International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Big Data Predictive Analysis for Type-2 Diabetes Based Heart Disease Using Feature Extraction and Classification by Machine Learning Architectures

Arvind Kumar Pandey*[1], Shreyanth S[2], Dr.J.Prabhakaran[3], Aniruddha Bodhankar[4] Avadhesh Kumar[5], Nayani Sateesh [6]

Submitted: 25/09/2022 **Accepted**: 23/12/2022

Abstract*:* Machine learning (ML), a branch of AI, enables computers to learn without being explicitly programmed. ML is widely applied in the healthcare industry to forecast a variety of chronic conditions. For improved clinical paths to prevent complications and postpone the onset of diabetes, earlier diabetes prediction is essential. This research propose novel technique in type 2 diabetes based heart disease detection in big data predictive analysis using machine learning method. Input data has been collected as type 2 diabetes and processed for noise removal and dimensionality reduction. Then the processed data features has been extracted for detecting the abnormality of type 2 diabetes using regression model based linear discriminant analysis. The extracted features shows the abnormal type 2 diabetes and for predicting heart disease by classifying the extracted data using VGG-16 Net_gradient NN. Experimental analysis has been carried out in terms of accuracy, precision, recall, F-1 score, RMSE and MAP for various diabetes dataset. Proposed technique attained accuracy of 96%, precision of 67%, recall of 79%, F-1 score of 63%, RMSE of 66% and MAP of 68%.

Keywords: *Machine Learning, big data, predictive analysis, type 2 diabetes, dimensionality reduction*

1. Introduction

Diabetes is a disease that develops when your blood glucose level is higher. The primary and first source of blood glucose is provided by the food you eat. In particular, a substance produced by the digestive system allows different sugars from food to enter your cells and be used as an energy source [1]. Sometimes a corpse doesn't use offence well, doesn't cause enough uproar, or both. Sometimes people refer to this condition as "borderline diabetes" or "a touch of sugar." These circumstances suggest that someone may not actually have diabetes or may have a milder disease, although each instance of the ailment is dangerous. Type 1, type 2, and developing diabetes are three most common types of diseases. At that time, the sugar energy supply is still in your blood and has not reached your basic unit of a living thing. HD refers to a wide range of illnesses that affect your heart. Term "heart disease" refers to a variety of infections, including blood vessel diseases including coronary artery infection, arrhythmias, and inherited heart defects (innate heart surrenders) [2]. Machine learning can be crucial in predicting the presence or absence of diabetes, heart infections, locomotor disorders, and more. Such information, if predicted well in advance, can provide professionals with critical information that will allow them to modify their therapy and decision-making for each patient. Diabetes is the most prevalent condition that can lead to death. In 2012, heart disease, kidney problems, and other causes of death contributed to an additional 2.2 million deaths, which brought the total number of deaths from diabetes-related causes to about 1.5 million. As of 2017, there were 8.8% of persons globally who had diabetes. By 2045, it is anticipated to reach 10%[3]. About 77 million Indians have high blood sugar, placing the country in second place for the number of diabetic patients worldwide (Saeedi et al., 2019). According to the National Diabetes Statistics Report 2020, 34.2 million Americans are thought to have high blood sugar. Only 26.9 million people in the population have diabetes, while another 7.3 million are undiagnosed with the condition (US Department of Health and Human Services, 2020). A doctor can identify diabetes manually or automatically

[1] *Assistant Professor, Department of Computer Science, Arka Jain University, Jamshedpur, Jharkhand, India. Email Id: arvind.p@arkajainuniversity.ac.in , Orchid Id- 0000-0001-5294-0190*
[2] *MTech in Data Science and Engineering, Birla Institute of Technology and Science, Pilani, Rajasthan, India. Email id: shreyanth0810@gmail.com Orchid Id: 0000-0002-9991-5491*
[3] *Associate Professor, Department of Kalasalingam Business School, Kalasalingam Academy of Research and Education (Deemed to be University) Email id: praba.psg@gmail.com, Orchid id: 0000-0001-9415-5203*
[4] *Assistant Professor, Department of Management, Dr. Ambedkar Institute of Management Studies & Research, Nagpur, Email Id: aniruddha.bodhankar16@gmail.com, Orchid Id: 0000-0001-8721-5714*
[5] *Professor, Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India Email Id: avadheshkumar@galgotiasuniversity.edu.in, ORCID id: 0000-0002-9469-9611*
[6] *Sr. Assistant Professor, Information Technology Department, CVR College of Engineering, Ibrahimpatnam, Hyderabad, India. Email id: n.sateesh@cvr.ac.in, Orchid Id: 0000-0001-8216-1809*

with the aid of a device. Both diagnostic methods have advantages and disadvantages. The biggest advantage of performing diagnoses manually is that an automatic instrument is not required [4].

