International Journal of Healthcare Information Systems and Informatics
Volume 16 • Issue 4

Diabetes Prediction Using Enhanced SVM and Deep Neural Network Learning Techniques:

An Algorithmic Approach for Early Screening of Diabetes:

P. Nagaraj, Kalasalingam Academy of Research and Education, Krishnankoil, India

P. Deepalakshmi, Kalasalingam Academy of Research and Education, Krishnankoil, India

ABSTRACT

Diabetes, caused by the rise in level of glucose in the blood, has many devices to identify it from blood samples. Diabetes, when unnoticed, may bring many serious diseases like heart attack and kidney disease. In this way, there is a requirement for solid research and learning model enhancement in the field of gestational diabetes identification and analysis. SVM is one of the powerful classification models in machine learning, and similarly, deep neural networks are powerful under deep learning models. In this work, the authors applied enhanced support vector machine and deep learning model deep neural network for diabetes prediction and screening. The proposed method uses a deep neural network obtaining its input from the output of enhanced support vector machine, thus having a combined efficacy. The dataset considered includes 768 patients' data with eight major features and a target column with result "Positive" or "Negative." Experiment is done with Python, and the outcome of the demonstration shows that the deep learning model gives more efficiency for diabetes prediction.

KEYWORDS

Classification Algorithm, Deep Learning Model, Deep Neural Networks, Diabetes, Machine Learning, Support Vector Machine

1. INTRODUCTION

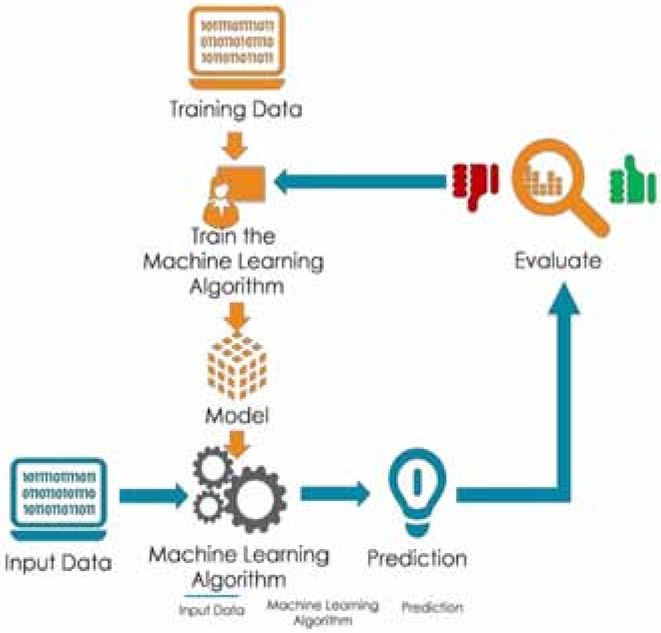
Information mining is a powerful way to drawing meaningful information from datasets containing massive embedded data. Data mining may be fruitfully exploited in hospitals where a huge volume of data exists. These hospital datasets often need to be clustered, or even further classified to gain a meaningful analysis of the underlying data. Soft computing techniques such as pattern recognition (PR) and machine learning (ML) have been used for identifying statistical parameters from the dataset. A World Health Organization (WHO) report claims that almost 422 million people worldwide are suffering from both TYPE-1 and TYPE-2 diabetes (https://www.who.int/health-topics/diabetes). Also, as released by the WHO, India's profile shows that 46% of the total population have prevailing diabetes and related risk factors.

DOI: 10.4018/IJHISI.20211001.oa25

Diabetes Mellitus (DM) is mainly sorted as three types: (a) diabetes mellitus or DM, (b) insulin resistance, and (c) the third one is gestational diabetes generally found amid pregnant ladies. DM is generally caused by high blood glucose levels and is a typical disease that influences those individuals who have imbalance in blood glucose levels, especially for pregnant women. In fact, past research has demonstrated that pregnant women with diabetes are increasingly inclined to have newborns with birth defects than those without diabetes. Kanguru, et al. (2014) notes that, in such situation, the child may be influenced by prevailing illness conditions, for example, coronary illness and spina bifida. The beginning stage for living admirably with DM is an early diagnosis. Hence, the primary aim of our work is to predict diabetic or non-diabetic using ML and deep learning algorithms. A major study of the classification systems gives high accuracy with high handling time, though few strategies have yielded low precision even with enormous dataset. Along these lines, our work aims for high accuracy and less processing time with immense dataset.

Figure 1 represents a schema of the ML model utilized in our work. Nowadays, many efficient analysis techniques are available at affordable cost. Data analytic approaches improve the disease detection accuracy in modern hospitals. DM, when detected early, can avoid serious complications and may also be managed via a healthy diet. The key objective here is to build a prediction system by combining the usefulness of ML and Learning Model algorithms.

Figure 1. Schema of a Machine Learning (ML) Model


Training Data
Train the Machine Learning Algorithm
Evaluate
Model
Input Data
Machine Learning Algorithm
Prediction
Input Data
Machine Learning Algorithm
Prediction

Following the introduction, Section 2 reviews the extant research in diabetes detection using ML and deep learning approaches. Section 3 offers insights into our proposed novel approach with details on the PIMA dataset vis-à-vis details on the correlation-based feature selection (CFS) subset evaluation algorithm. Section 4 shifts focus into the data visualization with details on the key analytic approaches employed, namely, enhanced support vector machine (ESVM) and deep neural network (DNN). Section 5 then presents the results and reflections. Finally, in Section 6, we offer concluding remarks with insights into future research directions.

2. RELATED WORKS

Numerous works have exploited ML, clustering, classification, information mining and learning applicable for diabetes discovery systems. Here, we survey the more prominent works in the extant literature with their concise propositions.

In an early work, Ensan, et al (2006) proposed the Fuzzy Clustering (FACT) method, which stipulates the quantity of fitting clusters to be dependent upon the data density. Even so, the proposed algorithm is insensitive to the initial number of clusters, while initial cluster numbers are often set to be less than the threshold number of clusters. Their strategy generates a number of clusters by making new clusters to focus on the detection of outliers. In their work, they demonstrated experimentally that the proposed heuristic algorithm exhibits a superior performance than the conventional K-means computation. The classification technique on ML namely SVM for DM detection is presented by Purnami, et al. (2009). Here, they proposed an improved version of SVM, namely Smooth SVM (SSVM) and Multiple Knot Spline SSVM (MKS-SSVM). These researchers applied both algorithms over the PIMA dataset with results indicating a higher accuracy for MKS-SSVM v. SSVM.

In time, Aishwarya et al. (2013) proposed a fuzzy logic-based application to analyze DM. In their model, two correlations among symptoms and diseases were generated, more specifically, an occurrence relationship and a confirmability relationship. The occurrence relationship confirms recurring appearance of a symptom; conversely, the confirmability relationship portrays the intensity of symptoms for the disease presence. Moreover, they proposed Fuzzy Logic with minimum and maximum relationship and made use of a real-world dataset of 40 patients to determine the fuzzy relationship when investigating DM. Kumari and Chitra (2013) has also correlated few ML and data mining techniques for the prediction of DM. In their work, they utilized SVM. Exploratory outcomes on a genuine diabetes dataset demonstrated that understandable SVMs provide a promising accuracy for DM prediction with investigational results showing an accuracy of 79.00%. It is noted that improving the efficacy of SVM classifiers may still be achieved by imparting the feature subset selection process into it.

