Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7729128

AN EFFICIENT HETEROGENEOUS ENSEMBLE SMOTE BASED LEARNING MODEL FOR DIABETES MELLITUS PREDICTION

Sandeep H[1]

[1]Research Scholar, Department of Information Science & Engineering, Don Bosco Institute of Technology, Bengaluru, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India, Email: sandeep.h12@gmail.com

Dr. B.K Raghavendra[2]

[2]Research Supervisor, Professor and Head, Department of Information Science & Engineering, Don Bosco Institute of Technology, Bengaluru, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India, Email: raghavendra.bk69@gmail.com

Abstract

The primary goal of the current work is to develop a heterogeneous ensemble model for the diagnosis of diabetes in patients using machine learning techniques. The problem of class imbalance is addressed by the proposed paradigm. Various sampling methods, like up-sampling, down-sampling, and the synthetic minority oversampling technique(SMOTE) are used to address the class imbalance problem. Different feature selection techniques, including Ranksum, Univariate Principle Component Analysis (PCA), Logistic Regression (ULOGR), Cross-Correlation Analysis (CRA), Gini Score, and Information Gain (IGFR) are used to identify the relevant features once the preprocessed data is retrieved. On the PIMA dataset, a variety of classification methods, notably LR, SVM, Naive Bayes, Bagging,Adaboost, and PNN are used to classify and predict if a sample is diabetic or not. The results showed that the MVE ensemble learning method combined with SMOTE sampled data yields the maximum performance with 95.81% accuracy and 0.94 as AUC.

Keywords: Machine Learning, SVM, Adaboost , NB, SMOTE.

1. INTRODUCTION

Due to the massive influx of clinical information, data mining on medical datasets is regarded as being among the most difficult fields in the health industry. In the area of clinical data retrieval, finding valuable relevant data and making a diagnosis of the illness have become critical issues. Diabetes is a metabolic condition brought on by hyperglycemia and is widely regarded as a chronic illness. According to a WHO estimate, diabetes claimed the lives of 1.2 million individuals in 2012 and affected 422 million people in 2014. Over the next ten years, it is anticipated that this figure will rise.

A solid CAD solution can be created by comprehending the nature of the data, the intricacy of the processing system, and the eventual demands. In this work, a reliable diabetes mellitus prediction system will be constructed, which is in line with such an optimistic design intention. Today's time constraints, hectic schedules, conserved lifestyles, and lack of physical activity have clearly led to a considerable number of health problems that have an impact on human existence and its quality of life. One of those serious health issues is Diabetic Mellitus, which has prompted both the medical and academic communities to work towards better and quick diagnosis methods [1-3][5].

Being a type of chronic disease, diabetes mellitus is also stated as a silent killer, because its symptoms emerges after a very long time, sometimes in the last stage. Functionally, diabetic mellitus suppresses or reduces the production of insulin in human body that eventually causes high-level of blood sugar. Subsequently, the increased blood-sugar gives rise to a number of health-complications including malfunctioning of the different organs like eyes, kidneys, and nerves. In major diabetic mellitus cases often called the polygenic disease, the exocrine gland doesn't produce hypoglycemic agent that eventually leads to multiple organ failure or malfunction. A recent study referred the statistics by the International Polygenic Disease Federation states that there are 382 million individuals living with polygenic disease globally. Researches reveal that by 2035, this will be doubled to 592 million [5]. In India as well, an total of 8.7% population in the age range of 20-70 years is suffering from diabetes.

To avoid such issues, in the last few years efforts have been made to achieve extremely precise and successful CAD solution for diabetic mellitus identification and treatment. Because of this purpose, in this research, the key focus is made on designing a novel and robust diabetic mellitus detection system for early and optimal diagnosis decisions.This work presents the development of a unique and reliable heterogeneous ensemble paradigm for the identification of diabetes mellitus.The suggested approach, which differs from traditional machine learning approaches, uses a distributed optimization paradigm to handle the class imbalance problem, which is followed by the selection of the most significant features as well as a hybrid feature assessment towards diabetic mellitus detection. Inevitably, a novel heterogeneous ensemble learning method for a two-class classification system is suggested, in contrast to conventional independent ML categorization algorithms.

