Prediction of diabetes using cost sensitive learning and oversampling techniques on Bangladeshi and Indian female patients

Badiuzzaman Pranto*, Sk. Maliha Mehnaz*, Sifat Momen* and Syed Maruful Huq †
*Department of Electrical and Computer Engineering, North South University
Dhaka, Bangladesh
†Department of Electrical and Computer Engineering, Presidency University
Dhaka, Bangladesh
Email: *{badiuzzaman.pranto,maliha.mehnaz, sifat.momen}@northsouth.edu, †maruf@pu.edu.bd

Abstract—**Diabetes is a major non-communicable disease that is responsible for many associated health risks and is rapidly increasing in low and middle income countries like Bangladesh. Class imbalance existing in datasets is a dire issue that can result the predictions of diabetes to be biased towards the majority class - thus reducing the reliability of machine learning models. Considering the associated risks of diabetes, a decrease in recall can result in life threatening consequences. In order to tackle this problem, a cost-sensitive learning and synthetic minority oversampling technique (SMOTE) have been applied on the PIMA Indian dataset. After that, the models have been tested on PIMA test set as well as on dataset collected from Kurmitola General Hospital (KGH), Dhaka, Bangladesh. Our results demonstrate that this proposed approach has successfully improved the reliability of the previous ML models to predict diabetes among Bangladeshi female population.**

Index Terms—**Diabetes Prediction, Imbalanced dataset, Cost-Sensitive Learning, SMOTE, Precision, Recall**

I. Introduction

Diabetes mellitus, or simply diabetes, is classified as a metabolic disorder that occurs when pancreas does not produce proper insulin which subsequently results in high amount of sugar in blood [1]. Severe complications such as diabetic peripheral neuropathy [2], diabetic nephropathy [3], diabetic retinopathy [4], kidney failure [5] and coronary heart diseases [6] can occur as a consequence of diabetes. About 366 million people are already affected with diabetes and this number has been predicted to increase up to 552 million by 2030 [7]. The number of diabetic patients have increased from 180 million to 422 million between 1980 and 2014 [8]. Long-term complications of diabetes can develop the chances of associated risks, eventually turning into a life-threatening issue. Almost 50% of all deaths are attributable to high blood glucose occurring before the age of 70 [9]. These premature deaths are noticed at a higher rate in the low and middle income countries [8] like Bangladesh. An estimated 10 million people in Bangladesh are already suffering from diabetes [10]. Around 40,142 people died from diabetes which constituted 5.09% of the total deaths according to the WHO report of 2017 [10]. Proper treatment from an early stage can, however, mitigate the malignancy of the disease. Hence, diabetes detection at an early age has now become a crucial factor to attenuate the effects of this disease.

Machine learning techniques have recently established itself to be immensely useful in the field of medical diagnosis [11]–[14]. These techniques require data to train their predictive models but the models intuitively has a tendency of biased prediction towards the majority class if the dataset is imbalanced [15]–[18]. In our earlier work, we used PIMA Indian dataset (imbalanced dataset) to train machine learning classifiers in order to predict diabetes among Bangladeshi female patients [19]. The dataset contained higher number of non-diabetic samples (majority class) compared to diabetic samples (minority class) rendering in undesirably high false negative predictions. This means many patients who have diabetes will be classified as non-diabetic resulting in life endangering consequences considering the associated risks of diabetes. Missing a diabetes diagnosis or predicting false negative is exacerbating since the patient could lose his/her life due to delay of treatment and medication. In this extended version, our objective is to reduce the number of false negative predictions. To achieve our goal, we applied cost-sensitive learning and synthetic minority over-sampling technique (SMOTE) and came up with an unbiased predictive model that can predict diabetes in Bangladeshi female population with higher reliability. The models are trained on the PIMA dataset and tested on Kurmitola General Hospital(KGH) dataset which was collected for this research work. Previously [19], it has been shown that the two datasets have similar statistical distribution - thus allowing reliable and highly accurate prediction of classes to be made using the KGH dataset from models that were trained on the PIMA dataset.

Rest of the paper is organized as follows: section II discusses the related work on diabetes detection followed by section III that explains the research methodology embraced in this paper. Sections III-D and III-E respectively describes how cost sensitive learning and SMOTE are applied in our research. Results are discussed in section IV and finally the

978-0-7381-4403-0/20/$31.00 © 2020 IEEE

paper is concluded in section V.

