IDMPF: intelligent diabetes mellitus prediction framework using machine learning

Leila Ismail and Huned Materwala
Intelligent Distributed Computing and Systems Research Laboratory, Department of Computer Science and Software Engineering, United Arab Emirates University, Al Ain, United Arab Emirates

Received 19 October 2020
Revised 3 March 2021
14 April 2021
Accepted 7 May 2021

Abstract

Purpose – Machine Learning is an intelligent methodology used for prediction and has shown promising results in predictive classifications. One of the critical areas in which machine learning can save lives is diabetes prediction. Diabetes is a chronic disease and one of the 10 causes of death worldwide. It is expected that the total number of diabetes will be 700 million in 2045; a 51.18% increase compared to 2019. These are alarming figures, and therefore, it becomes an emergency to provide an accurate diabetes prediction.

Design/methodology/approach – Health professionals and stakeholders are striving for classification models to support prognosis of diabetes and formulate strategies for prevention. The authors conduct literature review of machine models and propose an intelligent framework for diabetes prediction.

Findings – The authors provide critical analysis of machine learning models, propose and evaluate an intelligent machine learning-based architecture for diabetes prediction. The authors implement and evaluate the decision tree (DT)-based random forest (RF) and support vector machine (SVM) learning models for diabetes prediction as the mostly used approaches in the literature using our framework.

Originality/value – This paper provides novel intelligent diabetes mellitus prediction framework (IDMPF) using machine learning. The framework is the result of a critical examination of prediction models in the literature and their application to diabetes. The authors identify the training methodologies, models evaluation strategies, the challenges in diabetes prediction and propose solutions within the framework. The research results can be used by health professionals, stakeholders, students and researchers working in the diabetes prediction area.

Keywords Artificial intelligence, Machine learning, Intelligent agents, Prediction, Data analytics, Health informatics, eHealth, Diabetes mellitus type 2

Paper type Research paper

1. Introduction

Machine learning modeling is an intelligent way to extract the hidden relationship among different variables in a dataset. It has been used as a decision-support system for prediction in different applications' domains such as healthcare, education and industry [1–3]. Machine learning models can be classified into three main categories: (1) supervised learning, (2) unsupervised learning and (3) semi-supervised learning [4] (Figure S1 available at https://github.com/Dr-Leila-Ismail). The objective of a machine learning classification model is to predict the class of a given input data [5]. They are heavily used in healthcare for disease diagnosis and prognosis, fraud detection, drug efficiency and the development of a

The authors thank the anonymous reviewers for their valuable comments which helped them improve the content, quality and presentation of this paper.

Applied Computing and Informatics
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1108/ACI-10-2020-0094

nationwide prevention plan [6]. Diabetes disease has attracted a lot of attention lately due to its proliferation and dangerous consequences that may lead to death. Diabetes prediction is a classification problem, where the input features variables are the risk factors [7], and the aim is to classify an individual, based on class labels, as diabetic or non-diabetic [8].

Few machine learning prediction frameworks have been proposed in the literature for healthcare [9–12]. However, to the best of our knowledge, there is no comprehensive framework in the literature which depicts the process of diabetes data analytics from domain understanding to model deployment. In this paper, we propose an intelligent diabetes mellitus prediction framework (IDMPF) using machine learning models, as support for allied health professionals, consisting of doctors, dieticians, medical technologists, therapists and pathologists, for better diagnosis and prognosis of diseases, for better patient care. The framework helps stakeholders, such as insurance companies, pharmaceutical firms and the government to put in place a preventive plan and an effective healthcare strategy. IDMPF is based on the principles of data analytic lifecycle [13]. The proposed IDMPF is evaluated using the decision tree (DT)-based random forest (RF) and support vector machine (SVM) classification models, as they are the most used in the literature [8, 12, 14–29] from 2010 to 2019, as shown in Figure S2 (https://github.com/Dr-Leila-Ismail). Very few works compare RF and SVM [19, 20, 22]. While [19], and [20] do not report on the number of observations in the considered dataset, [22] uses a dataset consisting of 2500 observations. They do not consider the impact of an imbalanced dataset on the prediction results. In this paper, we evaluate the models in terms of accuracy, precision, recall, F-measure, receiver operating characteristic (ROC) curve, area under the ROC curve (AUC) and execution time using a dataset having 65,839 observations.