2. Related Work

Prabhu et al [3] designed a deep belief network (DBN) based diabetes mellitus prediction model. To enhance prediction accuracy, authors suggested pre-processing and data-normalization, especially for Pima Indian diabetic mellitus dataset. As machine learning classifier, authors applied NB, Decision Tree (DT), LR, Random Forest (RF), SVM and DBN, where the later (i.e., DBN) was found superior over other state-of-art techniques. Similarly, Barhate et al [4] applied KNN, LOGR, DT, RF, Gradient Boosting, SVM and Artificial Neural Network (ANN) algorithms to perform diabetic mellitus or Type II diabetic mellitus prediction. Results revealed that amongst the different machine learning models, RF based method exhibited the highest prediction accuracy of 79.7%. For both typical and non-typical instances, the authors proposed a diabetes mellitus prediction model using various ML algorithms like Bagging, LR, and RF.Authors in [6] too applied RF algorithm for diabetic mellitus classification. Woldemichael et al [7] too applied different ML algorithms such as J48, NB, back propagation ANN and SVM for diabetic mellitus classification.

To achieve better efficiency authors proposed 5-fold cross-validation assisted ANN for better accuracy. To further enhance classical ANN based prediction, Kathiroli et al [8] proposed a cascade correlation and ANN based diabetic mellitus prediction model, which classifies it as Type 1 and Type 2 diabetes mellitus. Unlike classical ANN learning, authors applied cascade correlation for ANN formation. In fact, in this approach, their proposed cascade correlation method increased the number of hidden layers sequentially so as to reduce the error soon, without imposing local minima and convergence problem. Here, the prime objective was to increase number of hidden layers so as to extract more deep features that as a result that could help enhance the prediction accuracy. As an explorative effort, Faruque et al [8] applied different machine learning algorithms like SVM, NB, KNN and C4.5 DT algorithms for diabetic mellitus detection. Authors found that the decision tree algorithm C4.5 performs better in terms of prediction accuracy. However, its efficacy over the complex and large non-linear inputs seems limited. Rubaiat et al [9] too applied multi-layer perceptron ANN (ANN-MLP) for diabetic mellitus detection or prediction over Pima Indian dataset. Before performing classification, to enhance computation as well as accuracy, authors applied k-means clustering algorithm.

As classifier, authors applied Random Forest, Logistic Regression and MLP-ANN algorithms. Authors identified MLP-ANN as superior model which could achieve the maximum classification accuracy of 77%. However, the performance observed can't be stated as an optimal solution towards diabetic mellitus detection purpose, which demands higher accuracy to ensure early and optimal diagnosis. The authors of [10] used an ML method to perform Spectral categorization of blood glucose levels.To perform classification, authors applied random forest algorithm (RF) followed by SVM. This approach yields prediction accuracy of 67.5%, which can't be recommended for real-world CAD solution. To further enhance the performance authors applied Principle Component Analysis (PCA) based feature selection followed by SVM based classification. Authors found that this approach performs better (accuracy 77.5%) than the classical RF based model (67.5%). Alassaf et al [19] developed machine learning assisted pre-emptive diagnosis model of diabetes mellitus detection and classification. authors applied different machine learning models like ANN, SVM, NB, k-NN algorithms for two-class classification. In their work, ANN based prediction was found superior over SVM, NB and k-NN with 77.5% prediction accuracy.

3. PROPOSED SYSTEM

The following four stages are included in the overall proposed diabetic mellitus prediction model in terms of functionality. Data Collection comes first followed by Pre-processing and Sampling, followed by pattern Evaluation and finally Two-Class Categorization.

.

