

Diabetes forecasting by analyzing electronic health record data

Vo Nhat Tan Huynh^{1,2}, Nhat Tan Le^{1,2,*}



Use your smartphone to scan this QR code and download this article

ABSTRACT

In recent years, the prevalence of diabetes has surged, presenting a significant challenge to health-care systems worldwide. To address this issue, innovative approaches are imperative to enhance prediction and management capabilities. Consequently, this study underscores the potential of utilizing Electronic Health Records (EHRs) for diabetes prediction through data mining methods. The primary objectives entail forecasting diabetes based on EHRs using Data Mining and Machine Learning techniques, followed by the analysis of crucial features associated with diabetes risk. To achieve these aims, the research leveraged EHR data sourced from The Korean Health Records of the Korean Genome and Epidemiology (KoGES), comprising five data files spanning five different time points. Machine Learning models were then employed for diabetes forecasting. These files were subdivided into four scenarios, each representing forecasting with varying numbers of time points (2, 3, 4, and 5), aimed at assessing diabetes prediction across multiple scenarios and comparing model performances. Then, Machine Learning models were applied to the processed data for forecasting across these scenarios. The final phase involved feature analysis utilizing the SHAP method. The results indicate that the models exhibited optimal performance in the scenario with five time points, yielding AUC-PR and Weighted F1 scores of 0.94 and 0.97, respectively, for the Random Forest model. Additionally, the model demonstrated commendable Precision and Recall metrics for each class, especially in diabetes class, with 0.86 and 0.75 respectively. The analysis of important features revealed associations with future diabetes risk. These outcomes suggest that the models were effectively constructed and trained, exhibiting relatively high predictive precision. However, certain challenges remain, including suboptimal model performance and the need to improve SHAP's interpretability concerning features of medium and small importance.

Key words: Diabetes forecasting, electronic health record, machine learning

¹Faculty of Applied Science, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, Dien Hong Ward, Ho Chi Minh City, Vietnam

²Vietnam National University Ho Chi Minh City, Linh Trung Ward, Ho Chi Minh City, Vietnam

Correspondence

Nhat Tan Le, Faculty of Applied Science, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, Dien Hong Ward, Ho Chi Minh City, Vietnam

Vietnam National University Ho Chi Minh City, Linh Trung Ward, Ho Chi Minh City, Vietnam

Email: lenhattan@hcmut.edu.vn

History

- Received: 19-08-2024
- Revised: 26-09-2024
- Accepted: 16-09-2026
- Published Online: 25-09-2026

DOI : <https://doi.org/10.32508/vnuhcmj-et.v9i3.1430>



INTRODUCTION

In the era of digital and technological breakthroughs, the healthcare systems are gradually escaping from its previous cumbersome and bureaucratic situation towards the digitalization of everything. In which, the electronic health records (EHR) is leveraged to streamline patient information management and enhance clinical decision – making, as well as forecasting the onset of some chronic diseases. One of the challenges in the contemporary healthcare is that the prevalence of people with chronic diseases is proliferating, in which diabetes stands out as a significant global health concern. Because it is not straightforward to detect any sign of this disease's onset, it can only be diagnosed by performing relevant tests but this also means that the patient at that time may already have obvious symptoms of the disease and there would be likelihood that it has progressed to Type 2 Diabetes. Thus, an innovative approach to forecast and manage the disease effectively is called for. And in

this circumstance, the intersection of EHR data mining and diabetes forecasting appears as a promising field in healthcare research.

An Electronic Health Record (EHR) is a digital collection of patient information encompassing medical history, demographics, laboratory outcomes, medication records and other vital health data.... In comparison with cumbersome paper health records, EHR allows us to retrieve information smoothly and can be accessed by numerous departments simultaneously. Besides, EHR is able to enhance accuracy, efficiency and the quality of healthcare information. And above all, this system prioritizes privacy and security of patients, mitigating medical malfunctions. A typical EHR system includes all patient information, from their health status to their medical history. Personal data in an EHR can be divided into the following groups: socio-demographics, medical history, medication history, immunization records, laboratory results, vital signs... Apart from the above groups, there are other categories of data in EHR, all of them can

Cite this article : Huynh V N T, Le N T. Diabetes forecasting by analyzing electronic health record data. VNUHCM J. Eng. Technol. 2026; 9(3):3380-3388.

Copyright

© VNUHCM Journal. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

give healthcare providers insights into a patient's status, as well as personal information¹.

