

Dual Edge Classifier Based Support Vector Machine (Desvm) Classifier for Clinical Dataset

S. Kavipriya, T. Deepa

Abstract: Data mining is the progression of determining hidden information that are available in the existing data. Data mining discovers interesting, convenient relationships in huge volume of data. Many fields including medical field is using data mining for classifying the data. Classification is method which assigns a data in the collection to predict the objective class. Classifying a diabetic patient is tedious job in the current medical field. The main intention of this paper is to propose a novel classifier enhancing support vector machine to correctly classify the diabetic patients more accurately than the previous classifiers. Performance metrics such as sensitivity, specificity, rate of true positive and false positive, precision, accuracy and time taken for feature selection are used. In the proposed classifier threshold value is fixed for metric recall and true negative rate. The results are demonstrated with better performance.

Keywords: Classification, SVM, Gestational Diabetes, Prediction, and Accuracy.

I. INTRODUCTION

Data Mining (DM) occupies the space between the science of computer, optimization, and estimations, where it routinely appears in different disciplines. All things considered, DM is the path toward seeking idea about information from interchanging divergent perspectives. Here learning can suggest any sorts of consolidated or obscure information that are concealed as rough data. It can be a course action of rules made from available data accumulated from couple of patients of a particular infirmity and healthful people. These benchmarks perchance utilized for forecasting the health condition of new patients. By, all things considered, DM endeavours are portrayed into two classes: (i) descriptive, and (ii) predictive. Descriptive DM errands depict a objective data set in brief, enlightening, discriminative structures, where the predictive DM errands lead the acknowledgment and induction on present data to construct the future prediction. Supervised learning aspires for developing a predictive strategy from data when the feature of class is accessible. There exists two sorts of class features: (i) category based features, and (ii) numeric based features. The category based feature can take simply apparent characteristics, while the numeric based feature can choose vast quantity of characteristics.

Revised Manuscript Received on 30 March 2019.

* Correspondence Author

S. Kavipriya*, Research Scholar, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, Tamil Nadu, India

Dr. T. Deepa, Assistant Professor, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, Tamil Nadu, India

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

By and large, the predictive model is known as a classification model, and when a class include obvious cum numeric, then it is said as regression model. As it were, order is to find and forecast the patients a category, while regression forecast's are numeric esteem. Here, the term "supervised" implies the classification method knows about the prediction because the features of the class are resolved.

II. LITERATURE REVIEW

Brisimi. T. S et al.,2018 has proposed 2 methods namely K - LRT which was a test method following the ratio of likelihood, and a Joint Clustering Classifier which was utilized to detect the hidden clusters in patient history and it ensembles the classifier to all the cluster, furthered theory based sample was developed to provide guaranty for upcoming strategies. Wang. Y et al.,2017b proposed decision making strategy based on sharing scheme for the diabetes people of type-2 category, patients were taken care by the information gathered from the sample of previous history. For the classification purpose multilabel methodology was used to provide medications to the patients. Turksoy. K et al.,2017 proposed a classifier for finding the of faults in the concentration of glucose which were suggested the glucose sensor and a dynamic model based on nonlinear was developed, where they named the method as multivariable statistical monitoring methods. Lekha. S., 2018 explored the utilization of enhanced version of single dimensional convolution neural network method to adopt feature extraction concept with benchmark classification methods, where the results shows that there exist a significant level of reducing the limitations individually. Kijanka. P et al.,2018 presented an classical two dimensional fourier transformation concept to utilize the method of multi signal classification, and to give effective estimation to classification k-space cum space velocity curves were used. Rasti. R et al.,2018 presented a Computer-aided Diagnosis method which was fully based on convolution ensembling. This method aimed to detect the casual retina of diabetic patients by making a comparison with casual pathologies. Wang. D et al.,2017a presented a general methodology connecting feature extraction by pulse, and it extends dimension of feature from one dimension time series to two dimension time series matrix, where predictable wrist features relating to pulse were analyzed to enhance the results. Vyas. R et al.,2018 by utilizing Genetic Programming (GP) method demonstrated the based Symbolic Regression (SR) approach for predicting the disease. Moreira. W. L et al., 2018 proposed Artificial

Neural Network based classifier by making an analysis on model, evaluation of performance, and comparison of different classification techniques, namely radial basis cum neural network. This method has mainly focussed to detect all the cases of gestational diabetes which can efficiently find the risk level for conceived women and fetus. Ruiz. E et al.,2016 presented a classifier for assigning the correct meal and measuring the moment to the available insufficient glycaemia information, where many machine learning methodologies were analyzed to design the better classifier to provide increased accuracy. Kavipriya. S, Deepa. T., 2018 proposed a comprehensive feature selection based support vector machine classifier (CFS-SVM) for classifying clinical dataset where it utilizes the increased significant pattern for choosing the most editable features. It has performed the classification between the samples of different classes. Khine. M. L et al.,1999 identified the frequency of gestational diabetes among the teenager and reviewed the factors of risk that could easily find a subset for teenager patients which have great risk for getting gestational diabetes. Zhou. W et al.,2017 made a analysis and presented a unsupervised classifier fully depending on detecting the disease Posterior Cerebral Artery. It doesn't make a focus on training set, and suggested that the problem of class imbalance that can be avoided at a minimal level. Li. H et al.,2017 presented a method for classification which was fuzzy theory based and provided the configuration for data with unexpectancies in stochastic as well as fuzzy nature. The design for tuning the procedures were analyzed with the terms of probability performance measure, that is., for working with high complex scenarios. Lekha. S, Suchetha. M et al.,2018 proposed the enhanced version of deep learning based convolution neural network classification algorithm to solve the issues in perceptron multi layer. This method utilizes the support vector machine concept for enhancing the total performance detection applications in real time. Sisodia. D et al.,2018 made a study with the target to develop a method to prognosticate the similarity of diabetes among patients in terms of high level accuracy.

