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

Improved Feature Selection and Classification for Diabetes Mellitus Using Random Forest-Based U-Net Classifier

[1]P. V. Sankarganesh, [2]Dr. R. Priya

Submitted: 14/03/2024 **Revised**: 29/04/2024 **Accepted**: 07/05/2024

Abstract: In recent times, the prediction of diabetes has become a serious due to the presence of numerousattributes while collecting data. Therefore, many models are developed eradicate such problemto attain increased rate of accuracy during the prediction of diabetes. With varying individuals,the dataset changes dynamically and this leads to fluctuations in the prediction accuracy. The poor-quality outcomes result in poor accuracy and this affects the classification ability of the knowledge mining algorithms. In this paper, the develop a feature selection-based classificationmodelling using machine learning algorithm that aims at improving the rate of classification accuracy. The study uses random forest classifier as its feature selection tool and then the classification is conducted using deep U-net classifier. The simulation is conducted to test theefficacy of the model against various models over several available datasets. The results of simulation shows that the proposed method achieves higher rate of classification accuracy thanthe existing methods.

Keywords*: Prediction, diabetes, attributes, random forest, U-net.*

1. Introduction

High blood glucose levels are the primary cause of diabetes, widely known as the silent killer. Blood glucose levels that remain abnormally high over an extended period of time define diabetes [1]. High blood sugar levels occur when the body is unable to properly digest or metabolise carbohydrates. Blood glucose is the body basic fuel, and we get it primarily from the food we eat [2]. Pancreatic beta cells secrete insulin, a peptide hormone necessary for the body to utilise the glucose they create as fuel. However, if the body is unable to produce enough insulin, glucose will build up in the blood, resulting in a rise in blood sugar [3].

While there is currently no treatment for diabetes, adopting healthy habits can greatly improve one quality of life. Failure to timely deliver treatment can lead to complications with the kidneys, neurological system, eyes, lower limb amputation, and heart [4]. For this reason, it ispreferable to establish a diabetes prognosis as soon as possible so that the body systems can continue to operate normally [5]. The World Health Organization (WHO) estimates that by 2019 there will be more than 470 million people worldwide living with diabetes, and that number will nearly double to reach 700 million by 2045. A person can develop either type 1 or type 2 diabetes, or even prediabetes [6].

- ***Type 1 Diabetes***: Exogenous insulin injections may be required to keep blood sugar levels consistent if the pancreas is unable to produce enough insulin. This type of diabetes typically strikes people in their twenties and thirties.
- ***Type 2 Diabetes***: This individual, to be more specific, because the metabolic mechanism of the body is unable to properly break down the food being taken, the prevalence of sugar in the blood rises in those with this illness. This form of diabetes can also be genetically transmitted within families. This kind of diabetes typically affects adults between the ages of 45 and 60.
- ***Gestational Diabetes***: A rise in insulin production and other hormonal shifts during pregnancy play a role in the development of this form of diabetes.
- ***Prediabetes***: High blood sugar levels, but not high enough to be diagnosed as diabetes, characterise this illness, also known as borderline diabetes. Prediabetes is another name for borderline diabetes.

Various models [7] employ a subset of available machine learning methods to forecast future diabetes incidence rates. Some examples of such methods are decision trees, multi layer perceptrons, and random forests. The core tenet of the study area known as machine learning is the concept that computers may learn from examples and data to enhance their capacity to make predictions about the future. Because

[1]*Research Scholar, Vels Institute of Science, Technology & Advanced Studies (VISTAS). Chennai-117.*
sankarphd2017@gmail.com
[2]*Professor, Department of MCA, School of Computing Science, Vels Institute of Science, Technology & Advanced Studies (VISTAS). Chennai-117.*
priyaa.research@gmail.com

the logic is derived from the trained data and validated with the tested data, programmers are unnecessary. It a type of artificial intelligence in which past information is used to aid in forecasting the future. It may be categorized into two classes, and the study elaborate on both of them below.

Supervised Learning: Pretrained models might serve as a road map for further exploration. Predictions are made by first training a new model on the provided datasets or model, and thenusing that model to re-train the original dataset or model.

Unsupervised Learning*:* Learning is primarily accomplished through observation. The computer analyses the dataset for patterns and correlations, and then sorts the data into meaningful categories based on those patterns and relationships.

