Classification of Diabetes using Multilayer Perceptron

1st Sivasankari.S S
Department of ECE
HKBK College of Engineering
Bangalore, India
yes.sivasankari@gmail.com

2nd J. Surendiran
Department of ECE
HKBK College of Engineering
Bangalore, India
surenjaya1981@gmail.com

3rd N. Yuvaraj
R&D, ICT Academy
IIT-Chennai, Research Park
Chennai, India
yraj1989@gmail.com

4th M. Ramkumar
Department of CSE
HKBK College of Engineering
Bangalore, India
Ramkumar.cs@hkbk.edu.in

5th C.N. Ravi
Department of CSE
CMR Engineering College
Hyderabad, India
mail2cnravi@gmail.com

6th R.G Vidhya
Department of ECE
HKBK College of Engineering
Bangalore, India
vidhya50.ece@gmail.com

***Abstract*—The breakthroughs in public healthcare infrastructure have resulted in a large influx of highly sensitive and critical healthcare information. The application of sophisticated data analysis techniques can aid in the early detection and prevention of a variety of fatal diseases. Diabetes can cause heart disease, renal disease, and nerve damage, all of which are life-threatening complications of the disease. The goal of this work is to identify, detect, and forecast the emergence of diabetes in its earliest stages by employing machine learning techniques and algorithms. When it comes to diabetes classification, an MLP is used. The experimental evaluation was carried out using the PIMA Indian Diabetes dataset. According to the study findings, MLP outperforms the competition in terms of accuracy, with an accuracy rate of 86.08%. Following this, a comparison of the suggested technique with the existing state of the art is carried out, proving the flexibility of the proposed approach to a wide range of public healthcare applications.**

Keywords—Fatal Diseases, Diabetes Mellitus, Prediction and Classification.

I. INTRODUCTION

Diabetes is referred as diabetes mellitus (DM) and Hypoglycemia, often known as low blood sugar, in which the body is unable to synthesise glucose in the bloodstream. Type 1 diabetes is a prevalent type, while Type 2 diabetes is the second most common [1]. Type 1 is more likely to develop in children than in adults. The immune system in a people attacks the pancreas with type 1, causing it to generate antibodies that cause the organ to malfunction and cease generating insulin. It is possible to develop substantial issues and nervous system, despite the fact that it is less severe than type 1 diabetes [2].

Diabetes mellitus is a key topic of medical study because of the disease enormous social impact and the massive amount of data that is generated as a result of this impact [3]. Machine learning approaches in DM diagnosis, management, and other clinical administration elements are certainly of major concern. An ensemble strategy for diabetes classification utilising machine learning has been proposed in this article, which builds on a range of methodologies that have been explored previously [4]-[8].

Obstacles such as gestational diabetes mellitus and obesity occur during pregnancy and have a long-term influence on the health of both the mother and the child. In non-pregnant women, because of the relationship between the microvascular block risk and the level of glucose in the blood, several criteria for recognising diabetes have been developed. As a result of the improvement in living standards, diabetes has become increasingly prevalent. The only way to avoid its problems is to discover and diagnose them early [9].

There has been a significant amount of research on disease prediction, including diagnosis, prediction, classification, and treatment. Various machine learning (ML) [10] – [15] methods have been developed and applied to the identification and prediction of diseases throughout the last few years. The implementation of ML with conventional approaches has resulted in significant improvements. Machine learning (ML) has proved its ability to deal with enormous numbers of variables in an efficient and effective manner while simultaneously developing robust prediction models. Supervised machine learning methods are used to investigate the independent terms and variables of the dependent term.

Biomedical datasets are transformed into useable information through this process, which allows for top-tier clinical research while also enhancing patient care. As previously noted, there have been numerous improvements in machine learning approaches as a result of the requirement to classify diabetic patients. The Pima Indian diabetes dataset available to the public. There is a total of 769 data points, 500 of which are free of diabetes and 268 of which are positive for the presence of diabetes.

