IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 11, Issue 1 - V11I1-1376)
Available online at: https://www.ijariit.com

Diabetes Prognosis Using Machine Learning

Anshika Sharma
anshika.sharma.cse.2021@miet.ac.in
Meerut Institute of Engineering and Technology, Meerut, Uttar Pradesh

Divasha alag
divasha.alag.cse.2021@miet.ac.in
Meerut Institute of Engineering and Technology, Meerut, Uttar Pradesh

Atharva
atharva.y.cse.2021@miet.ac.in
Meerut Institute of Engineering and Technology, Meerut, Uttar Pradesh

Aditya Pratap Singh
aditya.singh.cse.2021@miet.ac.in
Meerut Institute of Engineering and Technology, Meerut, Uttar Pradesh

ABSTRACT

Diabetes is a prolonged disorder brought on by above-normal blood glucose levels, leading to symptoms like frequent urination, thirst, and hunger. It can May cause significant complications, such as blindness, kidney failure, heart failure, and stroke. The pancreas usually produces insulin to help cells absorb glucose for energy, but this process fails in diabetes. Machine learning offers tools for early diabetes prediction. Various algorithms, such as KNearest Neighbors, Logistic Regression, Random Forest, and Decision Tree, are evaluated to select the most accurate model for diagnosis.

Keywords: *Diabetes in Pregnant Women, Machine Algorithms, Linear Regression*

INTRODUCTION

Diabetes is increasingly becoming one of the most prevalent health challenges, affecting not only adults but also a significant number of younger individuals. To grasp the intricacies of diabetes and its development, it is essential to first understand the body's functioning in the absence of the condition. Glucose, commonly referred to as blood sugar, is derived from the consumption of carbohydrate-rich foods, which serve as the body's primary energy source. Carbohydrates are found in a variety of food items, including bread, cereals, pasta, rice, fruits, dairy products, and starchy vegetables. When these foods are consumed, the body metabolizes them into glucose, which is then transported through the bloodstream.

Glucose plays a vital role in the body. Some of it is directed to the brain, aiding cognitive functions, while the rest is delivered to body cells for energy. Excess glucose is stored in the liver for future energy requirements. However, the process of utilizing glucose requires the presence of insulin, a hormone produced by the beta cells of the pancreas. Insulin acts as a key, allowing glucose to pass from the bloodstream into cells by unlocking their doors. When the pancreas produces insufficient insulin (a condition known as insulin deficiency) or when the

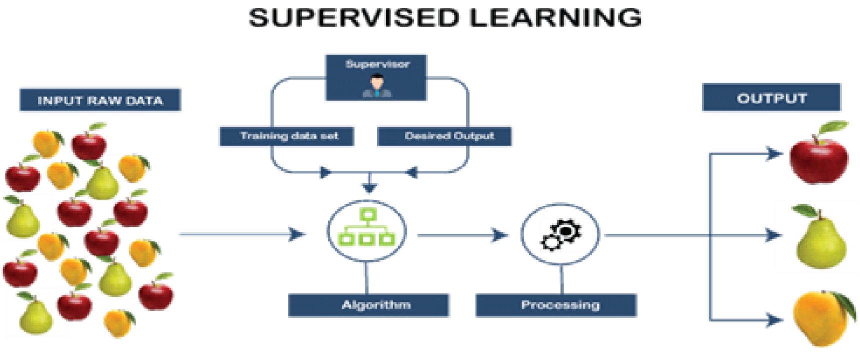
body's cells fail to effectively utilize the insulin produced (referred to as insulin resistance), glucose accumulates in the bloodstream, leading to hyperglycemia—a hallmark of diabetes. Diabetes Mellitus, therefore, refers to elevated blood glucose levels, which may also manifest as sugar in the urine. Machine learning, a subset of artificial intelligence, encompasses various approaches for enabling systems to learn and improve from data without explicit programming. Among the core types of machine learning is supervised learning, which focuses on training algorithms using labeled datasets. In this context, the model learns to map input data to corresponding outputs by identifying patterns in the training data. Once trained, the model can predict outputs for new, unseen inputs with high accuracy.