III. DUAL-EDGED SUPPORT VECTOR MACHINE (DESVM)

Let $Z = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ be a elements of a set, with $x_i \in X_N \subset R^{N+1}$, $y_i \in Y = \{\pm 1\}$, and $z_i = (x_{i+1}; y_{i+1})$. Let $f(x)$ be a dual-edged classifier to such an extent that produces $h(x) = \text{sign}(f(x))$.

Let $EP = \{z_i \in Z | y_i = +1\} \neq \emptyset$ and $EN = \{z_i \in Z | y_i = -1\} \neq \emptyset$ be the arrangements of preparing designs for the positive class and the negative class, separately. Let $FP = \aleph EP$ and $FN = \aleph EN$ be the quantity of positive and negative samples, where $N = FP + FN$. Based on matrix table, the most usually utilized metric for the assessment of the speculation execution of the classifier $f(x)$ on a test set $(x \times y)$ is the exactness, signified by fully conditioning, which figures the extent of occurrences that are accurately classified by the model, that is,

$$Ac(x, y) = \frac{TP + TN}{FP + FN} \quad (3.1)$$

Utilizing these metrics, the two classes namely positive and negative, it has a similar preference with the end goal of classification. Subsequently, precision can be considered as false when a class is considered of more noteworthy importance than alternate class. In this manner, the cost for wrong classifications in the preference class is more notable than the cost for wrong classifications in alternate class. Also, precision can be considered as false when probabilities of classes contrast extraordinarily (imbalanced datasets), since the metrics neglect to think about effort for wrong classifications, and therefore exceptionally it delicate to the tendency between classes [He. H, Garcia. E. A, 2009].

In this way, different measures of evaluation must be considered. By considering exactness rates on EP and EN independently, the recall and specificity metrics, meant by Re and Sp separately, are characterized as takes after:

$$Re(x, y) = \frac{TP}{FP} \quad (3.2)$$

$$Sp(x, y) = \frac{TN}{FN} \quad (3.3)$$

The review measure is the extent of positive cases that are accurately recognized (TPR). Then again, specificity is the divisions of effectively recognized cases among all examples that are negative (TNR). The two metrics are measures of arithmetic. Consequently, another articulation for exactness can be determined as takes after the calculating the $Ac(x, y)$ as below:

$$Ac(x, y) = \frac{FP.Re(x, y) + FN.Sp(x, y)}{FP + FN} \quad (3.4)$$

Calculation is done as a weighted arithmetic mean of recall and specificity, where the related weights are the quantity of positive occurrences and negative examples, separately. For certain situation the set EP is viewed as a "preference" with the end goal of classification, at that point the $Re(x, y)$ measure is more illustrative than the $Sp(x, y)$ measure, and $Ac(x, y)$ is never again a sufficient metric. In this way, another way to deal with acquire a classifier when one class is viewed as a "preference" is presented: For a settled test, it is necessary to consider a set y contained in $(x \times y)$, when looking for a robust classifier, the search is limited to classifiers that hold review measures at a predefined level. Inside this arrangement of classifiers, it is looked for so as to boost specificity. It is significant that, as a rule, it isn't conceivable to discretionarily get incredible recall and specificity esteems, that is, the limit of any classifier can't build the quantity of the genuine positives without additionally expanding the quantity of FP.



Inside the arrangement of classifiers $F(E)$, a swapping must be found between $Re(b)$ a moving back capacity of b , and $Sp(b)$, an expanding capacity of b , keeping in mind the end goal to boost classification. By accepting that the positive class is of a higher priority than the negative class, then $\beta = \max \{b \in R, Re(b) = 1\}$ is the best dual-edge for the preparation set Z , which prompts the classifier $f_\beta(x) = \langle x, E \rangle - \beta$. Nevertheless, $Sp(\beta)$ can be subjectively little if the occasion $p1$ is an exception. A better approach to evade this issue is to make an arrangement in classifiers $F(E)$ is to settle a limit $0 \leq E \leq 1$, with the end goal that $Re(b) \geq E$ to ensure a base genuine TP in the positive class and to expand $Sp(b)$. Henceforth, the accompanying equation is considered with subject to $Re(b) \geq E, 0 \leq E \leq 1, f_b \in F(E)$.

