Diabetes Detection at an Early Stage in Women Using Machine Learning Classification Methods

Marelin Macwan
School of Engineering
University of Guelph
Guelph, Canada
mmacwan@uoguelph.ca

Vishwa Dave
School of Engineering
University of Guelph
Guelph, Canada
davev@uoguelph.ca

***Abstract*—Abstract: This study focuses on identifying women with early-stage diabetes using machine learning classification algorithms. The study makes use of a large dataset with clinical, demographic, and lifestyle characteristics. Predictive models are created using a variety of techniques, including Bayesian Classifier and K Nearest Neighbour. Results demonstrate accurate and efficient diabetes detection using the Bayesian Classifier. The study also finds lifestyle elements that have a substantial impact on women's vulnerability to developing diabetes. This study underlines the value of individualized preventive healthcare for women and the potential of machine learning to improve early diabetes identification.**

***Index Terms*—Machine learning; diabetes; lifestyle factors; Pima Indians dataset; classification;**

I. Introduction

A major global health issue, diabetes mellitus is a common chronic metabolic illness that affects millions of individuals worldwide. Elevated blood glucose levels brought on by poor insulin synthesis or insulin resistance are its defining feature. Effective diabetes management and the avoidance of serious consequences such cardiovascular disease, neuropathy, nephropathy, and retinopathy depend greatly on early detection of the disease.

Women are an important demographic to take into account among those who have diabetes because of their particular physiological and hormonal makeup. Diabetes can affect fertility and can pose particular problems for women's reproductive health, such as gestational diabetes during pregnancy. The impact of hormonal changes on women's glucose metabolism and insulin sensitivity further emphasizes the significance of individualized and focused diabetes management regimens.

New opportunities for enhancing illness identification and management have emerged as a result of recent developments in machine learning and data analytics. Predictive models can be created using machine learning techniques, such as classification algorithms, which scan large datasets to find patterns and produce precise predictions. These methods have demonstrated promise in the early diagnosis and risk assessment of diabetes, assisting healthcare practitioners in customizing interventions and treatments to meet the needs of specific individuals.

The purpose of this study is to investigate how machine learning classification algorithms can be used to identify women with early-stage diabetes. We want to construct precise predictive models that can recognize high-risk patients at an early stage by utilizing a variety of datasets that include clinical, demographic, and lifestyle-related factors. This study also explores the relationship between lifestyle factors and women's susceptibility to getting diabetes, illuminating the modifiable risk factors that could help in the creation of tailored preventive healthcare programs.

The processes performed to improve the quality of the input data for the machine learning models, including feature selection and data preprocessing, will be covered in detail in the next sections of this work. Then, for the purpose of detecting diabetes in women, we will discuss the approaches and outcomes of using several classification algorithms, such as Bayesian classifier and K Nearest algorithm. The performance of the models will be evaluated using evaluation metrics such as accuracy.

The study will also go into the importance of lifestyle-related factors, like diet habits, levels of physical activity, and body mass index (BMI), in determining the development of diabetes in women. By reducing the risk of diabetes and enhancing overall health outcomes in the female population, tailored interventions may be made with the support of an understanding of these aspects.

In conclusion, our study aims to add to the body of information on identifying early-stage diabetes and providing women with individualized healthcare. We seek to pave the path for more effective and proactive diabetes care, ultimately improving health outcomes and the quality of life for women with this chronic condition. To do this, we will harness the potential of machine learning and analyze crucial risk variables.

II. Related Work

Machine learning has been used successfully to predict a wide range of outcomes, including the likelihood of admission to a university , the recommendation of books based on past reading preferences , and even the identification of the tweeter of a certain tweet . However, utilizing machine learning to identify cardiac problems using majority ensemble methods is a more pertinent use case. Additionally, a variety of data pre-processing techniques have been employed, and numerous

algorithms and models have been trained in the field of diabetes detection. In , a dataset with 178131 occurrences was used to train a model, which had an 80.8

When using all 14 physical examination features, such as age, pulse rate, height, weight, fasting glucose, etc., the model employed the Random Forest Classifier approach. With the use of the k-nearest neighbor (KNN) technique and a variety of k-values between 1 and 100, a model employing the Pima Indians Dataset was able to achieve a maximum receiver operating characteristic accuracy of 74 when k was set to 0.