2. Literature review and critical analysis

RF is a DT-based model [30] that uses a tree structure to define the sequences of decisions and the corresponding outcomes [31]. Each risk factor (feature) is represented by a node in the tree (Figure 1 (a)) where the model decides to select a particular branch and traverse down the tree. A node without further branches is called a leaf node that represents the class label, i.e. positive (diabetic) or negative (non-diabetic). DT uses a greedy algorithm for the selection of a risk factor to split the tree. The risk factor having the highest information gain is selected for splitting. The information gain for a feature is calculated using Eqn (1).

$$\text{Info Gain}_{\text{Feature}} = H_{\text{class}} - H_{(\text{class}|\text{feature})} \tag{1}$$

where H_{class} represents the base entropy calculated using Eqn (2) and $H_{\text{class}|\text{feature}}$ represents the conditional entropy calculated using Eqn (3).

$$H_{\text{class}} = \sum_{\forall\, \text{class} \in \text{set of classes}} P(\text{class}) \log_2 P(\text{class}) \tag{2}$$

$$\begin{aligned} H_{(\text{class}|\text{feature})} &= \sum_{f} P(f) H(\text{class}|\text{feature} = f) \\ &= \sum_{\forall f \in \text{feature}} P(f) \sum_{\forall\, \text{class} \in \text{set of classes}} P(\text{class}|f) \log_2 P(\text{class}|f) \end{aligned} \tag{3}$$

where P(class) is the probability of the number of observations in the given class compared to the total number of observations and f is the set of values for a feature.

SVM [32] creates a decision boundary known as hyperplane that separates the observations into positive (diabetic) and negative (non diabetic) classes. Figure 1 (b) shows the SVM hyperplane that separates the positive and negative classes for two different