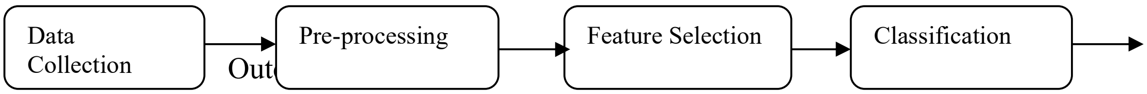

Fig 1. A schematic of the Proposed system

Considering the data under study and its element wise information, we observed that the prevalence of diabetes was merely (approximate) 9%, signifying the skewness of the class priors. Alternatively, the class-priors in the dataset that was taken for consideration were heavily biased in favor of the Non-Diabetic class.In this scenario, developing a system with roughly 90% of positive class outputs may eventually result in any prediction output being positive, increasing the likelihood of a false alarm.

Unfortunately, no significant researches address such limitations. Considering such limitations, in this paper we used pre-processing methods such data sub-samplingto address the issue of skewed class problem. Here, we have used sub-sampling techniques to select a representative sample for the majority class (non-diabetic subject). The sampling technique can actually solve the class imbalance issue that prevents ML models from doing effective training and delivering improved precision [1]. In order to address the aforementioned issues, we used three different types of sampling techniques: up-sampling, down-sampling, and SMOTE.

4.1 SMOTE

SMOTE selects the minority sample instances randomly and then looks for K nearest instances in the minority neighbors, once the instances are selected they connect with the feature space instances and form a line segment between the same.

The SMOTE algorithm is described in detail below:

1. Look for the k-nearest neighbors from minority class.
2. Select the new instances using the formula

New instances = original samples + difference * gap (0,1).

3. Add the new instances to the minority samples set. The new synthetic instances generated might be accepted or rejected based on two conditions:

- ***Compute the distance 'dm1' from synthetic samples to Minority samples Sm1.***
- ***Compute the distance 'dm2' from synthetic samples to Majority samples Sm2.***
- ***Then the select the min{dm1,dm2} and accept the synthetic samples.***

4. Finally, a new balanced dataset is created.

5. Feature Selection

The pre-processing and sub-sampling phase of work yields four distinct set of data, Up-Sampled data, Down-Sampled data, SMOTE data and Original data. This is the matter of fact that amongst these data (within each sub-sampled data), there can be certain insignificant elements that don't have significant impact on final prediction. Additionally, such insignificant data-elements could worsen the prediction performance. Alleviating such limitations demand, certain feature selection technique(s).Thus, we applied different types of feature selection methods like CRA, PCA, GSFR and IGFR.

5.1 CRA

A mathematical technique called cross-correlation analysis is used to show how closely two variables are related or correlated. It also identifies or indicates the association's strengths and direction. The intensity of the connection or association is often within a range, from plus or minus, 1 to 0.The value closer to 1 signifies higher association. In our proposed model, we applied Pearson correlation based CRA assessment. Mathematically, Pearson correlation is presented as per (4).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4}$$

Those feature instances with correlation coefficient higher than 0.5 were retained while the remaining were dropped from further computation. To be noted, these all feature selection methods were applied on each sample (i.e., original data, up-sampled data, down-sampled data and SMOTE data).

5.2 PCA

We applied PCA as one of the feature selection methods for diabetic mellitus prediction. To accomplish the above, we got the principal component and eigenvalues of the covariance for every selected feature and related occurrence. Here, we calculated the separation between each feature occurrence and the average principal component (here, 0.5).The cases with lower Eigen distances were kept since it was important for prediction whereas the instances with higher distances were eliminated. By using the aforementioned PCA feature selection strategy, we were able to keep a set of important characteristics for use in two-class categorization.

5.3 GSFR

The degree of impurity in a data set is determined by the Gini Score, commonly referred as the Gini Index, using (5).

$$m(s) = \sum_{i \neq j} \widehat{P_{si}}\,\widehat{P_{sj}} = 1 - \sum_{j} \widehat{P_{sj}} \tag{5}$$

The GSFR strategy is primarily used to generalize the differential defects, which describe the variability of a distribution relating to two class labels (i and j). It is also referred to as the expected rate of failure when a class label is randomly selected from the feature group. This feature selection method is a suitable solution to the problem because the imperfection criterion for this entropy-based method considerably surged at the same frequency.In our proposed GSFR model, we apply (6), to measure probability of a feature variable to remain in the feature subset.

$$GSFR(S) = 1 - \sum_{i=1,\ldots,m} P_i^2 \tag{6}$$

In above equation (6), the parameter P_i signifies the probability that a tuple in feature set S belongs to the class C_i. We obtained the value of P_i as per $|C_i, S|/|S|$.