The exploration at the relationship of EHR data mining and diabetes forecasting is not only a success in technology but also a critical milestone toward proactive and preventive healthcare. Thanks to the ability to predict the onset or progression of diabetes through data analytics, patients now are able to actively participate in their own healthcare journey, which makes them more responsible for their health. In the context, when the global medicine is moving towards 4P healthcare system (predictive, preventive, personalized, participatory), this research not only addresses the immediate ordeals of diabetes management but also aligns with the broader target of transforming healthcare into a more patient-centric and preventive system.

RELATED WORKS

With the outstanding features of Machine Learning, it is not perplexing to understand why this tool is employed in many fields, such as: automation, engineering..., especially in healthcare. In this realm of medicine, Machine Learning is utilized to forecast some diseases, or improve the performance of patient care... based on an immense amount of data from EHR. Let's delve into the forecasting aspect of Machine Learning. There has been a lot of scientific research on this method just the past few years:

Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records

In 2019, a group of researchers from many famous universities in countries such as: New Zealand, Vietnam, Singapore, UK... , enhanced the prediction of Type 2 Diabetes Mellitus onset by utilizing a wide and deep learning model - a subset of ML, combining a generalized linear model with a deep feed - forward neural network. The study focused on leveraging EHR to forecast diabetes onset, from which improving the quality and efficiency of healthcare².

The proposed method involves training a wide range of models with a logistic loss function through stochastic gradient descent. Employing public hospital records from Practice Fusion EHR for the US population, the research encompasses de-identified EHR for 9,948 patients, with 1,904 diagnosed with Type 2 Diabetes Mellitus. Forecasts for diabetes in 2012 are based on a dataset from 2009 - 2011. Addressing class imbalance, Synthetic Minority Oversampling Technique (SMOTE) is applied during cross - validation

training folds, creating synthetic samples for the minority class. SMOTE ratios of 150% and 300% are used, leading to diabetes: non-diabetes distributions of 1:2 and 1:1 in the training set, respectively².

In terms of the outcome, the final ensemble model, without SMOTE, achieves an accuracy of 84.28%, an area under the receiver operating characteristic curve (AUC) of 84.13%, sensitivity of 31.17%, and specificity of 96.85%. Employing SMOTE at 150% and 300% does not improve AUC (83.33% and 82.12%, respectively) but enhances sensitivity (49.40% and 71.57%, respectively) with a moderate decrease in specificity (90.16% and 76.59%, respectively)².

Chronic disease prediction using common data model

This study analyzed data from 790,822 individuals aged 20 and older, each with at least a year of hospital records from Ajou University Hospital in Suwon, South Korea, spanning from 1999 to 2020. The primary objective was to predict the risk of developing chronic diseases as these individuals age. Notably, the study excluded patients already diagnosed with chronic conditions such as type 2 diabetes, high blood pressure, hyperlipidemia, and cardiovascular disease. This exclusion criterion was crucial for focusing the research on identifying new chronic disease risks, rather than tracking the progression of pre-existing conditions³.

In terms of methods, they used a combination of 4 Ensemble Learnings: Logistic Regression, Random Forest, Gradient Boosting, and XGBoost. These algorithms were used to classify labels and explain the effects of variables on labels in predictive research³.

And at the end result, XGBoost presented the highest predictive performance for the 4 diseases (diabetes, hypertension, hyperlipidemia, cardiovascular disease) of 80% or more (0.84 to 0.93 in AUC standards) showing the best performance³.

Machine learning for the prediction of new-onset diabetes during 5-year follow-up in non-diabetic patients with cardiovascular risks

The data used in this study were obtained retrospectively from the EMR database of Korea University Guro Hospital (KUGH), and the protocol was approved by the Medical Device Institutional Review Board at KUGH (#KUGH 13017). The initially acquired subjects in the study comprised 52631 individuals (426182 visits during the study period) who visited the cardiovascular center of KUGH from January 2004 to December 2008⁴.

In this study, a total of 8454 patients who had no history of diabetes and were treated at the cardiovascular center of Korea University Guro Hospital were enrolled. All subjects completed 5 years of follow up. The prevalence of type 2 diabetes mellitus during follow up was 4.78% (404/8454). A total of 28 variables were extracted from the EMRs. In order to verify the cross-validation test according to the prediction model, logistic regression (LR), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and K-nearest neighbor (KNN) algorithm models were generated. The LR model was considered as the existing statistical analysis method⁴.