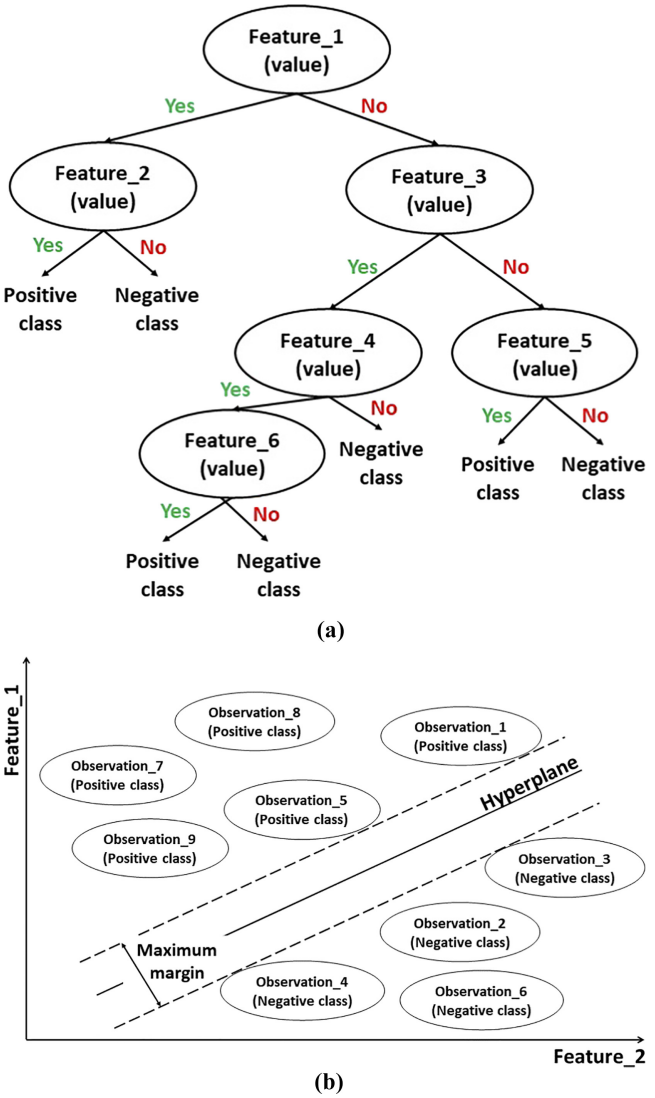

Figure 1.
Classification models: (a) decision tree (DT) and (b) support vector machine (SVM)

features. We evaluate the SVM model using its different kernels: linear, polynomial, radial basis function (RBF) and sigmoid. We obtain the hyperplane using Eqn (4).

$$\text{Minimize } \emptyset(w) = \frac{1}{2}||w||^2, \quad \text{s.t.,} \quad c_i(wf_i + b) \geq 1 \tag{4}$$

where f_i are the features, c_i are the class labels, w is the normal of the hyperplane, and b is the bias.

A literature survey was carried out (Table S1 available at: https://github.com/Dr-Leila-Ismail) to compare the performance of RF and SVM for diabetes prediction. Mostly the studies use a dataset with less than 10,000 observations. Only three works evaluate the models in terms

of F-measure. F-measure is important in the case of an imbalanced dataset (very common healthcare sector). This is because, F-measure reveals how much the model is correctly classifying the minority class, which cannot be detected by accuracy [33]. The present work proposes IDMPF, as a support system for accurate diabetes prediction. The study evaluates IDMPF using the RF and SVM models in terms of accuracy, precision, recall, F-measure, ROC curve, AUC and execution time using the UCI diabetes dataset having 12 features and 65,839 observations [34].

3. The proposed intelligent diabetes mellitus prediction framework (IDMPF)

A framework for diabetes prediction in terms of stages is presented to describe the characteristics of the data used in diabetes prediction and how this data fits within the framework. The proposed IDMPF is based on the data analytics lifecycle which depicts the process of data collection, organization and analysis to extract correlations, hidden patterns and other invaluable information [13]. Figure 2 presents the stages of IDMPF.

3.1 Domain understanding

(1) Understand the diabetes problem. For instance, type 1, type 2, or gestational diabetes [35].

(2) List the potential risk factors by consulting an expert and surveying the literature [36].

(3) State the objective of the prediction model, i.e., binomial classes (diabetic/non-diabetic) or multiple classes (non-diabetic/pre-diabetic/diabetic) prediction, prediction for men and/or women and comparison of diabetes prevalence between different age groups.

3.2 Data collection

(1) Collect data from an online public data repository such as UCI machine learning repository [37], request it from a critical care database, such as MIMIC [38] and/or create it using patients' medical data after consent. This process can be automated by developing an intelligent agent. The inclusion of the risk factors in the dataset should be verified.

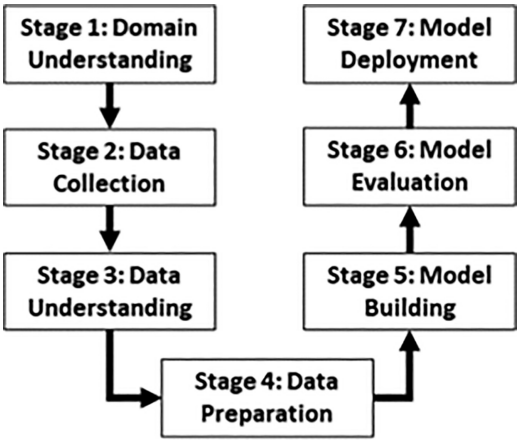

Figure 2.
Stages of our proposed IDMPF

3.3 Data understanding

(1) Aggregate the dataset if it is divided into multiple files. For instance, one file can contain the demographic data of the patients such as age, gender, ethnicity, education level and marital status, while another file can contain the medication and laboratory data such as BMI, cholesterol level, blood pressure and pulse rate [39].