$$\max_{b \in R} Sp(b) \quad (3.5)$$

It ought to be demonstrated that the proposed approach can be seen as hypothesis testing from concluding value, that is, the esteem $(1 - E)$ in the DESVM method is like the essentialness level α in hypothesis testing. Therefore, given E esteem with the aim of lying between 0 and 1, where the classifier $f \in F(E)$. Considered is as per the following:

$$f(x) = f_{bE}(x) \quad (3.6)$$

$$f_{br}(x) = \sum_{i=1}^N \gamma_i y_i K(x_i, x) - b_E \quad (3.7)$$

Given an arrangement of classifiers $F(E) = \{f_b(x) = \langle x, W_0 \rangle - b\}$, where b belong to the set R on a preparation set Z , the impact of the edge on the execution of the DESVM can be broke down: for $0 \leq E \leq E' \leq 1$, the $b_{E'}$ dual-edge for E' and confirms that $Re(b_{r'}) \geq E' \geq E$, and therefore $Sp(b_{r'}) - Sp(b_r)$ where b_r is used in the calculation for r . Furthermore, since $Sp(b)$ is an expanding capacity of b , then $b_E \geq b_{E'}$ and, since $Re(b)$ is a moving back capacity of b , then $Re(b_E) \leq Re(b_{E'})$. Therefore, if $0 \leq E \leq 1$ then $Re(b_E)$ is an expanding capacity of E , and $Sp(b_E)$ is a moving back capacity of E on the preparation set Z . This outcome can be seen perfectness in the experimentation given in the section 6.

IV. ABOUT THE DATASET

PIMA Indian diabetes dataset was used to evaluate the proposed classifiers performance with the existing classifier FELM [Nahato. K. B et al.,2016]. PIMA dataset is multivariate in nature and it contains 768 instances with 9 attributes including the class label. 268 instances are identified as the patients with gestational diabetes. The attribute contains mixture of integer and real numbers. The attribute information are given in Table. 4.1.

Table 4.1. PIMA Dataset's Attribute Information

S. No	Name of the Feature	Descriptive Information	Domain Range	Zero or One
1	Preg	Number of times pregnant	0 to 5	111
2	Glu	Concentration of plasma glucose in 2 h of oral glucose tolerance test	0 to 199	5
3	Bp	Pressure of diastolic blood (units: mm Hg)	0 to 122	35
4	Skin	Thickness of triceps skin fold (units: mm)	0 to 99	227
5	Insulin	2-h serum (units: mu U/ml)	0 to 846	374
6	IBM	Index of body mass (units: kg/mt sq)	0 to 67	11
7	FDP	Function of diabetes pedigree	0.078 to 2.42	-
8	Age	Age (units: years)	21 to 81	-
9	Class	Label of class	0 or 1	NI L

V. PERFORMANCE METRICS

MATLAB R2013a is used for evaluating selected clinical dataset. The benchmark metrics sensitivity, specificity, rate of true positive and true negative, precision and accuracy were utilized for evaluating the performance of the proposed work. The metrics are calculated by the values of True Positives (TP), False Negatives (FN), True Negatives (TN) and False Positives (FP). TP indicates the correctly identified instances that who are affected by gestational diabetes. If suppose the patients are not classified correctly, then it becomes FN. Healthy instances that are identified correctly by the classifier falls in TN, else it becomes FP.

VI. RESULTS AND DISCUSSIONS

The overall performance analysis of the FELM and the proposed CFS based SVM (CFS-SVM) in terms of TP, TN, FP, FN, sensitivity and specificity is presented in Table-6.

Table-6.1: Performance Analysis of TP, TN, FP, FN, Sensitivity and Specificity

Metrics →	TP		TN		FP		FN		Sensitivity		Specificity	
	FELM	Proposed (DESVM)	FELM	Proposed (DESVM)	FELM	Proposed (DESVM)	FELM	Proposed (DESVM)	FELM	Proposed (DESVM)	FELM	Proposed (DESVM)
80-20	199	222	49	33	10	3	10	10	95.22	95.69	83.05	91.67
70-30	186	220	53	30	20	5	9	13	95.38	94.42	72.60	85.71
60-40	178	206	52	41	21	9	17	12	91.28	94.50	71.23	82.00
50-50	177	208	44	42	22	8	25	10	87.62	95.41	66.67	84.00

Table-6.2: Performance Analysis of TPR, FPR, Precision, Measure and Accuracy

Metrics →	TPR		FPR		Precision		Accuracy	
	FELM	Proposed (DESVM)	FELM	Proposed (DESVM)	FELM	Proposed (DESVM)	FELM	Proposed (DESVM)
80-20	95.22	95.69	16.95	8.33	95.22	98.67	92.54	95.15
70-30	95.38	94.42	27.40	14.29	90.29	97.78	89.18	93.28
60-40	91.28	94.50	28.77	18.00	89.45	95.81	85.82	92.16
50-50	87.62	95.41	33.33	16.00	88.94	96.30	82.46	93.28

As far as inferences from the results are concerned, the accuracy of the proposed DESVM is improved and the time taken for feature selection is reduced. It is to be noted that the existing and proposed classifiers are allowed to train first and tested next. Also the performance analysis in terms of TPR, FPR, precision and Accuracy is portrayed in Table-6.2.

Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN + FP}{TN + FP + TP}$
Rate of True Positive	$\frac{FP}{TN + FP}$

(TPR)	
Rate of False Positive (FPR)	
Precision	$\frac{TP}{TP + FP + TN}$
Accuracy	$\frac{TP + TN + FP + FN}{TP + TN + FP + FN}$

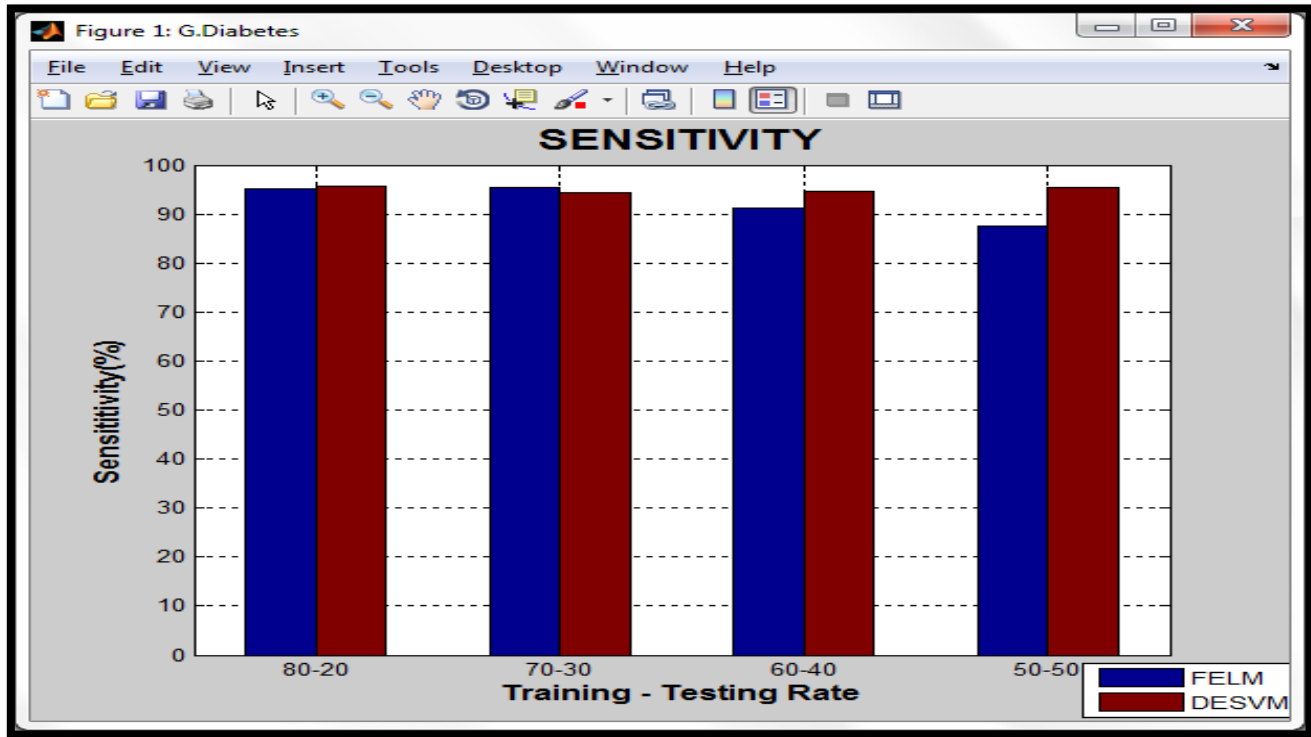


Fig.6.1 Performance Comparison of Sensitivity for FELM Vs R-ABC with varying Training and Testing Rate

Sensitivity is the calculation of actual positives ratio which are identified correctly as positives by the classifier. From the Fig 6.1 it is clearly evident that the proposed classifier DESVM performs better in identifying the positives than the

FELM (Fuzzy-sets and Extreme Learning Machine) [Kindie Biredagn Nahato et al.,2016]. The result values of Fig.6.1 are predicted in Table-6.1.

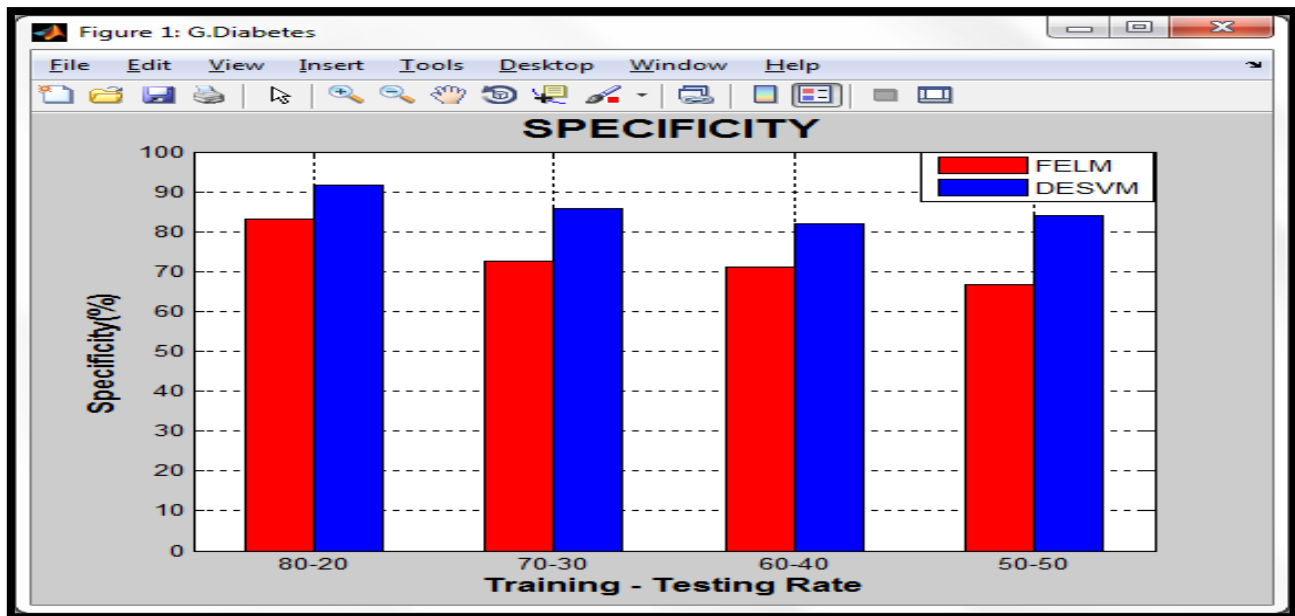


Fig.6.2 Performance Comparison of Specificity for FELM Vs R-ABC with varying Training and Testing Rate