5.4 IGFR

Information gain is mathematically defined as per (7).

$$IG(t) = -\sum_{i} \Pr(c_i) log \Pr(c_i) + \Pr(t) \sum_{i} \Pr(c_i|t)\, log \Pr(c_i|t) + \Pr(t) \sum_{i} \Pr(c_i|t)\, logPr(c_i|t) \tag{7}$$

Using this approach we suggested for feature selection, a term-goodness parameter to evaluate the importance of a case or attribute value for the prediction of diabetes mellitus. By determining if a term is included in the data corpus or feature set, we were able to calculate the number of bits required for diabetes mellitus prediction.Thus, performing above stated feature selection methods, we obtained the different feature set to be further used for classification. Interestingly, we focused on discovering the most effective or most suitable computing platform for the identification and classification of diabetes mellitus in this study.In order to achieve this, we evaluated each of the aforementioned features in terms of their potential to produce better precision.In addition, we developed a new feature vector by concatenating different features (SPA, ULOGR, CRA, PCA, GSFR and IGFR) and named it as “All Matrix” (AM). Thus, the total number of features prepared for diabetic mellitus prediction were seven.

6.HeterogeneousDeep Ensemble Learning

As stated, the problem of diabetic prediction is a two-class classification where the machine learning model is supposed to predict each patient or subject as “Diabetic” or Non-Diabetic”. Noticeably, authors have used a particular algorithm for machine learning as a solitary classifier in the most common ways to classify diabetes mellitus; nevertheless, Optimization techniques behave significantly, or even various individuals' responses to the identical input,indicating a variety of performances. It would be inappropriate to generalize one method as the best one in this situation. The concept of ensemble learning can be very important in preventing such problems.Employing multi-party consensus concept where an ensemble learning model could employ multiple base classifiers to label each subject as “Diabetic” or Non-Diabetic” can be a viable solution.

The HEL model are given as follows:

1. Logistic Regression (LOGR)
2. Bagging Classifier
3. AdaBoost Classifier
4. Support Vector Machine (SVM)
5. Probabilistic Neural Network
6. Navie Bayes

6.1 Logistic Regression

The regression method known as logistic regression (LR) is frequently used in text categorization and data analysis. Regression is carried out using LOGR over the multiple independent factors and the dependent variable in our proposed diabetes mellitus categorization problem. This is the way the regression analysis segregates each sample as either diabetic or not. In order to execute linear regression over the input features in our suggested LOGR approach, we used (8).

$$\mathrm{logit}[\pi(x)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \ldots . + \beta_m X_m \qquad (8)$$

Here (8), the logit method is used, with the independent parameter being x_i and the dependent parameter being logit[(x)]. The above-mentioned statistical framework uses the logit method to transform the binary outputs between 0 to 1 to -∞ to +∞. The total number of independent variables is shown in equation (9) in accordance with the input Pima Indian diabetes dataset. Thus, it is possible to determine if an individual is "Diabetic" or "Non-Diabetic" based on the parameter (9).

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \ldots + \beta_m X_m}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \ldots + \beta_m X_m}} \qquad (9)$$

6.2 Bagging Classifier (BC)

By applying basic classifiers one at a time to arbitrary subsets of the initial dataset, a bagging classifier, an ensemble meta-estimator, aggregates the individual estimates into a final prediction. By including randomization in the creation of the meta-estimator, it is widely used to lower the variable resistance of a black-box estimator. Voting or bagging reduces variance, which suggests overfitting by average; nevertheless, bagging increases bias, which is mitigated by lowering variance.

6.3 AdaBoost Classifier (AC)

Machine learning ensemble approaches use the AdaBoost algorithm, also known as adaptive boosting, as a boosting technique. Adaptive boosting refers to the process of reassigning weights to each instance, with heavier weights going to instances that were mistakenly categorized. In supervised learning, boosting is used to reduce noise and variable problems.