(2) Refer to the disease coding system (e.g. the International Classification of Diseases (ICD)-9 [40]) if the risk factors are represented by codes.

(3) Decide on the class labels, based on the expert's advice or domain understanding, for each observation in the dataset in case they are not mentioned. For instance, observations having fasting plasma glucose level <100 mg/dl can be labeled as a non-diabetic class, a level between 100-125 mg/dl can be labeled as a pre-diabetic class and level >125 mg/dl can be labeled as a diabetic class [18].

3.4 Data preparation

3.4.1 Feature selection.

(1) Exclude the features that do not contribute to diabetes to avoid overfitting the model at its building stage. For instance, features, such as data sequence number, hospital ID, time and date should be removed.

(2) Use all the features (risk factors) available in the dataset or select a subset of features by applying feature selection algorithms [41], or taking an expert's advice, or using a hybrid approach. Ideally, researchers should evaluate several feature selection algorithms or a combination of these algorithms along with the classification model and then select the features which provide the highest accuracy, F-measure and AUC.

3.4.2 Data preprocessing.

(1) Remove the outliers for better accuracy [42] using manual visualization of the data plot or machine learning [43].

(2) Normalize the numerical features having varying ranges to avoid bias [42]. For example, the model could be biased toward plasma glucose's range of 44-199 compared to BMI's range 18.2–67.1.

(3) Identify the missing values (no value or zero) in the dataset, based on domain understanding. For example, if an observation has the value 0 for BMI, then it could be a missing value as BMI cannot be 0, whereas a value of 0 for age could represent a newborn.

(4) Treat the missing values by removing the corresponding observations or adding synthetic values, using statistical or machine learning approaches [44].

(5) Balance the imbalanced dataset by over-sampling, under-sampling, or a hybrid approach [45]. Ideally, evaluate different approaches with the classification model and then select the approach providing the highest accuracy, F-measure and AUC.

3.5 Model building

(1) Split the dataset for model training (building) and validation, by dividing it into 70% and 30% respectively, or using the k-fold cross-validation technique [46].

(2) Develop the model using the preprocessed dataset.

3.6 Model evaluation

(1) Use the validation dataset to evaluate the developed model.

(2) Select the evaluation metrics [33] to analyze the performance of the developed model. The most commonly used metric is accuracy.

(3) Evaluate the complexity of the developed model by measuring the execution time.

(4) Evaluate F-measure and AUC which are useful in case the dataset is imbalanced.

3.7 Model deployment

(1) Apply the developed model to predict diabetes.

(2) Re-develop the model based on updated and/or new data (go to 3.5).

The use of a systematic experimental methodology, depicted by the above stages, to the problem of diabetes prediction is necessary for the best prediction results as oversight of a step can lead to inaccurate results. For instance, if the dataset is imbalanced, the model might be very accurate but will not be able to detect the minority class, which could life-threatening in the case of a diabetic minority. Table 1 compares the work in the literature, on machine learning-based prediction framework for healthcare and diabetes in particular and our work.

4. Performance analysis

4.1 Experimental environment

To evaluate the performance of RF and SVM for the prediction of type 2 diabetes, the proposed framework was using an imbalanced UCI dataset with parameters presented in Table 2. The performance of the classifiers with and without feature selection was judged before and after data balancing, and using correlation attribute evaluator [47] for feature selection as it improves the accuracy for diabetes prediction [22]. The data balancing techniques used in the experiments were adopted from [45], namely: Random over-sampling (RO), Synthetic minority oversampling technique (SMOTE), Borderline SMOTE (B_SMOTE), Borderline SMOTE-SVM (B_SMOTE-SVM), Adaptive synthetic sampling (ADASYN), K-means SMOTE (k_SMOTE), Random under-sampling (RU), Near miss (NM), Tomek links (TL), Edited nearest neighbors (ENN), Repeated ENN (R-ENN), All k-NN, Instance hardness threshold (IHT), One-sided selection (OSS), Neighborhood cleaning rule (NCR), SMOTE + ENN, SMOTE + Tomek links (SMOTE_TL). All the experiments are performed using Python 3.8 [48].