Key algorithms in supervised learning include:

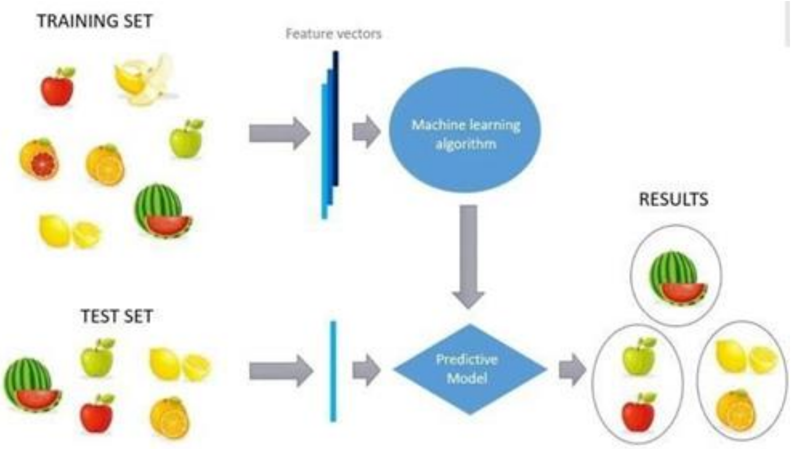
Linear Regression: A method used to estimate the relationship between a dependent variable and several independent variables, typically applied in forecasting continuous outcomes.

Logistic Regression: Primarily used for binary classification tasks, this algorithm estimates probabilities to determine the class of an input.

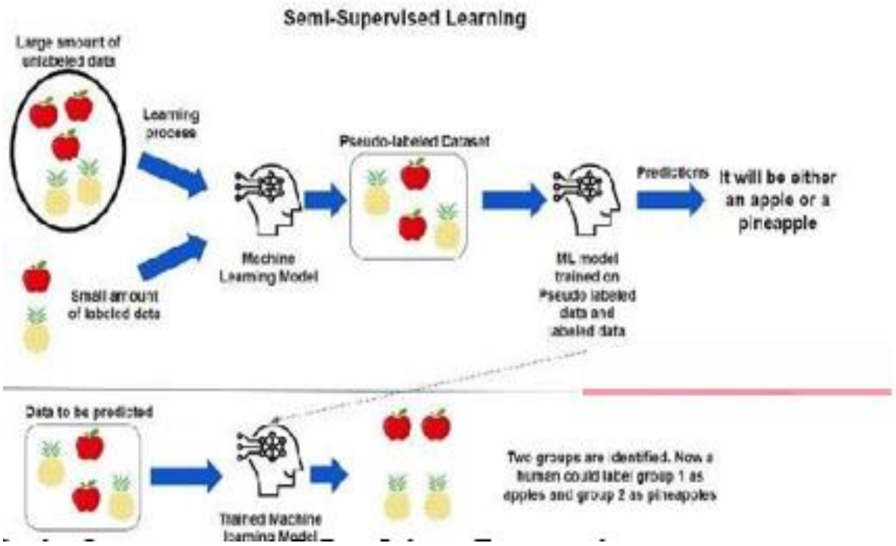
Naïve Bayes Classifier: Built upon Bayes' theorem, this algorithm is particularly effective for text classification and other problems involving categorical data. Supervised learning techniques are extensively utilized in healthcare for predictive analytics, aiding in early detection, diagnosis, and personalized treatment plans, including for conditions like diabetes.



Unsupervised Learning:



Unsupervised learning involves a set of techniques designed to analyze data without labeled responses. Here, the goal is not to make predictions but to uncover hidden patterns or structures within the data. This method is particularly useful when we have only a set of features measured on multiple observations but no corresponding output variables. It focuses on answering questions like: "Can the data be visualized in an insightful way?" or "Are there subgroups or clusters within the dataset?" Clustering algorithms are a key part of unsupervised machine learning. A few of the widely adopted algorithms include:
1) K Means Clustering Algorithm 2) Apriori Algorithm.
Semi-Supervised Learning:
In semi-supervised machine learning, the model benefits from both guided instruction and selfdirected learning. This method, as suggested by its name, combines features of both supervised and unsupervised learning. It employs a small amount of labeled data, as seen in supervised learning, alongside a larger quantity of unlabeled data, similar to unsupervised learning. Initially, the model is trained with labeled data, and the partially trained model is subsequently used to generate pseudolabels for the unlabeled data, promoting further training.


