Original Article

An Analysis of various Machine Learning Techniques for Predicting Diabetes in its Early Stages

Durga P[1], Sudhakar T[2]
[1] CSE, VIT-AP University, Amaravati, AP, India.
[2] CSE, VIT-AP University, Amaravati, AP, India.
Email: pdurga593@gmail.com[1],sudhakar.t@vitap.ac.in[2]
DOI: 10.47750/pnr.2022.13.S01.238

Abstract

Chronic metabolic disease diabetes is analyzed based on high glucose levels in the blood, these levels become more serious to coronary heart, blood vessels, eyes, kidneys, and nerves. The most prevalent type of disease, known as type 2, generally affects most adults when the required insulin is not produced in the body. Diabetes that affects human health is type-1 diabetes. The other name for type-1 diabetes is insulin-structured diabetes. This disease causes small illnesses to the pancreas and reduces the generation of insulin gradually. Access to affordable medications, such as insulin, is essential for those with diabetes to survive. Making predictions from clinical data is one of these challenges. In information technology, gadget mastering is a developing scientific discipline that deals with the methods through which machines learn from experience. After the analysis of several Machine Learning (ML) techniques, this study aims to develop a machine that can accurately detect diabetes in a patient early on. Additionally, this project is pursuing a suggestion for a powerful method for the early identification of diabetic disease.

Keywords: Machine Learning Models, Disease Prediction, Random Forest, Logistic Regression.

INTRODUCTION

Diabetes Mellitus (DM), sometimes known as diabetes, is one of the unpredictable diseases that may occur due to the lack of insulin [1]. Insulin is one of the significant hormones which is generated by the pancreas and permits the cells to obtain glucose from food sources to provide energy to humans [2]. To increase and maintain the high glucose levels in the blood hyperglycemia is used. The two significant factors that show huge impact are: the body cannot generate the required insulin in the blood cells and it is not effectively responding to insulin. To get the energy to the human the blood glucose the required insulin has to generate. If the body doesn't use glucose for generating energy then this causes hyperglycemia. This may affect the fitness of the human. The other side effects are such as sudden heart attacks, blindness, issues with kidneys, etc.

DM is one of the most common diseases that millions of people are affecting with this disease globally. This may be increased very sharply in coming years. DM is divided into various types that most widely affect human health. The most important types of diabetes are called type 1 diabetes (T1D) and type 2 diabetes (T2D), based on insulin levels in the blood. T2D is one of the default types of diabetes and 90% of people are affected by this diabetes. T1D is one of the types of diabetes and only adults may affect by this disease. Type-2 diabetes is brought on by the body's ineffective utilization of insulin. Due to physical sluggishness and being overweight, this happens. Only pregnant women can get gestational diabetes. Due to these, it will show a great impact on the pregnancy and the fitness of the child. After the baby is born, this type of diabetes disappears, or it can cause type-2 diabetes. The main factors that contribute to T2D are lifestyle, Other types of DM include gestational diabetes, endocrinopathies, MODY (maturity Onset Diabetes of the Younger), neonatal, mitochondrial, and pregnant diabetes. These categories are based on the profile of insulin secretion and time of onset. Polyuria, polydipsia, and extreme weight loss are symptoms of DM. Blood glucose levels affect the diagnosis (fasting plasma glucose = 7. 0 mmol/L.

OBJECTIVE

The goal of this research is to create a system that can conduct early diabetes prediction for a patient with greater accuracy by merging the findings of several Machine Learning (ML) approaches.

LITERATURE SURVEY

When predicting glucose concentration (BG) levels using time-series data from diabetes type 1 patients, Jinyu Xie et al. [3] examined the effectiveness of several widely used ML models in contrast to a traditional Autoregression with External inputs (ARX) model (T1D). Methods: ML-based regression models, such as a foundational LSTM network and a Temporal Convolution network, are built using DL models. The average Root Mean Square Error (RMSE), temporal gain (TG) for early identification, and the neutralized strength of the 2nd order variations (ESOD) of the anticipated time-series data are the metrics for assessing the accuracy of predictions. and indicate the danger of fabricated metrics on hypo/hyperglycemia activities. The ARX version performed with the minimum median RMSE for both recursive and direct procedures and the second greatest common TG under the straightforward approach. It also had a greater general normalized ESOD than some other models.