But most DM data has a high complexity and ambiguity components, which makes it difficult to measure and manage with conventional methods. The classifier approach is among the mostwell-known and potent methods in the world of data processing. It can be obtained. The revolutionary impact of this technology on standard diagnostic procedures has aided in the treatment of a wide range of complex disorders including diabetes [8].

In this paper, develop a feature selection-based classification modelling is developed using machine learning algorithm that aims at improving the rate of classification accuracy. The goal is to improve the level of accuracy required for classification. The deep U-net classifier employs therandom forest classifier to select features for classification, and then feeds those features on tothe subject of the study.

2. Related Works

Diabetes is not contagious and can have devastating consequences on health, but it can cause major issues in the short and long term. The effects of diabetes on an individual health, finances, and emotional state, as well as those of their loved ones, were examined in a report published by the World Health Organization [11]. About 1.2 million lives were lost due to the health catastrophe, according to the data that has been gathered so far. Nearly 2.2 million peopleglobally lost their lives due to diabetes-related complications such as cardiovascular disease and other diseases.

Persistently elevated blood sugar levels are a major risk factor for diabetes [12], often known as a chronic disease. Within the framework of this study, decision-support system using Decision Stump and Ada Boost algorithm is the literature review is performed on the topic of classifiers. Ada Boost is a computational method for validating correctness, and it has been tested with the help of the foundation classifiers Naive Bayes, Support Vector Machine, and Decision Tree. Accuracy Boosting using Ada Boost, Ada Boost is a performance-enhancing tool. Ada Boost 80.72% accuracy with a choice stump as the basic classifier is particularly impressive when compared to the accuracy.

Machine learning [13] is one area where AI is having an impact since it creates calculations that can quickly take in examples and select criteria based on data. Information mining pipelines now include machine learning methods so that they can be used with traditional metric-based procedures. This has allowed for data to be kept distinct from the learning process. Nearly a thousand individual electronic health records were used for the study. Logistic Stepwise component choice regression has been used to predict the occurrence of retinopathy, neuropathy, or nephropathy.The study has used Random forests (RF) to deal with missing data and have linked the right ways to handle class differences. Age, length of the deciding period, BMI, glycated haemoglobin (HbA1c), hypertension, and smoking status are also taken into account. The improved complexity desire models even yielded a number as precise as 0.838. In order to create unique models with evident clinical relevance, a wide range of parameters was selectedfor every conceivable complexity and time frame.

In this work [14], several alternative categorization strategies are applied to a dataset includinginformation about the Pima Indians. In order to reach a diagnosis, both positive and negative diabetes signs are compared and contrasted. Compared to other algorithms, MLP provides superior performance and accuracy for detecting diabetes using a data mining technique such as WEKA.

Abnormally high or low blood glucose levels could have serious short- and long-term effects. Predictive modellingshows that notifying people ahead of time of projected variations in blood glucose allows them to take preventative steps. The study offers a solution that uses a basic physiological blood glucose model evolution to generate informative highlights for a support vector regression presentation that has already been customized for the individual with diabetes. In order to create support vector regression that has already been constructed with tolerant-specific data, this model is employed. Better than diabetes specialists, this new model can predict blood sugar levels, and it might be used to prevent nearly a quarter of hypoglycemic episodes if administered 30 minutes beforehand. Although the current comparison only has a 42% degree of accuracy, patients who respond to hypoglycemia warnings will not be negatively

affected by intervention because the majority of false alarms occur within hypoglycemic zones [15].

A growth to 380 million is predicted for the following 20 years.There is a pressing need in modern society for a thorough analysis of a classifier that can detect diabetes at the lowest feasible cost and with the highest possible efficiency because diabetes isso prevalent and crippling. The Pima Indian Diabetic Database, which is maintained by the machine learning research unit at the University of California, Irvine, is presently utilized as the gold standard to validate the accuracy of information mining predictions for the classification of data relevant to diabetes. Within the framework of the proposed method, a support vector machine (SVM) is used to analyse diabetes. The machine learning seeks to understand how a therapy dataset rich in characteristics might be used to categories diabetic problems. The research concluded that support vector machines are useful for diagnosing diabetes [16].