Semi-Supervised Learning
Large amount of unlabeled data
Learning process
Small amount of labeled data
Machine Learning Model
Pseudo-labeled Dataset
ML model trained on Pseudo labeled data and labeled data
Predictions
It will be either an apple or a pineapple
Data to be predicted
Trained Machine learning Model
Two groups are identified. Now a human could label group 1 as apples and group 2 as pineapples

Reinforcement machine learning:

Reinforcement learning algorithms identify optimal actions by engaging in a process of trial and error. These algorithms determine the next action based on behaviors learned through experience, enabling them to identify strategies that maximize long-term rewards.

LITERATURE REVIEW

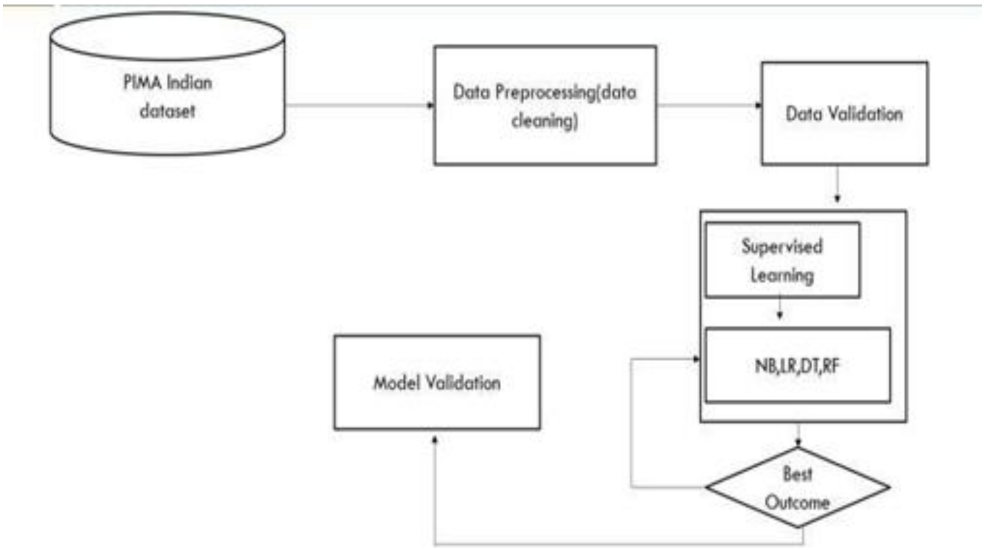
The study of related work highlights outcomes derived from diverse healthcare datasets, where analyses and predictions have been conducted using a range of methodologies and approaches. Researchers have developed and implemented various predictive models employing data mining techniques, machine learning algorithms, or a combination of these methods.

Kalyankar et al. (2017): Kalyankar, G. D., and colleagues (2017) proposed a system for predictive analysis of diabetic patient data, utilizing machine learning techniques combined with Hadoop for efficient data processing. Recent studies have explored varied approaches to diabetes prediction. Anand, A et al. (2015) developed a prediction system based on lifestyle factors, while Dr. Saravana Kumar N M and colleagues (2015) utilized Hadoop and MapReduce frameworks to analyze diabetic data for disease type and risk assessment. Built on a Hadoop framework, it offers a cost- effective solution suitable for healthcare organizations.

Aiswarya Iyer (2015): Aiswarya Different researchers have explored various classification approaches for diabetes prediction. Aiswarya Iyer (2015) evaluated Naïve Bayes and Decision Tree algorithms, comparing their effectiveness in pattern detection within diabetes data. K. Rajesh et al. (2012) implemented the C4.5 Decision Tree algorithm to identify data patterns and improve classification accuracy. Humar Kahramanli et al. (2008) combined artificial neural networks with fuzzy logic for their prediction model.

Diabetes diagnosis traditionally depends on physical exams and chemical tests. Georga E et al. (2009) explored data mining techniques for glucose prediction through the METABO system. Breault JL highlighted diabetes as a major health concern in the US, noting the importance of patient registries for research. Jayalakshmi T developed an artificial neural network-based classification method. Vijayan V's team proposed a machine learning diagnostic approach, while Pal CJ and Witten IH examined key data mining methodologies that generate knowledge representations.