According to the research history, a variety of ML algorithms is used on this dataset for the aim of disease classification, none of which has reached an accuracy of more than 76%, according to the research history. As a result, we came up with the idea of improvising as a group rather than individually. The subject of this research is machine learning models, and it investigates their performance, theory, and attributes in greater depth.

The classification approach has been adopted by scientists in place of the regression strategy for making disease predictions. Its performance has been assessed using the accuracy, precision, recall, and F1-score of the aforementioned algorithm as measures of its effectiveness.

II. RELATED WORKS

Healthcare professionals rely on correct diagnosis in order to perform their duties effectively. Diagnosing a patient type of diabetes necessitates a variety of different tests, making it one of the most difficult duties faced by medical practitioners. Considering too many variables at the same time when diagnosing has the potential to produce erroneous conclusions. As a result, determining the kind of diabetes that someone has is incredibly difficult to accomplish. Recent years have seen a significant impact on the healthcare industry, particularly as a result of machine learning techniques. Numerous investigations have been conducted to determine the classification of diabetes.

Qawqzeh et al. [15] came up with a logistic regression for the classification of type 2 diabetes. It was built using data from 459 training, and 128 testing. When their proposed system was used to categorise 552 people as non-diabetic, they were able to attain an accuracy rate of 92%. The proposed method, on the other hand, has not been evaluated in comparison to existing methods.

Pethunachiyar [16] developed a technique for categorising patients with diabetes mellitus by employing a support vector machine (SVM). According to him, linear SVM is more efficient than any of the other models that have been discussed thus far. Despite this, there is no comparison of current state-of-the-art technology, and there are no specifics about parameter selection.

Diabetes was classified using and naive Bayes (NB) and SVMs classification. These conclusions were drawn from the PIMA Indian Diabetes Study, which used data from the PIMA Indian Diabetes Study. The current as well as the level of accuracy obtained, are notably absent from the document [17].

Choubey et al. [18] conducted a study in which they compared several diabetes classification techniques. The datasets utilised were a local diabetes dataset and PIMA Indian datasets. Feature engineering was carried out using PCA and LDA, both of which were shown to be beneficial in boosting the accuracy of the classification method and removing undesired features from the dataset.

A machine learning paradigm that they developed was used to identify and predict diabetes by Maniruzzaman et al. [19]. They employed four machine learning techniques for the classification of diabetes: naive Bayes, decision trees, AdaBoost, and random forests, among others. Additionally, they used three alternative partition techniques in addition to the 20 trials they conducted. The researchers used data from the National Health and Nutrition Survey (NHNS) for both diabetic and non-diabetic patients to put their innovative technique through its paces.

Ahuja et al. [20] conducted an examination of ML algorithms, including neural networks, deep learning, and multilayer perceptrons, on the PIMA dataset for diabetic classification. When compared to the data, MLP was determined to be superior to other classifiers. According to the authors, fine-tuning and efficient feature engineering can help to increase the performance of MLP algorithms. Recent research by Mohapatra et al. [21] has demonstrated the use of MLP to classify diabetes.

D.Saravanan [22] proposed an ensemble method for the prediction of type 2 diabetes in order to improve accuracy. The PIMA dataset from the University of California, Irvine Machine Learning Repository was used in this work. The bootstrap approach with cross-validation was used to train the four base learners of the stacking ensemble, which were used to train the stacking ensemble four base learners. However, neither variable selection nor a comparison of the current state of the art are mentioned, though.

Kumari et al. [23] created a soft computing-based diabetes prediction system based on an ensemble of three commonly used supervised machine learning algorithms. It was discovered that they had used PIMA datasets in their investigation. When they compared their system performance to that of state-of-the-art individual and ensemble approaches, they discovered that it outperformed them by 79%.