The purpose of the study by Bum Ju Lee et al. [4] was to evaluate the relationship between type 2 diabetes and the HW phenotype in Korean adults and to evaluate the predictive power of various phenotypes, including combinations of individual anthropometric measurements and TG levels. This retrospective go-sectional observation included 11 937 topics between November 2006 and August 2013. They took anthropometric measurements, fasting plasma glucose, and TG levels. They employed binary logistic regression (LR), using HW and individual anthropometric data, to look at statistically significant differences between healthy patients and those who have type 2 diabetes. Two machine learning algorithms, NB and LR, were utilized to assess the predictive power of various phenotypes in order to produce more accurate prediction results. A tenfold go-validation technique has been used in all prediction studies. The presence of HW was the factor most substantially associated with type 2 diabetes among all the factors ($p < 0.001$, adjusted odds ratio (OR) = 2.07 [95% CI, 1.72-2.49] in men; $p <$ zero.001, adjusted OR = 2.09 [1.79-2.45] in women).

In order to predict the early onset of diabetes patients, Tuan Minh le et al. [5] conducted device research to learn about versions. It is a distinct wrapper-based function option that optimizes the Multilayer Perceptron (MLP) to reduce the number of necessary input attributes using Adaptive Particle Swam Optimization (APSO) and Grey Wolf Optimization (GWO). Additionally, results were compared between this method with a variety of conventional system learning approaches, including LR, SVM, DT, KNN, NBC, and RFC. Results of the proposed approach indicate that less functionality is now desired, not the best. Good predictive accuracy, although, might be achieved (97% for APGWO - MLP and 96% for GWO - MLP).

M. S. Islam et al., [6] The best risk indicators were identified using two innovative feature extraction methods, which were then used in conjunction with an ML pipeline to make long-term predictions of type 2 diabetes. The San Antonio Heart Research, a longitudinal clinical study, was used to assess the proposed procedures. In determining whether a person will get type 2 diabetes within the next 7-8 years or not, their suggested model was 95.94% accurate.

S. Perveen et al. [7] A novel estimated approach was created to estimate a person's 8-year likelihood of developing Type 2 Diabetes Mellitus using NDDM as a component with HMM (T2DM). The proposed technique is examined using actual clinical data acquired from CPCSSN. The outcomes showed that the suggested method, which utilized the intermittent EMR data that was available, could successfully approximate and enhance predictive performance.

N. Fazakis et al., [8] a framework for monitoring users' health, wellbeing, and functional capacity that is worker-centric, IoT-enabled, and equipped with AI technologies. In this direction, the created system focuses on predicting the risk of developing diabetes and applies, evaluates, and incorporates particular KDD process components. Particularly, several Supervised ML models are considered for dataset construction, feature selection, and classification. The proposed model achieved the AUC of 0.884 which is better compared with existing ML models.

M. Shokrekhodaei et al., [9] proposed the lightweight approach with different wavelengths to increase the performance based on sensitivity and selection of glucose with unique solutions. To reduce the errors multiple wavelengths are used to analyze the components in blood and tissues. The proposed approach performance is increased to 99% based on the classification done.

U. Ahmed et al., [10] presented a diabetes prediction model using a combined machine learning approach. The data will be analyzed using SVM and ANN models to determine whether the diagnosis of diabetes is positive or negative. The results of these types serve as a function of entering a member of this unusual type, and the logical deviation determines whether the diagnosis of diabetes is true or false. The combined version is stored in the cloud storage system for later use. A combined model determines whether a patient has diabetes based on their current medical history. The proposed integrated ML model outperforms the previously described method with a predictive value of 94.87.