II. Related Work

Many researchers have conducted experiments to detect diabetes using PIMA and other notable datasets.

Li et al., for example, conducted experiments on the PIMA Indian dataset and BUPA liver disorders dataset to detect diabetes and liver disorder among patients [18]. The research obtained an accuracy of 83.57% and 86.36% after balancing and extension by support vector machine (SVM) on PIMA and BUPA dataset respectively.

In another work, a dataset from National Health and Nutrition Examination Survey was collected by a group of researchers to develop a machine learning based system to predict diabetic patients [20]. Their proposed system showed that the logistic regression based feature selection and random forest classifier combined together gave the highest accuracy of 94.25% with 0.95 AUC for K10 protocol.

Zahirnia and colleagues used cost-sensitive learning on the PIMA Indian dataset and dataset of Tabriz, Iran [21]. They evaluated the performance of proposed methods that can minimize the total feature and misclassification costs and concluded that cost-sensitive attribute selection algorithm using histograms (CASH) proposed by Weiss et al. [22] performed better in comparison to the other methods.

Faniqul et al. acquired a dataset of 520 instances from Sylhet Diabetes Hospital, Bangladesh to find the best algorithm for accurate prediction of diabetes [23]. They performed 10-fold cross validation, evaluated several machine learning classifiers and found that random forest performed best with an accuracy of 97.4%.

A group of researchers took a dataset from Tehran lipid and glucose study (TLGS) and implied probabilistic neural network (PNN), naïve bayes, and decision tree(DT) classifiers along with SMOTE in order to predict diabetes [24]. ROC convex haul (ROCCH) was used to compare the performance of the classification models and it was observed that PNN and DT gave the highest accuracy of 78.5% and 79.4% respectively.

This research work demonstrates comparative analysis on a dataset collected from a diagnostic lab in Kashmir valley in order to predict diabetic patients with and without SMOTE [25]. It was concluded that the decision tree combined with SMOTE gave better results in detecting possible diabetic patients with an accuracy of 94.7013%.

As observed, many remarkable performances have been obtained using machine learning techniques in order to detect diabetes pursuing diverse approaches and many are still continuing. The key difference lying in this research that distinguishes from the existing ones is that, the work aims to find the best performing classifier with least false negative prediction to detect diabetes among Bangladeshi female while the classifiers are actually trained on the PIMA Indian dataset. As per our knowledge, research on such purpose has not been conducted before.

III. Methodology

Figure 1 shows the methodology embraced in our research. The classifiers were applied on both PIMA and KGH dataset. Summary of both dataset are provided in Section III-A and in Table I. Both the dataset are first pre-processed in order to make them conducing for machine learning algorithms to be applied on them. Following that, the PIMA dataset was split into training and test sets. After that, the classifiers were trained by the PIMA train set and the models were tested on both PIMA test set and the KGH dataset. Performance of the classifiers have been evaluated and a comparative analysis is shown with and without the application of cost-sensitive learning and SMOTE.