Contribution of this research is as follows:

1. To propose novel method in type 2 diabetes based heart disease detection in big data predictive analysis using machine learning technique
2. the processed data features has been extracted for detecting the abnormality of type 2 diabetes using regression model based linear discriminant analysis.
3. The extracted features shows the abnormal type 2 diabetes and for predicting heart disease by classifying the extracted data using VGG-16 Net_gradient neural network.

2. Literature Review

Heart and diabetes issues have historically been the two most common causes of death worldwide. Furthermore, it is a problem that requires a solution today to predict the same thing or simply to suggest a slight chance of it. Machine learning has paved its role in the medical industry by assisting in decision-making and forecasting by training over vast amounts of data that already exist in the form of datasets. According to the study in [5], cardiovascular disorders are substantially correlated with the severity and mortality of COVID-19, but diabetes mellitus and hypertension are only weakly related. According to [6], the research focuses on the application of neural network technology to analyse and demonstrate more accurate diabetes prediction. Early decision is possible using a reasonably sound computational method, as stated by [7]. In this study, diabetes is predicted early on using machine learning techniques. This section discusses a few works that are very closely connected. For the purpose of predicting diabetes, many research papers have made use of the Pima Indians Diabetes Dataset (PIDD). Weka tool and ML techniques were used by [8]. Researcher methods can be generically categorised into four categories: ML, data mining, hybrid methods, and genetic or NN methods. Deep learning techniques were applied to electrocardiogram (ECG) information in work [9] to detect diabetes. They specifically utilized CNN and LSTM, and subsequently support vector machines were used to extract features. They discovered a very high accuracy of 95.7% as a result. In order to forecast diabetes, author [10] applied 3 ML approaches to PIDD: decision tree (DT), naive based (NB), and SVM. The accuracy of Naive Bayes classifier was determined to be 76.30%. Work [11] used updated kNN and logistic regression data mining approaches to reliably forecast up to 95.42% of a person's probability of getting type 2 diabetes. The adjustment was carried out by empirically choosing the first seed point's value. In order to forecast diabetes, author [10] applied 3 ML approaches to PIDD: decision tree (DT), naive based (NB), and SVM. The accuracy of Naive Bayes classifier was determined to be 76.30%. Work [11] used updated kNN and logistic regression data mining approaches to reliably forecast up to 95.42% of a person's probability of getting type 2 diabetes. The adjustment was carried out by empirically choosing first seed point's value. By running 100 runs and choosing smallest value of "inside cluster sum of squared errors," initial seed point was determined [12]. Among the three methods, DT was discovered to deliver the best results. Work [13] used a hybrid method that first applied genetic algorithm (GA) for feature selection and then RBFNN for classification. They discovered that the hybrid approach outperformed RBFNN on its own [14].

3. System Model

This section discuss novel method in type 2 diabetes based heart disease detection in big data predictive analysis using machine learning method. Input data has been collected as type 2 diabetes and processed for noise removal and dimensionality reduction. Then processed data features has been extracted for detecting the abnormality of type 2 diabetes using regression model based linear discriminant analysis. The extracted features shows the abnormal type 2 diabetes and for predicting heart disease by classifying the extracted data using VGG-16 Net_gradient neural network. Proposed architecture is shown in figure-1.