N. E. Costea et al., [11] compared the experimentally achieved findings for the prediction of diabetes using three machine learning methods. The three algorithms under consideration are NB, RF, and SVM. The process involves evaluating the algorithms' performance while considering various metrics to evaluate various approaches and improve accuracy. They discovered that Random Forest and Support Vector Machine (SVM) achieved accuracy levels above 80%.

A. Yahyaoui et al. [12] proposed a DS based on ML methods to predict diabetes. They contrasted common ML methods with DL models. SVM and RF have been utilized for the traditional machine learning technique. To predict and recognize diabetic patients, they employed a fully convolutional neural network (CNN) for DL. The Pima Indians Diabetes database, which comprises 768 samples and 8 features for each, is used to test the suggested methodology. 268 people who had 500 samples had diabetes, according to the results. DL, SVM, and RF had accuracy rates of 76.81 percent, 65.38 percent, and 83.67 percent, respectively. According to the research observations, RF was more reliable than DL and SVM approaches for predicting diabetes.

METHODOLOGY

Disease prediction using Machine Learning

Machine learning techniques are frequently employed to forecast diabetes and produce better outcomes. One of the common machine learning techniques in the medical industry that has excellent classification capability is the decision tree (DT). Many DTs are produced from the random forest. The neural network is an ML technique that has recently gained popularity and offers superior performance in several areas.

i. Models for Supervised Learning and Prediction

Algorithms are used to construct predictive models and are supervised during learning. A prediction model projects the values that are missing using other dataset values. Using a set of input data and a set of output data, a supervised learning technique builds a model that can make precise predictions about how a new dataset would react. DT, ANN, Instance-based Learning, Bayesian Method, and Ensemble Methods are a few examples of supervised learning techniques. These machine learning techniques are highly sought after. [13]

ii. Descriptive models and unsupervised learning

Unsupervised learning is used to create descriptive models. In this model, the inputs are known, but the output is not. Transactional data is where unsupervised learning is most commonly applied. The clustering techniques used in this method include k-Means clustering and k-Medians clustering. [13]

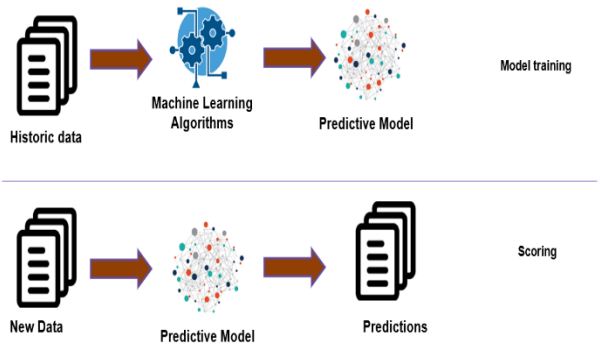
iii. Semi-supervised Learning

The semi-supervised learning technique on the training dataset uses labeled and unlabeled data. Techniques for classification and regression fall under semi-supervised learning. Regression methods include linear regression and logistic regression, for instance. [13]

DATASET

PIMA and Indian diabetes datasets were downloaded for this work via Kaggle. We have 768 observations and 8 variables in the PIMA and There are 17 attributes and 953 observations in this diabetes dataset, which includes an outcome of 0 (or) 1. A score of 0 indicates that a person does not have diabetes, whereas a score of 1 indicates that they do. The diabetes dataset for Pima Indians is split into 80% and 20%. Eighty percent of the data are from training, and 20 percent are from testing. In this study, we use the Python programming language to create classification techniques like the Random Forest tree and the Logistic Regression algorithm**.**

Fig 1: Role of ML in Disease Prediction



PROPOSED SYSTEM

In the suggested method, supervised learning from machine learning is employed to produce precise results for chronic conditions like diabetes. There are two sections to the dataset. 1):- Training part 2) Testing portion. The model is created by applying the algorithm to the training set, which is then used to produce the model for the test, after which its performance is calculated. In the outlined study, we used Python programming to apply the Logistic Regression and Random Forest techniques to the Pima Indian diabetes dataset.