More recently, Kumar, et al. (2014) proposed using Genetic Algorithm (GA) to locate a proper component subclass with SVM classifier on various datasets to progress the characterization precision. The GA-based classifier seems to improve the boundary esteems for SVM, acquiring the ideal subset of features. An identification of outcomes with existing SVM methodology showed that the suggested strategy yielded an accuracy of 83.00%. As well, Jhaldiyal and Mishra (2014) used Principal Component Analysis (PCA) and a SVM for the prediction of diabetic patients. Investigational outcomes from the study exhibited that the previous level can be amended upon as they had a classification accuracy of almost 93.66%.

Vispute, et al. (2015) considered the risk of DM by order procedures. In their work, four ML procedures, namely Decision Tree (DT), Artificial Neural Networks (ANN), Logistic Regression (LR) and Naive Bayes (NB), were evaluated. These researchers created a web application with PHP as front end and MySQL as backend, in which they utilized the ROC curve method for the DM forecasting. The data are fed into the application which then displays the output with actual v. forecasted figures. Finally, their experiments proved that Random Forest (RF) accomplishes great accuracy. About the same time, Lingaraj, et al. (2015) experimented with DM detection using the WEKA tool via the NB classifier. The data used for the study were collected from Indian Hospitals with 1865 instances comprising blood test and urine test attributes. They experimented with 10-fold cross-validation and compared the results. Though the authors have done 10-fold validation, the accuracy achieved via the computational model is only 84.89%, which is very minimal and needs to be further explored.

Lately, Perveena, et al. (2016) employed DT J48 for DM detection, dependent on risk factors. In their system, they demonstrated that Adaboost outperforms in efficiency than bagging and DT J48. Owing to the absence of base learners in the ensemble framework, this work however delimits itself in various performance measures, thereby providing a research gap to be filled by the addition of base learners to an ensemble of classifiers. Vashi and Mishra (2016) reviewed DM prediction via a survey where they discussed more about causes of diabetes and types of diabetes involved in the DM detection. The authors focused more on classification and clustering techniques for DM detection, aside from neural networks. However, no empirical data were reported in the Vashi-Mishra work.

VrushaliBalpande & Wajgi (2017) stuided classification techniques for the DM detection by applying classification algorithms, DT and NB classifier in WEKA tool. They concluded that the DT method achieves better performance than NB classifier. Anjali (2017) use neural networks (NNs) and supervised learning methods to predict DM, its overall survivability in comorbidity. The researcher proposed a methodology based on PCA to decrease the dimension of extracted features with NN as the classifier. The accuracy result was 92.2%. The performance of the work can be still increased by tuning the parameters of the NN used. With the use of PCA, the need for data regularization tends to create additional overhead to the algorithm. Edla, et al (2017) proposed a diabetes decision support system (DSS) which is constructed by utilizing radial basis function NN classifier (RBFNN). RBFNN is a three-layer neural network, in which the first layer is used to handle contributions of the model, the second layer is the concealed layer made out of several non-direct RBF actuation units, and the last layer is the yield layer of neural organization classifier. Gaussian capacities are utilized to execute enactment works in RBFNN. RBFNN utilizes the PIMA Indian dataset for building a DM expectation model that accomplishes 73.91% accuracy, 81.33% sensitivity, and 60% specificity. The limitation on this methodology is the tradeoff that exists between the number of hidden layers and the accuracy.

Cui, et al. (2018) suggested a hybrid estimate model that also uses PCA to the original dataset and then used C4.5 algorithms for constructing the classifier model. The classification exactness of their work was only about 89.0% which can be further elevated by using suitable feature selection methods. Haritha, et al. (2018) discussed another data mining approach for DM detection combining classification techniques and association rule mining techniques. The authors assessed the classification execution strategy by utilizing the KNN classifier. Initially, the authors used 10 features to prepare the KNN analysis and accomplished 61.9% accuracy. Later, the authors chose 6 features for training by using PSO techniques and were able to improve the accuracy to 88.5%. Rashid and Abdullah (2018) discussed Type-1 and Type-2 DM detection using firefly and cuckoo search algorithms for selecting attributes from the Indian PIMA dataset. From the selected optimal features using firefly algorithm, they applied KNN classifier for the DM prediction while for the optimal features identified using cuckoo search, they used Fuzzy-KNN classifier. From their experimental study, they concluded that cuckoo-Fuzzy KNN achieved the best classification accuracy. Improvement in the learning rate of optimization algorithm can also be made.

In separate attempts, Zhang, et al. (2018) discussed a DM diagnosis via the application of hybrid ant colony, GA and NNs. The author considered the Indian PIMA dataset, in which feature selection is done using ant colony and GA. Then, they applied back propagation NNs to the selected features. They experimentally proved that they achieved good accuracy via this model. The central processing unit (CPU) time taken becomes a rising issue when hybridizing optimization algorithms, which needs to be addressed in similar applications. Kadhm, et al. (2018) discussed the DM detection problem by feed forward neural networks (FFNN). Here, the author used the Indian PIMA diabetes dataset and applied a two-layer feed forward NN algorithm. They proved experimentally that more than 82% accuracy can be attained via the FFNN model. Wu, et al (2018) proposed a novel model for the discovery and forecast of DM type-2 utilizing K-Means Clustering and LR model calculations. The proposed technique guarantees the enhancement in expectation precision which comprises of both group and class strategies. The proposed techniques upgrade precision by 3% in anticipating diabetes. The limiting factor of this algorithm is the implementation issues with voluminous data. Sisodia., et al (2018) used a classification algorithm to builds up a diabetic forecast model to analyze diabetes at an underlying stage. They utilized three AI models, specifically SVM, DT, NB techniques on the PIMA Indian DM dataset for building a diabetic expectation model. The performance of these three techniques is assessed utilizing various measures such as accuracy, precision, recall, F1-score, and the recipient working characteristics curve. The end-results demonstrated that the NB algorithm performs better than the DT and SVM. The accuracy of the NB algorithm is 76.3% which is higher than the DT with 73.82% precision and SVM with 65.1% accuracy.

Even more lately, Perveen, et al. (2019) introduced an insightful SVM model for diagnosing DM. These authors viewed DM as a significant medical problem worldwide and uncovered that 80% of DM complexities can be interfered with whenever these have been identified at a beginning phase. In the proposed model, various data mining and ML algorithms have been assessed for the DM forecast. For example, they proposed the SVM model with an extra module for turning the "discovery" model of an SVM into a justifiable portrayal. This novel framework gives a choice on SVM grouping with conspicuous precision. Still, improvement can be made in the aspect of performance improvement for all sampling cases. In contrast, Zhu, et al, (2019) proposed an altered K-Means for extracting strident data and PCA for the characterization of the dimensional datasets. Their work uses clustered instances in the primary stage and distinguished highlight subcategory in the subsequent stage to identify the best classifier for predicting DM. Test results imply the avalanche SVM along with PCA subset recognized to have provided the classification accuracy of SVM to a high of 81.28%. The lack of mechanism for the selection of the number of principal components will be a limiting factor for algorithms that include PCA in its working.