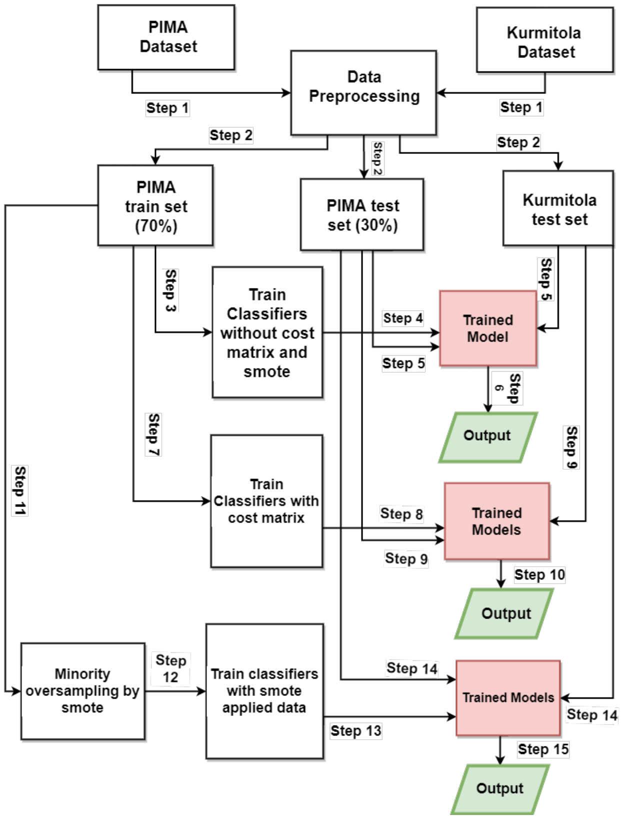

Fig. 1. Workflow of research methodology

A. Dataset Acquisition and Description

PIMA Indian dataset [26] is a popular dataset for diabetes prediction using machine learning techniques. The dataset was obtained from the National Institute of Diabetes and Digestive and Kidney diseases, India. In addition to this, more data were collected from the Kurmitola General Hospital (KGH), Dhaka to test the performance of the classifiers on Bangladeshi female population. The PIMA dataset has been used to train since the distribution of the PIMA dataset and that of KGH dataset are found similar [19]. A team of intern doctors were approached

TABLE I
SUMMARY OF THE DATASET

Dataset	**Number of instances**	**Number of features**	**Diabetic**	**Non-diabetic**	**Diabetic:Non-diabetic (ratio)**
PIMA dataset	768	8	268	500	0.35 : 0.65
PIMA training set	537	4	187	350	0.35 : 0.65
PIMA test set	231	4	81	150	0.35 : 0.65
KGH dataset	181	4	50	131	0.28 : 0.72

to collect the dataset by arranging a short discussion with the patients. This process took approximately twenty-one days during November, 2019. Distribution of both the PIMA dataset and KGH dataset are shown through violin plots in Figures 2 and 3.

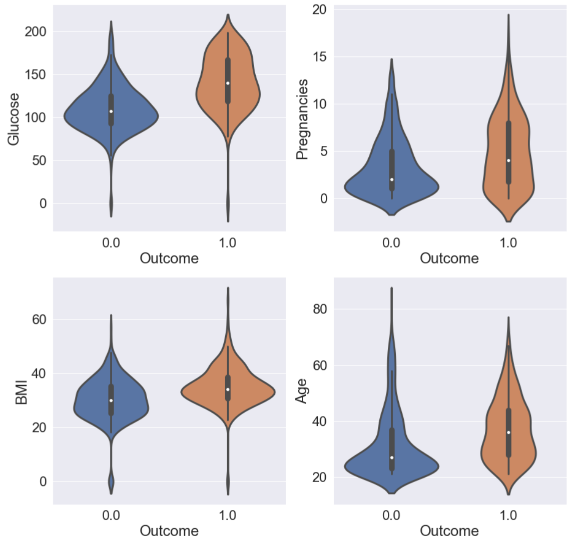

Fig. 2. Violin plot of PIMA Indian dataset

B. Data pre-processing

Pre-processing has been carried out so that the dataset is more conducive for the machine learning algorithms to be applied on them. It was observed that the PIMA Indian dataset had eight features while the KGH dataset had four features in total. Based on the common set of features available between the PIMA dataset and KGH dataset, evaluation of the classifiers was proceeded with the following four common features:

- Number of pregnancies
- Blood sugar level
- Body mass index (BMI) and
- Age.

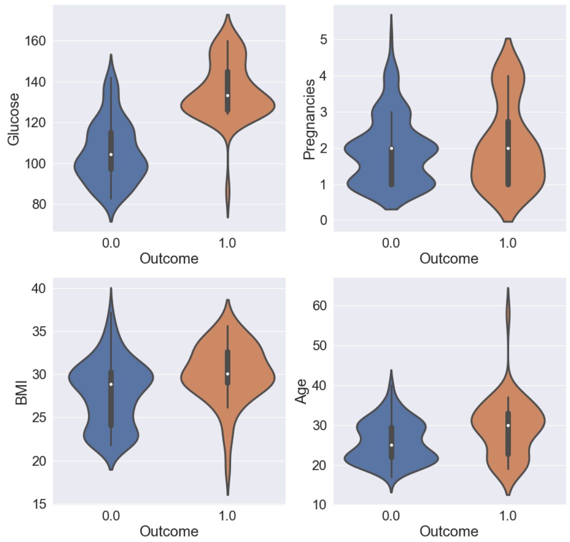

Fig. 3. Violin plot of Kurmitola Hospital dataset from Dhaka

After that, the numerical units of the features on both the PIMA dataset and KGH dataset were matched. The PIMA dataset was then divided into 70% training set and 30% test set. Three-fold rotational cross-validation was performed on the training set to build the model. It was observed in both dataset, some features contained large values compared to others, which would allow one feature to dominate over other features and this would eventually lead to misclassifications [16]. In order to tackle this issue, the numerical values of each feature were normalized (see equation 1).