Fig. 2 Shows an architecture schematic for the diabetes prediction model. There are four distinct modules in this model. These modules consist of

i. Dataset Collection
ii. Data Pre-processing
iii. Model Building
iv. Performance Analysis

i. Dataset Collection

Data collection and interpretation are covered in this module to identify patterns and trends that may be used to forecast outcomes and assess them. So here is a description of the data set. There are 17 attributes and 953 observations in this diabetes dataset.

ii. Data Pre-processing

Inconsistent data are handled in this model stage to produce more precise and accurate findings. Missing values can be found in this dataset. Because certain attributes, such as age, skin thickness, blood pressure, and BMI, cannot have zero values, we imputed missing values for these attributes. The dataset is then scaled to normalize each value.

iii. Model Building

The creation of a model for diabetes prediction occurs at this important stage. In this, we've developed a number of machine learning strategies for diabetes prediction. K-NN, LR, DT, and RF Classifier are some of these technologies.

iv. Performance Analysis

The performance analysis is mainly focused on calculating by using a confusion matrix.

Confusion Matrix

A confusion matrix is used to assess the effectiveness of the classification models for a specific test data set. It cannot be determined until the test data's true values are known.

True Negative: If the actual value is negative, then the predicted value is also negative.

True Positive**:** If the actual value is positive, then the predicted value is also positive.

False Negative**:** This error, which is also known as a Type-II error, occurs when the model projected that the result would be no while it was really positive.

False Positive**:** A Type-I error is when the model predicted "Yes," but the actual outcome was "No."

Classification Accuracy**:** This is a crucial factor in determining how accurate a problem's classification is. It specifies how frequently the model predicts the right result. The number of accurate predictions made by the classifier divided by the total number of predictions made by the classifiers can be used to compute it. The following is the formula:

$$\textbf{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Precision**:** It might be expressed as the fraction of correctly predicted positive classifications that actually happened or as the number of accurate outputs the model produced. It can be determined by applying the formula below:

$$\textbf{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Recall**:** As our model correctly predicted, it is regarded as the positive classes that are not included in the total. There has to be a big recall.

$$\textbf{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

F-measure**:** It is challenging to compare two models if one has a high recall and a poor precision. F-score can therefore be used for this purpose. This score enables us to assess recall and precision simultaneously. If the recall and precision are equal, the F-score is at its highest. Using the formula below, it can be calculated:

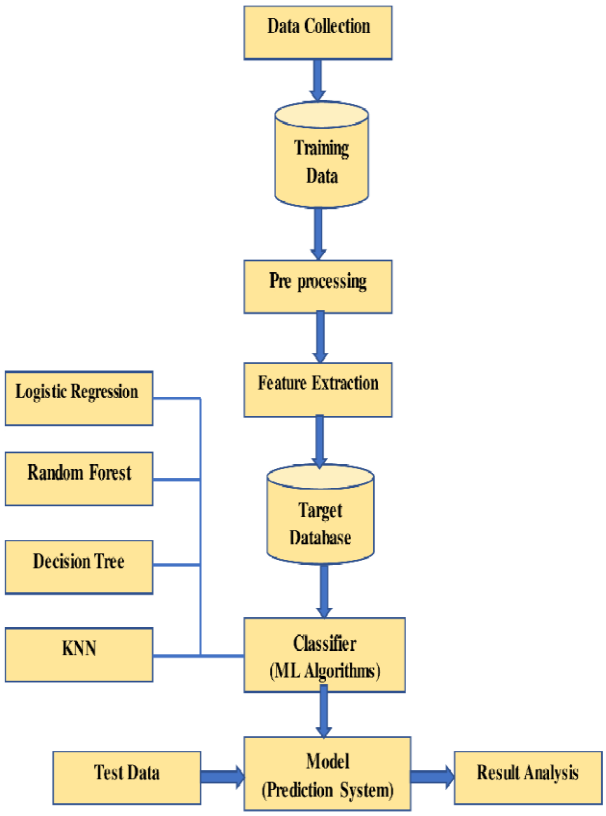
$$\textbf{F1} - \textbf{Score} = 2 * \frac{(\mathrm{Precision} * \mathrm{Recall})}{(\mathrm{Precision} + \mathrm{Recall})}$$