In order to forecast diabetes in its early or onset stage, Islam et al. [24] used a combination of techniques. Training methods included cross-validation and percentage splits, which were both employed in this study. They collected data from 529 Bangladeshi patients, both diabetic and non-diabetic, using questionnaires administered at a hospital in the nation. The experimental results reveal that the random forest algorithm outperforms them all by a significant margin. Although there is no comparison to the present state of the art, there is no clear reporting of the accuracy that was attained in this study.

The use of ML approaches to predict early and ongoing diabetes mellitus in females has been demonstrated in several studies. [25] They employed typical machine learning methods to construct a framework for predicting diabetes in order to better understand the disease.

Using machine learning models for diabetes prediction that were published between 2010 and 2019, Hussain and Naaz [26] did a comprehensive evaluation of the literature on this topic. They evaluated the algorithms based on the Matthews correlation coefficient and discovered that naive Bayes and random forests outperformed the other algorithms in terms of overall performance.

III. Proposed Method

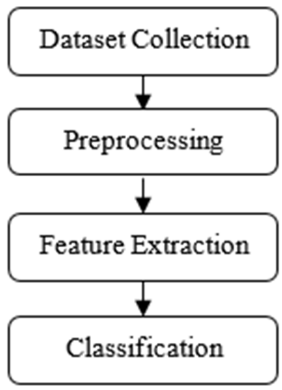

Fig. 1. Proposed architecture for diabetes classification

A variety of machine learning techniques have been applied to make decisions in the proposed diabetes classification system. For classification, we used and fine-tuned MLP for the first time, owing to its good performance in healthcare, notably in the prediction of diabetic complications and the illustration is given in Figure 1.

A. Classification

The study makes improvements to three commonly used cutting-edge strategies for identifying diabetes that were previously published. For the most part, the proposed ways of putting a person into one of the two categories of diabetes are compared and contrasted against one another. The following are the various diabetic treatment options that have been offered.

Multilayer Perceptron: Our experimental setup incorporates a well-tuned multilayer perceptron for diabetes classification that was developed by our team of researchers. As indicated in Figure 2, a classification method is represented as a network of numerous layers that are connected to one another. Perceptrons are the fundamental building blocks of our paradigm, and they are linear combinations of input and weights. When developing Algorithm 1, we employed a sigmoid unit as the activation function.

The suggested method is comprised of three major phases, as follows: The weights are initially established, then the output layer (δk) output is calculated by applying the sigmoid activation function to the weights and layer outputs. Second, the inaccuracy of all concealed units is determined at the hidden layer level (δh). Finally, in order to reduce network errors, all network weights (wi, j) are updated in a backwards manner. A step-by-step summary of the entire diabetes classification procedure is provided by Algorithm 1 for diabetes classification.

Algorithm 1 Classification using MLP

Input: Total patients, skin thickness, BP, age, BMI, HB weight,
Output: A Trained Model
Method: Initialize weights with a random number
while (Weights ≤ Threshold) **do**
  for Training samples, **do**
    Input samples is set as input to the network
    Compute the required output;
  end for
  for output *k*, **do**

$$\delta k = Ok(1 - Ok)(tk - Ok)$$

  end for
  for hidden layer *h* **do**

$$\delta_h = O_h(1\text{-}O_h)\ \Sigma_{k\in output} w_{h,k}\delta_k$$

  end for
  Update weight of each network $w_{i,j}$;

$$wi,j = wi,j + \Delta wi,j;\ \Delta wi,j = \eta\delta jxi,j$$

End

Because there are eight different variables in the multilayer perceptron classification model, eight neurons of the multilayer perceptron classification model. When computing the weights and inputs in the middle layer, the sigmoid unit will be employed as a calculator. At the end of the process, the output layer will calculate the results.

The hidden layers are used to compute the output data, starting with the values and weights of the input layer as a starting point for computation. Sigmoid activation functions are applied to each unit in the hidden layer in order to reduce the vast amount of data to a more manageable range between 0 and 1. This computation can be used by every middle layer. It is also treated with the same technique as the input layer, yielding results that can be used to create predictions regarding the development of diabetes.

