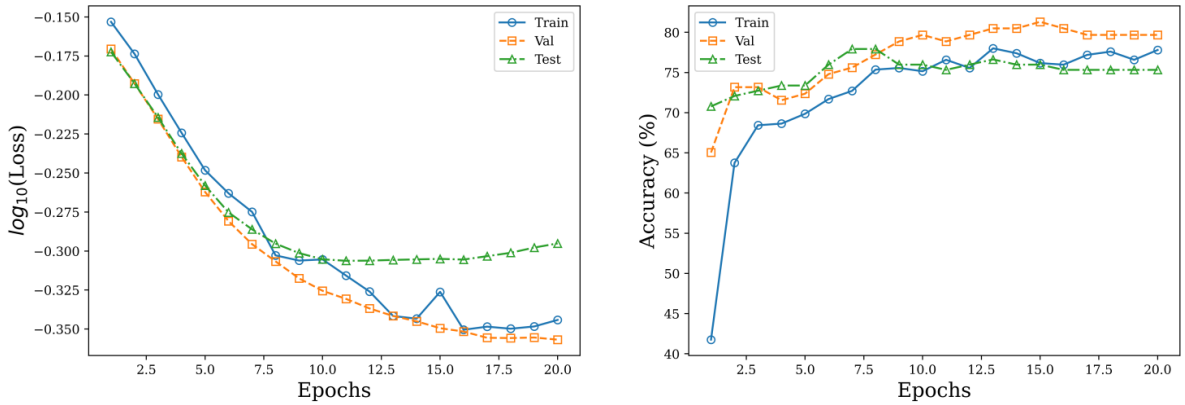
3.1. Preliminary Investigation and Base Evaluation

Logistic regression was employed to construct the predictive baseline for diabetes risk. The R^2 value for the model of age and blood glucose was 0.0954. The interpretable DT classifier had a baseline accuracy of 72.77%. Feature importance analysis of the DT model showed that glucose at 0.50 and BMI at 0.28 were the leading features reducing Gini impurity.

3.2. Deep Learning Performance

The main subject of this investigation, an 8-32-16-1 MLP structure, showed better ability in generating high-level feature representations, overcoming the linear limitations of the basic models to form more reliable classification boundaries.

3.2.1. Training Mechanisms and Convergence



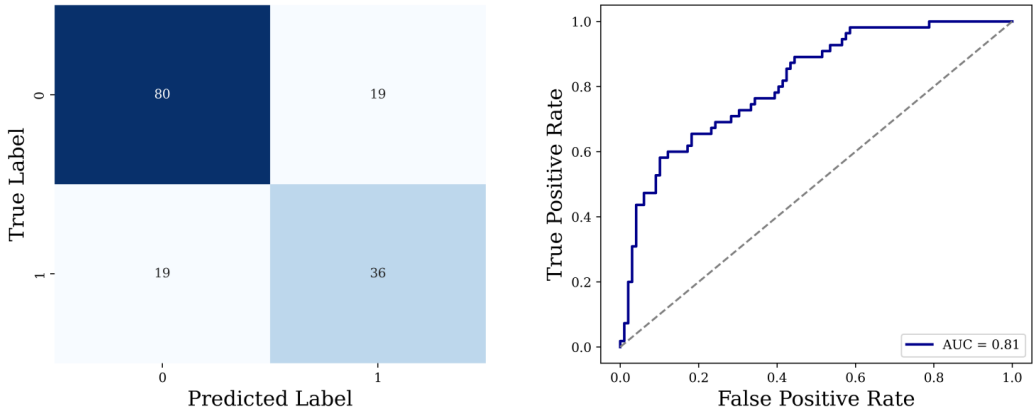
(a) (b)

Fig. 1 (a) Loss curve showing the rapid convergence of binary cross-entropy loss on a logarithmic scale for the proposed 8-32-16-1 MLP model during training, validation and testing. (b) Accuracy progression over 20 epochs of classification performance for training, validation and testing sets of the model (Photo/Picture credit: Original).

The MLP was optimised for 20 epochs with the Adam optimiser. A relatively sharp drop in the loss confirmed that the 8-32-16-1 MLP structure could effectively learn the non-linear patterns in the PIDD features. Fig. 1 shows the training dynamics. As shown in Fig. 1 (a), the loss curve logarithmically declines rapidly in the first 5 epochs. After approximately 15 epochs, all the training, validation, and testing losses have converged. Therefore, it can be observed that both dropout at 0.2 and L2 regularisation at 10^{-4} have successfully reduced overfitting. Fig. 1 (b) shows that the corresponding accuracy also rises relatively steadily and is around 81% at validation.

3.2.2. Classification Robustness and Separation

The MLP obtained the optimal decision boundaries by minimising the binary cross-entropy loss, which contributed to an equal predictive ability for both healthy and diabetic classes. The classification metrics are shown in Fig. 2.



(a) (b)

Fig. 2 Classification performance metrics of the 8-32-16-1 MLP on the test set. (a) Confusion matrix shows the distribution of true and predicted classes, which has a relatively balanced predictive ability. (b) The Receiver Operating Characteristic (ROC) curve indicates the discriminatory power of the model (Photo/Picture credit: Original).

As shown in Fig. 2 (a), the confusion matrix indicates that the MLP correctly identifies 80 healthy individuals as true negatives and 36 diabetic patients as true positives, and has a balanced misclassification rate of 19 false positives and 19 false negatives. In Fig. 2 (b), it can be seen that the ROC curve has an AUC of 0.81.

