Analysis of Feature Extraction Algorithms for Artificial Intelligence Based Diabetes Detection

G R Ashisha[1,2], Sai Kiran Oruganti[3]

[1] Postdoctoral Researcher, Lincoln University College, Malaysia; [2] Assistant Professor, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India; [3] Associate Professor, Lincoln University College, Malaysia

Email ID: pdf.ashisha@lincoln.edu.my grashisha27@gmail.com

Abstract: Diabetes is among the fastest-growing health issues worldwide, hence necessitating proper predictive models. This research focuses on the development of an AI-based predictive model for diagnosing diabetes based on the Diabetes 130-US hospitals dataset. Various preprocessing techniques are conducted, including data cleaning, encoding, and imputation of missing values. To balance the classes in the data set, the Synthetic Minority Over-sampling Technique (SMOTE) is employed. The feature extraction process is done by applying the Random Forest technique, while the selected features are classified through the XGBoost classifier. Model performance is analyzed with the help of accuracy, precision, recall, and F1 score metrics. As per experiment findings, performance of the model gets improved due to feature extraction. The value of accuracy is raised from 78% to 79% due to feature extraction. Likewise, there is improvement in the case of precision, recall, and F1 score as well. Proposed methodology makes prediction more reliable with less computational cost. It will be helpful for healthcare professionals in diabetes detection at an early stage.

Keywords: Feature Extraction; Diabetes Prediction; Machine Learning; SMOTE; Random Forest.

Introduction

The problem of diabetes has taken the status of a significant health issue at the global level and has been affecting millions of people worldwide by leading to many issues including heart disease, neuropathy, and renal failure. As per latest surveys [1], over 500 million adults worldwide have been reported to be suffering from diabetes, and the chances of growth in numbers exponentially are very high in the upcoming years. This highlights the significance of timely detection and treatment of diabetes. Though traditional methods of detection have been clinically accurate, these are invasive and time consuming. In relation to the development of Machine Learning (ML) and Artificial Intelligence (AI), many researches [2] have applied data-driven techniques to predict the early stage of diabetes. ML algorithms can identify valuable patterns from large amounts of data in clinical domains and make decisions accordingly. Yet, current ML models encounter multiple problems such as high dimensionality, class imbalance, missing data, and inclusion of irrelevant/redundant features. Specifically, class imbalance is another issue in the dataset for predicting disease, as the classifiers usually favor majority classes, which leads to the failure to predict the presence of diabetes in patients. Moreover, ineffective feature extraction techniques do not enable the selection of the most useful features, which results in

the poor performance of classifiers. It was demonstrated that the use of effective feature selection and extraction techniques is necessary for predicting diabetes through identifying the risk factors [3]. In addition, feature engineering techniques assist in representing raw data in an appropriate way.
In order to tackle these problems, this study presents a framework based on machine learning approach which incorporates data preprocessing, balance of classes through SMOTE, and feature extraction via Random Forest method. The features extracted from the data set are used for training an XGBoost classifier for precise prediction of diabetes.

Related work

Machine Learning in Diabetic Prediction in Healthcare is important to mention that machine learning in predicting the occurrence of diabetes among patients has been a subject of considerable interest during the last decade. Earlier works were mainly concerned with statistical models including logistic regression and decision trees as means of defining factors leading to diabetes. With an increase in the number of data in healthcare, the tendency changed towards machine learning
Many studies have targeted on utilizing different ML algorithms like SVM, ANN, and ensembles to enhance prediction accuracy. In [3], researchers developed a machine learning algorithm for predicting diabetes through multiple classifiers, obtaining fair accuracy but without any efficient feature selection methods. In [2], an AI approach based on deep learning models was utilized, resulting in better accuracy but at the cost of extensive computing power. In [4], a novel method was developed by combining feature selection with machine learning algorithms.

Key Contribution

In this paper, a new approach is introduced for diabetes prediction using ML algorithms with special attention to feature extraction and efficiency of models. The main contributions made in this study are outlined below

- A framework for early diagnosis of diabetes using clinical and health-related databases is built by means of ML algorithms.
- Feature selection for prediction is performed using the algorithm based on Random Forests, which makes it possible to perform reduction of data dimensionality.
- The proposed method allows one to effectively decrease the number of input features while preserving high accuracy.

Method, Experiments and Results

The system is developed based on a well-defined framework for machine learning algorithm of diabetes prediction. First, the diabetes data set is preprocessed by dealing with the issue of missing values and normalizing the input variables. Further, Random Forest-based feature extraction is carried out in order to extract those features that are more relevant. It is done to reduce redundancy, eliminate noise, and reveal hidden structures present in the dataset. The consequence of the feature extraction process is the dimensionality reduction of data and hence its interpretability as well as computational efficiency.

Following feature extraction process, the new feature set is used for training the ML model. In the current research, the choice of XGBoost classifier for prediction is justified by its computational efficiency and accuracy. The sequence of actions included in the workflow of the proposed model is shown in Fig. 1.