Logistic Regression (LR)

LR is the model, the dependent variable can only have the values "0" and "1," which represent consequences like pass/fail, win/lose, alive/dead, or healthy/ill. LR is used in the majority of medical fields, social sciences, and machine learning. For

example, the Trauma and Injury Severity Score (TRISS), which is often used to predict death in wounded patients, was first created using Logistic Regression (LR). It has developed a tonne of diagnostic tools for determining the seriousness of a patient. The approach may be used in engineering as well, especially when figuring out how likely it is for a certain system, process, or object to fail. In marketing applications, it is widely used to predict a customer's propensity to purchase a product or end a subscription. A commercial application assesses the probability that a customer will default on their mortgage interest. It can predict someone's probability to pursue economics as a field. In order to analyze sequential data in natural language, conditional random fields are a logistic regression extension. In this study, LR and seven patient factors were used to determine if diabetes existed or not.

Fig 2: System Architecture



RESULTS & DISCUSSION

Experiments are conducted by using a python programming language with the help of powerful libraries functions. Pima Indians Diabetic Information and the Diabetes Dataset 2019 were the 2 resources we used to conduct the trials. The Pima Indians Diabetes Database contains eight attributes and 768 observations of data. 2019 Diabetic Dataset comprises 953 observations and Seventeen attributes, Table 1 shows the parameters: sensitivity or recall, accuracy, precision, and F1-score. Apply the confusion matrix to the observed results. Despite using various ML algorithms on the dataset, the accuracy results are as follows. The maximum accuracy, 97 percent, is provided by logistic regression.

Table I: Diabetes dataset test results

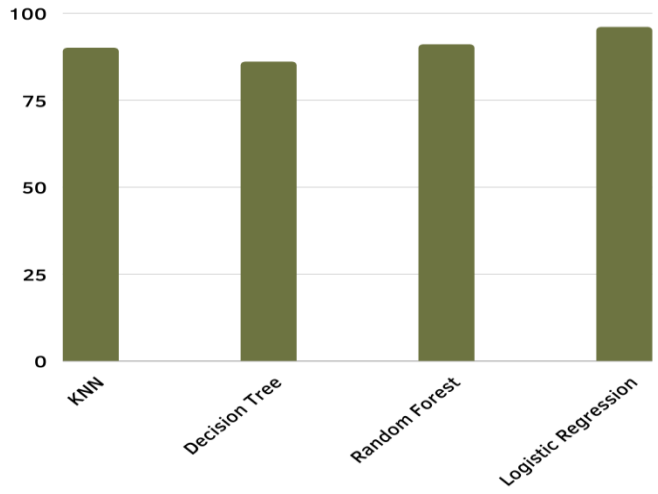
Algorithms	Recall	Accuracy	Precision	F1-Score
LR	**0.96**	**0.97**	**0.95**	**0.96**
RF	0.78	0.91	0.78	0.79
DT	0.81	0.86	0.79	0.77
KNN	0.89	0.90	0.99	0.97

Accuracy, F1-Score, Precision, and Recall are the many performance metrics being compared. Visualizing these accuracy helps us comprehend the differences between them. Table 2 displays the outcomes of the algorithm on the Diabetes dataset for 2019. The outcomes demonstrate that the LR model was the most effective in this situation.

Table II: Compares the accuracy of diabetes datasets used in this study and the PIMA diabetes dataset

Algorithms	Accuracy with PIMA Data set	Accuracy of the Diabetes Dataset Used in This Article
Logistic Regression	76%	97%
Random Forest	72%	91%
Decision Tree	74%	86%
KNN	72%	90%

Fig 3: Comparison of various ML algorithms based on accuracies



The comparison study of the machine learning approach on the Diabetes Dataset 2019 is shown in Fig. 3 where X-axis represents accuracy and Y- axis represents ML algorithms. Out of which the LR proved the best accuracy.