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{(X_{\text{max}} - X_{\text{min}})} \tag{1}$$

C. Machine learning classifiers

After completing the data pre-processing stage, data is ready to be fed to the machine learning classifiers. Decision tree, K-Nearest neighbor, Naïve bayes and Random forest classifiers were evaluated in our work. The classifiers, Decision tree, K-Nearest neighbor and Random forest require hyper-parameter tuning. The tuning was performed for each classifier with a three-fold cross validation on training set. Brief descriptions of the classifiers are provided as follows.

1) Decision Tree: Decision tree is a machine learning algorithm that generates a tree structure from root node to leaf node where each node makes decision based on if/else condition [15], [16]. This algorithm chooses the root node by calculating high information gain and entropy using equations 2 and 3 and respectively. It was observed that the tree performs best at depth = 2 on this particular dataset. The decision tree for this research work was generated by using classification and regression trees (CART) algorithm [27].

$$Entropy(S) = -(P_{\oplus} log_2 P_{\oplus} + P_{\ominus} log_2 P_{\ominus}) \tag{2}$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{3}$$

2) K-Nearest Neighbor: The K-nearest neighbor (KNN) is a machine learning algorithm which is mainly used for classification problems [16]. For every query data, the KNN calculates its distance from all other training data-points. From there, the K nearest neighbors are selected of the particular query data-point. Once the neighbors are identified, the algorithm conducts a simple voting between the neighbors from which, the majority class is labeled as the predicted class. It is observed that $K = 16$ provides the best performance for this dataset. Euclidean distance (equation 4) metric is used to measure the distances from the query data point to the training data points.

$$\text{d(x,y)} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4}$$

3) Random forest: The random forest algorithm is an ensemble approach of decision trees. The algorithm of random forest trains T numbers of decision trees usually via bagging and then considers majority voting among them [15]. We used a total of 800 decision trees (i.e. $T = 800$) on the training set where all the 800 individual trees were generated by CART algorithm.

4) Naïve Bayes: Naïve bayes follows the bayes' theorem to predict the likelihood of an event depending on the prior knowledge [15], [16]. It pursues the conditional independence in which all the attributes are independent to each other given the value of the output class. Equation 5 shows how naïve bayes classifier calculates probability of a test instance to belong to a particular class.

$$\begin{aligned} P(Y|X) &= \frac{P(X|Y) \times P(Y)}{P(X)} \\ &= \frac{P(X_1, X_2, X_3 \ldots X_n|(Y))}{P(X)} \end{aligned} \tag{5}$$

Here $X =< X_1, X_2, X_3 \ldots X_n >$ is the feature space and Y is the output class that the classifier wants to predict.

D. Cost sensitive learning

A misclassification occurs when the predicted output of a particular input does not match the actual output. In a binary classification problem, two types of misclassification can occur: false negative and false positive. Machine learning cost insensitive classifiers assume that, all kinds of misclassifications incur the same amount of cost [17]. But this in not always a pragmatic approach when dealing with situation where it could cost an individual heavily due to misclassification (e.g. in healthcare). Missing a diabetes patient i.e. giving false negative prediction is more costly than giving a false positive prediction because the patient can suffer for a life-time if he/she is not diagnosed for diabetes at an early stage. Cost-sensitive learning is a type of learning that takes such misclassification costs into consideration and penalize the classifiers differently for each types of misclassifications [28].

$$\bar{M} = \frac{1}{n} \sum_{i=1}^{n} \delta(h(x_i, y_i)) \tag{6}$$

where,

$$\delta(h(x_i, y_i)) = \begin{cases} 0 & \text{if, } h(x_i) = y_i \\ 1 & \text{if, } h(x_i) \neq y_i \end{cases}$$

Here, $h(x_i)$ is the predicted output, y_i is the actual output and $\bar{M}$ is the average misclassification.