Specificity is the measure of classifier’s ability to identify negative results. From the Fig. 6.2 it can be observed that the proposed mechanism DESVM does not work better in terms of specificity than the FELM (Fuzzy-sets and Extreme

Learning Machine) [Nahato. K. B et al.,2016]. This is due to the degree of relevance mismatch. The result values of Fig.6.2 are predicted in Table-6.1.

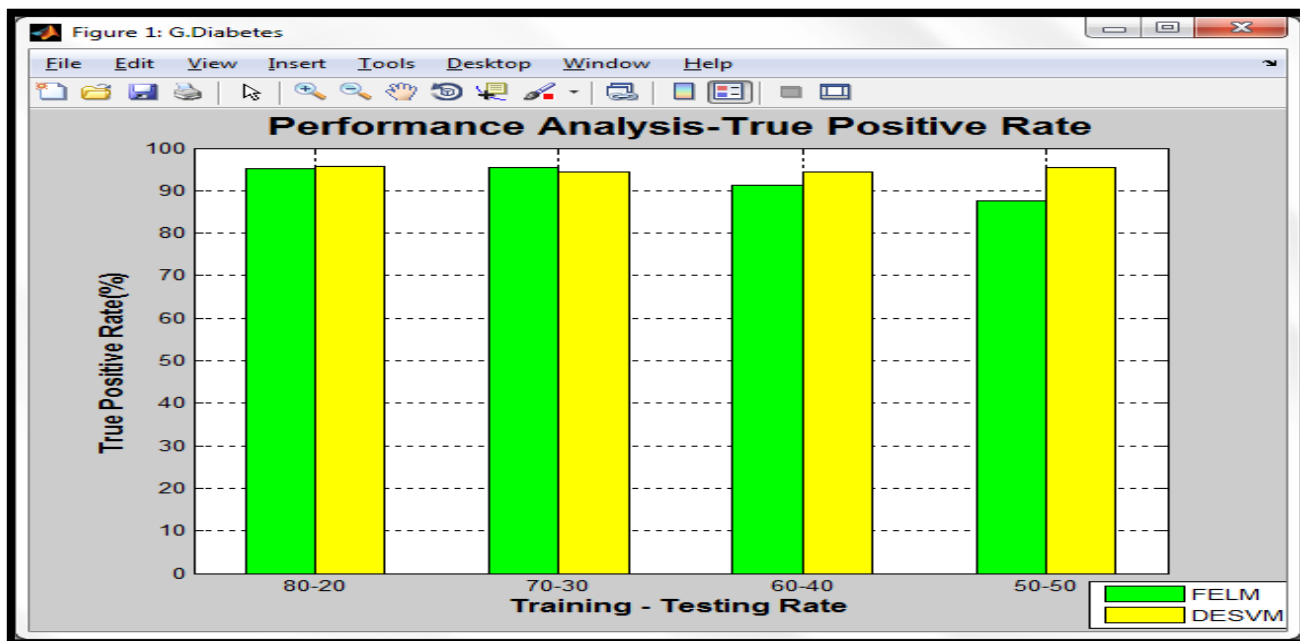


Fig.6.3 Performance Comparison of True Positive Rate for FELM Vs R-ABC with varying Training and Testing Rate

True Positive Rate (TPR) refers to the positives that were correctly labelled by the classifier. From the Fig.6.3, it is evident that the proposed DESVM produces better TPR than

the FELM (Fuzzy-sets and Extreme Learning Machine) [Nahato. K. B et al.,2016]. The result values of Fig.6.3 are predicted in Table-6.2.

