Automated Detection of Type 2 Diabetes with Imbalanced and Machine Learning Methods

G. Anirudh and Upasana Talukdar

1 Introduction

Diabetes is one of the most common chronic diseases affecting around 415 million people around the world. Early diagnosis and prediction of diabetes can suppress its effects and can prevent long-term complications. In the past few years, literature reported many works on the prediction of diabetes using machine learning algorithms, tested on PIMA dataset,[1] one of the most widely used diabetes datasets in literature [1–3]. However, such datasets are imbalanced. Class imbalance problem can be defined as having an unequal distribution of the data. Such a problem poses a challenge in detecting and extracting diabetic patterns. Because of the dominance of one class, existing machine learning algorithms may fail to detect diabetic cases accurately. Nnamoko and Korkontzelos [3] proposed a two-step data pre-processing approach on PIMA Dataset, where the first step identified the outliers using the Interquartile Range (IQR) algorithm and the second step employed Synthetic Minority Oversampling Technique (SMOTE).

This paper aims to find the best machine learning model for predicting diabetes with an imbalanced source. In this process, this research work presents rigorous experimentation in three categories: category 1: experiments with classification algorithms, category 2: experiments with ensemble methods, and category 3: experiments with imbalanced data pre-processing (different undersampling, oversampling, and

[1] https://www.kaggle.com/uciml/pima-indians-diabetes-database.

G. Anirudh (✉)
Department of Data Science and Analytics, Central University of Rajasthan, Ajmer, Rajasthan, India
e-mail: ganirudhani90@gmail.com

U. Talukdar
Department of Computer Science and Engineering, Indian Institute of Information Technology Guwahati, Guwahati, India

R. Doriya et al. (eds.), *Machine Learning, Image Processing, Network Security and Data Sciences*, Lecture Notes in Electrical Engineering 946,
https://doi.org/10.1007/978-981-19-5868-7_3

combination techniques) and classification algorithms. Undersampling, oversampling, and combination are the techniques to adjust the class distribution of data. Undersampling down-sizes the majority class by removing observations, oversampling over-sizes the minority class by adding observations, while in combination methods, the data is oversampled and then the transformed data is undersampled.

The performance of the solutions has been evaluated using six different metrics: $F1$-score, Precision, Recall, Area Under Receiver Operating Characteristic curve (AUROC), Area Under Precision-Recall curve (AUPR), and Classification Accuracy (Accuracy). Experimental results show that the amalgamation of imbalanced data pre-processing methods improves the performance of traditional machine learning classifiers achieving the best accuracy as 98.49%. The results are compared with the existing methods in the literature. The proposed model yields better performance in terms of accuracy as compared to all other existing methods. Besides, we examined the validity of our proposed model in other domains (not related to healthcare) with the credit card dataset that exhibits high-class imbalance.

2 Related Works

The health sector has been showing impeccable growth in terms of technology, with the use of machine learning and deep learning. Few notable contributions are, detection of lung cancer [4, 5], dermatoscopic melanocytic skin lesion segmentation [6], lung segmentation [7, 8], and diabetes detection [1–3]. One common problem with methodologies for dealing with such data is the class imbalance. Literature reported many ways to tackle the class imbalance problem in various domains. Common approaches for handling class imbalance are undersampling and oversampling techniques or a combination of both.

Undersampling: Different methods under undersampling techniques can be categorized as *a. Methods that select the samples to keep*: Near Miss [9], and Condensed Nearest Neighbor Rule [10], *b. Methods that select the samples to delete*: Tomek Links [11], and Edited Nearest Neighbors [12], *c. Combinations of keep and delete methods*: One-Sided Selection [13], and Neighborhood Cleaning Rule [14].

Oversampling: In the similar way, oversampling methods can be categorized into different methods: a. Synthetic Minority Oversampling Technique [15], b. Borderline-SMOTE [16], c. Borderline-SMOTE SVM [17], d. Adaptive Synthetic Sampling (ADASYN) [18], where (b), (c) and (d) are extensions of (a).