3.3. Structural Interpretability and Physiological Consistency

A weight heatmap is used to show the internal logic of the network in a quantitative way, and it is consistent with the clinical theories of diabesity. While glucose and BMI remain the dominant drivers of the model's firing patterns, the heatmap in Fig. 3 also reveals significant synergistic interactions involving age and insulin across the 32 hidden units.

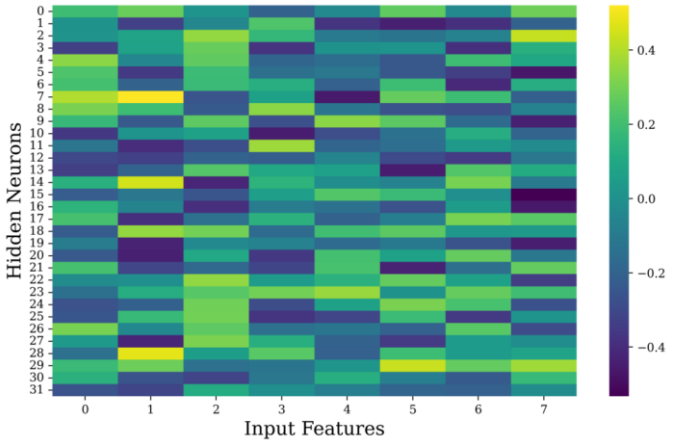


Fig. 3 Weight heatmap of the first hidden layer $(W^{(1)} \in \mathbb{R}^{32\times8})$ in the 8-32-16-1 MLP architecture. The intensity of the colours is the strength of the synaptic connection between one of the 8 input physiological parameters and a specific neuron. Therefore, neurons with higher intensity have been given more weight in the learning process.

Based on the results of the qualitative analysis of the heatmap, columns for glucose and BMI have a significantly higher weight intensity in most of the neurons compared to the other variables. Consequently, this model shows that the two are the main factors for diabetes risk. Fig. 3 is a visualisation of the abstract numerical weights in a spatial form, and by connecting advanced neural computation with physiological causality, it has provided the necessary glass-box transparency for clinical trust.

3.4. Discussion and Clinical Relevance

This study has shown that non-linear models are required for the risk assessment of T2DM. Some of the main deficiencies have still been identified. A sample size of 761 is too small, and since the PIDD is particular to a certain group, it may not be widely applicable. Recently, research on the PIDD has shown that although neural network structures such as GRU or MLP have good internal validity, their performance varies when faced with real-world Electronic Health Records (EHRs) from different demographic groups [4, 8]. Therefore, many more ethnic datasets need to be obtained, and improvements to the structure must be made for wider clinical application. Given the above shortcomings, the following studies should focus on external validation across multi-center datasets to ensure the model's cross-ethnic robustness, and subsequent rounds can use Automated Machine Learning (AutoML) for hyperparameter optimisation instead of manual tuning [10].

4. Conclusion

A multi-stage analytical framework for the predictive modeling of diabetes risk is used in this paper, and the Pima Indians Diabetes Dataset is employed. By transitioning from foundational linear correlation analysis to a tailored multi-layer perceptron architecture, deep learning has transcended the limitations of traditional statistical models in capturing metabolic interactions.

First of all, the preliminary statistical analysis reinforced the link between aging and glycemic decline, yet confirmed that age alone was an insufficient predictor. The 8-32-16-1 MLP engine showed better classification accuracy and discriminative power than the baseline decision tree. The integration of weight heatmap analysis and feature importance ranking injected interpretability into the neural

network. By quantifying the dominant role of glucose and BMI, this research bridges the gap between black box computation and clinical causality.

In short, this glass box method is a high-reliability and open-display approach to conduct early diabetes detection. Dropout regularisation and weight visualisation are feasible ways to expand the application range of current medical informatics. The results underscore that both the combination of deep neural network features and physiological interpretability are crucial for personalised risk assessment for Type 2 Diabetes Mellitus.

References

[1] Thameem F, Farook V S, Bogardus C, Prochazka M. Analysis of PBX1 as a candidate gene for type 2 diabetes mellitus in Pima Indians. Biochimica et Biophysica Acta - Molecular Basis of Disease, 2001, 1537(3): 231-236.

[2] International Diabetes Federation. IDF Diabetes Atlas. 11th ed. Brussels: International Diabetes Federation, 2024.

[3] Li X, et al. Performance of diabetes risk prediction models: A systematic review and meta-analysis. Scientific Reports, 2025.

[4] Shams M Y, et al. A systematic review of machine learning models for predicting type 2 diabetes mellitus using electronic health records. Asian Journal of Electrical Sciences, 2025.

[5] Tang Z W. Research on ensemble learning model based on XGBoost and its application. Changsha: Changsha University of Science & Technology, 2021.

[6] Su H M. Research on classification method of gene expression data based on fuzzy support vector machine. Changsha: Hunan University, 2009.

[7] Lu C C. Research on diabetes decision algorithm based on deep learning. Hangzhou: Zhejiang Sci-Tech University, 2018.

[8] Shams M Y, Abd El-Hafeez T, El-Henawy I M. A novel RFE-GRU model for diabetes classification using PIMA Indian dataset. Scientific Reports, 2025, 15(1): 1234.

[9] National Institute of Diabetes and Digestive and Kidney Diseases. Pima Indians Diabetes Database. Kaggle, 1990. Accessed Apr. 1, 2026.

[10] Zhao Y Q. Research on gestational diabetes prediction method based on historical diagnostic data. Changsha: Changsha University of Science & Technology, 2022.