The Logistic Regression (LR) method is provided here with features that played a key influence in prediction. The total importance of each character having a significant part in diabetes has been plotted, with the X-axis representing the relevance of each feature and the Y-axis representing the names of the features constituted in fig 4.

Fig 4: Graph Representation of Diabetes Prediction with LR

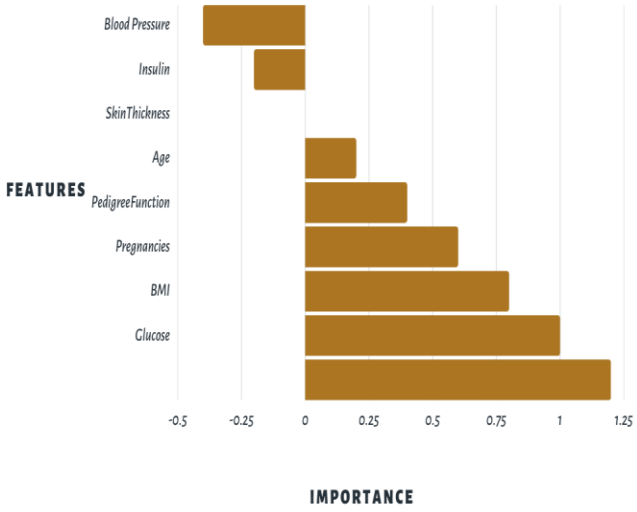
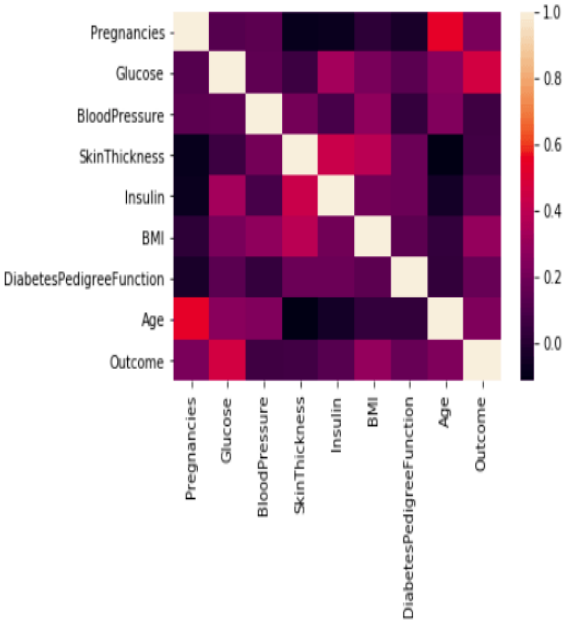


Fig 5: Predicting Diabetes with Logistic Regression



CONCLUSION

Machine Learning (ML) has the potential to dramatically change the prediction of diabetes risk with the application of cutting-edge computational techniques and the availability of a substantial number of epidemiological and genetic diabetes risk datasets.

Effective diabetes therapy depends on early detection. The classification of the dataset used in this study was carried out using a variety of methodologies, with Logistic Regression achieving the highest accuracy of 96 percent. In this study, an ML technique for predicting diabetes levels was described. The approach might also help scientists develop a precise and practical tool that would be available to doctors to help them decide how serious the sickness is. This study can be broadened to determine the likelihood that people without diabetes will develop the disease in the ensuing years.