Combination of Undersampling and Oversampling: In the Combination family, we combine oversampling methods and undersampling to make it more effective. A few examples of effective combinations are: (i) SMOTE and Tomek Links [19], and (ii) SMOTE and Edited Nearest Neighbors [20].

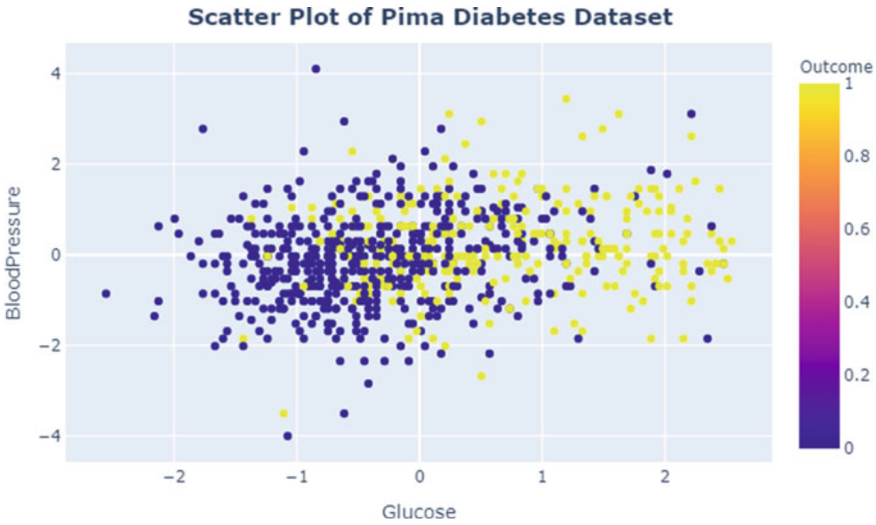


Fig. 1 Scatter plot of PIMA dataset

3 Materials and Methods

3.1 Dataset Description

PIMA dataset[2] is used in this paper for analysis. It has a total of 768 samples with 9 features. The class ratio of diabetic to non-diabetic is 0.34:0.66 (see Fig. 1). The yellow dots and the blue dots in the scatter plot represent the diabetic cases and non-diabetic cases.

3.2 Feature Engineering

The data contains 0 as a measurement for certain features. The pregnancy column in the dataset containing 0 indicates that the woman is 0 times pregnant. Age and Diabetes Pedigree Function are continuous attributes. Apart from Outcome, Age, Diabetes Pedigree Function, and Pregnancies, the rest of the features containing 0 are assumed to be missing observation. The assumed missing values are replaced with the median since the median is not affected by extreme values.

[2] https://www.kaggle.com/uciml/pima-indians-diabetes-database.

3.3 Experimental Setup

This paper reported rigorous experimentation to tackle the class imbalance problem and adopted various state-of-the-art methodologies to attain better performance in the prediction of diabetic cases. The experiments were conducted in three different categories.

- **Category 1—Experiments with traditional machine learning algorithms**: In this category, five different supervised classification algorithms are employed on our dataset, which includes Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbour (KNN), and Deep Neural Network (DNN).
- **Category 2—Experiments with ensemble machine learning algorithms**: In this category, five different ensemble algorithms are applied to our dataset, which includes Bagging, Random Forest, AdaBoost, Gradient Boosting, and XGBoost.
- **Category 3—Experiments with imbalanced data pre-processing and traditional machine learning methods**: Here, to tackle the class imbalance problem different undersampling techniques, oversampling techniques, and a combination of both have been employed before feeding it to the machine learning algorithms. The undersampling techniques like Random Undersampling (RU), Near miss-1 [9], Near miss-2 [9], Tomek Links [11], Edited Nearest Neighbors (ENN) [12], and One-Sided Selection [13] are employed. On the other hand, oversampling techniques like Random Oversampling (RO), Synthetic Minority Oversampling Technique (SMOTE) [15], Borderline SMOTE-1 [16], Borderline SMOTE-2 [16], SVM-SMOTE [17], and Adaptive Synthetic Sampling (ADASYN) [18] are employed in this study. From the experiments with undersampling and oversampling methodologies, the two best oversampling and undersampling methods are picked up. The combination of the two methods is then investigated.