4.2 Experiments

The dataset were preprocessed by removing the irrelevant features such as encounter id, patient number, admission type id, discharge deposition id, hospital time in and time out, and payer code, and remove the feature "weight" as it has 100% missing values. The resultant dataset includes race, gender, age, diagnosis 1, diagnosis 2, diagnosis 3 and diabetes medication. Diagnosis 1, 2 and 3 represent the results of the primary, secondary and additional secondary diagnoses respectively. A class label for diabetes was created based on the diabetes medication feature. The class value is set to 1, i.e. diabetic, if the corresponding value in the diabetes medication column is "yes", else it is set to "0", i.e. non-diabetic. We remove all the observations having missing values. For diagnoses 1, 2 and 3, we extracted the ICD-9 code values of the diseases that are risk factors of type 2 diabetes such as obesity, hypertension and cardiovascular disease. A column for each risk factor is added. The value

Work		[9]	[10]	[11]	[12]	Present study
The objective of the prediction framework		Prediction of short- and long-term; Treatment response in initially antipsychotic-naïve; Schizophrenia patients	Classification of sleep stages	Prediction of heart disease	Prediction of diabetes	Prediction of diabetes
Domain understanding		✗	✗	✗	✗	✓
Data collection		✓	✓	✓	✓	✓
Data understanding		✓	✗	✗	✗	✓
Data preparation	Feature selection	✓	✓	✓	✓	✓
	Data normalization/ standardization	✓	✗	✓	✗	✓
	Treating missing values	✓	✗	✓	✓	✓
	Data balancing	✗	✓	✗	✗	✓
Model building		✓	✓	✓	✓	✓
Model evaluation		✓	✓	✓	✓	✓
Model deployment		✓	✓	✓	✓	✓

Note(s): ✓ → considered; ✗ → not considered

Table 1. Work on machine learning-based framework for healthcare

Dataset	Description	Features	# Positive class observations (diabetic)	# Negative class observations (non-diabetic)
UCI	From 1999-2008 clinical care outcomes of male (30922) and female (34917) patients, Caucasian, Asian, African American, Hispanic and other races, between 0-100 years old (on average [50-60) years), from 130 US hospitals	*Categorical* – age, race *Binary* – alcohol consumption, blood pressure, blurred vision, cholesterol, gender, heart disease, obesity, pregnancy and uric acid	51,034 (77.5%)	14,805 (22.5%)

Table 2. Characteristics of the preprocessed UCI diabetes dataset used in the experiments

for every observation for each risk factor is set to "1" if the disease appears in either diagnosis 1, 2 and 3, otherwise, it is set to "0".

The 70% of the dataset was used for training and 30% for testing. The study evaluates the data balancing techniques for different values of involved parameters and select the parameters that result in the highest AUC value. The accuracy and the F-measure are calculated using Eqs (5) and (6) respectively. The execution time is calculated by adding the model training and testing times.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$F - \text{measure} = \frac{2(\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})} \tag{6}$$

where *TP* (True Positive) represents the number of observations, in the positive class, that are classified as positive, *TN* (True Negative) represents the number of observations, in the negative class, that are classified as negative, *FP* (False Positive) represents the number of observations, in the negative class, that are classified as positive, and *FN* (False Negative) represents the number of observations, in the positive class, that are classified as negative. The values of recall for the positive and negative class are calculated using Eqs (7) and (8) respectively and the values of precision for the positive and negative class are calculated using Eqs (9) and (10) respectively.