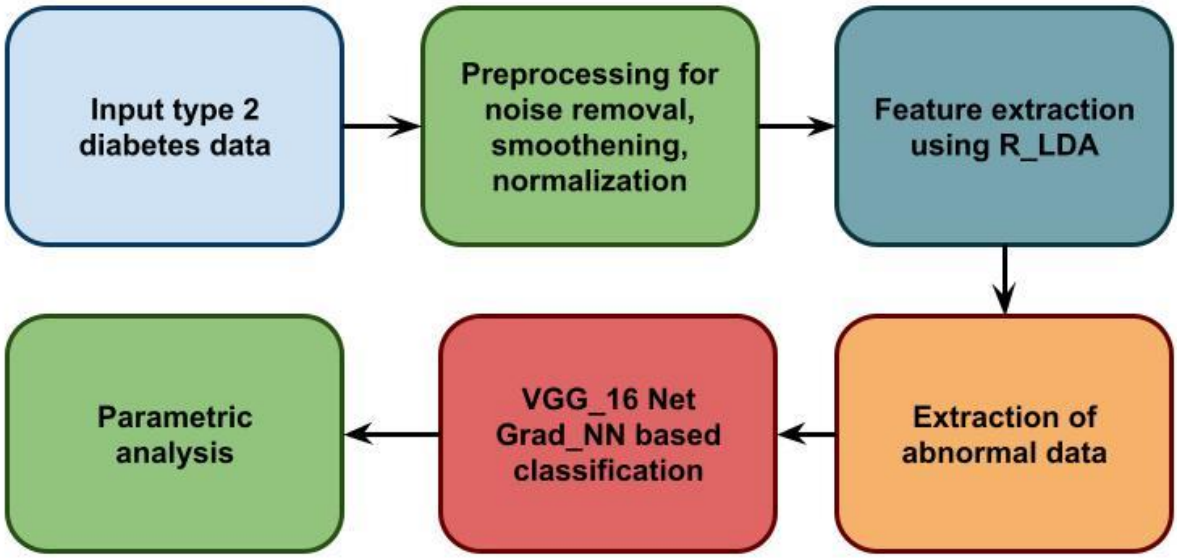

Fig. 1: Overall proposed architecture

Data pre-processing is a crucial step to take before developing ML models in order to get better outcomes. The obtained dataset was preprocessed using several statistical libraries, an Integrated Development Environment called Spyder, and the Python (3.9.1) programming language. Techniques like resampling and discretization were used. Interquartile range approach was used to replace outlier with viable sampling values after the outlier was detected using a boxplot. Before creating the machine learning models, data transformation was done to make the data more effective. Additionally, the dataset has been cleaned of duplicate, inconsistent, and corrupted data utilising a variety of data exploration and analysis approaches.

Regression Model Based Linear Discriminant Analysis

Assume that the data [Y1,Y2,..., Yn] are independent and that Y is a binary response variable with Yi = 1 for the presence of the character and Yi = 0 for the absence of the character. Let I represent the success probability. Additionally, think of the collection of explanatory variables x = (x1, x2,..., xp) as being either discrete, continuous, or a combination of both. The logistic function for I is then provided by equation (1). (2)

$$\text{logit}\,(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{i,pi} \quad (1)$$

where

$$\pi_i = \frac{\exp\,(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{i,p})}{1+\exp\,(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{i,p})} = \frac{\exp\,(x_i'\beta)}{1+\exp\,(x_i'\beta)} = \Lambda(x_i'\beta) \quad (2)$$

Here, I represents the likelihood that a sample falls into a certain category of the dichotomous answer variable, sometimes known as the "success probability," and it is obvious that $0 \le \pi i \le 1$. The logistic cdf is represented by $\Lambda(.)$ where $\lambda(z) = e^2/(1+e^{-z}) = 1/(1+e^{-z})$ and β^s is a vector of parameters that need to be estimated. The odds ratio or relative risk is the expression $\left(\frac{\pi_i}{1-\pi_i}\right)$. Think about the logistic model that uses the logistic function of eq. (3) as the sole predictor variable, X.

$$\pi(X) = \frac{\exp\,(X_i\beta)}{1+\exp\,(X_i\beta)} \quad (3)$$

Finding estimates that yield a value near to one for all subjects π(X) who have diabetes and a figure close to zero for everyone else is our goal. The likelihood function is defined mathematically by eq (4)

$$L(\beta_0,\beta_1) = \prod_{i:y_l=1}\pi(x_i)\prod_{r':y_\ell=0}\left(1-\pi\left(x_{p'}\right)\right) \quad (4)$$

In order to optimise this likelihood function, the estimations are selected. In order to construct and apply the log-likelihood function for estimate, we take the logarithm on both sides. To determine if any subset of estimates β is zero, we employed the likelihood ratio. Assume that p and r stand for the complete model's and the reduced model's respective numbers of βˆ eq(5) outputs the likelihood ratio test statistic

$$\Lambda^* = -2\left[l\left(\hat{\beta}^{(0)}\right) - l(\hat{\beta})\right] \quad (5)$$