6.4 Probabilistic Neural Network

PNN represents a variant of feedforward neural network, often used for classification problems. In our proposed PNN model, the parent probability distribution function (PDF) of each subject-class or patient is approximated by means of a Parzan window and a non-parametric function. Subsequently, employing PDF of each subject class label, the likelihood of a new input is obtained as per Bayes rule. Here, the use of Bayes rule helps in assigning class the highest posterior probability to the new input. This method reduces probability of mis-detection or mis-classification significantly. Structurally, our proposed PNN model was derived from Bayesian network in conjunction with Kernel Fisher discriminant analysis [1].

Structurally, it is designed as a multi-layered feedforward network with four layers; input later, pattern layer, summation layer and output layer. Here, the first layer estimates the distance from the input vector to the training input vector. This as a result generates a vector where each element signifies how close the input is to the training input. Similarly, the second layer adds or performs summation of the contribution of each class-label of the input and eventually generates its output as a probability vector. Eventually, a transfer function is applied on the output of the second layer and thus selects the maximum of the probability vector and generates 1 for positive class (i.e., Diabetic Mellitus) and 0 for negative (i.e., Non-Diabetic" or other chronic disease.The mathematical model for each Gaussian centered on the corresponding class 1 and class 2 point be $x^{(p)}$ and $y^{(q)}$ (it signifies feature vector for N-dimensional vector) for any input vector x be (20) and (21).

$$g_1(x) = \left[\frac{1}{\sqrt{2\pi\sigma^{2^N}}}\right] exp\left\{-\frac{-\left\|x - x^{(p)}\right\|^2}{(2\sigma^2)}\right\} \tag{20}$$

$$g_2(y) = \left[\frac{1}{\sqrt{2\pi\sigma^{2^N}}}\right] exp\left\{-\frac{-\left\|y - y^{(q)}\right\|^2}{(2\sigma^2)}\right\} \tag{21}$$

6.5 Support Vector Machine

SVM is a standard algorithm that belongs to supervised learning techniques [6].SVM classifier dominates when compared to other classification techniques. For a given dataset, SVM builds a predictive model that classifies new samples by using hyperplane [3] to separate the data samples into different classes.The main aim is to draw the hyperplane in such a way that it maintains the maximum distance between the hyperplane and data samples of either class.

SVM is classified into two type'slinear and non-linear hyperplanes [5].The region between the two hyperplanes is called a margin with a halfway maximum distance and these hyperplanes are defined by the equations w.x - b=1 if data points lie above or on the boundary with class label =1 and w.x-b= -1 if data points lie below or on the boundary with class label = -1. The distance between the two hyperplanesis $2/\| w \|$ which can be maximized by minimizing $\| w \|$. A non-linear SVM model can be built by minimizing the following equation, which is described as:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(\vec{w}\cdot\vec{x}_i - b)\right)\right] + \lambda\|\vec{w}\|^2 \tag{2}$$

Where max ( ) method is equal to zero if x_i lies on the correct margin side and λ determines the trade-off between increased margin size andx_i

6.6 Naïve Bayes

NB classifier belongs to condition probability, which is based on Bayes theorem [10], it involves self-standing assumptions with independent features. Bayes theorem [4] is defined mathematically as

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}\cdot \tag{3}$$

Where,

P (A) is prior probability the hypothesis that B belongs to some class

B is the evidence (data sample) of unknown class label

P (A/B) is the posterior probability which the hypothesis holds for given observed data B

P (B/A) is the likelihood

6.7 Maximum Voting Ensemble (MVE)

In this paper, we have applied above discussed 6 machine learning models as the base classifier to constitute a n advanced MVE with Bagging Technique[1]. To form ensemble structure all classifiers are run over the same dataset and predicts each subject as “Diabetic” and “Non-Diabetic” and thus labels them as “1” and “0”, respectively. Thus, obtaining class outputs (i.e., label) of each subject or patient we obtained the maximum voting score. The category with the best score 1 or 0—is deemed to have the higher score in the end.In other words, amongst the 10 base-classifiers, if 6 classifiers predict a subject or patient as “Diabetic” with label 1, then the subject is finally predicted as “Diabetic”. Similarly, if minimum 6 classifiers predict a subject as “Non-Diabetic” and label it as “0”, that subject is finally predicted as “Non-Diabetic”. Thus, based on consensus each subject in the data is classified as diabetic or non-diabetic.