Diabetic disease is one of the most significant infections in the medical profession, and this study uses Least Square Support Vector Machine (LS-SVM) and Generalized Discriminant Analysis (GDA) [17] to make a diagnosis. To this end, we have proposed a new framework for course learning that makes use of least-squares support vector machines and generalized discriminant analysis. The study splits framework into two parts. We began by using generalised discriminant analysis to identify the characteristics of healthy and diabetic data that were statistically distinct from one another. Second, we ranked the diabetes dataset using the LS-SVM method. Although LS-SVM achieved an accuracy of 78.21% in clustering with 10- cross validation. The GDA–LS-VM framework achieved an order exactness of 82.05% with 10-crease across approval. Using the arrangement precision method, the k-fold validation approach, we investigate the robustness of the proposed design. The accuracy rate of 82.05% that was achieved with the usage of this method is highly encouraging when compared to other, more conventional approaches to classification.

3. Proposed Method

In this section, the proposed model is discussed with illustrations provided in Figure 1.

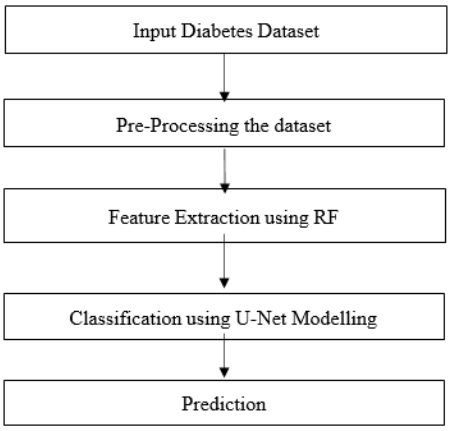

Fig 1: Proposed Model

Feature Extraction:

In the proposed method, feature selection plays a critical role. The goal of feature selection is to reduce the rate of dataset dimensional by eliminating unnecessary features with proper selection of original feature according to predetermined criteria. Feature selection using initialset of features allows for this.

Lets assume N is the number of features in some hypothetical feature collection ($n1$, $n2$, $n3$,... nk). Feature selection refers to the process of deciding which K features would be most useful to incorporate into a certain programme. To choose features, one must first generate a subset, evaluate it, and then take the necessary steps to pause and seek out validation techniques. The selecting process consists of these steps.

When someone refers to a random forest, they are referring to a very large set of decision treesthat have been gathered together. Each tree makes a prediction using some subset of the data, and the community then decides which one is the most reliable. Another technique that can be used to limit the likelihood of over fitting is to average the outputs of each decision tree.

The algorithmic process of a random forest looks like this:

Step 1: Select random samples from the input datasets

Step 2: Construct the decision tree for the selected random sample

Step 3: Make predictions for each decision tree

Step 4: Undergo voting on the predicted results

Step 5: The results with highest votes are considered as the final predicted value.

The dataset is refined by picking the most important characteristics to use in our analysis, and then any

outlying data points are eliminated. Outlier refers to a value that is significantly different from the normal values. It is possible to calculate what values constitute outliers usingthe following formula:

$$p(x)=\begin{cases} x & \text{if } q_1=\dfrac{3*IQR}{2}\le x\le q_3+\dfrac{3*IQR}{2} \\ 0 & \text{if } q_1\ne\dfrac{3*IQR}{2}\le x\le q_3+\dfrac{3*IQR}{2} \end{cases}$$

where

$P(x)$ - outlier rejection,

x - feature vector of the input instances in a n-dimensional space, and

$q1$ - first quartile,

$q3$ - third quartile, and

IQR - inter quartile attribute range.

After the incorrect data was corrected, the blanks were subsequently populated with the proper values. Predictions about individual patients may be off if the dataset has many blanks, whichit does. Using a mean filter, which is a method for data analysis,the missing infer were informed. In addition, the emergence of outliers is not a side effect of utilising a mean-centered technique for imputation of missing values.

$$p(x)=\begin{cases} mean(x) & if\ x=\dfrac{null}{missed} \\ x & if\ x\ne\dfrac{null}{missed} \end{cases}$$ where

$q(x)$ - mean imputation and, it is estimated as the number of times the feature vector appears inthe given space.

The study considers the dataset to a 10-fold cross-validation method, with each fold acting as both a training set and a testing set in turn, in accordance with standard data preparation protocols. In doing so, the guarantee the highest degree of precision in our data set. The remaining dataset, denoted *K*-1, will be used for training, while the remaining dataset, denoted *K*, will be utilized for testing. Then, the study adjusts the parameters of our *k*-nearest-neighbor algorithm, random forest, decision trees, and neural network until it is optimal.