Therefore, because inner product captures similarity between two vectors, provides a measure of that similarity. We may create a kernel matrix using equation (6) by computing kernel of two matrices, $\boldsymbol{X_1} \in \mathbb{R}^{d\times n_1}$and $\boldsymbol{X_2} \in \mathbb{R}^{d\times n_2}$

$$\mathbb{R}^{n_1\times n_2} \ni \boldsymbol{K}(\boldsymbol{X_1},\boldsymbol{X_2}) := \boldsymbol{\Phi}(\boldsymbol{X_1})^{\top}\boldsymbol{\Phi}(\boldsymbol{X_2}) \quad (6)$$

$$\mathbb{R}^{n\times n} \ni \boldsymbol{K}_x := \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) = \boldsymbol{\Phi}(\boldsymbol{X})^{\top}\boldsymbol{\Phi}(\boldsymbol{X}) \quad (7)$$

where Φ(X) := [φ(x1), . . . , φ(xn)] ∈ R t×n is pulled data. d_B in feature space is given by eq. (8):

$$\mathbb{R} \ni d_B := \phi(u)^{\top}\Phi(S_B)\phi(u)$$
$$\overset{(a)}{=} \boldsymbol{\theta}^{\top}\boldsymbol{\Phi}(\boldsymbol{X})^{\top}\big(\phi(\boldsymbol{\mu}_1) - \phi(\boldsymbol{\mu}_2)\big) \quad (8)$$
$$\big(\phi(\boldsymbol{\mu}_1) - \phi(\boldsymbol{\mu}_2)\big)^{\top}\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{\theta},$$

For j-th elass (here $j \in \{1,2\}$), we have by eq. (9):

$$\boldsymbol{\theta}^{\top}\boldsymbol{\Phi}(\boldsymbol{X})^{\top}\phi(\boldsymbol{\mu}_j) \overset{(95)}{=} \sum_{i=1}^{n}\theta_i\phi(\boldsymbol{x}_i)^{\top}\phi(\boldsymbol{\mu}_j)\left(\phi\big(\boldsymbol{x}_i^{(j)}\big) - \phi(\boldsymbol{\mu}_j)\right)^{\top}\Big)$$
$$\overset{(98)}{=} \frac{1}{n_j}\sum_{i=1}^{n}\sum_{k=1}^{n_j}\theta_i\phi(x_i)^{\top}\phi\big(x_k^{(j)}\big)$$
$$\overset{(86)}{=} \sum_{i=1}^{n}\sum_{k=1}^{n_j}\theta_i k\big(x_i, x_k^{(j)}\big) = \theta^{\top}m_j \quad (9)$$

dW in feature space is by eq. (10), (11)

$$\mathbb{R} \ni d_W :=$$
$$\phi(\boldsymbol{u})^{\top}\Phi(\boldsymbol{S}_W)\phi(\boldsymbol{u}) \overset{(\alpha)}{=} \big(\sum_{\ell=1}^{n}\theta_\ell\phi(\boldsymbol{x}_\ell)^{\top}\big)\Big(\sum_{j=1}^{c}\sum_{i=1}^{n_j}\Big(\phi\big(x_i^{(j)}\big) - \phi\big(\boldsymbol{\mu}_j\big)\Big)\Big(\sum_{k=1}^{n}\theta_k\phi(\boldsymbol{x}_k)\Big) =$$
$$\sum_{j=1}^{c}\sum_{\ell=1}^{n}\sum_{i=1}^{n_j}\sum_{k=1}^{n}\big(\theta_l\phi(x_\ell)^{\top}\big(\phi\big(x_i^{(j)}\big) - \phi(p) \quad (10)$$
$$\sum_{j=1}^{c}\sum_{i=1}^{n}\sum_{i=1}^{c}\sum_{i=1}^{n}$$

$$\left(\theta_k k\left(x_i, x_i^{()\prime}\right) - \frac{1}{n}\sum_{i=1}^{n}\theta_i k\big(\boldsymbol{x}_i, x_i^{(j)}\big)\right)$$
$$\left(\theta_k k\big(x_k, x_i^{(i)}\big) - \frac{1}{n_j}\sum_{x=1}^{n}\theta_k k\big(x_k, x_i^{(j)}\big)\right)$$
$$= \sum_{j=1}^{c}\sum_{i=1}^{n}\sum_{i=1}^{n}\sum_{i=1}^{n}$$
$$= \sum_{j=1}^{c}\Big(\sum_{i=1}^{n}\sum_{i=1}^{n_j}\sum_{k=1}^{n}\Big(\theta_t\theta_k k\big(x_t, x_i^{(j)}\big)k\big(x_k, x_i^{(j)}\big)\Big)$$
$$+\frac{\theta_i\theta_k}{n_j^{\top}}\sum_{k=1}^{n=1}\sum_{i=1}^{nj\left(x_t, x_i^{(j)}\right)k\left(x_k, x_i^{(j)}\right)}\Big) \overset{(e)}{=} \sum_{j=1}^{c}\Big(\theta^{\top}K_jK_j^{\top}\theta -$$
$$\theta^{\top}K_j\frac{1}{n_j}11^{\top}K_j^{\top}\theta\Big) \quad (11)$$