$$\text{Recall (positive class)} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{Recall (negative class)} = \frac{TN}{TN + FP} \tag{8}$$

$$\text{Precision (positive class)} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Precision (negative class)} = \frac{TN}{TN + FN} \tag{10}$$

4.3 Experimental results analysis

Figure 3 shows the accuracy, precision, recall and F-measure of RF and SVM models with and without feature selection algorithm, before and after data balancing. The precision, recall and F-measure values are presented for the diabetic class (+), non-diabetic class (−) and their weighted averages (A). We present the results for the data balancing techniques that have the highest F-measure value among those which have an AUC value greater than 0.5. Before data balancing, RF outperforms SVM in terms of accuracy and F-measure, meaning that the DT is

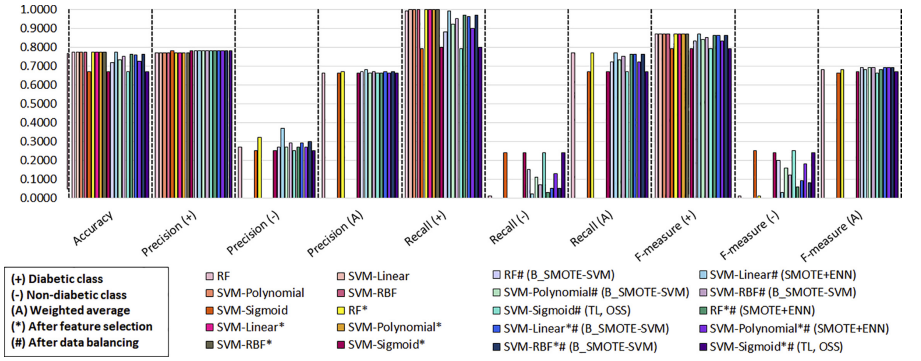

Figure 3.
Performance of the classification models with and without feature selection before and after data balancing

more suitable for diabetes prediction, which is consistent with the literature [19,20]. SVM-linear, polynomial and RBF kernels have higher accuracy than SVM-sigmoid. However, they cannot detect the minority non-diabetic class using the imbalanced UCI. The relative performance of RF and SVM does not change before and after feature selection. The selected features in our experiments, i.e. age, blood pressure, cholesterol, gender and obesity, are the same as the ones in the literature, as shown in Table S2 (https://github.com/Dr-Leila-Ismail). After data balancing, the SVM-linear kernel outperforms the other models under study in terms of accuracy without feature selection, but after feature selection, RF yields the highest accuracy. Moreover, after data balancing SVM models with linear, polynomial and RBF kernels can predict the non-diabetic minority class. Figure 4 shows ROC and AUC for the developed models with and without feature selection before and after data balancing. It shows that before data balancing the SVM models with linear, polynomial and RBF kernels have an AUC of 0.5, with and without feature selection, revealing that the model is randomly assigning all the observations to the majority diabetic class. However, after data balancing the AUC of the models under study are greater than 0.5, revealing a detection of the minority class.

Table 3 shows our experimental results on the execution time of the models with and without feature selection, before and after data balancing. It shows that the execution time of the models decreases after feature selection.