In order to determine a suitable cost for false negative prediction such that the false positive prediction does not increase as well, a graph has been plotted in Figure 4 with false negative cost against the F1-score. It is observed that for KNN, decision tree, naïve bayes and random forest, the false negative cost 3, 3, 5 and 2 provided the highest F1-score. During this process, the false positive cost is always constant at 1.

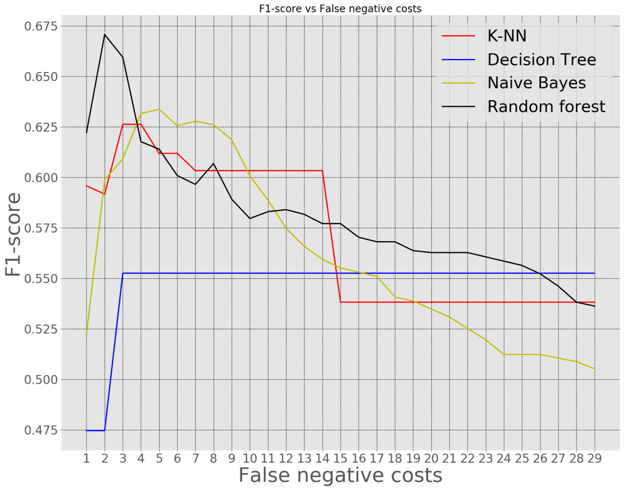

Fig. 4. F1-score against false negative costs

TABLE II
EXPERIMENTAL RESULTS WITHOUT APPLYING COST SENSITIVE LEARNING AND SMOTE

Dataset	Classifier Name	Accuracy	Precision	Recall	F1-score	AUC
PIMA Testset	KNN	0.76	0.50	0.66	0.57	0.80
	Decision Tree	0.73	0.38	0.64	0.47	0.71
	Naive Bayes	0.72	0.47	0.58	0.52	**0.83**
	Random Forest	**0.78**	**0.55**	**0.69**	**0.62**	0.80
KGH Testset	KNN	**0.81**	0.34	0.94	0.50	0.76
	Decision Tree	0.79	**1.00**	0.26	0.41	0.71
	Naive Bayes	0.78	0.20	**1.00**	0.33	0.81
	Random Forest	0.77	0.42	0.64	**0.51**	**0.84**

TABLE III
EXPERIMENTAL RESULTS AFTER APPLYING COST SENSITIVE LEARNING; (↑) INDICATES IMPROVEMENT

Dataset	Classifier Name	Accuracy	Precision	Recall	F1-score	AUC
PIMA Testset	KNN	**0.78** ↑	0.50	0.84 ↑	0.63 ↑	0.72
	Decision Tree	0.50	0.39 ↑	**0.96** ↑	0.55 ↑	0.62
	Naive Bayes	0.68	0.50 ↑	0.86 ↑	0.63 ↑	0.73
	Random Forest	0.77	**0.61** ↑	0.74 ↑	**0.67** ↑	**0.76**
KGH Testset	KNN	0.70	0.47 ↑	**0.70**	**0.56** ↑	**0.70**
	Decision Tree	0.53	0.36	**0.86** ↑	0.50 ↑	0.63
	Naive Bayes	**0.75**	0.53 ↑	0.66	0.59 ↑	0.72
	Random Forest	**0.75**	**0.54** ↑	0.68 ↑	**0.60** ↑	**0.73**

TABLE IV
EXPERIMENTAL RESULTS AFTER APPLYING SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE); (↑) INDICATES IMPROVEMENT

Dataset	Classifier Name	Accuracy	Precision	Recall	F1-score	AUC
PIMA Testset	KNN	0.73	0.57 ↑	0.66	0.61 ↑	0.71
	Decision Tree	0.73	0.59 ↑	0.55	0.57 ↑	0.68
	Naive Bayes	0.72	0.56 ↑	0.58	0.57 ↑	0.68
	Random Forest	**0.79**	**0.68** ↑	**0.68**	**0.68** ↑	**0.76**
KGH Testset	KNN	0.77	0.58 ↑	0.56	0.57 ↑	0.70
	Decision Tree	0.67	0.43	0.52 ↑	0.47 ↑	0.63
	Naive Bayes	**0.79**	**0.83** ↑	0.30	0.44 ↑	0.68
	Random Forest	**0.79** ↑	0.60 ↑	**0.68** ↑	**0.64** ↑	**0.75**