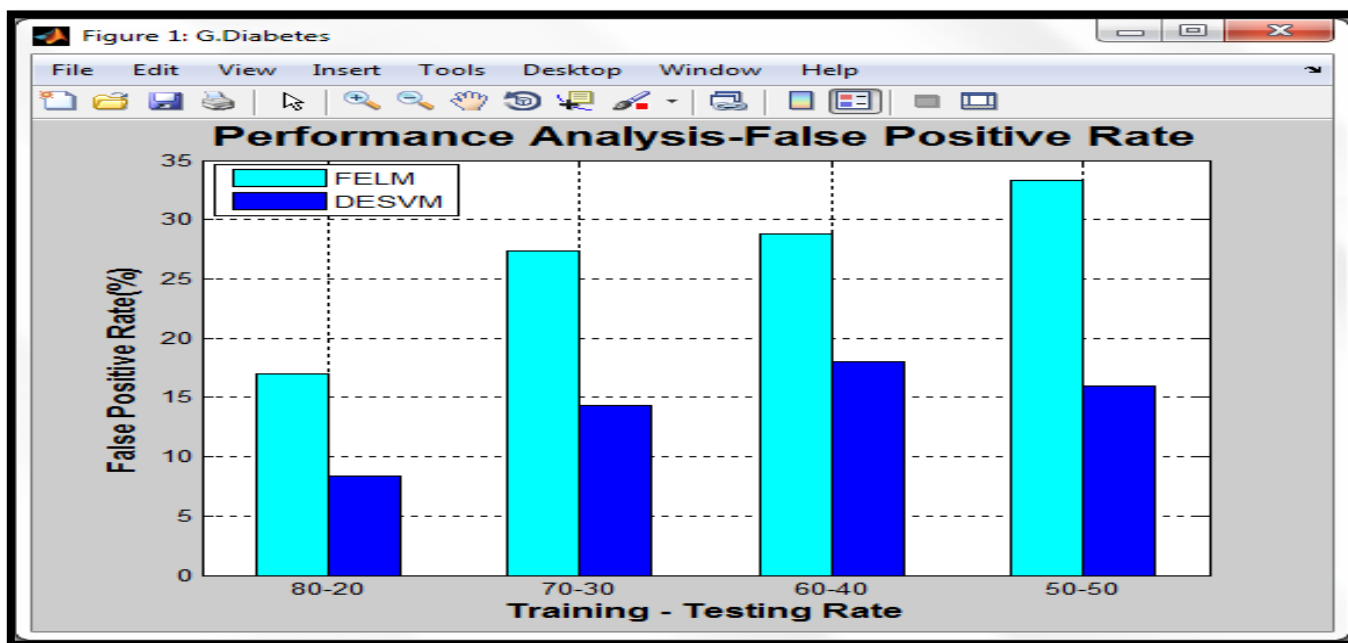


Fig.6.4 Performance Analysis Comparison of False Positive Rate for FELM Vs R-ABC with varying Training and Testing Rate

False Positive Rate (FPR) refers to the negatives that were incorrectly labelled as positive. From the Fig. 6.4, it is clear that DESVM attains the certain degree of relevance mismatch and results in producing a little bit of increase in the false positive rate in certain Training and Testing Rate

when comparing with FELM (Fuzzy-sets and Extreme Learning Machine) [Nahato. K. B et al.,2016]. The result values of Fig.6.4 are predicted in Table-6.2.

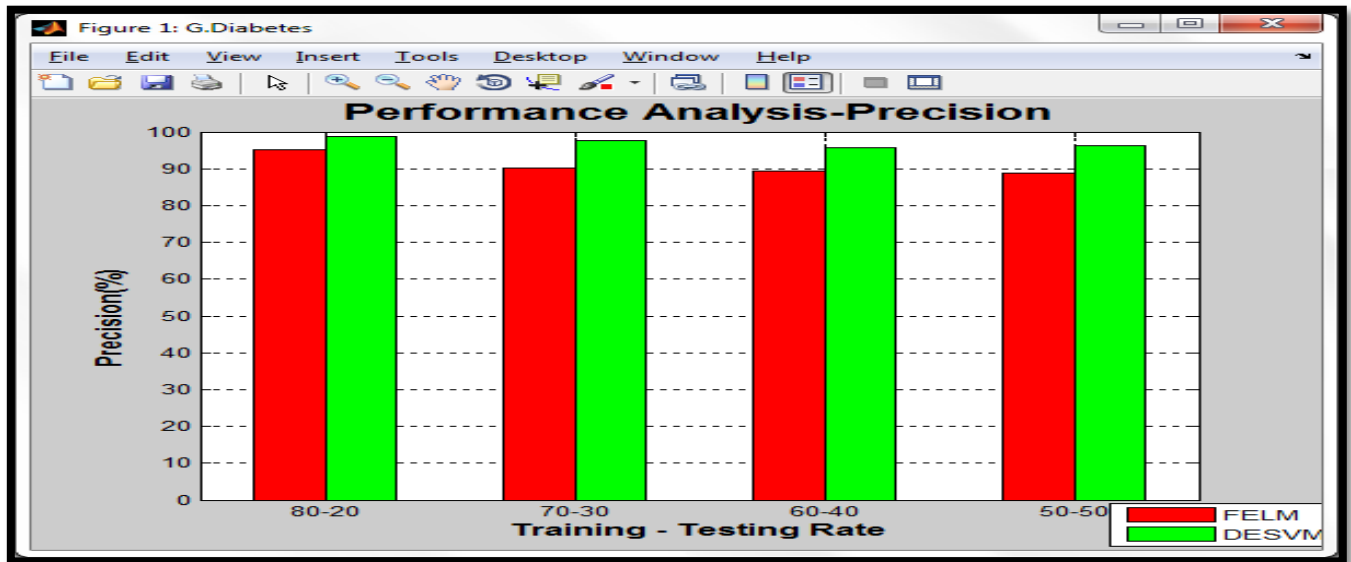


Fig.6.5 Performance Analysis Comparison of Precision for FELM Vs R-ABC with varying Training and Testing Rate

Precision is the measure of accurately predicted positive values to the total predicted positive values. From Fig 6.5 it is clear that, DESVM predicts the accurate positives than the

FELM (Fuzzy-sets and Extreme Learning Machine) [Nahato. K. B et al.,2016]. The result values of Fig.6.5 are predicted in Table-6.2.