REFERENCES

1. Z. Punthakee, R. Goldenberg, and P. Katz, “Definition, Classification, and Diagnosis of Diabetes, Prediabetes and Metabolic Syndrome,” Can. J. Diabetes, vol. 42, pp. S10–S15, 2018.
2. M. N. Piero, “Diabetes mellitus – a devastating metabolic disorder,” Asian J. Biomed. Pharm. Sci., vol. 4, no. 40, pp. 1–7, 2015.
3. J. Xie and Q. Wang, "Benchmarking Machine Learning Algorithms on Blood Glucose Prediction for Type I Diabetes in Comparison With Classical Time-Series Models," in IEEE Transactions on Biomedical Engineering, vol. 67, no. 11, pp. 3101-3124, Nov. 2020, doi: 10.1109/TBME.2020.2975959.
4. B. J. Lee and J. Y. Kim, "Identification of Type 2 Diabetes Risk Factors Using Phenotypes Consisting of Anthropometry and Triglycerides based on Machine Learning," in IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 1, pp. 39-46, Jan. 2016, doi: 10.1109/JBHI.2015.2396520.
5. T. M. Le, T. M. Vo, T. N. Pham, and S. V. T. Dao, "A Novel Wrapper–Based Feature Selection for Early Diabetes Prediction Enhanced With a Metaheuristic," in IEEE Access, vol. 9, pp. 7869-7884, 2021, doi: 10.1109/ACCESS.2020.3047942.
6. M. S. Islam, M. K. Qaraqe, S. B. Belhaouari, and M. A. Abdul-Ghani, "Advanced Techniques for Predicting the Future Progression of Type 2 Diabetes," in IEEE Access, vol. 8, pp. 120537-120547, 2020, doi: 10.1109/ACCESS.2020.3005540.
7. S. Perveen, M. Shahbaz, T. Saba, K. Keshavjee, A. Rehman and A. Guergachi, "Handling Irregularly Sampled Longitudinal Data and Prognostic Modeling of Diabetes Using Machine Learning Technique," in IEEE Access, vol. 8, pp. 21875-21885, 2020, doi: 10.1109/ACCESS.2020.2968608.
8. N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis and K. Moustakas, "Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction," in IEEE Access, vol. 9, pp. 103737-103757, 2021, doi: 10.1109/ACCESS.2021.3098691.
9. M. Shokrekhodaei, D. P. Cistola, R. C. Roberts and S. Quinones, "Non-Invasive Glucose Monitoring Using Optical Sensor and Machine Learning Techniques for Diabetes Applications," in IEEE Access, vol. 9, pp. 73029-73045, 2021, doi: 10.1109/ACCESS.2021.3079182.
10. U. Ahmed et al., "Prediction of Diabetes Empowered With Fused Machine Learning," in IEEE Access, vol. 10, pp. 8529-8538, 2022, doi: 10.1109/ACCESS.2022.3142097.
11. N. E. Costea, E. V. Moisi, and D. E. Popescu, "Comparison of Machine Learning Algorithms for Prediction of Diabetes," 2021 16th International Conference on Engineering of Modern Electric Systems (EMES), 2021, pp. 1-4, doi: 10.1109/EMES52337.2021.9484116.
12. A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," 2019 1st International Informatics and Software Engineering Conference (UBMYK), 2019, pp. 1-4, doi: 10.1109/UBMYK48245.2019.8965556.
13. B. Nithya and Dr. V. Ilango,” Predictive Analytics in Health Care Using Machine Learning Tools and Techniques”, International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.

AUTHORS PROFILE

Mrs. P Durga (Putta Durga) earned her undergraduate degree from Chadalawada Ramanamma Engineering College, Tirupati, JNTUA, and her Master’s degree in Computer Science and Engineering from SRK Institute of Technology, Vijayawada, JNTUK, Andhra Pradesh, India. She is having four years of experience in teaching and presently she is pursuing a Ph.D. degree in the School of Computer Science & Engineering (SCOPE), Vellore Institute of Technology (VIT–AP), Amaravathi, Andhra Pradesh, India. *Her current research interests* are in the *area* of Decision Making, Deep Learning, Digital Image Processing, and Machine Learning. She had authored one book chapter entitled “a state of the art review on machine learning algorithms for solving classification problems”

Dr.T.Sudhakar Received his B.E and M.E. degrees in computer science and engineering discipline in the years 2002 and 2005 respectively. He received his Ph.D. degree from the Faculty of Information and Communication Engineering, Anna University, Chennai in 2019. He is currently working as an Associate Professor, at the School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India. He has 17 years of teaching experience and 8 years of research experience. He has published a number of research articles in various reputed international journals and conferences. His research interests include Cryptography, Network security, Design of security protocols, Algorithms and Cyber Security.