7.Results and Discussion

Here, the fundamental goal is to find the best computing framework for predicting diabetes mellitus.Considering the dataset named Pima Indian we designed overall prediction model could address all at hand issues such as data imbalance, lack of suitable feature selection and prediction generalization. Though, Pima Indians diabetic dataset comprises different features however, limited number of samples and even uneven sample distribution could limit its suitability towards machine learning based prediction. Exploring in depth we found that almost 9% of the subjects or the samples were positive while rest were negative and, in this case, (often called skewed class problem or unbalanced data) merely training a machine learning model can give false prediction results or the prediction output more inclined towards negative class. To alleviate such problem, we applied sub-sampling concepts.

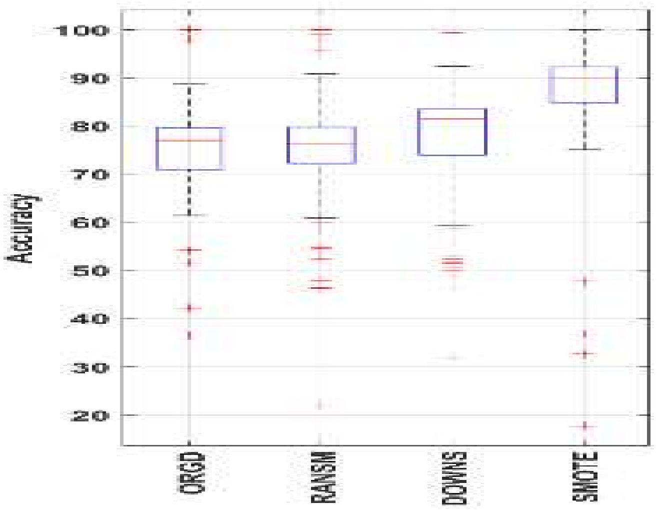

Fig. 1 Accuracy (%) using various sampling methodologies

The precision, F-measure, and AUC performance for various sampling strategies are shown in Figures 1 through 3.Observing results (Fig. 1-Fig 3), it can easily be found that the accuracy with SMOTE sampling method is the highest (90%), while the other samples such as random samples and down-sampled data yields accuracy of 78% and 84% accuracy, respectively. Interestingly, the original sample (without sub-sampling) yielded the accuracy of 78%. These results affirm the factuality of [7][12] whose prediction accuracy were less than 78%. Similarly, F-Measure with original data was 0.84, followed by down-sampled data (0.81), SMOTE (0.78) and random sampled data (0.76). Remarkably, SMOTE data's AUC efficiency is superior (0.92) to other chosen data (i.e., Original data (0.80), Random data (0.82) and Down sampled data (0.85)).

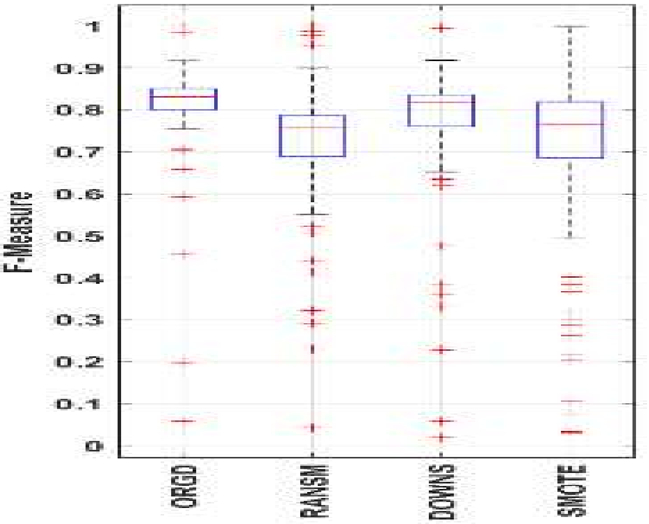

Fig. 2 F-Measure performance with the different sampling techniques

The advantage of SMOTE over other sampled data is supported in light of the aforementioned findings. The accuracy effectiveness using various feature selection techniques is shown in

Figure 4. For the present work, a variety of feature selection techniques are implemented, including SPM, ULOGR, CRA, PCA, GSFR, and IGFR.IGFR.