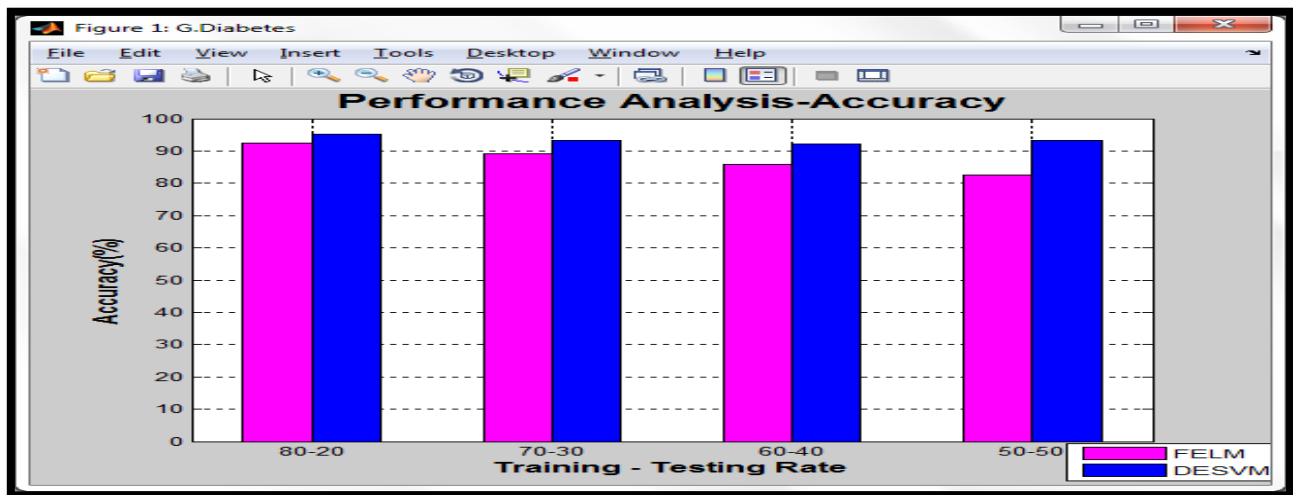


Fig.6.6 Performance Analysis Comparison of Accuracy for FELM Vs R-ABC with varying Training and Testing Rate

Accuracy is the measure of ratio of correctly predicted observation to the total observations. The accuracy result is portrayed in Fig.6.6 and it illustrates that the proposed DESVM harvests better

accuracy than FELM (Fuzzy-sets and Extreme Learning Machine) [Nahato. K. B et al.,2016]. The result values of Fig.6.6 are predicted in Table-6.2.

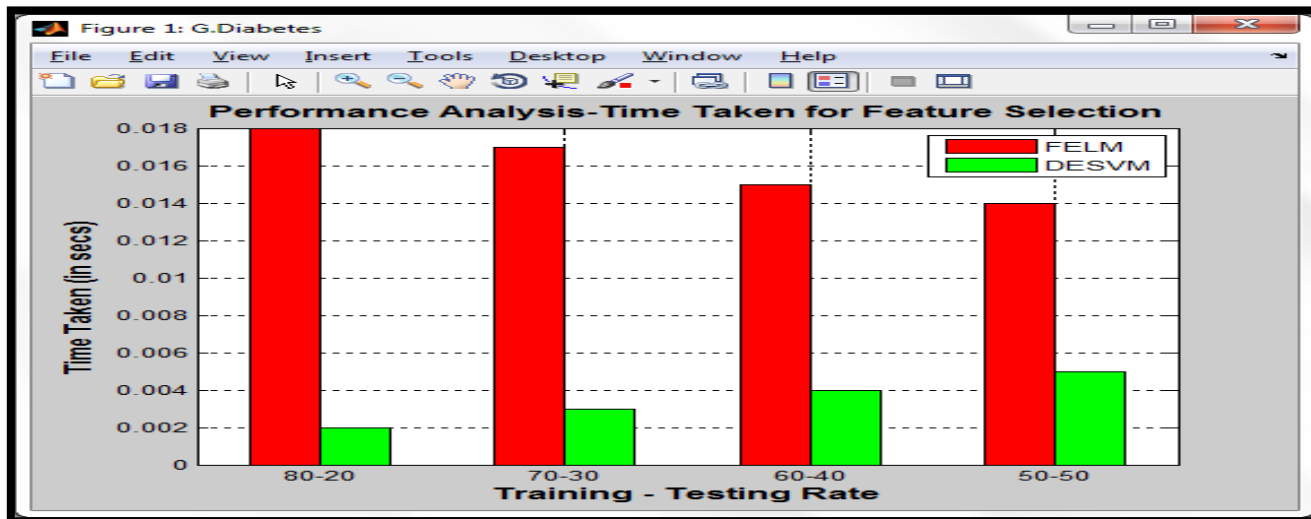


Fig.6.7 Performance Analysis Comparison of Elapsed Time for PCA-SVM Vs R-ABC with varying Training and Testing Rate

Time taken is the measure of finding how much period the algorithm takes for feature selection. Fig 6.7 shows that DESVM took less time for feature selection than the FELM (Fuzzy-sets and Extreme Learning Machine) [Nahato. K. B et al.,2016].