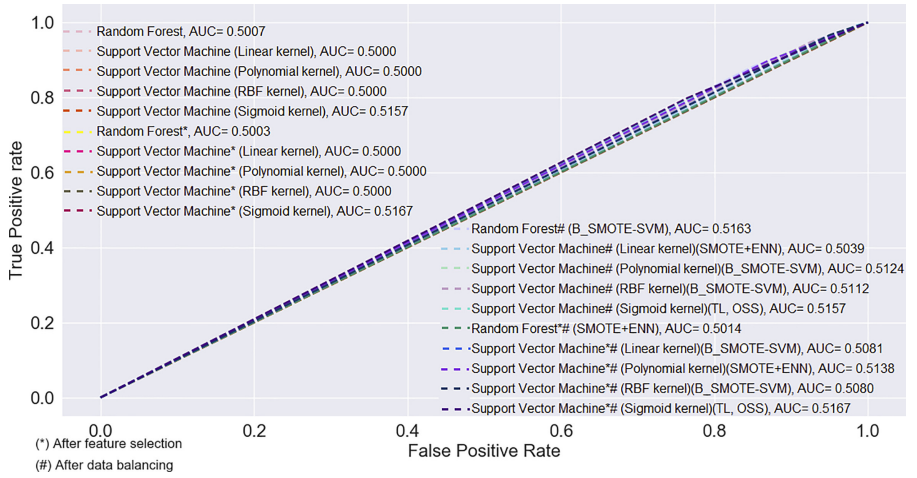

Figure 4.
ROC curve and AUC of the classification models with and without feature selection before and after data balancing

	Execution time (seconds)			
	Before data balancing		After data balancing	
	Without feature selection	With feature selection	Without feature selection	With feature selection
RF	1.74	0.77	46.01	8.76
SVM – linear	66.66	56.49	20.1	10.04
SVM – polynomial	78.58	64.17	257.94	9.26
SVM – RBF	163.1	114.97	572.02	457.08
SVM – sigmoid	115.7	99.89	65.61	66.29

Table 3.
Execution times of the classification models

5. Conclusions and summary

Being a global crisis it is crucial to predict the prevalence of diabetes in an individual to reduce the risk of complications and to save lives. The paper evaluates recent works on diabetes prediction that have used DT-RF and SVM models. In addition, different machine learning-based prediction frameworks for healthcare and diabetes in particular were analyzed. The proposed framework (IDMPF) is the result of a critical analysis of machine models in the literature and our implementation of RF and SVM for diabetes prediction. The performance of the models in terms of accuracy, precision, recall, F-measure, ROC curve, AUC and execution time was evaluated. In addition, challenges involved in diabetes prediction are highlighted to guide future research. The present study will help allied health professionals and researchers in the field of diabetes prediction. For an imbalanced dataset, data balancing techniques could help to detect the minority class. However, the performance of the models is data-driven and dependent on the features being used, and therefore, cannot be generalized. The IDMPF is evaluated using the most two used classification models in the literature. A larger spectrum of models will be considered in our future work.

References

1. Meherwar F, Maruf P. Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl. 2017; 9: 1-16.
2. Prenkaj B, Velardi P, Stilo G, *et al.* A survey of machine learning approaches for student dropout prediction in online courses. ACM Comput Surv. 2020; 53: 1-34.
3. Angelopoulos A, Michailidis ET, Nomikos N, *et al.* Tackling faults in the industry 4.0 era-a survey of machine-learning solutions and key aspects. Sensors. 2020; 20: 1-34.
4. Ayodele TO. Types of machine learning algorithms. New Adv Mach Learn. 2010; 3: 19-48.
5. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: a review of classification techniques. Emerg Artif Intell Appl Comput Eng. 2007; 160: 3-24.
6. Leila I, Materwala H, Karduck P, Adem A. Requirements of health data management systems for biomedical care and research: scoping review. J Med Internet Res. 2020; 22. doi: 10.2196/17508.
7. Ismail L, Materwala H, Al Kaabi J. Association of risk factors with type 2 diabetes: a systematic review. Comput Struct Biotechnol J. 2021; 19: 1759.
8. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Comput Sci. 2018; 132: 1578-85. doi: 10.1016/j.procs.2018.05.122.
9. Ambrosen KS, Skjerbæk MW, Foldager J, *et al.* A machine-learning framework for robust and reliable prediction of short-and long-term treatment response in initially antipsychotic-naive schizophrenia patients based on multimodal neuropsychiatric data. Transl Psychiatry. 2020; 10: 1-13.
10. Phan H, Andreott F, Cooray N, *et al.* Joint classification and prediction CNN framework for automatic sleep stage classification. IEEE Trans Biomed Eng. 2018; 66: 1285-96.
11. Haq AU, Li JP, Memon MH, *et al.* A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mob Inf Syst. 2018; 2018: 1-21.
12. Songthung P, Sripanidkulchai K. Improving type 2 diabetes mellitus risk prediction using classification. In: International Joint Conference on Computer Science and Software Engineering (JCSSE). 2016. pp 1-6.
13. Pouchard L. Revisiting the data lifecycle with big data curation. 2015.
14. Karegowda AGVP, Jayaram M, Manjunath A.. Rule based classification for diabetic patients using cascaded K-means and decision tree C4.5. Int J Comput Appl. 2012; 45. doi: 10.5120/6836-9460.
15. Nai-arun N, Moungmai R. Comparison of classifiers for the risk of diabetes prediction. Procedia Comput Sci. 2015; 69:132-42. doi: 10.1016/j.procs.2015.10.014.