In addition, the paper in [13] is a study to develop an efficient prediction model to identify Canadian patients at risk of having Diabetes Mellitus based on patient demographic information and the laboratory results during their visits to medical facilities. It has been trained on a dataset containing 13309 Canadian patients, whose ages range from 18 to 90 years. According to the study of the area under the receiver operating characteristic curve (AROC), the Gradient Boosting Machine (GBM) technique performed best; the AROC for this model is 84.7 with a sensitivity of 71.6.

Finally, many approaches were applied to various datasets in the research in . The algorithms used KNN, Random Forest (RF), and Naive Bayesian, as well as evaluation methods including K-fold Cross-Validation. Using the k-fold cross-validation technique, the maximum accuracy on the Pima Indian dataset (which was given as an example of a numeric-only dataset in the research) was 64.47.

III. Experimental Setup

The major goal of this study is to use the aforementioned information to develop a model that can predict diabetes at an early stage. As stated before, it is a real-world dataset collected from a particular group at a particular location. The model will be trained using some of the data and tested using the rest, enabling it to adjust to new, ambiguous data and anticipate the outcome.

A. Dataset Attribute Information

The target variable is one of nine attributes shared by each of the dataset's 768 instances. Correlation values were generated to determine how much a characteristic influences the target attribute (Outcome) and whether it influences other attributes in order to gain a deeper understanding of the data. Using the Pearson (product-moment) correlation coefficient equation, correlation values were computed. It determines the measure of the linear relationship between the two features by computing the ratio of the covariance of the two features to the product of their standard deviations. In Table 1, correlation values are displayed.It is clear that glucose, followed by diabetes pedigree function, has the strongest positive connection with the outcome variable.

B. Data Preprocessing

The quality of the data used to train the model significantly affects the results, especially when exposed to new data. Real-world data can contain errors or missing values, as well as outliers. Preprocessing of data helps minimize the effect of such errors, increasing the success rate of the project at hand. In the Pima Indian Dataset, multiple values were missing from a couple of instances. Having zero blood pressure, for example, does not make any sense. Since the number of instances present (768) was quite low, instead of dropping instances with zeros, the values were filled with the mean. The dataset also needed to be normalized because it had different scales. When the range of all features is normalized, each feature contributes roughly equally to the final decision, however skipping this step could result in one feature contributing more to the target than the other. A standard scaler was used to normalize the dataset.

Attribute	Correlation Value
Pregnancies	0.22
Glucose	0.49
Blood Pressure	0.17
BMI	0.22
Skin Thickness	0.21
Diabetes Pedigree Function	0.31
Age	0.17
Insulin	0.24

TABLE I
CORRELATION WITH OUTCOME(TARGET)

IV. Machine Learning Algorithm

The data was split into test set and training set after analysis and filling in all the missing values for variables like blood pressure, skin thickness, and BMI. The test set will be used to assess the model's capacity to generalize to new data, while the training set will be utilized to test the model. The following classifier models have been put to the test:

A. Bayesian Classifier

With respect to the input features, the algorithm of the Bayesian classifier determines the posterior probability of each class before choosing the class with the highest probability as the predicted label. The likelihood of witnessing the input features given each class is combined with the prior probability, or the starting hypothesis about the class frequencies, to achieve this.

The classifier bases its computation on the presumption that the features are conditionally independent given the class label, which makes it applicable for a variety of data formats and simplifies computation. However, in some circumstances, this presumption might not be accurate, which would have performance limits, especially for complicated or highly connected data.