VGG-16 Net_gradient neural network:

The pooling layers in VGG-16 are all 2×2 pooling layers with a stride size of 2, while convolutional layers are all 3×3 convolutional layers with a stride size of 1 with same padding. The VGG-16 input image size is 224×224 by default. The size of the feature map is halved after each pooling layer. The 7×7 with 512 channels feature map, which is enlarged into a vector with 25,088 (7×7×512) channels, is final feature map before completely connected layers. In order to assure correctness of model feature extraction as well as to realise model's lightweight and accelerate training, we will merge the original VGG-16 with the full convolution model and minimise the model's parameters as well as the number of layers of the complete connection layer. Our model combines the conventional full CNN model with the original VGG-16 model.

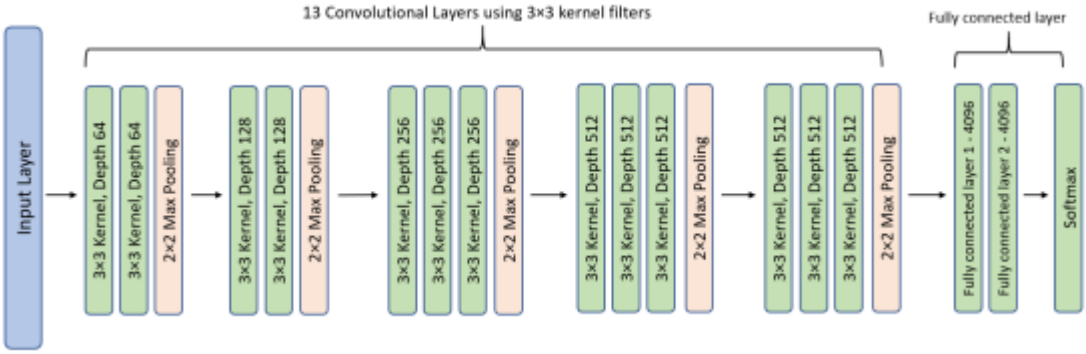

Fig. 2. VGG-16 model architecture – 13 convolutional layers and 2 Fully connected layers and 1 SoftMax classifier

VGG-16 - In their 2014 publication, "Very Deep Convolutional Network for Large Scale Image Recognition," Karen Simonyan and Andrew Zisserman introduced VGG-16 architecture.

$$R^{m,j} = \Delta\mathbf{u}^{m/+j} - \Delta_j\mathbf{u}^{mJ}$$
$$r_i^{m,j} = \Delta_j v_i^{m/j+j} - \Delta_j v_i^{mj}$$
$$d^{m,I} = \mathbf{u}^{\mathrm{m}/+1} - \mathbf{u}^{\mathrm{m}|} = \sum_{j=1}^{l}\Delta\mathbf{u}^{\mathrm{m}/+j}$$
$$= \sum_{j=1}^{t}\Delta j\mathbf{u}^{mj} + \sum_{j=1}^{l}R^{m,j}.$$
$$h_i^{m,I} = \mathbf{v}_i^{mj+1} - \mathbf{v}_i^{\mathrm{m}j} = \sum_{j=1}^{l}\Delta_j\mathbf{v}_i^{m/+j} = \sum_{j=1}^{l}\Delta_j\mathbf{v}_i^{\mathrm{m}J} +$$
$$\sum_{j=1}^{l}r_i^{m,j}. \quad (12)$$
$$\psi^{m,}, j = G^{\mathrm{mith}\ +Lj} - G^{mj,j},$$
$$m \in \mathbb{N}, j = 1,2,\ldots J, l = 1,2,\ldots J, i = 1,2,\ldots,n.$$

Moreover, we observe that by eq. (13)

$$\|\psi^{m,t,j}\| = \|G^{mJ+l,j} - G^{mJ,j}\| \leq$$
$$\max_{1\leq i\leq n}|g'(t_i)|\|\mathbf{x}^j\|\sum_{i=1}^{n}\|h_i^{m,l}\| \quad (13)$$
$$\leq \max_{1\leq i\leq n}|g'(t_i)|\|\mathbf{x}^j\|\sum_{i=1}^{n}\sum_{k=1}^{l}\|\Delta_k\mathbf{v}_i^{mj+k}\|$$
$$\leq C_5\eta_m,$$