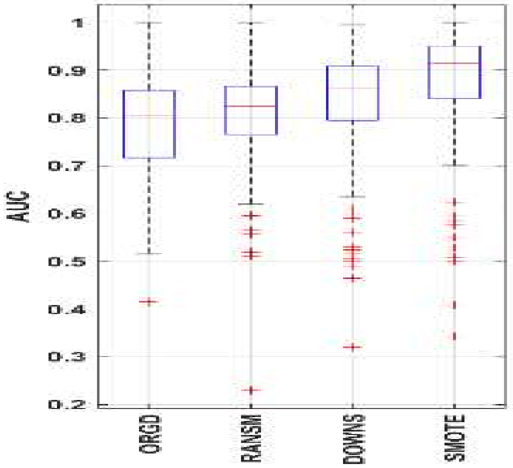

Fig. 3 Performance of AUC using various sampling algorithms.

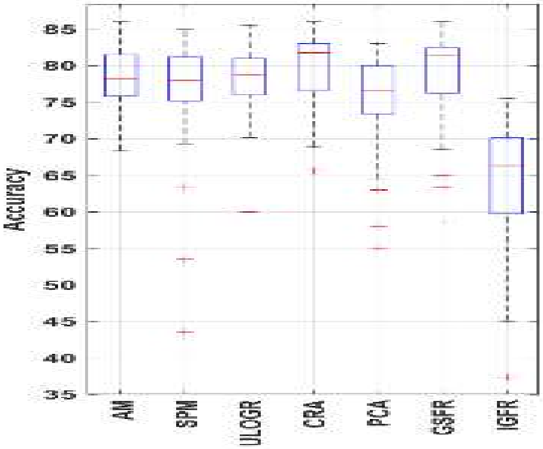

Fig. 4 Accuracy results using various feature selection strategies.

A comparative result analysis of Machine learning Ensemble classifiers such as NB,BC,,PNN,SVM with MVE are given in Table 1 to Table 3.

TABLE I AUC WITH ORIGINAL DATA

AUC performance with Original Data							
Data	Feature Selectio	SVM	NB	BC	AdaBoost	PNN	MVE
ORGD	CRA	0.8	0.8	0.8	0.8	0.8	0.8
ORGD	PCA	0.75	0.75	0.75	0.75	0.75	0.75
ORGD	GSFR	0.8	0.8	0.8	0.8	0.8	0.8
ORGD	IGFR	0.6	0.6	0.6	0.6	0.6	0.6

TABLE II AUC WITH SMOTE-SAMPLED DATA

AUC with SMOTE Data							
Data	Feature Selectio	SVM	NB	BC	AdaBoo st	PNN	MVE
SMOTE	CRA	0.93	0.92	0.91	0.89	0.88	0.93
SMOTE	PCA	0.91	0.91	0.89	0.91	0.83	0.91
SMOTE	GSFR	0.93	0.89	0.9	0.87	0.84	0.94
SMOTE	IGFR	0.8	0.79	0.62	0.7	0.82	0.8

TABLE III ACCURACY (%) WITH ORIGINAL DATA

Accuracy (%) with Original Data (%)							
Data Data	Featur e	SVM	NB	BC	AB	PNN	MVE
ORGD	CRA	84.41	74.02	64.28	74.02	75.32	75.97
ORGD	PCA	79.87	72.72	74.67	73.37	75.97	72.72
ORGD	GSFR	83.11	70.12	80.51	72.07	79.87	70.77
ORGD	IGFR	69.93	59.47	62.74	66.66	65.35	66.23