IV. Results and Discussion

In order to put the proposed diabetes classification system to the test, the PIMA Indian Diabetes dataset is employed. A comparison study is also carried out, utilising the most up-to-date computational techniques, which are also included in the package. On the basis of the experimental results, the suggested method outperforms the currently available algorithms in terms of performance. Included in this section are sections devoted to defining the dataset, performance indicators, and conducting a comparison study.

A. Dataset

This study made use of the PIMA Indian Diabetes dataset. Create an intelligent model for predicting if a person has diabetes based on some of the metrics contained in this dataset using data from this dataset. When it comes to the classification of diabetes, it is a binary classification problem to be solved. The variables are shown in Table 1.

TABLE I. Description of variables in the dataset.

Attributes	Range	Mean	Std. deviation
Pregnancies	0–17	3.946	3.454
Glucose	0–199	124.025	32.800
Blood pressure	0–122	70.828	19.783
Skin thickness	0–99	21.013	16.298
Insulin	0–846	81.795	117.875
BMI	0–67	32.800	8.077
Diabetes Pedigree Function	0.078–2.4	0.482	0.338
Age	21–81	34.030	12.095
Outcome	Y/N	0.359	0.492

As shown in Figure 4, the dataset contains 768 records of female patients over the age of 21 who are either healthy or diabetic, as determined by their blood glucose levels. The distribution of feature values is depicted in Figure 5. For the outcome of the target variable, there are only two potential values: 0 and 1. This dataset was used in order to get an accurate diagnosis of type 2 diabetes. PIMA Indian determines whether or not a user is at risk of acquiring diabetes in the next four years based on their behaviour.

To compare the proposed classification system with existing methodologies, this research employs the same experimental design as the PIMA Indian dataset from the previous paper. The performance measures that were employed and the outcomes that were obtained for classifying or predicting are explained further below. It is also possible to make a comparison with previous investigations.

The performance is evaluated using three extensively used performance measures: recall, precision, and accuracy. Recall, precision, and accuracy are three frequently used, measures to evaluate the performance of the presented techniques. Individuals who are not diabetic are appropriately identified as such in TP, whereas diabetic patients are correctly identified as such in TN. The patient has diabetes, as revealed by the FN, but a favourable result is indicated. The patient FP also suggests that he or she is a healthy individual who has been classified as having diabetes, which is a positive sign. The classification models were trained and evaluated using a 10-fold cross-validation approach, which was developed by the authors.