Most recently, Sivakumar, et al. (2020) endeavor to discover results for diagnosing the DM infection via assessing the helpful examples present in the information to give convenient treatment to the patients utilizing distinctive AI models. They acquired 76.3% and 75.7% precision for effectively ordered examples by AI classifiers, for example, the NB calculation and RF algorithm. Srivastava, et al. (2020) intended a model that uses the Fuzzy C-Means Clustering (FCM) algorithm with SVM for forecasting diabetes. The model conquered an 84.24% accuracy. Their use of a Fuzzy C-means clustering algorithm for disposing of undesired data is the additional advantage of this work but these authors have not compromised on the computational trade off that arises due to FCM. An automated algorithm to decide the number of cluster specifications becomes mandatory for this algorithm.

As shown in Table 1, the research gaps identified from the literature review shows explicitly that an algorithm to enhance SVM which includes a feature selection process can be formulated. Also, the literature review suggests the non-applicability of optimization algorithms and PCA due to constraints such as time consumption and data standardization. Altogether, it is apparent that a combination of Enhanced SVM (ESVM) with tuned parameters of instance (C), loss function (ε), penalty parameter (bit) and Deep Neural Network (DNN) is non-existing. As such, we prefer to explore a hybrid algorithm using this enhanced combination with reduced penalty parameters and loss function, speculating that this will have a meaningful insight in the field of DM prediction to provide accurate characterization.

3. OUR PROPOSED NOVEL APPROACH

In this work, the SVM-NN classification algorithms were considered for applications on the PIMA Indians Diabetes Database. Of the two, for a more accurate prediction, the NN was preferred. Figure 2 shows the overall architecture of our proposed novel approach.

We focus on the PIMA dataset from uci.edu (https://www.kaggle.com/uciml/pima-indians-diabetes-database) as it has previously been studied widely. As summarized in Table 2, the dataset comprises 768 samples with selective attributes. As a pre-processing step, we used correlation as correlation is very helpful in imputing missing values in the dataset. The casual relationship that exists between the available data can also be estimated with the help of correlation. The pre-processed features will then be used in subsequent phases.

A model built for the prediction of certain factors must be tested explicitly on known and unknown values. In building up the prediction system, this prima facie is accomplished in our work by categorizing the available dataset via certain rules of thumb. The projected method utilizes the ESVM

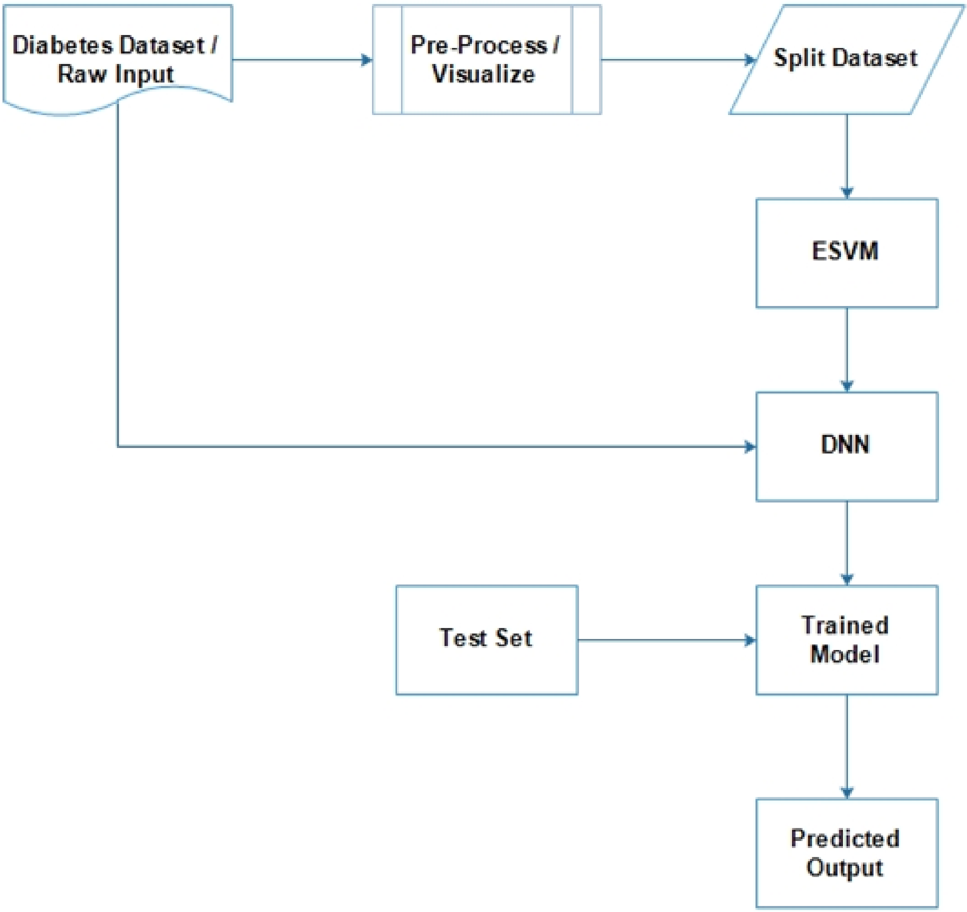
Table 1. Research Gaps

S. No	Authors'	Methodology	Research Gap identified
1	Perveena.S et.al [18]	Ensemble of classifiers such as Adaboost, J48	absence of base learners in ensemble framework
2	Sisodia, D., & Sisodia, D. S. [23]	Naive Bayes classification algorithm	Unavailability of feature selection algorithm
3	Purnami, S.W., et.al [20]	Smooth Support Vector Machine	Tradeoff between accuracy and time remains unsolved
4	Ensan, F et.al [5]	FACT: A new Fuzzy Adaptive Clustering	Minimization of quadratic error.
5	R. Aishwarya et.al [21]	SVM with Principle Component Analysis	Lack of interpretability due to use of PCA algorithm
6	Haritha, R et al [6]	Cuckoo-Fuzzy KNN	Improvement in the learning rate of optimization algorithm
7	Rashid, T. A., & Abdullah, S. M. [22]	Artificial Bee Colony, Genetic Algorithm, and Neural Network	CPU time taken when hybridizing optimization algorithms.
8	Zhang et. al [29]	Feed- Forward Neural Network	Use of activation function to improve diabetes prediction
10	Kumari, V. A., & Chitra, R [16]	SVM	Use of feature subset selection process
11	Zhu et.al [30]	PCA and K-means techniques	Lack of mechanism for Selection of number of principal components
12	Kumar G et al. [12]	Genetic algorithm with SVM	Premature convergence due to GA
13	Srivastava, A. K et.al [25]	FCM with SVM	Need for specifying number of clusters.
14	Anjali K et.al [1]	PCA, neural Network and cultural algorithm	Need for data regularization
15	Perveen S et al [19]	SVM	Performance improvement for all sampling cases
16	Wu H, et al [28]	K-Means & Logistic Regression	Implementation issues with voluminous data
17	Sivakumar. S, et al [24]	Naïve Bayes and random forest	Increased misclassification rate
19	Edla, et al [4]	RBFNN	Tradeoff between number of hidden layer and accuracy.

and deep learning technique, as will be elaborated later herein. Using the processing efficiency of the ESVM and deep learning technique, the process of training will also be performed for the training dataset selected. In order to find the efficacy of the proposed algorithm, it was further evaluated for DM prediction on test data. To enhance the user experience, these operations have been integrated in the form of a User Interface (UI) as a flowthrough application.

Put simply, we loaded the dataset, then pre-processed it to eliminate the null values. Following this, we visualized the pre-processed data as correlation matrix and histogram plots prior to applying the SVM and DNN algorithms to generate efficient predictions.