In terms of the end result, all predictive models maintained a change within the standard deviation of area under the curve (AUC) <0.01 in the analysis after a 10-fold cross-validation test. Among all predictive models, the LR learning model showed the highest prediction performance, with an AUC of 0.78. However, compared to the LR model, the LDA, QDA, and KNN models did not show a statistically significant difference⁴.

Prediction model for hypertension and diabetes using Korean Public Health Examination Data (2002-2017)

The research used the National Health Insurance Corporation database of Korea and checked whether hypertension or diabetes occurred from that date according to the number of health checkups over the past 10 years⁵.

And technically, two models were utilized in this study: XGBoost and Logistic Regression⁵.

Even though they do not use any special Machine Learning techniques, the model performance is very good. Specifically, Predicting hypertension (accuracy, 0.828 vs. 0.628; F1-score, 0.800 vs. 0.633; AUC, 0.828 vs. 0.630); diabetes mellitus (accuracy, 0.707 vs. 0.575; F1-score, 0.663 vs. 0.576; AUC, 0.710 vs. 0.575) and both diseases (accuracy, 0.950 vs. 0.612; F1-score, 0.950 vs. 0.614; AUC, 0.952 vs. 0.612). It was found that health check-ups had a great influence on the occurrence of hypertension and diabetes, and screening frequency was more important than other factors in the crucial variables⁵.

METHODOLOGY

Dataset and data extraction

After grasping knowledge and referencing from previous related research projects, The Korean Health Records dataset from the research work of The Korean Genome and Epidemiology (KoGES) is leveraged for

the study⁵. KoGES is an extensive cohort study initiative aimed at investigating the reciprocal relationship between gene - environment factors in prevalent chronic diseases, including Type 2 Diabetes Mellitus, hypertension, metabolic syndrome, obesity and cardiovascular disease in Korea. The dataset encompasses epidemiological information from a randomly selected sample of 1,000 Koreans. Each individual underwent testing and surveys on five occasions, each time is three year apart, spanning a period of 15 years, and then the results were saved into five data files corresponding to the order of surveys during that period, for example the first survey and testing results were saved into the first file, then the second test results were saved into the second file... and so on until there are five data files⁶. Nevertheless, these datasets are all raw and not clean with so many missing values, some features cannot be used for forecasting Type 2 Diabetes. Thus, before training and testing by machine learning, they needed to be cleaned and processed according to the purpose of using the dataset in the previous year to predict the diabetes outcome in the future.

In The Korean Health Records dataset, 42 features related to Socio - Demographics, Laboratory results, Medical history, Psychographics...were extracted to be trained by Machine Learning model. (In Table 1).

Diabetes forecasting scenarios

To build the best dataset as a basis for model training. The initial five data files, or five time points, will be divided into four different scenarios to predict diabetes. Specifically as follows:

- The forecasting case with 2 files includes dataset combinations: 1-2, 2-3, 3-4 and 4-5. These four combinations will then be combined into a single data-frame for prediction, meaning using the first file data to predict the second file in each combination.
- The forecasting case with 3 files includes the dataset combinations: 1-2-3, 2-3-4 and 3-4-5. These three combinations will also be combined into a single data-frame, using data from the first two time points to predict the third time point in each combination.
- The forecasting case with 4 files includes the combinations: 1-2-3-4 and 2-3-4-5. Just like the two cases above, a single data frame is combined to use data from the first three time points to predict the fourth time point in each combination.
- The forecasting case with 5 files includes only one combination: 1-2-3-4-5. A single data frame is combined to use data from all four first files to predict the fifth file.

Table 1: List of the features in the dataset⁶ (Source: from authors)

Group	Numbers	Features
Lab test results	9	HBA1C, glucose, creatinine, AST, ALT, total cholesterol, HDL, triglyceride, insulin.
Medical history	12	Hypertension, hyperlipidemia, family history of hypertension and diabetes, menarche start age, pregnancy, childbirth, menopause.
Socio - Demo	4	Gender, age, marital status.
Psychological	10	Drink status, frequency of drink (rice wine, Sake, Soju, beer, spirits), smoke, secondhand smoke, frequency of exercise.
Health & Body Parameters	7	Height, weight, waist, hip, pulse, systolic and diastolic blood pressure.

The purpose of these steps is to create multiple scenarios and use them to evaluate which scenario will produce the best performance.

Machine learning models

Random Forest

Random Forest is a machine learning model widely used in classification and prediction. Instead of using just one decision tree, it uses a “forest” of decision trees. Each tree is trained on a small portion of the data and uses random features. Then, the final prediction is calculated based on the majority from the member trees. This will enhance accuracy and reduce the risk of over-fitting compared to using a single tree⁷.