4 Experimental Results and Discussions

4.1 Category 1: Experiments with Traditional Machine Learning Methods

Five supervised classification algorithms were applied that include Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbour (KNN), and Deep Neural Network (DNN), with class weights as 0.34 and 0.66 for class 0 and class 1, respectively.

In the KNN, $k = 7$ is taken as with $k = 7$ kNN performed best. For DNN, the architecture is built with 3 layers with 5, 8, and 1 unit of nodes, and Rectified Linear Unit (ReLU) is used as activation function We have used Adam optimizer, with batch size 32, and the number of epochs 20.

Table 1 Performance of traditional machine learning algorithms

Algorithm	AUROC	AUPR	$F1$-score	Precision	Recall	Accuracy
Logistic Regression (LR)	0.85	0.73	0.72	0.8	0.80	0.79
K-Nearest Neighbour (KNN) (7)	0.75	0.64	0.58	0.70	0.71	0.70
Support Vector Machine (SVM)	0.85	0.72	0.72	0.79	0.79	0.79
Decision Tree (DT)	0.65	0.65	0.57	0.58	0.56	0.68
Deep Neural Networks (DNN)	0.77	0.46	0.46	0.74	0.72	0.71

Table 2 Performance of ensemble methods

Algorithm	AUROC	AUPR	$F1$-score	Precision	Recall	Accuracy
Bagging	0.78	0.67	0.51	0.72	0.72	0.72
Random Forest	0.82	0.72	0.64	0.75	0.76	0.75
AdaBoost	0.77	0.64	0.60	0.73	0.74	0.74
Gradient Boosting	0.80	0.68	0.64	0.75	0.76	0.75
XGBoost	0.82	0.71	0.65	0.76	0.77	0.76

The performance of different classification algorithms has been illustrated in Table 1. It is seen that, in terms of all the six evaluation metrics, LR performed best while SVM performed second best. DNN also gave a comparable performance in terms of AUROC, Precision, Recall, and Accuracy. However, in terms of AUPR and $F1$-score, DNN performed worst. Besides that, DNN requires further fine-tuning of hyper-parameters and implementing it after an appropriate pre-processing technique is computationally expensive, requires a large amount of memory, and computational source than LR, KNN, SVM, and DT. Hence, we eliminated DNN from the list of classification algorithms for further analysis.

4.2 Category 2: Experiments with Ensemble Machine Learning Methods

To evaluate the performance of ensemble methods on imbalanced data, this study includes experiments with five different ensemble algorithms: Bagging, Random Forest, AdaBoost, Gradient Boosting, and XGBoost. The results are given in Table 2 in terms of all the six evaluation metrics. It is seen from the results that XGBoost performed best in terms of all the evaluation metrics except AUPR, while Random Forest performs best in terms of AUPR and second best in terms of remaining metrics.

4.3 Category 3: Experiments with Imbalanced Data Pre-processing and Traditional Machine Learning Methods

This category of experimentation includes the amalgamation of imbalanced data pre-processing and machine learning methods. A variety of undersampling and oversampling techniques were examined to process the data. The undersampling techniques like RU, Near miss-1, Near miss-2, Tomek Links, ENN, and OSS are employed. On the other hand, oversampling techniques like RO, SMOTE, Borderline SMOTE-1, Borderline SMOTE-2, SVM-SMOTE, and ADASYN are applied in the study. The two best oversampling and undersampling methods are picked up, and their combination is investigated.