Figure 2. Overall Architecture of Proposed work



3.1 Dataset Details

Table 2 highlights the characteristics of PIMA dataset. We did not consider those missing and noisy attributes embedded in the dataset.

Table 2. Dataset Characteristics

Data set	PIMA
Number of samples	768
Feature Attributes	8
Output classes	2
Total number of feature attributes	9
Missing attribute status	None
Noisy attribute status	None

Table 3 details the feature attributes of PIMA dataset along with and the corresponding symbols. The dataset is ready to use as the feature values are converted to numerical values, for example, the class value is converted to 0 or 1 if tested positive v. negative.

Table 3. Diabetes Description

Features description	Features Symbol
Number of times pregnant	Preg
Plasma glucose concentration 2 hours in an oral glucose tolerance test	Plas
Diastolic blood pressure (mm Hg)	Pres
Triceps skin fold thickness (mm)	Skin
2-Hour serum insulin (mu U/ml)	Insu
Body mass index (weight in kg/ (height in m) ^2)	Mass
Diabetes pedigree function	Pedi
Age (years)	Age
Class variable (Positive or Negative)	Class

3.2 Correlation-Based Feature Selection (CFS) Subset Evaluation Algorithm

This algorithm looks for a subset of features that function admirably together by having the highest correlation. The score of the subset is called merit, which can be seen in the yield. It assesses the value of a subset of qualities by considering the individual prescient capacity of each element alongside the level of repetition between them. The wellness esteems for every selected subset can be estimated by using t-test significance function in the following equation

$$t = \sqrt{\frac{\left(n-2\right)}{1-r^2}} \quad (1)$$

where *n* is the number of instance and *r* is the correlation coefficient

Algorithm 1: (Correlation-based feature Selection)

Input: List of Capabilities X = {X_i: i = 1 to n}

Output: S – Choose highlights (subset of X)

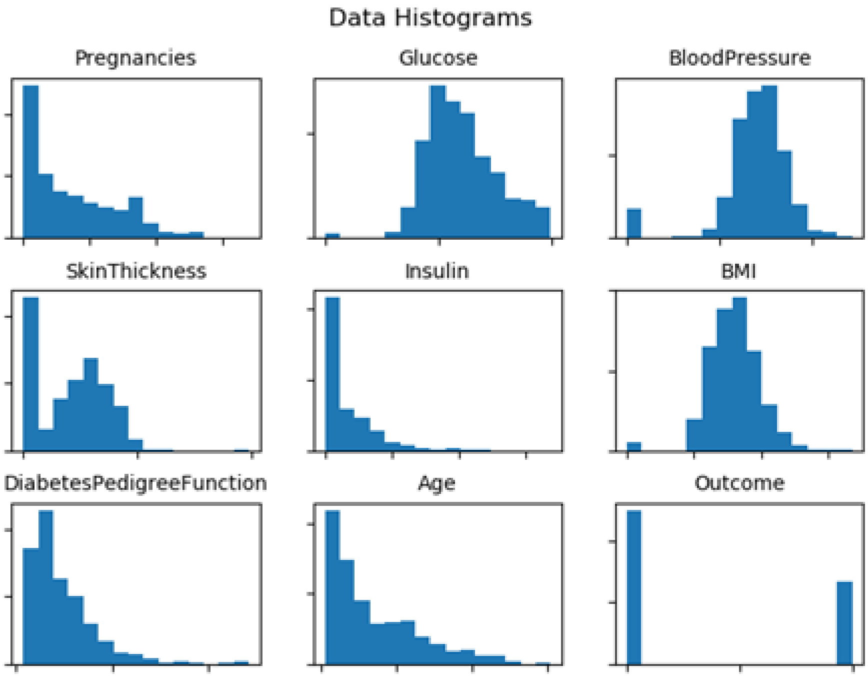
1. Instate the multitude by utilizing mapping capacity and produce the mapping variables as V_i
2. Characterize switch likelihood q
3. While (halting standards met)
4. Create M number of subset X
5. Assess wellness esteems for every subset selected by using equation (1)
6. Locate the best arrangement subset p* in the fundamental set
7. Generate subset i.e. {X = X_1 U X_2 U X_3 ……} using mapping capacity and variables
8. Compute the presentation proportions of the arrangement p* (feature selection) on the test set
9. Return the element subset of the property p*

We perform feature selection using the aforementioned algorithm considering the attributes, Plasma glucose concentration, Body mass index (kg/m^2), Diabetes pedigree function, Age (years) and Class Variable (nominal) - tested _positive v. tested_negative

4. DATA VISUALIZATION

Data from the PIMA dataset are used here. Eight predominant indicators, including pregnancies, Glucose, Blood pressure, Skin thickness, insulin, Body Mass Index (BMI), diabetes pedigree function, and age, served as key indicators as visualized in Figure 3 with column plot as histogram values with the outcomes indicated accordingly.

Figure 3. Data Visualization of PIMA dataset



4.1 Enhanced SVM (ESVM)

SVM is one of most commonly used supervised learning techniques, applied usually for both classification and regression problems. The algorithm works in such a way that each data is plotted as a point in n-dimensional space with the feature values representing each coordinate. Figure 4 shows the binary classification model, in which C1 and C2 are two different classes. SVM uses the best margin to classify the output.

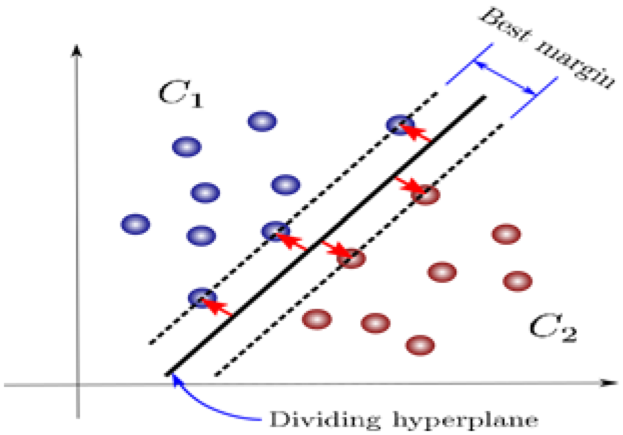
Accordingly, the SVM-based strategy was applied to locate a suitable classifier utilizing the gaussian function on a set of training datasets.

Instance Learning:

The issue of numerous instance learning is to become familiar with a model that can recognize a set of given positive and negative bits of occurrences. Each bit contains many instances. It accepts that a bit is positive only if it has at least one positive case; conversely, a bit is said to be negative if all cases are negative.

Given *n* bit parameters P_1, P_2, P_3…. P_n, there are k_i instances in each bit P_i, 1 £ i £ m with a label for each bit. Without the loss of simplification, each bit parameter Pi has a label Y_I in the range of

Figure 4. Support Vector Machine Classification



{-1, 1}. If the label of a bit parameter is negative, the labels of all instances in the bit are negative as given by the below equation:

$$\sum_{i^\circ I}^{n} Y = \frac{\mathrm{Y}i+1}{2} \geq 1\text{------} \qquad (2)$$

For all I, such that $\mathrm{Y}_{\mathrm{I}} = 1$.

Instance Feature Selection:

We now show the different instance learning model issues. In spite of the fact that different occasion learning models are used, each instance will be in the bit boundary. Hence, it is critical to change the continuous information into numerical highlights that encourage the utilization of multiple learning systems. As such, a nonlinear model is used to lay out auxiliary data of each instance to a support component vector.