Classification:

The semantic information becomes strong, but the spatial resolution decreases, as you move deeper into a network. U-Net is a model that use the skip connection for fusing the feature maps belonging to various levels. Such process is achieved without compromising the data semantic quality or ability to detect spatial features.

In this paper, the study uses U-Net channel attention model to classify the diabetes features, fed its trained model into a U-Net classification, and then analyzed voting outcomes of majority voting for classifying the final results in order to increase the accuracy of classification.

Figure 2 provides a illustration representation of the classification using U-Net model.

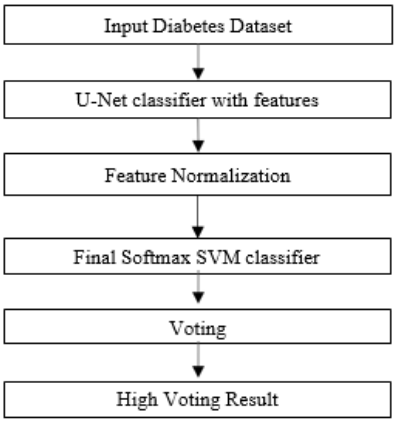

Fig 2: Algorithm Network Framework

U-Net Model

The U-Net, because of its symmetrical form makes it an ideal basis for a model of semantic segmentation. The layers of the U-Net architecture are as follows: input, convolutional, pooling, transposed convolutional, activation function, and output. Convolution is performed by the convolution layer using multiple convolution kernels of size 3×3 and a step size of 1, with the resultant feature map. As part of the convolution procedure, each piece of input data is assigned the same importance (this is referred to as weight sharing).

As a result, the training parameters are more intuitive and the computation is sped up. In addition to its

global benefits, convolution can also boost the transmission of signals within a neural network by enhancing their local perception. Activation is the process of making a neural network more suitable for nonlinear mapping, which increases the model expressiveness. The degree of non linearity in the network is increased to achieve this. ReLU is the name given to the activation function used in the U-Net model, and its definition is as follows.

$$\mathrm{ReLU}\ f(x) = \max(0, x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

ReLU is able to produce only positive results for inputs larger than zero, while having no effect on inputs smaller than zero. A term for this is one-sided suppression. This allows for faster network training by converting dense features to a sparser feature space of high-dimensional using linear differentiability.

The transposed convolution uses the process of up sampling; by performing many upsampling methods, the original dimensions of the feature map can be recovered. The details of classes with other features can be combined through the use of simultaneous upsampling and downsampling of feature channel dimensions. What we do around here is fusion. The final step involves applying a 1×1-parameter convolution kernel to the feature vectors in order to map them to the desired class distribution. Optimal performance of the loss function can be obtained by adjusting the neural network weights, which is within the realm of possibility. Once the predicted value has been made, this function can be used to show the degree of similarity between the two. What other network loss functions are represented by in U-Net are the border weights.

$$E = \sum_{x \in X} w(x) \log p_{\ell(x)}(x)$$

where $p\ell(x)$ - Soft max loss function, and

$\ell:\Omega \rightarrow \{1,\ldots,k\}$ - pixel point label value.

Classifier

The logistic regression layer used in U-Net uses Soft max function for increasing the accuracy of classification based on the regression principle, where the probabilistic loss function is considered as the loss function. Thus the feature space contents are standardized, and the classification results are presented as probabilities.

Assume that there are N distinct clusters in the dataset. The output of the last convolution layer, Y, is represented by the notation $Y=(y1,y2,...,yN)^T$, while the result of the Soft max computation is represented by the notation $s=(s1,s2,...,sN)^T$.

$$S_j = \frac{\exp(y_j)}{\sum_{j=1}^{N} \exp(y_j)}$$

SVM performance is far higher than that of Softmax. The SVM requires the introduction of a kernel function to transfer non-linearly distinct characteristics to a high-dimensional feature space, so transforming the feature data into a linearly distinguishable format. Since this is the case, the SVM is able to achieve its objective of producing linearly distinct feature data. The core idea behind support vector machines is a search for the best possible categorization hyperplane. New information that is not captured by the support vectors has no effect on the stability of this hyperplane. When employing Softmax, however, the decision plane shifts every time new samples are taken.

Channel Attention Module

There have been encouraging outcomes from applying the channel attention mechanism to theclassification of diabetes dataset.