XGBoost

XGBoost is a famous machine learning algorithm, commonly used for classification and prediction problems. Both XGBoost and Random Forest are ensemble algorithms, but instead of using independent decision trees, XGBoost uses sequential decision trees, meaning each tree is built based on errors of the previous tree. This model is generally known for its high performance and ability to handle large datasets, and also supports multiple configurations to optimize model accuracy⁸.

Data oversampling

There is a huge imbalance between normal (majority) and diabetic (minority) classes in the labeled feature, it can be seen clearly in Table II. Such an imbalance will cause the model to be biased and will mainly make predictions biased towards the majority class, whereas the context of this study is to optimize the prediction and classification of patients with diabetes.

Table 2: Number of observations in each case (Source: from authors)

Case	Number of observations		
	Normal (0)	Diabetes (1)	Total
2 files	3143	229	3372
3 files	2165	160	2325
4 files	1269	103	1372
5 files	497	38	535

Thus, the study used Oversampling technique, Borderline SMOTE, to address this phenomenon.

Borderline SMOTE, a variation of SMOTE, addresses a key weakness of its predecessor, particularly regarding outlier minority samples. Unlike traditional SMOTE, Borderline SMOTE focuses solely on generating synthetic samples for the minority class that reside near the boundary separating the minority and majority classes. It identifies “borderline” instances within the minority class, which are those close to the boundary and susceptible to misclassification. Through oversampling these borderline instances, Borderline SMOTE aims to rebalance class distribution and enhance the effectiveness of machine learning models⁹.

Evaluation method

Evaluation forecasting metrics

Afterwards, the result of the testing process is unfolded in a classification metrics table with noteworthy parameters helping to assess the performance of the model. Specifically:

- Accuracy: The proportion of correctly predicted observations (both positive and negative) to the total ob-

servations¹⁰.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (01)$$

$$When \begin{cases} TP : \text{True Positive} \\ TN : \text{True Negative} \\ FP : \text{False Positive} \\ FN : \text{False Negative} \end{cases}$$

- Precision: The proportion of correctly predicted positive observations to the total predicted positive observation¹⁰.

$$Precision = \frac{TP}{(TP + FP)} \quad (02)$$

- Recall: This parameter is also known as sensitivity or true positive rate, is the proportion of correctly predicted positive observations to the all observations in actual class – YES¹⁰.

$$Recall = \frac{TP}{TP + FN} \quad (03)$$

- F1-score: The harmonic mean of Precision and Recall, providing a balance between Precision and Recall¹¹.

$$F1 \text{ score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (04)$$

- Weighted F1-score: This metric calculates the F1-score for each class independently and then takes the average across all classes, but weighted by the number of true instances for each class. It gives more weight to classes with more instances. Thus, this metric gives insights into the overall performance of the model on all classes, accounting for any class imbalances¹¹.
- AUC-PR: Another valuable tool for evaluating binary classification models, particularly in highly imbalanced datasets where positive instances are scarce compared to negative samples, is the Area Under the Precision-Recall Curve (AUC-PR). The Precision-Recall curve depicts the relationship between Precision and True Positive Rate (TPR) across various decision thresholds. AUC-PR quantifies the area under this curve, ranging from (0, 0) to (1, 1), offering a comprehensive assessment of the model's classification performance across all decision thresholds. It provides insight into the model's effectiveness in accurately detecting positive cases¹².

Feature analysis

The SHAP (SHapley Additive exPlanations) method, pioneered by Lundberg and Lee in 2017, is a robust technique for model interpretability. Drawing inspiration from Lloyd Shapley's cooperative game theory, it offers a fair approach to distributing rewards among

players. SHAP values are determined by assessing the average impact of each feature across all possible permutations. This involves evaluating every conceivable arrangement of features and their influence on the model's prediction. By averaging these contributions across permutations, SHAP provides a balanced and easily understandable evaluation of each feature's significance in predicting the model's outcome¹³. This article will focus on analyzing the features of laboratory test results, here is SHAP value formula¹³:

$$\phi_i = \sum_{S \subseteq \{1, \dots, p\}, i \in S} \frac{|S|! (p - |S| - 1)!}{p!} [val(S \cup \{i\}) - val(S)] \quad (05)$$

Where:

$p!$ is number of ways to form a coalition of p players.

$|S|$ is number of players in coalition S .

$|S|!$ is number of ways coalition S can form.

$(p - |S| - 1)!$ is number of ways players can join after player i joins.