The twofold classification strategy of SVM was initially recommended as it can manage precisely with complex nonlinear limit models. Yet, this way of computing parameters is computationally expensive (Wang, et al., 2019). Hence, an enhanced multi-instance algorithm dependent on the SVM algorithm, which is similar to instance feature selection by Wang, et al. (2019), is proposed to be applied when working with small sample instances, nonlinear and high dimensional design perception. The key objective is to find a discriminative function which can measure the instance parameters as per given imperative. In the structure, the label of a bit is controlled by the better instance in the bit. In equation (1), we know that there is just one tag in the bit that are negative, then the value of $\sum_{i^\circ I}^{n} Y = \frac{\mathrm{Y}i+1}{2} \geq 1$ SVM performs ordering of values by defining several hyperplanes in a multidimensional space, enabling the partition of instances between various class labels. SVM can also deal with unlimited features. By applying SVM to fit a model, we get to characterize these features appropriately. In order to avoid the intricacies in the conventional SVM, we propose an enhanced SVM by updating the Parameter C, epsilon (ε) and bit. Typically, SVM discovers an edge, which isolates all positive v. negative models.

C is the expense of accurately fitting the characterization. The C parameter trades off the misclassification of preparing models vs. the generalization of the choice surface. The low value of C settles on the choice surface smoothly, while a high C targets ordering all preparation examples accurately by giving the model an opportunity to choose more fitting examples as a help vector.

The parameter ε is the inhumane loss function needed to be selected for an estimation. ε affects the perfection of the SVM's reaction and it influences the quantity of help vectors, so both the intricacy and the predictability of the system anticipate its worth.

The parameter bit is referred to as the penalty parameter, permitting errors. The first plan for a most extreme edge classifier requires detachable class dispersion and does not permit errors for preparing points. What the penalty parameter really does is to restrict the focus on an inappropriate side of the choice limit delicately and fluently.

SVM can deal with enormous element spaces and manages huge datasets. Overfitting can be constrained by a delicate edge approach. In SVM, training is generally simple, yet picking a decent bit of work is uncertain. Enhanced SVM has a distinctive parameter setting to abstain from misclassifying designs.

The algorithmic structure of ESVM algorithm is as follows:

Algorithm 2: (Proposed ESVM Algorithm)

Input: Initial Data Set S

Output: The optimal parameters C_i, bit

```
Initialization: For $i \in I$, $s_i = S_I$ (for all parameters, use the parameter C to initialize the labels for all the cases in the parameter)
Repeat:
    Find the parameter $\varepsilon$ and bit according to the SVM Model
    Find $C_i = \varepsilon^T s_i$ – bit for all instance in a positive parameter
    Use $s_i$ = significance ($C_i$) to recalculate the labels of all instance in the positive parameter
    For (Each positive parameter $P_I$) do
        If $\Sigma_{i \in I}$ (si + 1) / 2 = = 0 then
            Calculate $i^* = \text{argmax}_{i \in I} C_i$
            $S_i^* = 1$
        End If
    End for
While The label of the instance changes from the previous round do
End While
Output ($C_i$, bit)
```

To improve the prediction accuracy, the SVM output and the raw input from the dataset are applied to the various stages of the DNN method.

4.2 Deep Neural Network (DNN) Classification

The DNN algorithm works in the flow of initializing network, forward propagation, back propagating error, training the network to predicting the values.

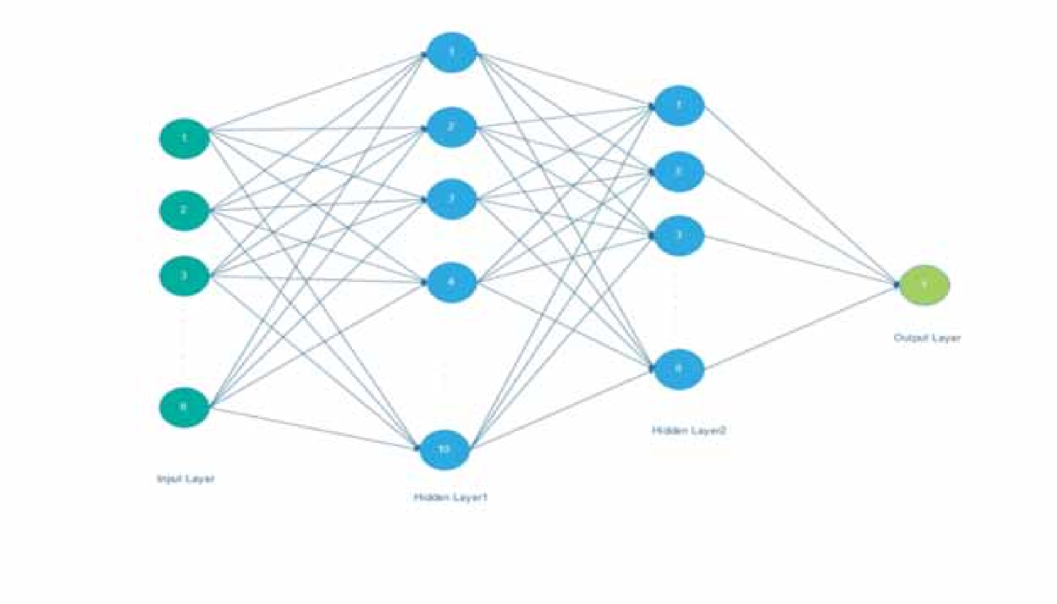
Figure 5 shows the architecture of DNN with two hidden layers that we consider for our problem. The number of hidden neurons should be less than twice the size of the input layer in order to get a smoother transition from the domain to the co-domain, thereby improving its generalization and reducing overfitting.

In the network initialization layer, input for the network is defined along with its activation function. Each neuron maintains weight. The input layer is given a row for each neuron from the training dataset. The next layer is the hidden layer. The third layer is the output layer that has one neuron for each class value. As it is sufficient to have one output neuron for a binary classification problem (to represent 0 or 1), in the proposed work, we initialized the input to DNN with input_dim, which is 8 in our dataset. The next (hidden) layer comprises two dense layers being assigned with neurons 10 and 8 respectively. The drop out ratio in the hidden layer is given 0.2 and the activation function used is 'sigmoid' in the output layer. Finally, we compiled the model with binary_crossentropy and 'adam'

Table 4. List of Notations Used

S.No.	Notation	Used in	Description
1	t	Equation No. (1)	t-test significance function for every selected subset
2	n	Equation No. (1)	Number of Instances used for t-test
3	r	Equation No. (1)	Correlation Coefficient
4	X	Algorithm No. (1)	List of Capabilities used for Feature Selection
5	S	Algorithm No. (1)	Correlation based Subset Selection
6	p and p*	Algorithm No. (1)	Presentation and Proportions of the best selection
7	M	Algorithm No. (1)	Number of Instances got in Subset selection
8	P_i and C_i	Algorithm No. (2)	Optimal parameters for Enhanced SVM
9	C	Algorithm No. (2)	Characterization Parameter
10	ε	Algorithm No. (2)	Inhumane loss function
11	Y_i	Algorithm No. (2)	Instance Learning Rate

Figure 5. Deep Neural Network (DNN) Architecture



optimizer. The output of the trained model is stored in our repository. The output result ‘0’ refers to No diabetes and ‘1’ refers to the presence of diabetes. This single input detection is done by DNN.