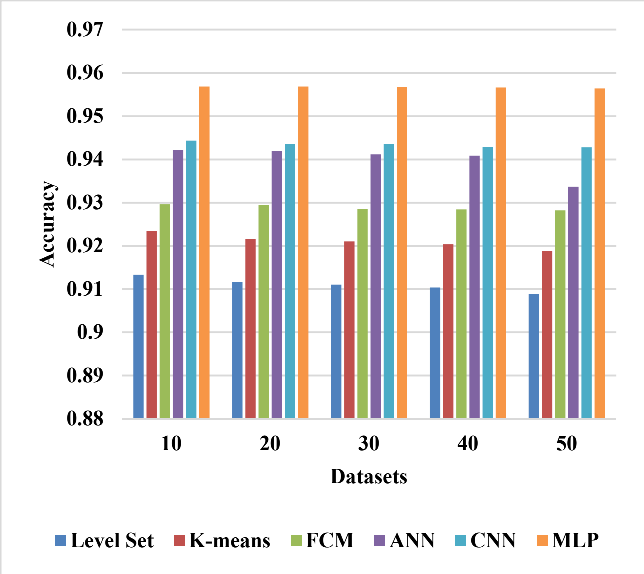

Fig. 2. Accuracy

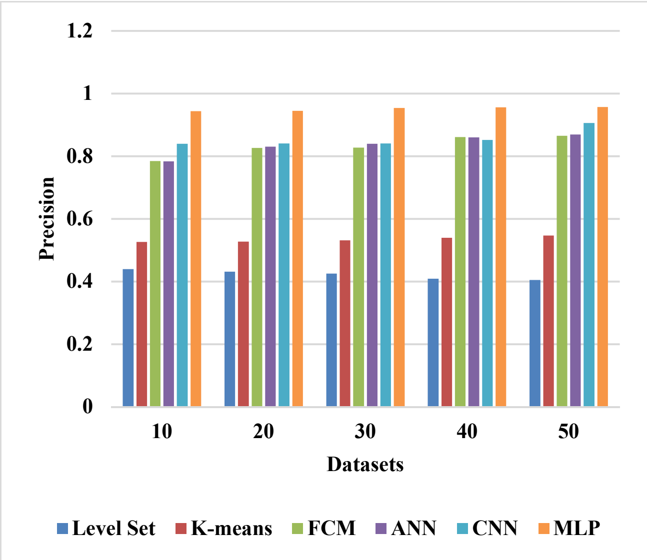

Fig. 3. Precision

The PIMA dataset is used to assess three state-of-the-art diabetic classification classifiers that have been developed recently. When comparing the accuracy of the fine-tuned MLP algorithm in Figure 2 to that of present systems, the highest accuracy of 86.083% can be observed. The precision rate has been increased than existing method which is shown in Figure 3. The comparison of recall rate with the existing method is shown in the figure 4.

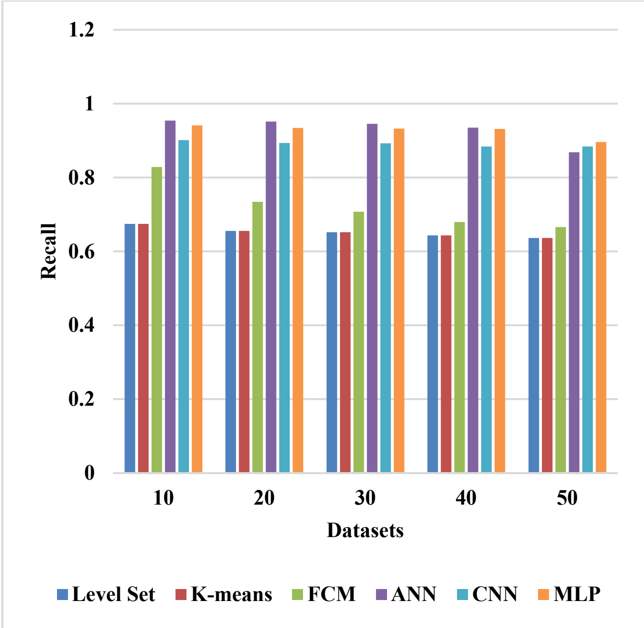

Fig. 4. Recall

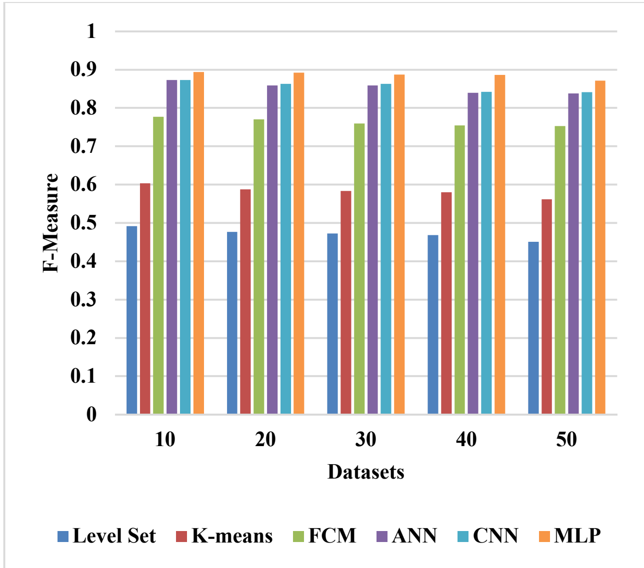

Fig. 5. F-Measure

In light of the findings, the calibrated MLP model might be utilised to accurately classify diabetes, as demonstrated. We believe that the classification approach we have described here will be beneficial to our hypothetical system in the future.