METHODOLOGIES



Data Processing: Data preprocessing involves preparing raw data to make it suitable for machine learning models. It is the first and critical step in creating a machine learning model, ensuring the data is properly formatted. The process of data preprocessing consists of the following steps:

Steps that are involved in Data Preprocessing are:

Clutching the Data - The first step in building a machine learning model is acquiring a data pool. A data pool consists of data related to a specific problem, formatted appropriately for the task at hand.

Importing the libraries - To perform data preprocessing using Python, several libraries must be imported. These libraries offer predefined functions to carry out specific tasks. Key libraries include:

Numpy - This library is used for performing mathematical operations in Python. It is fundamental for scientific computation and supports large, multidimensional arrays and matrices.

Matplolib - A Python library for 2D plotting. It helps in visualizing data, and the sub-library pyplot is typically used for plotting charts.

Pandas - One of the most widely used Python libraries for bringing and managing data pool. It is an open-source tool for data deception and review.

Importing the dataset - After acquiring the dataset, the next step is importing it into the working environment. The current directory must be set as the working directory before importing the dataset.

Handling missing data - Handling lacking data is pivotal because it can negatively affect the performance of the machine learning model. Missing values in a dataset need to be dealt with through imputation or removal methods.

Mapping categorical attributes - Machine learning models require numeral inlet, so categorical variables (such as "Country" or "Purchased") must be transformed into numerical representations. This encoding process helps prevent errors during model training.

Splitting the data pool - The data pool is divided into two subsets: a learning samples and a holdout set. This split helps assess the model’s behavior and ensures it generalizes well to unseen data.

Feature Scaling: Feature scaling standardizes the independent variables within a fixed range, ensuring that no feature dominates the model due to differing units or magnitude. This step is essential for algorithms sensitive to the scale of input data.

Classification: Supervised learning employs classification via the algorithm predicts the expected category for a given input. The training process involves a labeled data pool then assessed using new, previously unseen data. There are two types of classification: We can distinguish two kinds of Classifications:

Binary Classification: It focuses on sorting input into one of two possible classes. Predicting if a person is affected by a disease using medical data (Yes/No) is an example.

Multiclass Classification: In this case, the input is classified into one of several classes. For instance, predicting the species of a flower based on various features.

Algorithms:

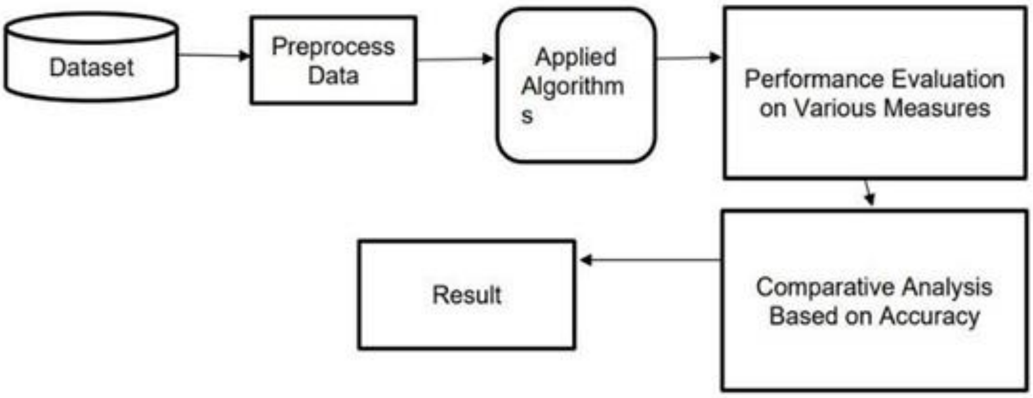
Naïve Bayes - The Naïve Bayes Classifier is based on Bayes' Theorem and is commonly used for classifying text, such as in email spam detection. It determines the likelihood that a data point belongs to a particular class and classifies it based on this probability.

Logistic Regression - Unlike linear regression, which deals with continuous data, logistic regression handles discrete data and is used for binary classification. It estimates the likelihood of an event taking place (1) or not taking place (0). It is often used in medical applications to predict conditions like cancer based on tumor size.

Decision tree – A predictive tree model is a tree-like diagram where each internal node reflects a feature, branches represent decision criteria, and the leaf nodes provide the predicted outcome.This algorithm can be used for both classification and regression tasks.