TABLE IV ACCURACY (%) WITH SMOTE SAMPLED DATA

Accuracy (%) with SMOTE sampled Data (%)							
Data Data	Feature Selecti	SVM	NB	BC	AB	PNN	MVE
SMOTE	CRA	89.53	93.86	88.12	90.25	0.88	91.1
SMOTE	PCA	92.78	92.06	90.14	89.05	0.83	91.07
SMOTE	GSFR	95.18	93.06	89.81	92.18	0.84	95.81
SMOTE	IGFR	84.14	87.38	80.14	80.14	0.82	87.14

With regard to the general performance, it can be seen that SMOTE tested data with GSFR properties coupled with MVE algorithms delivers the best results, with 95.81% accuracy and 0.94 as AUC.

8.CONCLUSION

The present work proposed a predictive model, which encompasses ensemble methodology for diabetic mellitus prediction.To address the key limitations such as data imbalance,low performance interms of accuracy and AUC. In this paper, different base classifiers were applied to alleviate the data imbalance problem and subsequent improvement in accuracy, to achieve this different sampling techniques were applied.SMOTE exhibited better accuracy and AUC performance.Subsequently, we applied suitable set of features for further computation and considered different algorithms like CRA, PCA,GSFRand IGFR were applied.

It's interesting to see that only the GFSR and CRA feature selection techniques were able to produce results that were satisfactory. A heterogeneous ensemble model was created utilizing SVM, NB,Bagging, Adaboost and PNN as the base classifier in order to retrieve the appropriate set of features from the various feature selection methods. To summarise, the MVE ensemble model was created by strategically using a total of 6 base classifiers. The entire performance evaluation showed that the MVE ensemble learning method combined with SMOTE sampled data with GFSR features provides the maximum performance with 95.81% accuracy and 0.94 as AUC.

Acknowledgment

We sincerely appreciate the support given to this research effort by the VTU, Belagavi. We also appreciate the helpful recommendations from Dr. Nataraj K R, Dean & Director (R&D), DBIT, Bengaluru.

References

1. Jimsha K Mathew and Sathyalakshmi S, ExpACVO-Hybrid Deep learning: Exponential anti corona virus optimization enabled hybrid deep learning for tongue image segmentation towards diabetes mellitus detection, Biomedical Signal Processing and Control, Vol. 83, January 2023, pp. 1-13

2. Shahid Mohammad Ganie and Majid Bashir Malik, An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators, Healthcare Analytics, Vol. 2, August 2022, pp. 1-14.

3. Sandeep Honnurappa and Bevoor Krishnappa Raghavendra. A Highly Robust Heterogenous Deep Ensemble Assisted Multi-Feature Learning Model for Diabetic Mellitus Prediction [J]. Int J Performability Eng, 2021, 17(11): 926-937.

4. Muhammad Waqar, Hassan Dawood ,Hussain Dawood , Nadeem Majeed,Ameen Banjar and Riad Alharbey"An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction," Hindawi Scientific Programming Volume 2021, Article ID 6621622, 12 pages.

5. M. T. Islam, M. Raihan, F. Farzana, M. G. M. Raju and M. B. Hossain, "An Empirical Study on Diabetes Mellitus Prediction for Typical and Non-Typical Cases using Machine Learning Approaches," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-7.

6. K. VijiyaKumar, B. Lavanya, I. Nirmala and S. S. Caroline, "Random Forest Algorithm for the Prediction of Diabetes," 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2019, pp. 1-5.

7. M. Habibullah, M. A. M. Oninda, A. N. Bahar, A. Dinh and K. A. Wahid, "NIR-Spectroscopic Classification of Blood Glucose Level using Machine Learning Approach," 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE), Edmonton, AB, Canada, 2019, pp. 1-4.

8. Ahmed Saad Hussein, Tianrui Li, Wondaferaw Yohannese Chubato, Kamal Bashir, "A-SMOTE: A New Preprocessing Approach for Highly Imbalanced Datasets by Improving SMOTE" International Journal of Computational Intelligence Systems 2019.