Channel attention is developed as a technique of extracting features in a tunable manner prior to the maximum pooling layer in order to construct a more effective feature map. The goal is to better direct the channel focus. Here is a comprehensive rundown of how the CAM functions: Combining the global average pooling F_{cavg} (1×1×C) with the global maximum pooling F_{cmax} (1×1×C) yields the following feature map: (H×W×C). The shared network two fully connected layers and activation layer then receive the inputs F_{cavg} and F_{cmax}, respectively. Finally, F_{cavg} and F_{cmax} are added together as they travel across the shared network.

The next step is to feed the outputs of this process into the sigmoid function, which produces the $M_c \in R^{1\times1\times C}$ channel attention map. Thereafter, the first fully connected layer is shrunk by $R^{1\times1\times C/r}$, and the second is expanded back to its original $R^{1\times1\times C}$ size. In Figure 3 can see the channel attention module is deficted.. The following formula is used for the purpose of identifying channelemphasis:

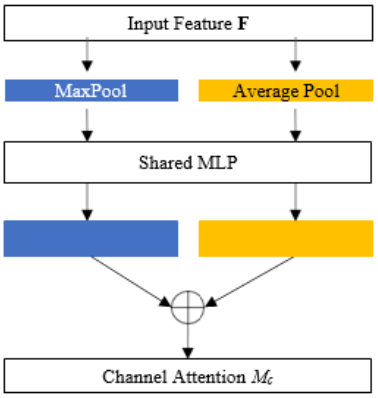

Fig 3: Channel attention module

Fusion

In order to categories the original samples, the network model selects the three layers of features that are the most informative, extracts those features, and feeds them into the SVM. Once this is complete, the three results are averaged to arrive at a consensus classification. In this study, we began by isolating experimental samples made up of dataset samples with all class labels. Initially, 60% dataset sample are used for training, 20% of the datasets for validation, and 20% of the datasets for testing. Features were voted on and chosen from Layers 2, 56, and 57 of the network models for use in the SVM classification.

4. Results and Discussion

A python simulator on an Intel Core i5-8500 CPU, where the learning rate, batch size extracted at each iteration, and optimal gradient clipping are just a few examples of hyper parameters for a deep learning network that can only be determined by empirical debugging. The model convergence to a solution is controlled by a number of hyper parameters, one of which is the learning rate. More precise results can be expected locally when the learning rate, loss function change, and network convergence are all slowed down. But if the learning rate is too high, it ispossible to find any local minima.

With such a rapid learning rate, the network gradients might quickly become unmanageable, hence the gradient threshold is typically adjusted to permit gradient clipping. So that any occurrences of local minima can be identified, this procedure is carried out. 1000 mini-batches are extracted throughout over 30 rounds of 30000 iterations. It is possible to detect gradients as small as 0.05 and as slow as 0.0001/s.

The PIMA Indians diabetes dataset is used which can be found in the UCI data repository after initially being made available to the public on Kaggle. There are 768 pregnant women represented in the database, 268 of them have diabetes and 500 is without any symptoms of diabetes. Eight patient-related variables and one class predictor for diabetes status made up the dataset nine total variables. The dataset contained some outliers and some missing values. It is possible to isolate the problematic data points by applying the method proposed. The missing values were imputed using a mean filter method, and then the dataset was returned toits usual consistency.

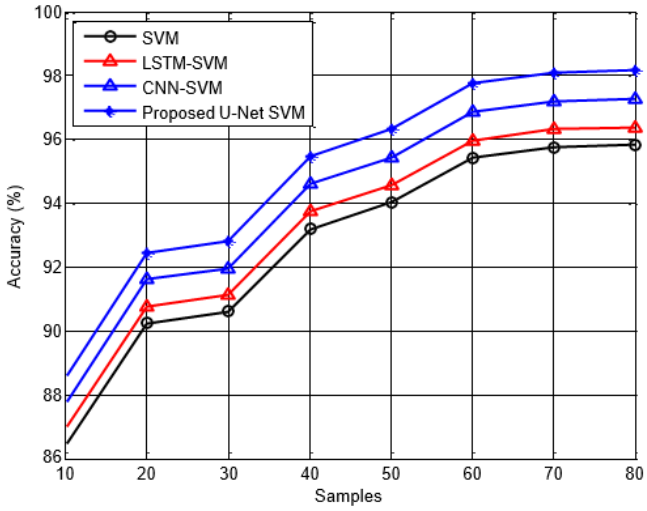

Fig 4: Accuracy

Figure 4 shows the results of accuracy between the proposed U-Net classifier with the existing models on diabetes prediction. The results of simulation shows that the proposed method achieves higher rate of classification accuracy than the other methods.