According to Figure 5, a proposed MLP algorithm outperforms with an accuracy of 86.083% when compared to the current state of the art (86.6% precision and 85.1% recall). These discoveries will have a significant impact on the planned hypothetical method for evaluating whether a patient has type 1 or type 2 diabetes.

V. Conclusions

In this paper, the authors proposed a model for supporting the healthcare business. The study developed an algorithm for the classification of diabetes that was based on MLPs. The primary purpose of the proposed system is to aid users in keeping track of their vital signs through the use of their mobile phones and other mobile devices. Users will be able to recognize their elevated risk of diabetes at an earlier stage as a result of the model projections about future blood glucose levels, which is an extra benefit. Diabetic patients are classified and predicted using MLP. The proposed methodologies are tested on the PIMA Indian Diabetes dataset, which is available online. In terms of accuracy, the two approaches outperform existing best practices by 86.083% and 87.26%, respectively, when compared to current best practices.

REFERENCES

[1] Miranda, M. A., Macias-Velasco, J. F., & Lawson, H. A. (2021). Pancreatic β-cell heterogeneity in health and diabetes: classes, sources, and subtypes. *American Journal of Physiology-Endocrinology and Metabolism*, *320*(4), E716-E731.

[2] Babalola, O. M., Garcia, T. J., Sefcik, E. F., & Peck, J. L. (2021). Improving Diabetes Education in Mexican American Older Adults. *Journal of Transcultural Nursing*, *32*(6), 799-809.

[3] Zaccardi, F., Kloecker, D. E., Buse, J. B., Mathieu, C., Khunti, K., & Davies, M. J. (2021). Use of metformin and cardiovascular effects of new classes of glucose-lowering agents: a meta-analysis of cardiovascular outcome trials in type 2 diabetes. *Diabetes Care*, *44*(2), e32-e34.

[4] Kannan, S., Idrees, B. A., Arulprakash, P., Ranganathan, V., Udayakumar, E., & Dhinakar, P. (2021). Analysis of convolutional recurrent neural network classifier for COVID-19 symptoms over computerised tomography images. International Journal of Computer Applications in Technology, 66(3-4), 427-432.

[5] Ojugo, A. A., & Ekurume, E. (2021). Predictive Intelligent Decision Support Model in Forecasting of the Diabetes Pandemic Using a Reinforcement Deep Learning Approach. *Int. J. Educ. Manag. Eng*, *11*(2), 40-48.

[6] Mohana, J., Yakkala, B., Vimalnath, S., Benson Mansingh, P. M., Srihari, K., ... & Sundramurthy, V. P. (2022). Application of Internet of Things on the Healthcare Field Using Convolutional Neural Network Processing. Journal of Healthcare Engineering, 2022.

[7] Khunti, K., Knighton, P., Zaccardi, F., Bakhai, C., Barron, E., Holman, N., ... & Valabhji, J. (2021). Prescription of glucose-lowering therapies and risk of COVID-19 mortality in people with type 2 diabetes: a nationwide observational study in England. *The Lancet Diabetes & Endocrinology*, *9*(5), 293-303.

[8] Syed, S. A., Sheela Sobana Rani, K., Mohammad, G. B., Chennam, K. K., Jaikumar, R., Natarajan, Y., ... & Sundramurthy, V. P. (2022). Design of Resources Allocation in 6G Cybertwin Technology Using the Fuzzy Neuro Model in Healthcare Systems. Journal of Healthcare Engineering, 2022.

[9] Greiver, M., Havard, A., Bowles, J. K., Kalia, S., Chen, T., Aliarzadeh, B., ... & de Lusignan, S. (2021). Trends in diabetes medication use in Australia, Canada, England, and Scotland: a repeated cross-sectional analysis in primary care. *British Journal of General Practice*, *71*(704), e209-e218.