The pre-processed data is then classified using traditional machine learning algorithms that include LR, KNN, SVM, and DT.

Undersampling Techniques The sampling strategy, one of the parameters in undersampling techniques, is defined as the ratio of the total number of samples in the minority class to the total number of samples in the majority class after re-sampling. However, for the present dataset, the minority class contains 268 instances, and the majority class contains 500 instances. Ideally, the denominator can take any value in the range of [268, 500]. The sampling strategy cannot be below 0.53 (268/500) for the current dataset. It can take any value in the range of [0.53, 1]. When the value is 1, the class ratio will be balanced to 0.5:0.5. However, doing so will reduce the number of samples from 768 to 536. But we aimed to remove only those samples which were affecting the models initially. Keeping these constraints in mind and the size of the data, we tuned this parameter to 0.625. The number of samples after down-sampling is 694, with the class ratio of diabetic to non-diabetic being 0.39:0.61. Table 3 presents the results when the data was pre-processed with undersampling methods. It was seen from the table that ENN performed best in terms of all evaluation metrics with all the machine learning algorithms, whereas Near miss-1 performed second best in terms of AUPR.

Oversampling Techniques In oversampling, a sampling strategy is defined as the ratio of the total number of samples in the minority class after re-sampling to the total number of samples in the majority class. The numerator can take any value in the range of [268, 500]. Therefore in our dataset, the parameter can take any value in the range of [0.53, 1]. Since the data is small, we compromised ourselves for maximum redundancy. This redundancy of information from the minority group will balance the instances of two classes in the dataset and thereby gives better results. We tuned the parameter to 1. All the minority instances will now be upsampled to the proportion of the majority class. The total number of samples after upsampling is 1000, with the class ratio of diabetic to non-diabetic being 0.5:0.5. Table 4 presents the results when the data is pre-processed with oversampling methods. It is seen from the table that RO gives the best performance in terms of all the evaluation metrics except Recall with all the machine learning algorithms, whereas SMOTE performed best in terms of Recall.

Table 3 Performance of different machine learning methods with undersampling techniques

Undersampling methods	ML algorithm	AUROC	AUPR	$F1$-score	Precision	Recall	Accuracy
RU	LR	0.86	0.80	0.66	0.76	0.76	0.76
	KNN (7)	0.82	0.72	0.59	0.70	0.70	0.70
	SVM	0.85	0.79	0.66	0.77	0.76	0.76
	DT	0.70	0.72	0.62	0.73	0.73	0.73
Near miss-1 (Neighbour = 3)	LR	0.81	0.85	0.75	0.73	0.73	0.74
	KNN (7)	0.75	0.81	0.69	0.70	0.69	0.69
	SVM	0.80	0.85	0.75	0.73	0.73	0.73
	DT	0.60	0.72	0.62	0.60	0.60	0.60
Near miss-2 (Neighbour = 3)	LR	0.82	0.79	0.70	0.69	0.69	0.69
	KNN (7)	0.79	0.83	0.74	0.73	0.73	0.73
	SVM	0.79	0.82	0.69	0.69	0.69	0.69
	DT	0.69	0.79	0.73	0.70	0.70	0.70
Tomek Links	LR	0.84	0.73	0.65	0.76	0.77	0.77
	KNN (7)	0.84	0.70	0.67	0.77	0.77	0.77
	SVM	0.85	0.74	0.63	0.75	0.75	0.75
	DT	0.72	0.71	0.65	0.74	0.73	0.73
ENN (Neighbours = 3)	LR	0.92	0.93	0.82	0.82	0.82	0.82
	KNN (7)	0.93	0.93	0.87	0.87	0.87	0.87
	SVM	0.92	0.93	0.83	0.83	0.83	0.83
	DT	0.82	0.87	0.84	0.82	0.82	0.82
OSS	LR	0.80	0.84	0.71	0.75	0.72	0.72
	KNN (7)	0.73	0.76	0.72	0.71	0.71	0.70
	SVM	0.73	0.78	0.72	0.7	0.7	0.69
	DT	0.72	0.78	0.72	0.72	0.72	0.72