E. Synthetic minority over-sampling technique (SMOTE)

SMOTE is an algorithm used for over-sampling the minority class proposed by Chawla et al [29]. This algorithm takes the minority class into consideration and over-samples the instances to match with the majority class. Doing so makes the dataset more balanced and thus reduces the biases the model otherwise exhibits. The minority samples are created as follows [29]–[32]. :

- SMOTE first considers the minority class and finds the k-nearest neighbors(by default $k = 5$).
- Afterwards, a neighbor is randomly selected and a synthetic sample is generated at a randomly selected point between the two samples in attribute space.

IV. RESULT

The evaluation results of the classifiers are summarized in tables II, III, and IV. Performances were evaluated in terms of accuracy, precision, recall, F1-score and area under the ROC curve.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{F1-score} = \frac{2 * recall * precision}{recall + precision} \tag{10}$$

Here,
TP = Actual class is positive and predicted positive.
TN = Actual class is negative and predicted negative.
FP = Actual class is negative but predicted positive.
FN = Actual class is positive but predicted negative.

It is evident from the tables that all classifiers performed well on both the PIMA test set and the KGH test set. The best results have been shown in bold and the values indicating better performance after application of cost sensitive learning and SMOTE are given with arrow signs. We can further conclude that after applying cost sensitive learning and SMOTE techniques, there have been significant improvements in recall which indicates that the classifiers are now producing less false negative predictions than before. Simultaneously, the F1-score and AUC exhibited remarkable improvement compared to previous results. The previous work solely demonstrated on the improved performance of the diabetic detection of Bangladeshi patients without considering the false negative predictions. This research work have reduced the false negative predictions and improved the F1-score satisfying the purpose of this research. It has been found the performance of the random forest has been more significant compared to other classifiers.

V. Conclusion

Diabetes is a long term issue that needs to be detected as well as controlled at a very early stage. The purpose of this research was to reduce the false negative predictions so that the misclassification of diabetic patients as non-diabetic reduces as well. In order to acquire the research goal, we have conducted a series of experiments and the results show that the research goal has been achieved. The biased behavior towards the false negative prediction has been reduced with application of cost-sensitive learning and SMOTE.

For future work, we aim to collect dataset with higher instances so that the proposed model can be further enriched to provide results with more confidence. Further extension of the work also includes the performance evaluation of complex classifiers such as deep neural network (DNN) and other deep learning techniques. The research idea can further be used in other types of disease predictions where data insufficiency is a major challenge.

References

[1] V. A. Kumari and R. Chitra, "Classification of diabetes disease using support vector machine," *International Journal of Engineering Research and Applications*, vol. 3, no. 2, pp. 1797–1801, 2013.

[2] M. Davies, S. Brophy, R. Williams, and A. Taylor, "The prevalence, severity, and impact of painful diabetic peripheral neuropathy in type 2 diabetes," *Diabetes care*, vol. 29, no. 7, pp. 1518–1522, 2006.

[3] E. Ritz and A. Stefanski, "Diabetic nephropathy in type ii diabetes," *American journal of kidney diseases*, vol. 27, no. 2, pp. 167–194, 1996.

[4] D. S. Fong, L. Aiello, T. W. Gardner, G. L. King, G. Blankenship, J. D. Cavallerano, F. L. Ferris, and R. Klein, "Retinopathy in diabetes," *Diabetes care*, vol. 27, no. suppl 1, pp. s84–s87, 2004.

[5] R. C. Atkins, E. M. Briganti, J. B. Lewis, L. G. Hunsicker, G. Braden, P. J. C. de Crespigny, G. DeFerrari, P. Drury, F. Locatelli, T. B. Wiegmann *et al.*, "Proteinuria reduction and progression to renal failure in patients with type 2 diabetes mellitus and overt nephropathy," *American journal of kidney diseases*, vol. 45, no. 2, pp. 281–287, 2005.