Table 5 shows the sample training result with DNN and epoch for the training used is 200 data elements.

Table 5. DNN Training Model

S.No.	Epoch	Training Loss	Accuracy
1	190	0.5898	0.9667
2	191	0.5896	0.9667
3	192	0.5877	0.9719
4	193	0.5863	0.9667
5	194	0.5818	0.9719
6	195	0.5803	0.9667
7	196	0.5794	0.9719
8	197	0.5787	0.9719
9	198	0.5772	0.9667
10	199	0.5771	0.9719
11	200	0.5767	0.9719

5. RESULTS & DISCUSSION

The PIMA Indian diabetes dataset is drawn from UCI repository (uci.edu) with the proposed analytic approach implemented in Python 3.6.4 with libraries Keras, TensorFlow, Scikit-Learn, Pandas, Matplotlib and other mandatory libraries. ML algorithm and deep learning algorithm is applied, more specifically, ESVM and DNN. We used these algorithms to aid in identifying DM cases. The result shows that deep learning is more efficient than ML algorithm. The proposed ESVM-driven DNN yields an accuracy of about 98.45 percentage.

The result of the experiments is shown below for the SVM algorithm. **Table 6** shows SVM precision metrics and other error values such as Root-mean-square error (RMSE), R-squared value, Mean Absolute Error (MAE), Mean Squared Error (MSE).

Table 6. SVM Metrics

Parameter	Value
MSE	0.322916667
MAE	0.322916667
R-SQUARED	0.476923077
RMSE	0.568257571
ACCURACY	98.45%

Figure 6 shows the resulting metrics arrived from SVM algorithm for diabetes prediction, which shows the accuracy around 98.45% via ML model.

Figure 7 shows the accuracy arrived from the DNN algorithm for training v. testing sets with DNN achieving more than 98% accuracy on both training and test sets.

Figure 8 shows the loss arrived from DNN algorithm for training and testing sets. As shown in the below figure, the regularization parameter reduces the loss and 200 epoch is executed to optimize the results.

Figure 6. SVM Result Metrics

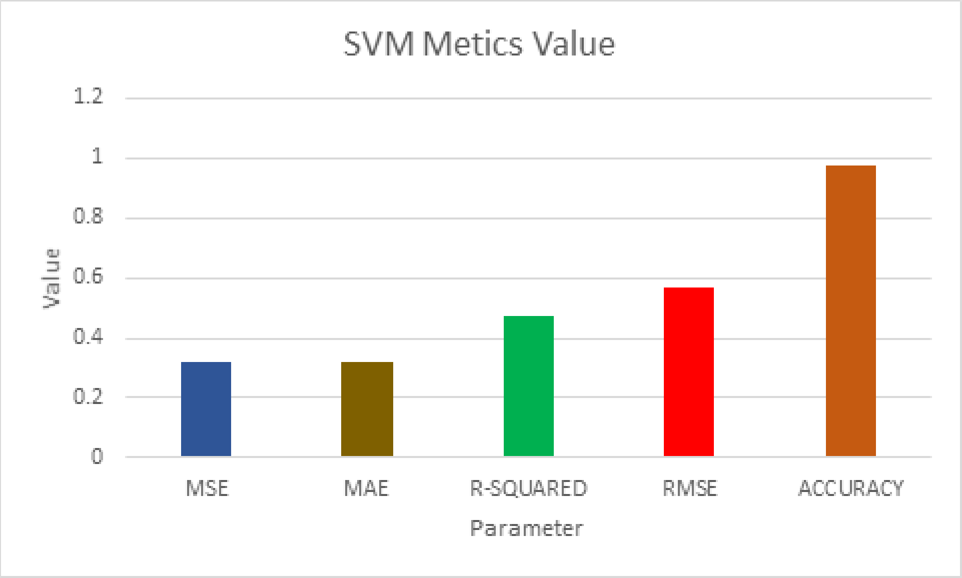


Figure 7. Training & Testing Accuracy for Neural Network

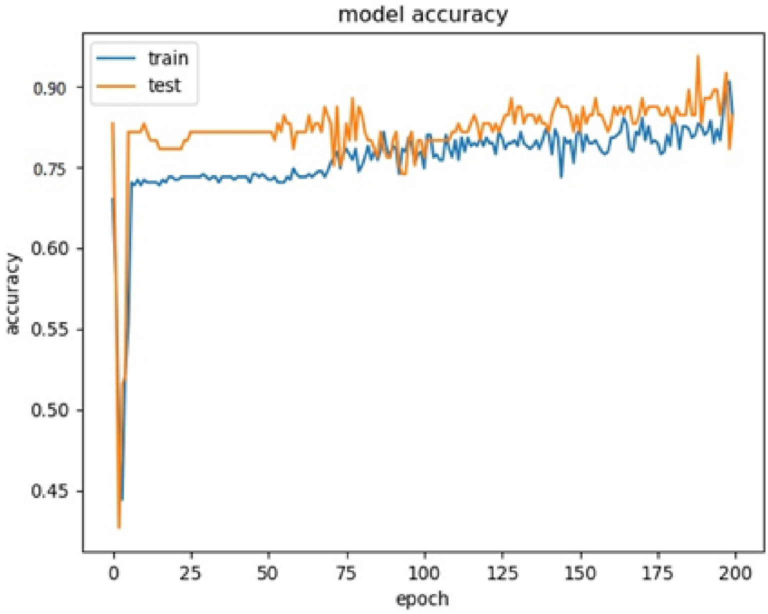
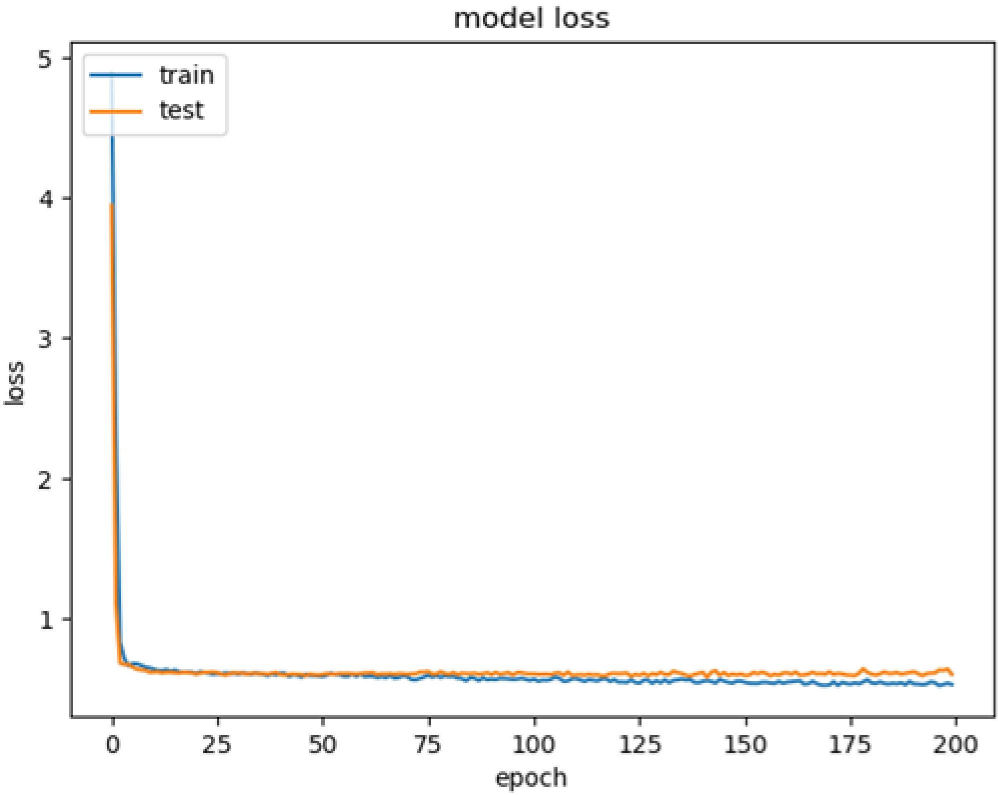


Figure 8. Training & Testing Loss for Neural Network



5. 1 ESVM v. DNN Performance Metrics

The confusion matrix which expresses the true and predicted condition in the same grid is shown in **Table 7**. As shown, the obtainable outcome of a nature task may be described as one of four categories.