Combination Methods After experimenting with different oversampling and undersampling techniques, we picked up the two best undersampling and oversampling methods in terms of AUPR and combined them. AUPR is not affected in the case of moderate to a high-class imbalance of the data and can also provide accurate predictions [21]. With undersampling, we observed that ENN (KNN and DT) surpassed all the remaining methods in terms of all the evaluation metrics. Apart from ENN, Near miss-1 (LR) performs second best and achieved greater than 80% with three classifiers in terms of AUPR. With oversampling, Random oversampling (KNN and DT) performs best in terms of AUROC, AUPR, $F1$-score, Precision, Accuracy, and better in terms of Recall. SMOTE (KNN) performs best in terms of Recall and second best in terms of AUPR. Hence, we made 4 combinations: RO + ENN, RO + Near miss-1, SMOTE + Near miss-1, and SMOTE + ENN. Figures 2 and 3 show the scatter plot after employing the aforesaid four combinations using the first two features. The yellow dots and the blue dots in the scatter plot represent the diabetic

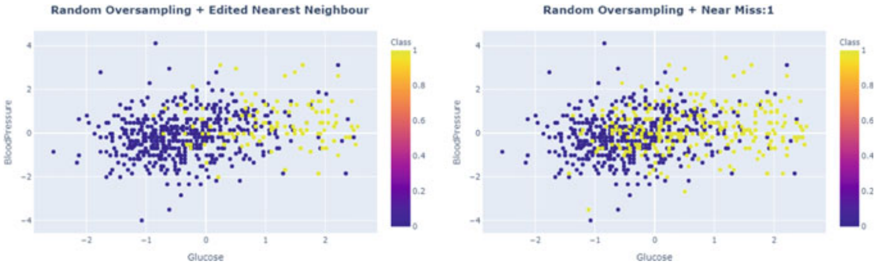
Table 4 Performance of different machine learning methods with oversampling techniques

Oversampling methods	ML algorithm	AUROC	AUPR	$F1$-score	Precision	Recall	Accuracy
RO	LR	0.81	0.76	0.73	0.74	0.74	0.74
	KNN (7)	0.91	0.91	0.8	0.81	0.79	0.78
	SVM	0.82	0.78	0.73	0.74	0.74	0.73
	DT	0.79	0.84	0.80	0.80	0.79	0.79
SMOTE	LR	0.82	0.77	0.73	0.75	0.74	0.73
	KNN (7)	0.86	0.82	0.77	0.69	0.88	0.75
	SVM	0.82	0.76	0.74	0.75	0.75	0.74
	DT	0.75	0.79	0.74	0.76	0.75	0.74
Borderline SMOTE-1	LR	0.76	0.69	0.65	0.67	0.67	0.66
	KNN (7)	0.84	0.78	0.78	0.80	0.76	0.76
	SVM	0.78	0.70	0.70	0.71	0.70	0.69
	DT	0.73	0.78	0.72	0.74	0.74	0.73
Borderline SMOTE-2	LR	0.78	0.70	0.68	0.65	0.71	0.69
	KNN (7)	0.81	0.74	0.75	0.76	0.73	0.73
	SVM	0.78	0.70	0.67	0.68	0.67	0.66
	DT	0.70	0.76	0.70	0.71	0.7	0.70
SVM-SMOTE	LR	0.80	0.74	0.71	0.73	0.73	0.73
	KNN (7)	0.85	0.82	0.77	0.78	0.76	0.76
	SVM	0.82	0.75	0.73	0.75	0.74	0.73
	DT	0.77	0.81	0.77	0.78	0.77	0.77
ADASYN	LR	0.81	0.75	0.71	0.71	0.71	0.70
	KNN (7)	0.79	0.74	0.74	0.74	0.72	0.71
	SVM	0.81	0.75	0.73	0.73	0.72	0.71
	DT	0.75	0.80	0.74	0.75	0.75	0.75

cases and non-diabetic cases, respectively. It is seen that the ratio of diabetic and non-diabetic cases improves as compared to Fig. 1.