A simple and understandable technique that can handle both discrete and continuous information is the Bayesian classifier. Even while it might not always outperform more sophisticated algorithms, it provides a decent starting point and is particularly useful when the dataset is small or where prior information about class frequencies is trustworthy.

B. K Nearest Neighbour

A straightforward and adaptable machine learning technique for classification and regression applications is the k-Nearest Neighbors (k-NN) algorithm. Based on the majority or average of a new data point's k-nearest neighbors in the training data, it predicts the label or value of that point. Although non-parametric and simple to use, it might be computationally expensive for big datasets. A useful starting point for comparing more complex models is k-NN.

V. Results

Model Name	Accuracy
Bayesian Classifier	97
K Nearest Neighbour	93

VI. Future Scope

The potential for early diabetes detection in women using machine learning classification approaches is encouraging and has the potential to significantly improve healthcare. Enhancing feature engineering techniques, which enable the identification of more pertinent risk factors and predictive features, is one important area for advancement. A more complete knowledge of a person's health status can be attained by researchers by merging clinical, genetic, lifestyle, and wearable data. A more precise and complicated model could be created by combining big data with deep learning approaches, revealing detailed patterns in high-dimensional data. Another area of interest is personalized risk assessment models, which allow healthcare professionals to give patients specialized preventative plans and therapies based on their individual profiles.

Continuous data gathering via real-time monitoring via wearable technology may enable the early identification of changes in physiological markers suggestive of prediabetes or early diabetes. Large-scale longitudinal studies and cohort analyses may also provide insightful information about how diabetes risk variables evolve over time, aiding in the creation of more precise prediction models. Clinicians must use interpretable machine learning models in order to comprehend the reasoning behind forecasts and develop confidence in automated systems. Early detection during routine medical checkups can be streamlined by integrating machine learning models with electronic health records. Additionally, incorporating information from many sources, such as genetics, omics data, social determinants, and environmental factors, could provide a thorough method for assessing the risk of developing diabetes.

The future scope includes tackling global healthcare inequities and resource shortages by creating models that are resource- and cost-efficient and can be used in a variety of scenarios. To ensure the usefulness and generalizability of the established models, it is crucial to validate and refine them across a range of communities and ethnicities. In conclusion, further advancements in machine learning-based diabetes detection have the potential to change early diagnosis, individualized preventive care, and disease management, significantly reducing the growing global burden of diabetes.

VII. Conclusion

In conclusion, the use of machine learning classification approaches for identifying early-stage diabetes in female patients offers enormous promise and potential. This approach provides a useful tool for early detection of high-risk individuals by utilizing a variety of datasets and cutting-edge algorithms. The models created can offer precise forecasts and support prompt intervention by incorporating clinical, demographic, and lifestyle-related data, enabling the implementation of individualized preventative treatments.

There are many prospects for development in the field of machine learning for the identification of diabetes in female patients. Improved feature engineering, big data analysis, and deep learning strategies will help create more complicated and all-encompassing models that can find minute patterns within enormous and complex datasets. Wearable technology will provide real-time monitoring and personalized risk assessment, enabling healthcare professionals to deliver personalised interventions that will improve patient outcomes and avert serious problems.

These models have the potential to address global healthcare inequities and resource shortages, making early diabetes detection accessible to a variety of demographics, as they are further validated and improved across broad groups. Furthermore, by combining electronic health records and interpretable models, clinical practice will be improved and adoption will become more widespread.

Diabetes diagnosis in women at an early stage using machine learning classification approaches offers a significant advancement in the quest for a healthy future. We can lessen the burden of diabetes and enhance the quality of life for countless women throughout the world by leveraging the power of data-driven insights and preventive healthcare measures. In order to ensure that these developments reach every corner of the global population and help women affected by this chronic condition have a better and healthier future, collaboration across disciplines, innovation, and inclusivity must be prioritized as this field of study continues to advance.

References

[1] American Diabetes Association. (2021). Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes. Diabetes Care, 44(Supplement 1), S15-S33.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.