Table 7. Confusion Matrix

		Predicted Condition	
	Total Data Set	**Prediction Positive**	**Prediction Negative**
True Condition	Condition Positive	True Positive (TP)	False Negative (FP)
	Condition Negative	False Positive (FP)	True Negative (TN)

While positive data correspond to precise information, negative data apparently allude to inaccuracy in the dataset. Accuracy is the quantity of effectively veracious anticipated diabetes instance classification, either for a free test set, or utilizing some variety of the cross-validation thought. and it is defined as,

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{3}$$

Sensitivity is the true positive rate. In the problem domain under consideration, it depicts what extent of data with diabetes are accurately distinguished as having diabetes, which is characterized as,

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4}$$

In general, Sensitivity demonstrates, how well the test predicts on the proper classification. Specificity is the true negative rate. It portrays what extent of data with diabetes are inaccurately recognized as having diabetes, which is characterized as,

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{5}$$

In general, Specificity gauges how well the test predicts the other classification.

The confusion matrix is shown in Table 8 and the evaluation metrics obtained by our proposed method is shown in Table 9.

Table 8. Confusion Matrix

Confusion Matrix	False	True
False	178	8
True	3	522

Table 9. Performance Metrics for the enhanced SVM and DNN

Performance Measures (Evaluation)	Reduced Data Set/ Percentage (%)
No. of Attributes Used	5
Sensitivity	99.43
Specificity	95.70
Positively Predicted Value	98.50
Negatively Predicted value	98.34
Accuracy	98.45

5. 2 Accuracy Comparison With Other Experiments

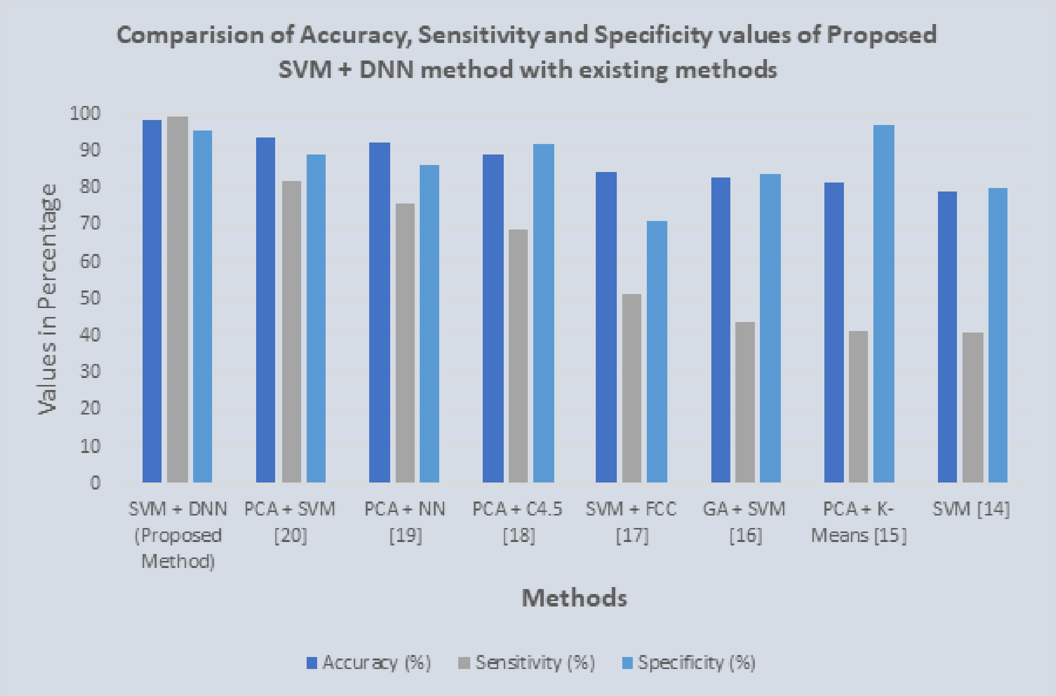
Figure 9 depicts the pictorial representation of the comparative analysis of accuracy, sensitivity, specificity and precall values of the proposed methodology with existing methodologies as given in the Table 10. From the comparative figures, it is apparent that the ESVM-DNN method yielded elevated values for all performance indicators.

Our research findings may now be viewed retrospectively. First, the least number of attributes selected via ESVM and DNN is 3 and that of extreme number of attributes is 5. Second, the maximum classification accuracy obtained by means of ESVM and DNN is 98.45%. Out of 768 instances, the ESVM-DNN approach selected 711 samples as correctly classified v. 57 samples to be identified as outliers. The outlier detection percentage is 7.42. Finally, pregnancies and age are considered as critical attributes for the PIMA diabetes dataset. Most importantly, in comparison with existing approaches, the ESVM-DNN method yielded elevated values for all key performance indicators, including accuracy, sensitivity and specificity measures.

Table 10. Accuracy Comparison

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)
SVM + DNN (Proposed Method)	**98.45**	**99.43**	**95.70**
PCA + SVM [12]	93.66	82.00	89.00
PCA + NN [1]	92.20	75.82	86.00
PCA + C4.5 [3]	89.00	68.90	92.00
SVM + FCC [25]	84.24	51.37	71.09
GA + SVM [15]	83.00	43.60	84.00
PCA + K-Means [30]	81.28	41.50	97.00
SVM [16]	79.00	40.86	80.20

Figure 9. Comparison of Performance Indicators



6. CONCLUSION AND FUTURE ENHANCEMENT

Medical data have important underlying pattern with hidden stories. It is critical to be able to unfold these stories by duly extracting the relevant data for analysis purposes via data mining and ML algorithms. In different areas of medicine, including medical image processing such as brain tumour, cancer disease detection, diabetes, liver disease and heart disease, or early detection of leukaemia and Parkinson disease identification, these techniques have been found to be useful.

This work aims to execute a forecasting model for the accurate and efficient estimation of DM. As discussed, an enormous proportion of the world population is at risk of having DM, making the emergence of unknown viruses and infectious diseases as in the era of COVID-19 even more dangerous for everyone living in this world. Consequently, in our proposed research, we have tried to apply powerful classifiers on the PIMA Indian dataset in order to demonstrate that ML and deep learning algorithm may be able to shorten the risk factors and improve the result regarding DM screening proficiency and accuracy. The results from our approach accomplished on the PIMA Indian dataset is higher than other proposed systems on an equivalent dataset utilizing ML and deep learning algorithms as shown in **Table 10**.