Table 5 presents the results of traditional machine learning algorithms pre-processed by the aforesaid combined methods. KNN with $k = 3$ is used, as with $k = 3$, the model achieves the highest performance.

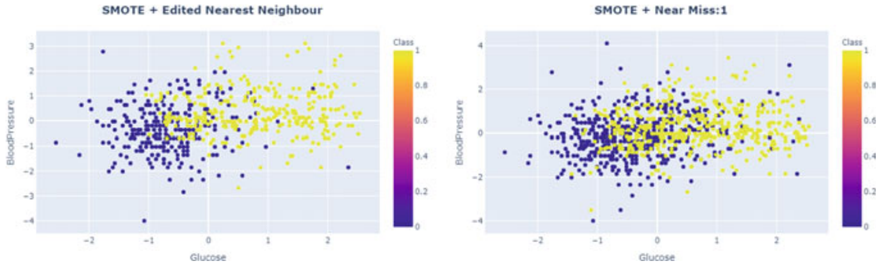
It is seen from Table 5 that SMOTE + ENN gives the best performance in terms of all evaluation metrics with all the classifiers, while SMOTE + ENN + KNN gives the highest performance. The combination (SMOTE + ENN) is further investigated with different ensemble machine learning methods (see Table 6). It is seen that the performance of the ensemble methods improves with the combination of these two imbalanced pre-processing techniques achieving the highest Accuracy of 96% with Random Forest.



(a) Scatter Plot with RO and ENN

(b) Scatter Plot with RO and Near Miss-1

Fig. 2 Scatter plot of Pima Indian dataset with RO and two best undersampling methods



(a) Scatter Plot with SMOTE and ENN

(b) Scatter Plot with SMOTE and Near Miss-1

Fig. 3 Scatter plot of Pima Indian dataset with SMOTE and two best undersampling methods

4.4 Comparison with Previous Studies

The comparison is also carried out with the state-of-the-art methods (see Table 7). The results present that our approach produced better accuracy as compared to past studies. Naz and Ahuja [2] obtained comparable accuracy with Deep Learning (98.07%) as compared to our work. However, Deep Learning is computationally extensive to train. Some of the studies listed in Table 7 have evaluated their performance in terms of other metrics. In particular, Nanni et al. [22] evaluated their performance in terms of $F1$-score, G-mean, and AUROC while Raghuwanshi and Shukla [23] presented in terms of G-mean and AUROC. In terms of $F1$-score and AUROC, our model performs best as compared to both studies. Zahirnia et al. [24] presented in terms of feature cost and misclassification cost and Wei et al. [25] evaluated with sensitivity, $F3$ and G-mean diabetes dataset.

Table 5 Performance of different machine learning methods with the combination of undersampling and oversampling techniques

Methods	ML algorithm	AUROC	AUPR	$F1$-score	Precision	Recall	Accuracy
RO + ENN (Sample size: 757)	LR	0.94	0.88	0.81	0.88	0.87	0.87
	KNN (8)	0.99	0.99	0.94	0.96	0.96	0.96
	SVM	0.93	0.88	0.80	0.87	0.87	0.87
	DT	0.93	0.91	0.90	0.93	0.92	0.92
RO + Near miss-1 (Sample size: 1000)	LR	0.84	0.83	0.75	0.76	0.76	0.75
	KNN (3)	0.88	0.89	0.82	0.84	0.81	0.80
	SVM	0.84	0.83	0.74	0.75	0.75	0.75
	DT	0.79	0.83	0.80	0.80	0.79	0.79
SMOTE + Near miss-1 (Sample size: 1000)	LR	0.81	0.80	0.69	0.70	0.70	0.70
	KNN (3)	0.85	0.84	0.78	0.78	0.75	0.75
	SVM	0.80	0.79	0.71	0.72	0.72	0.71
	DT	0.72	0.78	0.72	0.72	0.72	0.72
SMOTE + ENN (Sample size: 602)	LR	0.96	0.97	0.88	0.88	0.87	0.86
	KNN (3)	0.99	1	0.98	0.99	0.98	0.98
	SVM	0.96	0.97	0.90	0.89	0.88	0.88
	DT	0.95	0.95	0.93	0.92	0.92	0.91