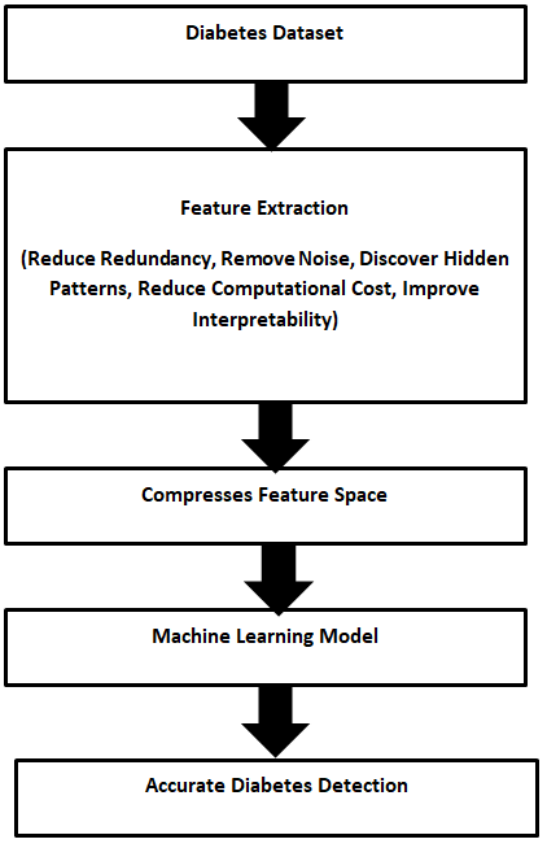

Figure 1. Workflow of developed method.

The experiment is performed on the Diabetes 130 US Hospitals data set that comprises 101,766 samples gathered from 130 clinical centers [5]. The data set contains roughly 50 attributes like laboratory parameters, treatment, and demographic information. The data set is separated into train and test sets in a standard way for the model performance assessment. The experiment is implemented using machine learning libraries, and the model performance is evaluated by means of standard metrics like accuracy, F1-score, precision and recall. The performance of the technique is analyzed prior to and after the process of feature extraction. The findings show that using feature extraction based on Random Forest leads to improved performance of the model. Firstly, accuracy becomes equal to 79.00% after feature extraction, while it was only 78.00% before. Secondly, the same can be said about precision, which reaches 77.56% after the process of feature extraction, whereas previously it was 77.00%. Also, recall improves from 75.00% to 75.23% (Table 1).

Table 1. Result of the proposed model

ML Models	Performance metrics							
	Accuracy (%) Before Feature Extraction	Accuracy (%) after Feature Extraction	Precision (%) Before Feature Extraction	Precision (%) after Feature Extraction	Recall (%) Before Feature Extraction	Recall (%) after Feature Extraction	F1-Score (%) before Feature Extraction	F1-Score (%) after Feature Extraction
XGBoost	78	79	77	77.56	75	75.23	74	75.10

Finally, the same applies to the F1-score, which equals 75.10% after feature extraction compared to 74.00% prior to it. This means that feature extraction is helpful in selection of the most informative features and thus helps to improve prediction performance. Moreover, reduced number of features leads to improved performance of calculations.

Discussions

The obtained experimental results clearly indicate the positive influence of Random Forest-based feature extraction on the working of the model predicting diabetes. Improvement of the model's performance using all of the evaluation metrics proves the positive role of the elimination of irrelevant and redundant features in enhancing the performance of the model. While the change of accuracy is quite small, the consistency of improvement of other parameters proves the stability of the suggested method. Such improvement of recall is especially important in healthcare application since it allows minimizing the risk of the missing positive cases.

Another notable point is the decrease in the number of features leading to higher computational efficiency. Reduction of the dimensionality of the feature space leads to lower complexity of the model. Additionally, the new algorithm gives the best compromise between the efficiency and the accuracy compared to other algorithms without the process of feature extraction. Even though there are some techniques that give more accurate results, they need a larger number of features and computational complexity. Thus, the results show that the feature extraction is an important component in improving the effectiveness of machine learning algorithms in medical problems, and the new technique can be implemented in clinical decision support systems.

Conclusions

The current paper considers the problem of early diabetes prediction through the application of machine learning. The problem is relevant due to the increasing spread of the disease and necessity of early diagnosing. The proposed approach includes the application of Random Forest for feature extraction and classification. According to the results of the experiments conducted on the chosen dataset, the approach provides appropriate accuracy and efficiency in features selection as compared to the basic models. At the same time, the methodology allows reducing the number of dimensions while

preserving the most important features. Thus, the paper proves that the machine learning helps to detect patients who are prone to developing the disease. Nevertheless, it has several limitations such as the use of relatively small dataset and dependence on some parameters in the model. Therefore, future research can be focused on the employment of the bigger dataset and hybrid models as well as deep learning algorithms for increasing the accuracy rate.

References

1. H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, K. Ogurtsova, B. B. Duncan et al., "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045", *Diabetes Research and Clinical Practice*, vol. 183, pp. 109119, 2022. https://doi.org/10.1016/j.diabres.2021.109119
2. P. B. Khokhar, C. Gravino and F. Palomba, "Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review", *Artificial Intelligence in Medicine,* vol. 164, pp. 103132, 2025. https://doi.org/10.1016/j.artmed.2025.103132
3. I. J. Kakoly, M. R. Hoque and N. Hasan, "Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms with Feature Selection Technique," Sustainability, vol. 15, no. 6, pp. 4930, 2023. https://doi.org/10.3390/su15064930
4. S. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104-116, 2017. https://doi.org/10.1016/j.csbj.2016.12.005
5. B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios and J. N. Clore, "Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records," BioMed Research International, vol. 2014, pp. 781670, 2014. https://doi.org/10.1155/2014/781670