$\frac{|S|! (p - |S| - 1)!}{p!}$ is weight.

RESULTS

After conducting testing and evaluation, the performance across all four scenarios demonstrated generally good and acceptable outcomes, as indicated in Table 3 and Table 4, despite the significant imbalance between the Normal and Diabetes classes within the dataset. Specifically, the Random Forest model, when applied the scenario with 5 files, exhibited the highest performance metrics, achieving a weighted F1 score of 97%, an AUC-PR of 94%. Conversely, the lowest performance was observed with the Random Forest classifier in the case with 4 files, which achieved respectable metrics of 91% weighted F1 score, but AUC-PR is only 40%. These results suggest that the approach holds promise as a solution for future diabetes prediction. However, further enhancements are necessary to bolster model performance.

Upon specific comparison of the scenarios, each demonstrates excellent forecasting performance. The study will prioritize three metrics: Precision and Recall, particularly within the diabetes class, and Weighted F1 scores. Across two tables, it is evident that the forecasting case with five time points demonstrates superior performance compared to other scenarios. Table I highlights the highest Precision, Recall, and weighted F1 score in the diabetes class, which are 0.86, 0.75, and 0.97 respectively.

The Random Forest model significantly outshines the XGBoost model, especially concerning Precision, Recall, Weighted F1 score, and AUC-PR metrics. Specifically, in the case with 5 files, both Random Forest

Table 3: The classification performance of Random Forest in 4 cases (Source: from authors)

Random Forest				
	Precision	Recall	Weighted F1	AUC-PR
2 files	Normal: 0.97	Normal: 0.97	0.94	0.63
	Diabetes: 0.57	Diabetes: 0.61		
3 files	Normal: 0.97	Normal: 0.96	0.94	0.55
	Diabetes: 0.53	Diabetes: 0.56		
4 files	Normal: 0.95	Normal: 0.96	0.91	0.40
	Diabetes: 0.44	Diabetes: 0.38		
5 files	Normal: 0.98	Normal: 0.99	0.97	0.94
	Diabetes: 0.86	Diabetes: 0.75		

Table 4: The classification performance of XGBoost in 4 cases (Source: from authors)

XGBoost				
	Precision	Recall	Weighted F1	AUC-PR
2 files	Normal: 0.97	Normal: 0.97	0.94	0.51
	Diabetes: 0.60	Diabetes: 0.52		
3 files	Normal: 0.96	Normal: 0.96	0.93	0.49
	Diabetes: 0.50	Diabetes: 0.50		
4 files	Normal: 0.96	Normal: 0.94	0.92	0.41
	Diabetes: 0.44	Diabetes: 0.52		
5 files	Normal: 0.97	Normal: 0.99	0.94	0.61
	Diabetes: 0.75	Diabetes: 0.38		

achieve very impressive AUC-PR and Weighted F1-score parameters, with 0.94 and 0.97 compared to only 0.61 and 0.94 in the XGBoost model.

In terms of analyzing the impact of test indicators on model's prediction. Figure 1 will show the features that have significant importance, mean high SHAP values, including HBA1C, Glucose (denoted as GLU0). T01, T02, T03 and T04 are the codes that indicate when these values are examined. For example, T01_HBA1C means this HBA1C value was sampled in the first survey, T02_GLU0 means this Glucose value was sampled in the second survey, etc.

And in Figure 2, it will illustrate the way how these features influence the model's output. Blue dots represent low SHAP values, while red dots represent high. And based on the figure, it can be seen that on the left side of the plot, there are a lot of blue dots, meaning that these values will make the model's prediction biased towards negative results (i.e. no disease). Meanwhile, the plot on the right has a dense concentration of red dots, and will make the prediction fall towards a positive result (i.e. diabetes). In short, when the HBA1C and Glucose features have higher SHAP values, the risk of diabetes also increases.

DISCUSSION

In summary, the current EHR is considered an extremely rich and useful source of data for exploitation and application in optimizing management in the health care sector in general, and also for prediction, such as preventing dangerous diseases in particular. The EHR itself is a collection of a large amount of medical information about patients, including information about medical history, symptoms, test results, information about used medications, information about medications, and other medical information treatments, and more. Machine learning algorithms and data analytics can be applied to the EHR to create predictive models. Specifically, algorithms can be trained on data from the EHR to predict the likelihood of disease, assess disease risk, predict treatment outcomes, or monitor the progression of a condition. In terms of model performance results, the case with 5 time points gives the most optimal results, because of the metrics it provides. Specifically, in Random Forest, the AUC-PR scores in each are very high, reaching the threshold of 0.94. Weighted F1 score is also very positive, reaching 0.97. Precision and Recall are quite impressive in both classes, especially in diabetes.