Nonetheless, there are some limitations with respect to our research. The primary key in finding the correct treatment for diabetes is to distinguish the disease in a screening stage. In the current work, a novel ESVM-DNN model was proposed for diabetes type forecast. The profound neural organization was pretrained with the quantity of epochs for the training phase was low, which guarantees that the technique can work quickly, even on any versatile stage. In DNN, the quantity of hidden layers should be not as much as double the size of the input layer. However, we utilized two hidden layers to perform DNN. In case it is diminished to one hidden layer, we can improve speculation.

Finally, we considered DM as it found to be a common and major disease amongst those residing in India. We considered PIMA Indian DM database, and tried to build a learning model via the ESVM and deep learning algorithm DNN to achieve a 98% accuracy in the case of DNN algorithms. To extend this work, we would recommend applying convolution models of DNN to yield further insights from the data analysis. Also, this research may be extended to have novel optimized feature selection methods applied before training the model for classification. Only with ongoing investigations and innovative algorithms can we reach greater success in our search for discovering hidden knowledge from valuable data embedded in key databases.

REFERENCES

Aishwarya, R., Gayathri, P., & Jaisankar, N. (2013). A method for classification using machine learning technique for diabetes. *IACSIT International Journal of Engineering and Technology*, *5*, 2903–2908.

Anjali, K. (2017). Khushbu Pawar diagnosis of diabetes mellitus using PCA, neural Network and cultural algorithm. *International Journal of Digital Application Contemp Res*, *5*(6), 115–125.

Balpande, V., & Wajgi, R. (2017). Review on Prediction of Diabetes using Data Mining Technique. *International Journal of Research and Scientific Innovation*, *4*, 43–46.

Cui, S., Wang, D., Wang, Y., Yu, P. W., & Jin, Y. (2018). An improved support vector machine-based diabetic readmission prediction. [PubMed]. *Computer Methods and Programs in Biomedicine*, *166*, 123–135. doi:10.1016/j.cmpb.2018.10.012

Edla, D. R., & Cheruku, R. (2017). Diabetes-finder: A bat optimized classification system for type-2 diabetes. *Procedia Computer Science*, *115*, 235–242. doi:10.1016/j.procs.2017.09.130

Ensan, F., Yaghmaee, M. H., & Bagheri, E. (2006). FACT: A new Fuzzy Adaptive Clustering Technique. The 11th IEEE Symposium on Computers and Communications. doi: doi:10.1109/ISCC.2006.73

Haritha, R., Babu, D. S., & Sammulal, P. (2018). A Hybrid Approach for Prediction of Type-1 and Type-2 Diabetes using Firefly and Cuckoo Search Algorithms. *International Journal of Applied Engineering Research: IJAER*, *13*(2), 896–907.

Hu, J., Wang, J., Lin, J., Liu, T., Zhong, Y., Liu, J., Zheng, Y., Gao, Y., He, J., & Shang, X. (2019). MD-SVM: A novel SVM-based algorithm for the motif discovery of transcription factor binding sites. [PubMed]. *BMC Bioinformatics*, *20*(7), 41–48. doi:10.1186/s12859-019-2735-3

Jhaldiyal, T., & Mishra, P. K. (2014). Analysis and Prediction of Diabetes Mellitus Using PCA, REP and SVM. *International Journal of Engineering and Technical Research*, *2*(8), 164–166.

Kadhm, M. S., Ghindawi, I. W., & Mhawi, D. E. (2018). An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach. *International Journal of Applied Engineering Research: IJAER*, *13*(6), 4038–4041.

Kanguru, L., Bezawada, N., Hussein, J., & Bell, J. (2014). The burden of diabetes mellitus during pregnancy in low-and middle-income countries: A systematic review. [PubMed]. *Global Health Action*, *7*(1), 23987. doi:10.3402/gha.v7.23987

Kumar, G. R., Ramachandra, G. A., & Nagamani, K. (2014). An efficient feature selection system to integrating SVM with genetic algorithm for large medical datasets. *International Journal of Advanced Research in Computer Science and Software Engineering*, *4*(2), 272–277.

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.

Lingaraj, H., Devadass, R., Gopi, V., & Palanisamy, K. (2015). Prediction of Diabetes Mellitus using Data Mining Techniques: A Review. Journal of Bioinformatics & Cheminformatics.

Perveen, S., Shahbaz, M., Keshavjee, K., & Guergachi, A. (2019). Metabolic syndrome and development of diabetes mellitus: Predictive modeling based on machine learning techniques. *IEEE Access. IEEE*, *7*, 1365–1375. doi:10.1109/ACCESS.2018.2884249

Perveena, S., Shahbaza, M., Guergachib, A., & Keshavjeec, K. (2016). Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. *Procedia Computer Science*, *82*, 115–121. doi:10.1016/j.procs.2016.04.016

Purnami, S. W., Embong, A., Zain, J. M., & Rahayu, S. P. (2009). A New Smooth Support Vector Machine and Its Applications in Diabetes Disease Diagnosis. *Journal of Computational Science*, *5*(12), 1003–1008. doi:10.3844/jcssp.2009.1003.1008

Rashid, T. A., & Abdullah, S. M. (2018). A hybrid of artificial bee colony, genetic algorithm, and neural network for diabetic mellitus diagnosing. *ARO-The Scientific Journal of Koya University*, *6*(1), 55–64. doi:10.14500/aro.10368

Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia Computer Science*, *132*, 1578–1585. doi:10.1016/j.procs.2018.05.122

Sivakumar, S., Venkataraman, S., & Bwatiramba, A. (2020). Classification Algorithm in Predicting the Diabetes in Early Stages. *Journal of Computational Science*, *16*(10), 1417–1422. doi:10.3844/jcssp.2020.1417.1422

Srivastava, A. K., Kumar, Y., & Singh, P. K. (2020). A Rule-Based Monitoring System for Accurate Prediction of Diabetes: Monitoring System for Diabetes. *International Journal of E-Health and Medical Communications*, *11*(3), 32–53.

Vashi, I., & Mishra, S. (2016). A Comparative Study of Classification Algorithms for Disease Prediction in Health Care. *International Journal of Innovative Research in Computer and Communication Engineering*, *4*(9).

Vispute, N. J., Sahu, D. K., & Rajput, A. (2015, December). A Survey on naive Bayes Algorithm for Diabetes Data Set Problems. *International Journal for Research in Applied Science and Engineering Technology*, *3*(12).

Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, *10*, 100–107. doi:10.1016/j.imu.2017.12.006

Zhang, Y., Lin, Z., Kang, Y., Ning, R., & Meng, Y. (2018). A feed-forward neural network model for the accurate prediction of diabetes mellitus. *International Journal of Scientific and Technology Research*, *7*(8), 151–155.

Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, *17*, 100179.

Nagaraj P.is working as an Assistant Professor in the department of Computer Science and Engineering, School of Computing, Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu. He is pursuing his PhD in the area of Health Care Recommender System in Data Analytics.

P. Deepalakshmi, PhD., is currently working as a Professor in Department of Computer Science and Engineering at Kalasalingam Academy of Research and Education (KARE), Virudhunagar, Tamilnadu, India. She is also serving as Dean, School of Computing. Her research interest includes Optimization Techniques, Network Routing, Distributed Computing, Network Security, Data Analytics, Machine Learning Techniques. She also takes care of KARE ACM student chapter as faculty mentor.