Table 6 Performance of ensemble methods with SMOTE + ENN

Algorithm	AUROC	AUPR	$F1$-score	Precision	Recall	Accuracy
Bagging	0.98	0.99	0.95	0.94	0.94	0.94
Random Forest	0.98	0.99	0.97	0.96	0.96	0.96
AdaBoost	0.98	0.99	0.95	0.93	0.93	0.93
Gradient Boosting classifier	0.98	0.99	0.95	0.94	0.94	0.94
XGBoost	0.98	0.98	0.94	0.93	0.93	0.93

5 Conclusion and Future Directions

The experiments portrayed in the paper proved that the imbalanced data processing methods lead to greater performance. To attain this, we investigated the effects of different imbalanced data processing methods and machine learning algorithms based on classification performance metrics. Results present that SMOTE + ENN gave the best performance on the PIMA Indian dataset. These results are also better as compared to the previous studies carried out on the Pima Indian dataset. However, not all the studies on diabetes prediction available in the literature are based on the same dataset, so we identified those with the same dataset and compared results. Future work would include investigation with different unsupervised methods and semi-supervised methods.

Table 7 Comparison with previous studies in terms of accuracy

Author/article	Year	Method	Best accuracy (%)
Kumari and Chitra [26]	2013	SVM	0.78
Iyer et al. [27]	2015	NB	0.79
Chen et al. [28]	2017	K-means and DT	0.90
Ramezani et al. [29]	2018	Logistic adaptive network-based fuzzy inference system	0.88
Haritha et al. [30]	2018	Firefly and cuckoo search algorithms	0.81
Zhang et al. [31]	2018	Feedforward NN	0.82
Nnamoko and Korkontzelos [3]	2020	C4.5 (IQRd + SMOTEd)	0.89
Naz and Ahuja [2]	2020	DL, ANN, SVM, and DT	0.98 (DL)
Maulidina et al. [32]	2021	Backward elimination and SVM	0.85
Our work		SMOTE + ENN + KNN	0.98

References

1. Benbelkacem S, Atmani B (2019) Random forests for diabetes diagnosis. In: 2019 international conference on computer and information sciences (ICCIS). IEEE, pp 1–4
2. Naz H, Ahuja S (2020) Deep learning approach for diabetes prediction using pima Indian dataset. J Diab Metab Disord 19(1):391–403
3. Nnamoko N, Korkontzelos I (2020) Efficient treatment of outliers and class imbalance for diabetes prediction. Artif Intell Med 104:101815
4. Sahu SP, Londhe ND, Verma S (2019) Pulmonary nodule detection in CT images using optimal multilevel thresholds and rule-based filtering. IETE J Res 1–18
5. Sahu SP, Londhe ND, Verma S, Singh BK, Banchhor SK (2021) Improved pulmonary lung nodules risk stratification in computed tomography images by fusing shape and texture features in a machine-learning paradigm. Int J Imaging Syst Technol 31(3):1503–1518
6. Singh L, Janghel RR, Sahu SP (2021) Slicaco: an automated novel hybrid approach for dermatoscopic melanocytic skin lesion segmentation. Int J Imaging Syst Technol
7. Sahu SP, Kumar R, Londhe ND, Verma S (2021) Segmentation of lungs in thoracic CTs using K-means clustering and morphological operations. In: Advances in biomedical engineering and technology. Springer, pp 331–343
8. Sahu SP, Agrawal P, Londhe ND et al (2017) A new hybrid approach using fuzzy clustering and morphological operations for lung segmentation in thoracic CT images. Biomed Pharmacol J 10(4):1949–1961
9. Mani I, Zhang I (2003) KNN approach to unbalanced data distributions: a case study involving information extraction. In: Proceedings of workshop on learning from imbalanced datasets, vol 126
10. Hart P (1968) The condensed nearest neighbor rule (corresp.). IEEE Trans Inf Theory 14(3):515–516
11. Tomek I et al (1976) Two modifications of CNN
12. Laurikkala J (2001) Improving identification of difficult small classes by balancing class distribution. In: Conference on artificial intelligence in medicine in Europe. Springer, pp 63–66
13. Kubat M, Matwin S et al (1997) Addressing the curse of imbalanced training sets: one-sided selection. In: ICML, vol 97. Citeseer, pp 179–186