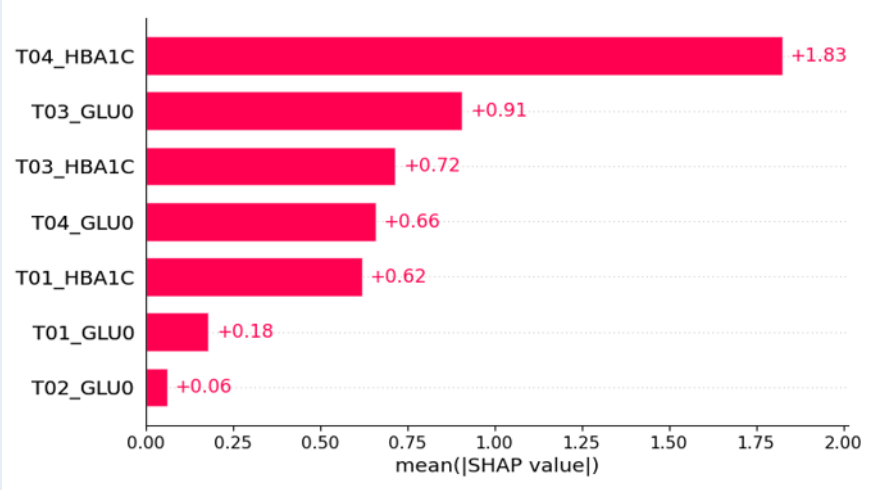


Figure 1: The mean SHAP values of the test indicator features have significant impacts on model’s prediction [Source: from authors]

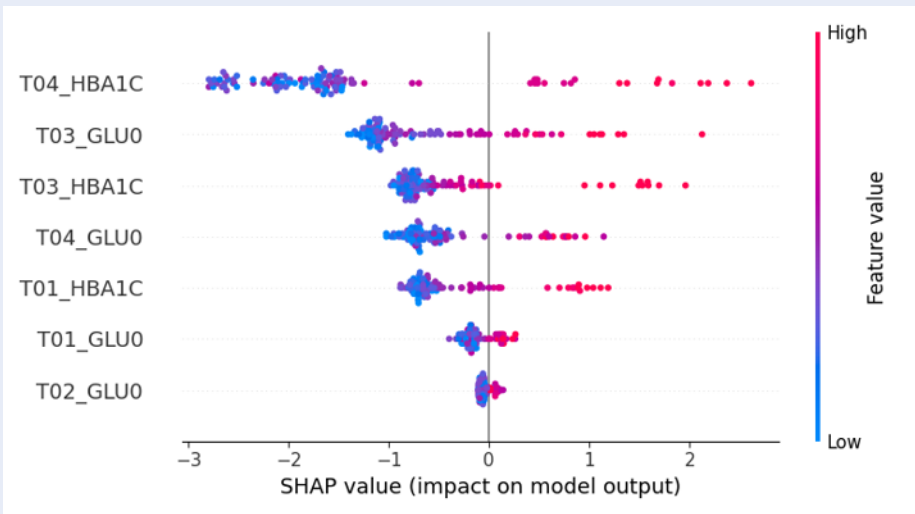


Figure 2: The plot shows how the features impact on prediction results of model [Source: from authors]

This shows that despite the imbalance, the model still can precisely predict a significant number of diabetes cases. The reason for this high performance partly lies in the number of features that ensure two factors: completeness and quality in this case. Not that more features will have better performance, but the quantity must be complete and of good quality so that the model can learn and find relationships between features effectively.

Regarding the feature analysis, Glucose and HBA1C are also the two popular methods to diagnose diabetes. This partly explains why SHAP is able to interpret very well the impact of these features on the

model’s prediction results. However, it is still necessary to further improve data processing, such as thoroughly removing outlier values to be able to interpret test indicator features with low SHAP values more clearly. Instead of focusing only on the features that have a big impact, that might ignore useful information from other features.

After analyzing all the strengths and weaknesses of the model, it is extremely necessary to propose directions for future improvement. As mentioned above, because the data processing is not optimized, there are still too many outliers that can cause noise and reduce the interpretability of SHAP method for features

with medium and low impact on the model's prediction. Apart from solving immediate problems, it is necessary to apply more advanced methods, like deep learning models such as: ANN, MLP... to predict diabetes.