Combining $f_j'(t)$'s Lipschitz continuity, (14), we have by eq. (15)

$$\left|f_j'(\mathbf{u}^{mJ+j}\cdot G^{mJ+j,j}) - f_j'(\mathbf{u}^{mJ}\cdot G^{mJ+j,j})\right|$$
$$\leq L\left|\mathbf{u}^{mJ+j}\cdot G^{mJ+j,j} - \mathbf{u}^{mJ}\cdot G^{mJ+j,j}\right| \quad (14)$$
$$\leq L\|d^{m,j}\|\|G^{mj+j,j}\| \leq LC_3\|d^{m,j}\|$$
$$\left|f_j'(\mathbf{u}^{mJ}\cdot G^{mj+j,j}) - f_j'(\mathbf{u}^{mj}\cdot G^{mj,j})\right|$$

$$\leq L\left|\mathbf{u}^{mJ} \cdot G^{mJ+j,j} - \mathbf{u}^{mJ} \cdot G^{mJ,j}\right| \quad (15)$$
$$\leq L\|\mathbf{u}^{mJ}\|\|\psi^{m,j,j}\| \leq LC_2\|\psi^{m,j,j}\|$$

where L > 0 is Lipschitz constant. By definition of R m, j , we see that by eq. (16)

$$R^{m,j} = \Delta_j \mathbf{u}^{mj+j} - \Delta_j \mathbf{u}^{mj}$$
$$= -\eta_m\left(f_j'(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,j})G^{mJ+j.j} - f_j'(\mathbf{u}^{mJ} \cdot G^{mJ\cdot j})G^{mJ.j}\right)$$
$$= -\eta_m\left[f_j'(\mathbf{u}^{mj+j} \cdot G^{mj+j,j})\psi^{m,j,j} + \left(f_j'(\mathbf{u}^{mJ+j} \cdot G^{mJ+j,j}) - f_j'(\mathbf{u}^{mJ} \cdot G^{mJ+j,j})\right)G^{mJ,j}\right. \quad (16)$$
$$\left. + \left(f_j'(\mathbf{u}^{mj} \cdot G^{mj+j,j}) - f_j'(\mathbf{u}^{mj} \cdot G^{mj,j})\right)G^{mj,j}\right].$$

4. Experimental Analysis

The Google Colab environment and important Python libraries are used to implement the suggested method.

Data Description: The total number of participants in this study is 952, including 372 women and 580 men who are 18 years of age or older. Participants were given a self-made questionnaire based on potential diabetes-causing factors, which is shown in Table 1. To verify the model's validity, the same tests were run on a different database called the PIMA Indian Diabetes database.

Table-1 Comparative analysis between proposed and existing technique based on various type-2 diabetes dataset

Parameters	**PIDD**	**SVM**	**BDPA_HD_MLA**
Accuracy	89	93	96
Precision	63	66	67
Recall	71	76	79
F1_Score	58	61	63
RMSE	63	63	66
MAP	65	67	68

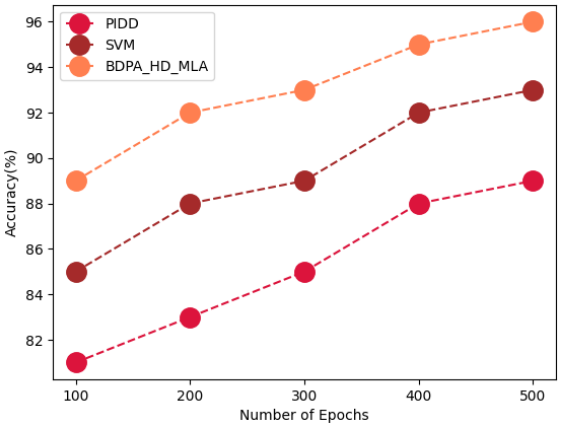

Fig.-2 Comparison of accuracy

The comparison between the proposed and existing techniques in terms of accuracy is shown in figure 2 above. The following is the official definition of accuracy: The total number of precise predictions matches the entire number of precise predictions. Comparison has been carried out based on number of users and proposed technique has attained accuracy of 96%, existing PIDD attained 89% and SVM attained 93%.