14. Laurikkala J (2001) Improving identification of difficult small classes by balancing class distribution. In: Conference on artificial intelligence in medicine in Europe. Springer, pp 63–66
15. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority oversampling technique. J Artif Intell Res 16:321–357
16. Han H, Wang WY, Mao BH (2005) Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing. Springer, pp 878–887
17. Tang Y, Zhang YQ, Chawla NV, Krasser S (2008) SVMs modeling for highly imbalanced classification. IEEE Trans Syst Man Cybern Part B (Cybern) 39(1):281–288
18. He H, Bai Y, Garcia EA, Li S (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE, pp 1322–1328
19. Batista GE, Bazzan AL, Monard MC et al (2003) Balancing training data for automated annotation of keywords: a case study. In: WOB, pp 10–18
20. Batista GE, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl 6(1):20–29
21. Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10(3):e0118432
22. Nanni L, Fantozzi C, Lazzarini N (2015) Coupling different methods for overcoming the class imbalance problem. Neurocomputing 158:48–61
23. Raghuwanshi BS, Shukla S (2019) Class imbalance learning using underbagging based kernelized extreme learning machine. Neurocomputing 329:172–187
24. Zahirnia K, Teimouri M, Rahmani R, Salaq A (2015) Diagnosis of type 2 diabetes using cost-sensitive learning. In: 2015 5th international conference on computer and knowledge engineering (ICCKE). IEEE, pp 158–163
25. Wei X, Jiang F, Wei F, Zhang J, Liao W, Cheng S (2017) An ensemble model for diabetes diagnosis in large-scale and imbalanced dataset. In: Proceedings of the computing frontiers conference, pp 71–78
26. Kumari VA, Chitra R (2013) Classification of diabetes disease using support vector machine. Int J Eng Res Appl 3(2):1797–1801
27. Iyer A, Jeyalatha S, Sumbaly R (2015) Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774
28. Chen W, Chen S, Zhang H, Wu T (2017) A hybrid prediction model for type 2 diabetes using k-means and decision tree. In: 2017 8th IEEE international conference on software engineering and service science (ICSESS). IEEE, pp 386–390
29. Ramezani R, Maadi M, Khatami SM (2018) A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. Alex Eng J 57(3):1883–1891
30. Haritha R, Babu DS, Sammulal P (2018) A hybrid approach for prediction of type-1 and type-2 diabetes using firefly and cuckoo search algorithms. Int J Appl Eng Res 13(2):896–907
31. Zhang Y, Lin Z, Kang Y, Ning R, Meng Y (2018) A feed-forward neural network model for the accurate prediction of diabetes mellitus. Int J Sci Technol Res 7(8):151–155
32. Maulidina F, Rustam Z, Hartini S, Wibowo V, Wirasati I, Sadewo W (2021) Feature optimization using backward elimination and support vector machines (SVM) algorithm for diabetes classification. J Phys Conf Ser 1821:012006