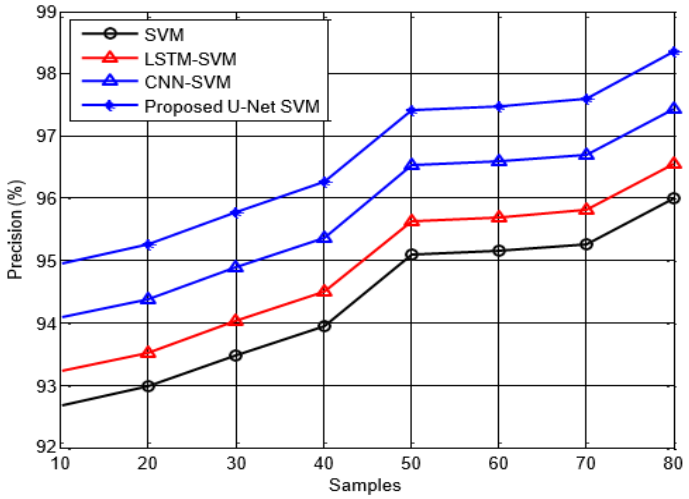

Fig 5: Precision

Figure 4 shows the results of precision between the proposed U-Net classifier with the existing models on diabetes prediction. The results of simulation shows that the proposed method achieves higher rate of precision than the other methods.

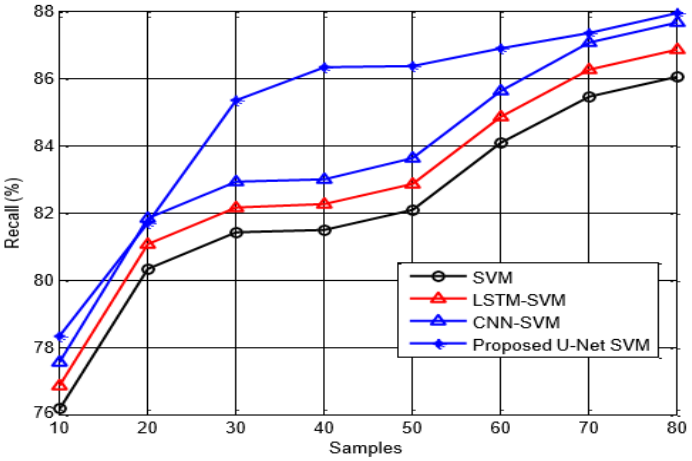

Fig 6: Recall

Figure 4 shows the results of recall between the proposed U-Net classifier with the existing models on diabetes prediction. The results of simulation shows that the proposed method achieves higher rate of recall than the other methods.

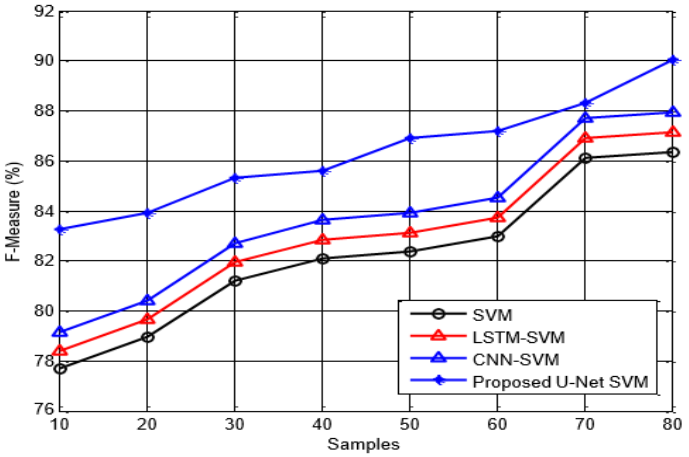

Fig 7: F-Measure

Figure 4 shows the results of F-Measure between the proposed U-Net classifier with the existing models on diabetes prediction. The results of simulation shows that the proposed method achieves higher rate of F-Measure than the other methods.