Random Forest - Random Forest is a classifier that contains a number of decision trees on various subsets of the given data pool and takes the average to improve the predictive accuracy of that dataset.

RESULTS AND DISCUSSIONS


Dataset
Preprocess Data
Applied Algorithm s
Performance Evaluation on Various Measures
Comparative Analysis Based on Accuracy
Result

1) Accuracy:

$Acc{=}TP\ {+}TNTP{+}\ TNFN{+}FP$

Accuracy reflects the overall performance of the classifier, indicating its ability to correctly classify the given samples.

2) Sensitivity: $Sn{=}\quad TP{+}FN$

This metric provides insight into how effectively the classifier identifies positive instances.

3) Specificity: $Sp{=}TN \qquad TN{+}FP$

While sensitivity focuses on identifying positive instances, specificity assesses how effectively the classifier identifies negative instances.

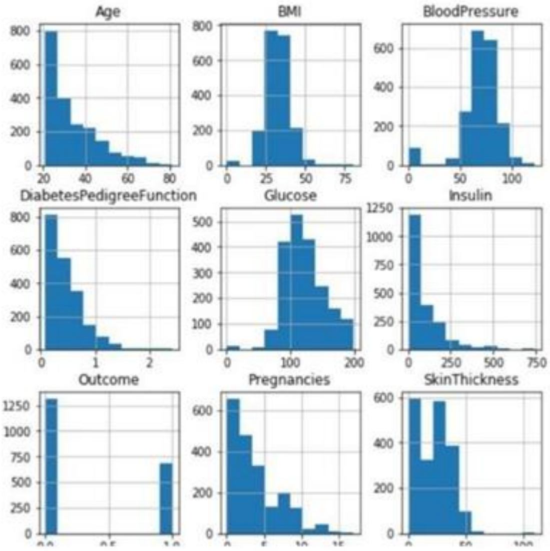
For evaluation purposes, the dataset is divided into two subsets: 80% is used for training, while the remaining 20% is reserved for testing the model's accuracy. Experimental Results:

Table 4.1 Classifiers and their Accuracies

Classifier	Accuracy
Naïve Bayes	77.93%
Logistic Regression	79%
Decision Tree	72.72%
Random Forest	77.27%

Table 4.2 Classifiers and their Precisions

Classifier	Precision	
Naïve Bayes	0.83	0.68
Logistic Regression	0.83	0.68
Decision Tree	0.81	0.60
Random Forest	0.83	0.67



Histogram: Examining the plots reveals how each feature and label are distributed across various ranges, highlighting the importance of scaling. Additionally, discrete bars indicate categorical variables, which require proper handling before applying machine learning algorithms. The outcome labels consist of two classes: 0, representing no disease, and 1, indicating the presence of disease.

CONCLUSION

This research presents a distinctive methodology for diabetes detection through machine learning algorithms. The study implements and analyzes four classifications

methods: Naive Bayes, Logistic Regression, Decision Tree, and Random Forest classifiers. The dataset underwent an 80-20 split, with 80% allocated for model training and 20% for testing purposes, ensuring reliable model evaluation.

The analysis revealed that among the tested models, Logistic Regression demonstrated superior performance with 79% accuracy. The remaining classifiers showed varying degrees of effectiveness: Naive Bayes achieved 77.93%, Random Forest reached 77.27%, and Decision Tree attained 72.72%. The consistently high-performance metrics, particularly from Logistic Regression, suggest these approaches offer promising capabilities for diabetes detection.

A comparative evaluation against previous research using identical data revealed superior results from our models. The enhanced accuracy and f-score metrics demonstrate meaningful progress in diabetes prediction methodology, suggesting an advancement over existing techniques in this field.

The research outcomes offer practical value for healthcare implementation. The machine learning approach presents a robust diabetes prediction system that can enhance diagnostic capabilities in clinical settings. Early and accurate identification of at-risk patients enables proactive medical intervention, potentially improving overall healthcare delivery and patient care standards.

Machine learning techniques enable efficient processing of complex medical data, facilitating rapid and precise analysis. This computational advantage is crucial in healthcare environments where prompt diagnosis and early preventive interventions can substantially impact patient outcomes.

This research validates machine learning's effectiveness for diabetes prediction, with our models achieving high accuracy rates. The results demonstrate the technology's value for early disease detection in healthcare settings. This work establishes a foundation for

continued advancement in AI-driven medical diagnostics, particularly for chronic condition management.