[10] Hannah, S., Deepa, A. J., Chooralil, V. S., BrillySangeetha, S., Arshath Raja, R., ... & Alene, A. (2022). Blockchain-Based Deep Learning to Process IoT Data Acquisition in Cognitive Data. BioMed Research International, 2022.

[11] Ferrari, M., Speight, J., Beath, A., Browne, J. L., & Mosely, K. (2021). The information-motivation-behavioral skills model explains physical activity levels for adults with type 2 diabetes across all weight classes. *Psychology, Health & Medicine*, *26*(3), 381-394.

[12] Shah, N., Karguppikar, M., Bhor, S., Ladkat, D., Khadilkar, V., & Khadilkar, A. (2021). Impact of lockdown for COVID-19 pandemic in Indian children and youth with type 1 diabetes from different socio-economic classes. *Journal of Pediatric Endocrinology and Metabolism*, *34*(2), 217-223.

[13] Mansi, I. A., Chansard, M., Lingvay, I., Zhang, S., Halm, E. A., & Alvarez, C. A. (2021). Association of statin therapy initiation with diabetes progression: a retrospective matched-cohort study. *JAMA internal medicine*, *181*(12), 1562-1574.

[14] Bouyahya, A., El Omari, N., Elmenyiy, N., Guaouguaou, F. E., Balahbib, A., Belmehdi, O., ... & Bakri, Y. (2021). Moroccan antidiabetic medicinal plants: Ethnobotanical studies, phytochemical bioactive compounds, preclinical investigations, toxicological validations and clinical evidences; challenges, guidance and perspectives for future management of diabetes worldwide. *Trends in Food Science & Technology*, *115*, 147-254.

[15] Qawqzeh, Y. K., Bajahzar, A. S., Jemmali, M., Otoom, M. M., & Thaljaoui, A. (2020). Classification of diabetes using photoplethysmogram (PPG) waveform analysis: logistic regression modeling. *BioMed Research International*, *2020*.

[16] Pethunachiyar, G. A. (2020, January). Classification of diabetes patients using kernel based support vector machines. In *2020 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-4). IEEE.

[17] Gupta, S., Verma, H. K., & Bhardwaj, D. (2021). Classification of diabetes using Naive Bayes and support vector machine as a technique. In *Operations Management and Systems Engineering* (pp. 365-376). Springer, Singapore.

[18] Choubey, D. K., Kumar, M., Shukla, V., Tripathi, S., & Dhandhania, V. K. (2020). Comparative analysis of classification methods with PCA and LDA for diabetes. *Current diabetes reviews*, *16*(8), 833-850.

[19] Maniruzzaman, M., Rahman, M., Ahammed, B., & Abedin, M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health information science and systems*, *8*(1), 1-14.

[20] Ahuja, R., Sharma, S. C., & Ali, M. (2019). A diabetic disease prediction model based on classification algorithms. *Annals of Emerging Technologies in Computing (AETiC), Print ISSN*, 2516-0281.

[21] Mohapatra, S. K., Swain, J. K., & Mohanty, M. N. (2019). Detection of diabetes using multilayer perceptron. In *International conference on intelligent computing and applications* (pp. 109-116). Springer, Singapore.

[22] D.Saravanan, J.Surendiran,"A new framework for video data retrieval using hierarchical clustering technique", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-6S3, September 2019.

[23] Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *2*, 40-46.

[24] Islam, M. M., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis* (pp. 113-125). Springer, Singapore.

[25] Malik, S., Harous, S., & El-Sayed, H. (2020, October). Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women. In *International Symposium on Modelling and Implementation of Complex Systems* (pp. 95-106). Springer, Cham.

[26] Hussain, A., & Naaz, S. (2021). Prediction of diabetes mellitus: comparative study of various machine learning models. In *International Conference on Innovative Computing and Communications* (pp. 103-115). Springer, Singapore.