CONCLUSION

In conclusion, all goals have been accomplished so far. The dataset was built and it is possible to find out which cases will give the best performance. Besides, successfully applying 2 Machine Learning models: Random Forest, XGBoost in predicting future diabetes, as well as being able to evaluate which model brings the most optimal results for the dataset. Through the results and images obtained, some insights on the data are also given. However, there are still some limitations that need to be improved. In general, the model can perform well under conditions of large imbalance in the dataset.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHOR CONTRIBUTION

Vo Nhat Tan Huynh: Methodology execution, data extraction and processing, machine learning model development, and writing - original draft preparation.
Nhat Tan Le: Supervision, overall research guidance, validation of the analytical results, and reviewing and editing of the final manuscript.

LIST OF ABBREVIATIONS

Abbreviation: Full Form
AUC-PR: Area Under the Precision-Recall Curve
EHR: Electronic Health Record
KoGES: Korean Genome and Epidemiology
ML: Machine Learning
SHAP: SHapley Additive exPlanations
SMOTE: Synthetic Minority Over-sampling Technique
TPR: True Positive Rate

ACKNOWLEDGEMENT

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

REFERENCES

1. Latha NA, Murthy BR, Sunitha U. Electronic health record. International Journal of Engineering. 2012;1(10):25–7.
2. Nguyen BP, Pham HN, Tran H, Nghiem N, Nguyen QH, Do TT, et al. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Computer Methods and Programs in Biomedicine. 2019;182. Available from: <https://doi.org/10.1016/j.cmpb.2019.105055>.
3. Lee C, Jo B, Woo H, Im Y, Park RW, Park C. Chronic Disease Prediction Using the Common Data Model: development Study. JMIR AI. 2022;1(1). Available from: <https://doi.org/10.2196/41030>.
4. Choi BG, Rha SW, Kim SW, Kang JH, Park JY, Noh YK. Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks. Yonsei Medical Journal. 2019;60(2):191–9. Available from: <https://doi.org/10.3349/ymj.2019.60.2.191>.
5. Jeong YW, Jung Y, Jeong H, Huh JH, Sung KC, Shin JH, et al. Prediction model for hypertension and diabetes mellitus using Korean public health examination data (2002–2017). Diagnostics (Basel). 2022;12(8):1967. Available from: <https://doi.org/10.3390/diagnostics12081967>.
6. Korea Disease Control and Prevention Agency, "2020," 30 August 2023. Available from: <https://www.kdca.go.kr/>.
7. Posonia AM, Vigneshwari S, Rani DJ. Machine learning based diabetes prediction using decision tree J48; 2020. Available from: <https://doi.org/10.1109/ICISS49785.2020.9316001>.
8. Wardani KD, Akbar M. Diabetes Risk Prediction using Feature Importance Extreme Gradient Boosting (XGBoost). Jurnal RESTI. 2023;7(4):824–31.
9. Xiaolong XX, Wen CHEN, Yanfei SUN. Over-sampling algorithm for imbalanced data classification. Journal of Systems Engineering and Electronics. 2019;30(6):1182–91. Available from: <https://doi.org/10.21629/JSEE.2019.06.12>.
10. Ruuska S, Hämäläinen W, Kajava S, Mughal M, Matilainen P, Mononen J. Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. Behavioural Processes. 2018;148:56–62. Available from: <https://doi.org/10.1016/j.beproc.2018.01.004>.
11. Wang R, Li J. Bayes test of precision, recall, and F1 measure for comparison of two natural language processing models. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. Available from: <https://doi.org/10.18653/v1/P19-1405>.
12. Khan SA, Rana ZA. Evaluating performance of software defect prediction models using area under precision-Recall curve (AUC-PR); 2019. Available from: <https://doi.org/10.23919/ICACS.2019.8689135>.
13. Aboze BJ. A Comprehensive Guide into SHAP (SHapley Additive exPlanations) Values; 2023. [Accessed: May 7, 2024]. Available from: <https://deepchecks.com/a-comprehensive-guide-into-shap-shapley-additive-explanations-values/>.