VII. CONCLUSIONS

In this phase of research, in order to reduce the complexity of feature selection and classification, a dual-edged support

vector machine (DESVM) classifier is proposed. Necessary updations are made to meet the objective of the research. With the help of edge esteem the classification is performed. Performance metrics such as sensitivity, specificity, true positive rate, false positive rate, precision, accuracy and time taken for feature selection are taken into account. The simulation results are presented that attained better performance in terms of the chosen performance metrics.

REFERENCES

1. Brisimi. T. S, Xu. T, Wang. T, Dai. W, Adams. W. G, Paschalidis. I. C, Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach, Proc of the IEEE, Vol. 106, No. 4, pp: 690-707, April 2018.
2. Sisodia. S, Dilip. D. S, Prediction Of Diabetes Using Classification Algorithms, Procedia Computer Science, Vol. 132, pp: 1578-1585, 2018.
3. He. H, Garcia. E.A, Learning From Imbalanced Data, IEEE Trans of Knowledge and Data Engineering., Vol. 21, No. 9, pp: 1263-1284, 2009.
4. Kavipriya. S , Deepa. T, Comprehensive Feature Selection Based Support Vector Machine Classifier (CFS-SVM) For Clinical Dataset, Journal of Theoretical and Applied Information Technology, Vol. 96, No 09, pp: 2665 – 2676, 2018.
5. Khine. M. L, Winklestein. A, Copel. J. A, Selective screening for gestational diabetes mellitus in adolescent pregnancies, Obstetrics & Gynecology, Vol. 93, No. 5, pp: 738-742, 1999.
6. Kijanka. P, Qiang. B, Song. P, Carrascal. C. A, Chen. S, Urban. M. W, Robust Phase Velocity Dispersion Estimation of Viscoelastic Materials Used for Medical Applications Based on the Multiple Signal Classification Method, IEEE Trans on Ultrasonics, Ferroelectrics, and Frequency Control, Vol. 65, No. 3, pp: 423-439, March 2018.
7. Lekha. S, Real-Time Non-Invasive Detection and Classification of Diabetes Using Modified Convolution Neural Network, IEEE Journ of Biomedical and Health Informatics, Vol. 22, No. 5, pp: 1630-1636, Sept. 2018.
8. Lekha. S, Suchetha. M, A Novel 1-D Convolution Neural Network With SVM Architecture for Real-Time Detection Applications, IEEE Sensors Journal, Vol. 18, No. 2, pp: 724-731, 15 Jan.15, 2018.
9. Li. H, Wang. Y, Zhan. G, Probabilistic Fuzzy Classification for Stochastic Data, IEEE Trans on Fuzzy Systems, Vol. 25, No. 6, pp: 1391-1402, 2017.
10. Moreira. M. L, W,Rodrigues. J. J. P. C,Kumar. N, Muhtadi. J. A, Korotaev. V, Evolutionary radial basis function network for gestational diabetes data analytics, Journ of Computational Science, Vol. 27, pp: 410-417, 2018.
11. Nahato. K. B, Nehemiah. K. H, Kannan. A, Hybrid Approach Using Fuzzy Sets And Extreme Learning Machine For Classifying Clinical Datasets, Informatics in Medicine Unlocked, Vol. 2, 2016, pp: 1-11.
12. Rasti. R, Rabbani. H, Mehridehnavi. A, Hajizadeh. F, Macular OCT Classification Using a Multi-Scale Convolutional Neural Network Ensemble, IEEE Trans on Medical Imaging, Vol. 37, No. 4, pp: 1024-1034, April 2018.
13. Ruiz. E. C, Saez. G. G, Rigla. M, Villaplana. M, Pons. B, Hernando. M.E, Automatic Classification Of Glycaemia Measurements To Enhance Data Interpretation In An Expert System For Gestational Diabetes, Expert Systems with Applications, Vol. 63, pp: 386-396, 2016
14. Turksoy. K, Roy. A, Cinar. A, Real-Time Model-Based Fault Detection of Continuous Glucose Sensor Measurements, IEEE Trans on Biomedical Engineering, Vol. 64, No. 7, pp: 1437-1445, 2017.
15. Vyas. R, Bapat. S, Goel. P, Karthikeyan. M, Tambe. S. S, Kulkarni. B. D, Application of Genetic Programming (GP) Formalism for Building Disease Predictive Models from Protein-Protein Interactions (PPI) Data, IEEE/ACM Trans on Computational Biology and Bioinformatics, Vol. 15, No. 1, pp: 27-37, 1 Jan.-Feb. 2018.
16. Wang. D, Zhang. D, Lu. G, Generalized Feature Extraction For Wrist Pulse Analysis: From 1-D Time Series To 2-D Matrix, IEEE Jour of Biomed and Health Informatics, Vol. 21, No. 4, pp: 978-985, 2017a.
17. Wang. Y, Li. P, Tian. Y, Ren. J, Li. J, A Shared Decision-Making System for Diabetes Medication Choice Utilizing Electronic Health Record Data, IEEE Jour of Biomedical and Health Informatics, Vol. 21, No. 5, pp: 1280-1287, 2017b.
18. Zhou. W, Wu. C, Chen. D, Yi. Y, Du. W, Automatic Microaneurysm Detection Using the Sparse Principal Component Analysis-Based Unsupervised Classification Method, IEEE Access, Vol. 5, pp: 2563-2572, 2017.