REFERENCES

[1] Kalyankar, G. D., Poojara, S. R., & Dharwadkar, N. V. (2017, February).

[2] Predictive analysis of diabetes patient data using machine learning and Hadoop frameworks. Proceedings of the 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) (pp. 619-624). IEEE.

[3] Anand, A., & Shakti, D. (2015, September). Prediction of diabetes based on lifestyle and individual health indicators. Proceedings of the 2015 1st International Conference on Next Generation Computing Technologies (NGCT) (pp. 673-676). IEEE.

[4] Nithya, B., & Ilango, V. (2017, June). Application of machine learning techniques for predictive healthcare analytics. Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 492-499).

[5] IEEE.

[6] Eswari, T., Sampath, P., & Lavanya, S. J. P. C. S. (2015). A predictive approach for diabetic data analysis in big data environments. Procedia Computer Science, 50, 203-210.

[7] Kumar, P. S., & Pranavi, S. (2017, December). A performance comparison of machine learning algorithms on diabetes datasets using big data analytics. In Proceedings of the 2017 International Conference on

[8] Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010, February). Use of association rules in the classification of type-2 diabetic patients.

[9] Proceedings of the 2010 Second International Conference on Machine Learning and Computing (pp. 330-334). IEEE.

[10] Aiswarya Iyer, S. Jeyalatha, and Ronak Sumbaly. "Using classification mining techniques for diabetes diagnosis." International Journal of Data Mining & Knowledge Management Process (IJDKP), 5(1), January 2015.

[11] Mani Butwall, Shraddha Kumar. "A data mining approach for diagnosing diabetes mellitus with the Random Forest Classifier." International Journal of Computer Applications, 120(8), 2015.

[12] Rajesh, K., & Sangeetha, V. (2012). Application of data mining techniques in diagnosing diabetes. International Journal of Engineering and Innovative Technology (IJEIT), 2(3), September 2012.

[13] Vijayan, V., & Ravikumar, A. (2014). A survey on data mining algorithms for predicting and diagnosing diabetes mellitus. International Journal of Computer Applications, 95(17).

[14] Georga E., et al. (2009). Using data mining techniques for blood glucose prediction and knowledge discovery in diabetic patients: The METABO diabetes management system. In 31st Annual International Conference IEEE EMBS (pp. 5633–5636). Minneapolis, Minnesota, USA.

[15] Breault, J. L., Goodall, C. R., & Fos, P. J. (2002). Data mining for diabetic data warehouse analysis. Artificial Intelligence in Medicine, 26(1-2), 37-54.

[16] Jayalakshmi, T., & Santhakumaran, A. (2010). A novel classification method using artificial neural networks for diabetes diagnosis. 2010 International Conference on Data Storage and Data Engineering. IEEE Computer Society (pp.

[17] 159–163).

[18] Yue, X., Wang, H., Jin, D., Li, M., & Jiang, W. (2016). Blockchain-based healthcare data gateway: Privacy risk control for healthcare intelligence. Journal of Medical Systems, 40, 1-8.

[19] Kowsher, M., Turaba, M. Y., Sajed, T., & Rahman, M. M. (2023). Prediction and treatment prognosis of type-2 diabetes using deep learning classifiers. arXiv preprint arXiv:2301.03093.

[20] Akula, R., Nguyen, N., & Garibay, I. (2019). Ensemble machine learning model for accurate type-2 diabetes prediction. arXiv preprint arXiv:1910.09356.

[21] Adler, A. (2021). Identifying key risk factors for diabetes and undiagnosed diabetes through machine learning. arXiv preprint arXiv:2105.09379.

[22] Bitencourt-Ferreira, G., & de Azevedo Jr, W. F. (2019). Machine learning to predict binding affinity. In Methods in Molecular Biology (Vol. 2053, pp. 251-273). Springer.

[23] Ha, S., Choi, S. J., & Choi, S. (2022). Risk prediction of sleep disorders using machine learning-based questionnaires: A development and validation study. Journal of Medical Internet Research, 24(9), e35807.

[24] Echouffo-Tcheugui, J. B., & Selvin, E. (2021). Machine learning in diabetes research: A review. Current Diabetes Reports, 21(9), 1-10.