Dự báo bệnh đái tháo đường dựa trên phân tích dữ liệu hồ sơ sức khỏe điện tử

Võ Nhật Tân Huỳnh^{1,2}, Nhật Tân Lê^{1,2,*}



Use your smartphone to scan this QR code and download this article

TÓM TẮT

Trong những năm gần đây, tỷ lệ mắc bệnh tiểu đường gia tăng mạnh mẽ, đặt ra một thách thức đáng kể đối với các hệ thống chăm sóc sức khỏe trên toàn thế giới. Để giải quyết vấn đề này, các phương pháp tiếp cận sáng tạo là yếu tố bắt buộc nhằm nâng cao khả năng dự đoán và quản lý bệnh. Theo đó, nghiên cứu này nhấn mạnh tiềm năng của việc sử dụng Hồ sơ Sức khỏe Điện tử (EHR) trong việc dự đoán bệnh tiểu đường thông qua các phương pháp khai phá dữ liệu. Mục tiêu chính của nghiên cứu là dự báo bệnh tiểu đường dựa trên dữ liệu EHR bằng các kỹ thuật Khai phá Dữ liệu và Học máy, theo sau là việc phân tích các đặc trưng quan trọng liên quan đến rủi ro mắc bệnh. Để đạt được mục tiêu này, nghiên cứu đã sử dụng dữ liệu EHR từ Hồ sơ Sức khỏe Hàn Quốc thuộc dự án Hệ gen và Dịch tễ học Hàn Quốc (KoGES), bao gồm năm tệp dữ liệu trải dài qua năm thời điểm khác nhau. Các mô hình Học máy sau đó được sử dụng để dự báo. Dữ liệu được chia thành bốn kịch bản, mỗi kịch bản đại diện cho việc dự báo với số lượng thời điểm khác nhau (2, 3, 4 và 5), nhằm đánh giá khả năng dự báo qua nhiều tình huống và so sánh hiệu suất giữa các mô hình. Giai đoạn cuối cùng bao gồm việc phân tích các đặc trưng thông qua phương pháp SHAP. Kết quả chỉ ra rằng các mô hình đạt hiệu suất tối ưu nhất ở kịch bản năm thời điểm, trong đó mô hình Random Forest mang lại chỉ số AUC-PR và Weighted F1 lần lượt là 0.94 và 0.97. Thêm vào đó, mô hình cũng thể hiện các chỉ số Precision và Recall rất đáng khen ngợi ở từng lớp, đặc biệt là ở lớp tiểu đường, với kết quả lần lượt là 0.86 và 0.75. Việc phân tích các đặc trưng quan trọng đã khám phá ra các mối liên hệ với nguy cơ mắc bệnh tiểu đường trong tương lai. Những kết quả này cho thấy các mô hình đã được xây dựng và huấn luyện hiệu quả, thể hiện độ chính xác dự báo tương đối cao. Tuy nhiên, vẫn còn tồn tại một số thách thức, bao gồm hiệu suất mô hình ở một số điểm chưa tối ưu và sự cần thiết phải cải thiện khả năng diễn giải của SHAP đối với các đặc trưng có mức độ quan trọng trung bình và nhỏ.

Từ khóa: Dự đoán bệnh tiểu đường, hồ sơ sức khỏe điện tử, học máy

¹Khoa Khoa học Ứng dụng, Trường Đại học Bách khoa, Đại học Quốc gia Thành phố Hồ Chí Minh (HCMUT), 268 đường Lý Thường Kiệt, phường Diên Hồng, Thành phố Hồ Chí Minh, Việt Nam

²Đại học Quốc gia Thành phố Hồ Chí Minh, phường Linh Trung, Thành phố Hồ Chí Minh, Việt Nam

Liên hệ

Ngày Tân Lê, Khoa Khoa học Ứng dụng, Trường Đại học Bách khoa, Đại học Quốc gia Thành phố Hồ Chí Minh (HCMUT), 268 đường Lý Thường Kiệt, phường Diên Hồng, Thành phố Hồ Chí Minh, Việt Nam

Đại học Quốc gia Thành phố Hồ Chí Minh, phường Linh Trung, Thành phố Hồ Chí Minh, Việt Nam

Email: lenhattan@hcmut.edu.vn

Lịch sử

- Ngày nhận: 19-08-2024
- Ngày sửa đổi: 26-09-2024
- Ngày chấp nhận: 16-09-2026
- Ngày đăng: 25-09-2026

DOI : <https://doi.org/10.32508/vnuhcmj-et.v9i3.1430>



Check for updates

Bản quyền

© Tạp chí ĐHQG Tp.HCM. Đây là bài báo công bố mở được phát hành theo các điều khoản của the Creative Commons Attribution 4.0 International license.

Trích dẫn bài báo này: Huỳnh V N T, Lê N T. Dự báo bệnh đái tháo đường dựa trên phân tích dữ liệu hồ sơ sức khỏe điện tử. *VNUHCM J. Eng. Technol.* 2026; 9(3):3380-3388.