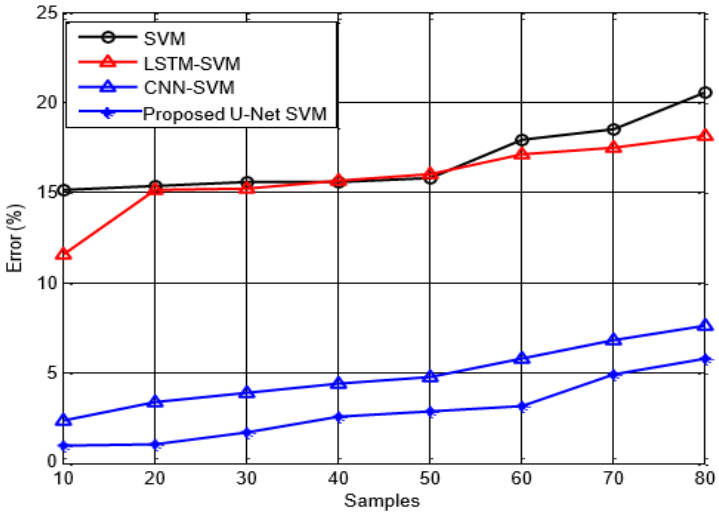

Fig 8: Classification Error

Figure 4 shows the results of classification error between the proposed U-Net classifier with the existing models on diabetes prediction. The results of simulation shows that the proposed method achieves higher rate of classification error than the other methods.

5. Conclusions

In this paper, feature-selection-based machine learning method for classificationmodelling, is developed with the end goal of improving the accuracy of classification. Because of this, data can be modeled more accurately. Following feature selection with the random forest classifier,classification is handled by the deep U-net classifier. The purpose of the simulation is to test the model accuracy and efficiency against competing models using publicly available data. The simulation results demonstrate that the proposed approach outperforms the state-of-the-art methods in terms of classification accuracy.

References

[1] Ramesh, J., Aburukba, R., & Sagahyroon, A. (2021). A remote healthcare monitoring framework for diabetes prediction using machine learning. *Healthcare Technology Letters*, *8*(3), 45-57.

[2] Jaiswal, V., Negi, A., & Pal, T. (2021). A review on current advances in machine learning based diabetes prediction. *Primary Care Diabetes*, *15*(3), 435-443.

[3] 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*.

[4] 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.

[5] Ravaut, M., Sadeghi, H., Leung, K. K., Volkovs, M., Kornas, K., Harish, V., ... & Rosella, L. (2021). Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. *NPJ digital medicine*, *4*(1), 1-12.

[6] Khaleel, F. A., & Al-Bakry, A. M. (2021). Diagnosis of diabetes using machine learning algorithms. *Materials Today: Proceedings*.

[7] García-Ordás, M. T., Benavides, C., Benítez-Andrades, J. A., Alaiz-Moretón, H., & García-Rodríguez, I. (2021). Diabetes detection using deep learning techniques with oversampling and feature augmentation. *Computer Methods and Programs in Biomedicine*, *202*, 105968.

[8] Nadeem, M. W., Goh, H. G., Ponnusamy, V., Andonovic, I., Khan, M. A., & Hussain, M. (2021, October). A fusion-based machine learning approach for the prediction of the onset of diabetes. In *Healthcare* (Vol. 9, No. 10, p. 1393). MDPI.

[11] Kokkinos, C. M., Tsouloupas, C. N., &

Voulgaridou, I. (2022). The effects of perceived psychological, educational, and financial impact of COVID-19 pandemic on Greek university students' satisfaction with life through Mental Health. *Journal of Affective Disorders*, *300*, 289-295.

[12] VeenaVijayan, V., & Anjali, C. (2015). Prediction and diagnosis of diabetes mellitus— a machine learning approach. Recent Adv. 2015.

[13] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, *15*, 104-116.

[14] Hina, S., Shaikh, A., & Sattar, S. A. (2017). Analyzing diabetes datasets using data mining. *Journal of Basic and Applied Sciences*, *13*, 466-471.

[15] Plis, K., Bunescu, R., Marling, C., Shubrook, J., & Schwartz, F. (2014, June). A machine learning approach to predicting blood glucose levels for diabetes management. In *Workshops at the Twenty-Eighth AAAI conference on artificial intelligence*.

[16] Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, *3*(2),1797-1801.

[17] Polat, K., Güneş, S., & Arslan, A. (2008). A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. *Expert systems with applications*, *34*(1), 482-487.