[6] S. H. Wild and C. D. Byrne, "Risk factors for diabetes and coronary heart disease," *Bmj*, vol. 333, no. 7576, pp. 1009–1011, 2006.

[7] D. R. Whiting, L. Guariguata, C. Weil, and J. Shaw, "Idf diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030," *Diabetes research and clinical practice*, vol. 94, no. 3, pp. 311–321, 2011.

[8] W. H. Organization *et al.*, "Global report on diabetes," 2016.

[9] G. Roglic *et al.*, "Who global report on diabetes: A summary," *International Journal of Noncommunicable Diseases*, vol. 1, no. 1, p. 3, 2016.

[10] A. Mohiuddin, "Diabetes fact: Bangladesh perspective," *International Journal of Diabetes Research*, vol. 2, no. 1, pp. 14–20, 2019.

[11] C. Salvatore, A. Cerasa, P. Battista, M. C. Gilardi, A. Quattrone, and I. Castiglioni, "Magnetic resonance imaging biomarkers for the early diagnosis of alzheimer's disease: a machine learning approach," *Frontiers in neuroscience*, vol. 9, p. 307, 2015.

[12] M. Motwani, D. Dey, D. S. Berman, G. Germano, S. Achenbach, M. H. Al-Mallah, D. Andreini, M. J. Budoff, F. Cademartiri, T. Q. Callister *et al.*, "Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis," *European heart journal*, vol. 38, no. 7, pp. 500–507, 2017.

[13] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in genetics*, vol. 9, p. 515, 2018.

[14] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[15] A. Géron, "Hands-on machine learning with scikit-learn, keras, and tensorflow, 2nd edition," September 2019.

[16] T. M. Mitchell *et al.*, "Machine learning," 1997.

[17] V. S. Sheng, C. X. Ling, A. Ni, and S. Zhang, "Cost-sensitive test strategies," in *Proceedings of the National Conference on Artificial Intelligence*, vol. 21, no. 1. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006, p. 482.

[18] D.-C. Li, C.-W. Liu, and S. C. Hu, "A learning method for the class imbalance problem with medical data sets," *Computers in biology and medicine*, vol. 40, no. 5, pp. 509–518, 2010.

[19] B. Pranto, S. Mehnaz, E. B. Mahid, I. M. Sadman, A. Rahman, and S. Momen, "Evaluating machine learning methods for predicting diabetes among female patients in bangladesh," *Information*, vol. 11, no. 8, p. 374, 2020.

[20] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Health Information Science and Systems*, vol. 8, no. 1, p. 7, 2020.

[21] K. Zahirnia, M. Teimouri, R. Rahmani, and A. Salaq, "Diagnosis of type 2 diabetes using cost-sensitive learning," in *2015 5th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE, 2015, pp. 158–163.

[22] Y. Weiss, Y. Elovici, and L. Rokach, "The cash algorithm-cost-sensitive attribute selection using histograms," *Information Sciences*, vol. 222, pp. 247–268, 2013.

[23] M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in *Computer Vision and Machine Intelligence in Medical Image Analysis*. Springer, 2020, pp. 113–125.

[24] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh, and D. Khalili, "The impact of oversampling with smote on the performance of 3 classifiers in prediction of type 2 diabetes," *Medical decision making*, vol. 36, no. 1, pp. 137–144, 2016.

[25] M. Shuja, S. Mittal, and M. Zaman, "Effective prediction of type ii diabetes mellitus using data mining classifiers and smote," in *Advances in Computing and Intelligent Systems*. Springer, 2020, pp. 195–211.

[26] PIMA, "University of california, irvine learning repository," https://www.kaggle.com/uciml/pima-indians-diabetes-database, October 06, 2016, last Accessed: 1 June 2020.

[27] D. Steinberg and P. Colla, "Cart: classification and regression trees," *The top ten algorithms in data mining*, vol. 9, p. 179, 2009.

[28] V. S. Ling, Charleroceedings, *Cost-Sensitive Learning*. Boston, MA: Springer US, 2010, pp. 231–235.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[30] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of artificial intelligence research*, vol. 61, pp. 863–905, 2018.

[31] J. Brownlee, "Smote for imbalanced classification with python," August 21, 2020, last accessed: August 29, 2020. [Online]. Available: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[32] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing*. Springer, 2005, pp. 878–887.