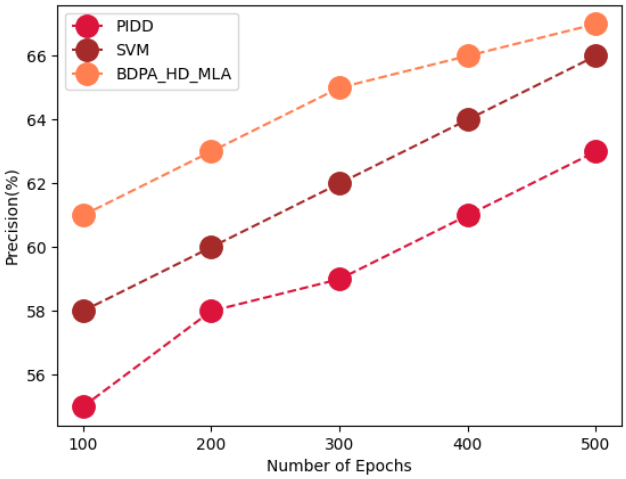

Fig.-3 Comparison of precision

The comparison of precision between proposed and existing techniques depending on number of epochs is shown in figure 3 above. Precision, or the calibre of an accurate prediction, is one measure of system performance. To calculate precision, divide the total number of real positive forecasts by the total number of accurate positive predictions. Proposed technique attained precision of 67%, existing PIDD attained 63% and SVM attained 66%.

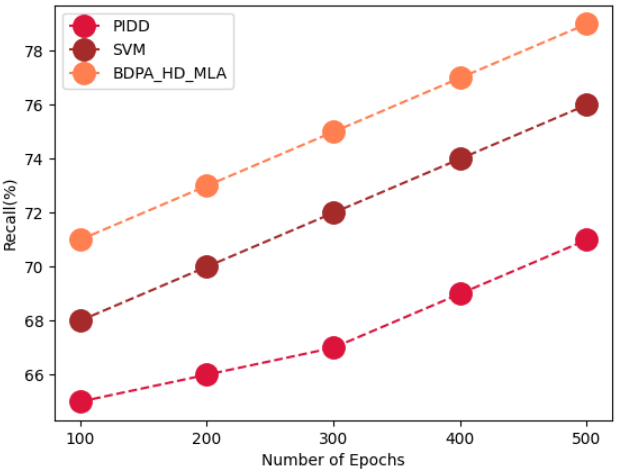

Fig.-4 Comparison of recall

In figure 4 above, recall for the proposed and existing strategies is contrasted based on the quantity of users. Recall is measured as the proportion of Positive samples that were correctly identified as Positive relative to all Positive samples. Proposed technique attained recall of 79%, existing PIDD attained 71% and SVM attained 76%.

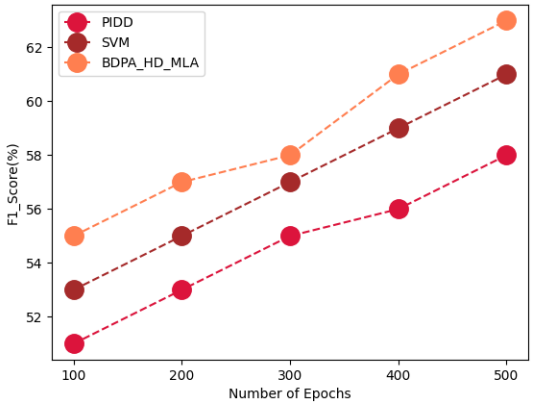

Fig.-5 Comparison of F-1 score

Comparison of F-1 score between proposed and existing techniques is shown in figure 5 above. F1 Score is evaluated as weighted average of Precision and Recall. As a result, when determining this score, both FP and FN are taken into account. Particularly if you have an uneven class distribution. Proposed technique attained F-1 score of 63%, existing PIDD attained 59% and SVM attained 61%.

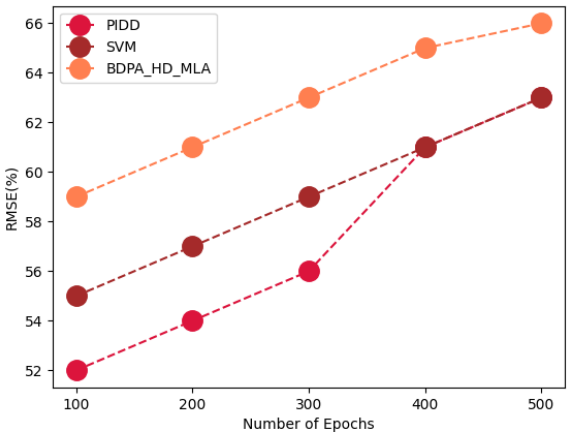

Fig.-6 Comparison of RMSE

From above figure-6 the comparison of RMSE between proposed and existing technique. It demonstrates the Euclidean separation between forecasts and observed true values. Proposed technique attained RMSE of 66%, existing PIDD attained 63% and SVM attained 63%.

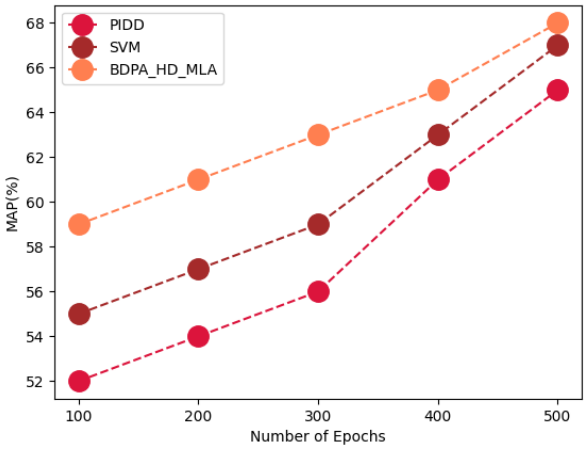

Fig.-7 Comparison of MAP

From above figure-7 the comparison of MAP between proposed and existing technique. Using a model and a prior probability or belief about the model, MAP entails computing a conditional probability of observing the data. For machine learning, MAP offers an alternative probability framework to maximum likelihood estimation. Proposed technique attained MAP of 68%, existing PIDD attained 65% and SVM attained 67%.

5. Conclusion

This research propose novel method in type 2 diabetes based heart disease detection in big data predictive analysis using machine learning method. Processed data features has been extracted utilizing regression model based linear discriminant analysis and classified using VGG-16 Net_gradient neural network. Categorization and prediction accuracy of the current approach is not very good. In this study, we suggested a diabetes prediction model that combines a few extrinsic factors that cause diabetes in addition to more common parameters like glucose, body mass index (BMI), age, insulin, etc. Compared to the old dataset, the new dataset improves classification accuracy. The proposed method achieved 96% accuracy, 67% precision, 79% recall, a 63% F-1 score, a 66% RMSE, and a 68% MAP.

References

[1] Ghojogh, B., Karray, F., & Crowley, M. (2019). Fisher and kernel Fisher discriminant analysis: Tutorial. *arXiv preprint arXiv:1906.09436*.

[2] Wu, W., Wang, J., Cheng, M., & Li, Z. (2011). Convergence analysis of online gradient method for BP neural networks. *Neural Networks*, *24*(1), 91-98.

[3] Hossain, M. E., Uddin, S., & Khan, A. (2021). Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes. *Expert Systems with Applications*, *164*, 113918.

[4] Nicolucci, A., Romeo, L., Bernardini, M., Vespasiani, M., Rossi, M. C., Petrelli, M., ... & Vespasiani, G. (2022). Prediction of complications of type 2 Diabetes: A Machine learning approach. *Diabetes Research and Clinical Practice*, *190*, 110013.

[5] Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A novel diabetes healthcare disease prediction framework using machine learning techniques. *Journal of Healthcare Engineering*, *2022*.

[6] Abdalrada, A. S., Abawajy, J., Al-Quraishi, T., & Islam, S. M. S. (2022). Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort

study. *Journal of Diabetes & Metabolic Disorders*, 1-11.

[7] Hosseini Sarkhosh, S. M., Esteghamati, A., Hemmatabadi, M., & Daraei, M. (2022). Predicting diabetic nephropathy in type 2 diabetic patients using machine learning algorithms. *Journal of Diabetes & Metabolic Disorders*, 1-9.

[8] Sampathkumar, A., Tesfayohani, M., Shandilya, S. K., Goyal, S. B., Shaukat Jamal, S., Shukla, P. K., ... & Albeedan, M. (2022). Internet of Medical Things (IoMT) and Reflective Belief Design-Based Big Data Analytics with Convolution Neural Network-Metaheuristic Optimization Procedure (CNN-MOP). *Computational Intelligence and Neuroscience*, *2022*.

[9] Kour, H., Sabharwal, M., Suvanov, S., & Anand, D. (2021). An assessment of type-2 diabetes risk prediction using machine learning techniques. In *Proceedings of International Conference on Big Data, Machine Learning and their Applications* (pp. 113-122). Springer, Singapore.

[10] Hassan, M. M., Billah, M. A. M., Rahman, M. M., Zaman, S., Shakil, M. M. H., & Angon, J. H. (2021, July). Early predictive analytics in healthcare for diabetes prediction using machine learning approach. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 01-05). IEEE.

[11] Sharma, A., & Mishra, P. K. (2022). Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. *International Journal of Information Technology*, *14*(4), 1949-1960.

[12] Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2021). Multiple disease prediction using Machine learning algorithms. *Materials Today: Proceedings*.

.
.