16. Perveen S, Shahbaz M, Gurgachi A, Keshavjee K. Performance analysis of data mining classification techniques to predict diabetes. Procedia Comput Sci. 2016; 82: 115-21. doi: 10.1016/j.procs.2016.04.016.

17. Zou Q, Qu K, Luo Y, *et al.* Predicting diabetes mellitus with machine learning techniques. Front Genet. 2018; 9. doi: 10.3389/fgene.2018.00515.

18. Yu W, Liu T, Valdez R, *et al.* Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak. 2010; 10. doi: 10.1186/1472-6947-10-16.

19. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak. 2011; 11. doi: 10.1186/1472-6947-11-51.

20. Kandhasamy JP, Balamurali S. Performance analysis of classifier models to predict diabetes mellitus. Procedia Comput Sci. 2015; 47: 45-51. doi: 10.1016/j.procs.2015.03.182.

21. Heydari M, Teimouri M, Heshmati Z, Alavinia SM. Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. Int J Diabetes Dev Ctries. 2016; 36: 167-73. doi: 10.1007/s13410-015-0374-4.

22. Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data. 2019; 6: 1-19.

23. Saravananathan K, Vekmurugan T. Analyzing diabetic data using classification algorithms in data mining. Indian J Sci Technol. 2016; 9: 1-6.

24. Nirmala Devi M, Balamurugan SA, Swathi UV. An amalgam KNN to predict Diabetes Mellitus. In: IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN). 2013.

25. Bashir S, Qamar U, Khan FH, Javed MY. An efficient rule-based classification of diabetes using ID3, C4.5 & CART ensembles. In: 12th International Conference on Frontiers of Information Technology. 2014. pp 226-231.

26. Wu H, Yang S, Huang Z, *et al.* Type 2 diabetes mellitus prediction model based on data mining. Informatics Med Unlocked. 2018; 10: 100-7. doi: 10.1016/j.imu.2017.12.006.

27. Tamilvanan B, Bhaskaran VM. An experimental study of diabetes disease prediction system using classification techniques. IOSR J Comput Eng. 2017; 19: 39-44. doi: 10.9790/0661-1901043944.

28. Wang C, Li L, Wang L, *et al.* Evaluating the risk of type 2 diabetes mellitus using artificial neural network: an effective classification approach. Diabetes Res Clin Pract. 2013; 100:111-18. doi: 10.1016/j.diabres.2013.01.023.

29. Selvakumar S, Kannan KS, Gothai Nachiyar S. Prediction of diabetes diagnosis using classification based data mining techniques. Int J Stat Syst. 2017; 12: 183-88.

30. Shaik AB, Srinivasan S. A brief survey on random forest ensembles in classification model. In: International Conference on Innovative Computing and Communications. 2019. pp 253-60.

31. EMC Education Services. Data science and big data analytics: discovering, analyzing, visualizing and presenting data. Wiley Publishing. 2015: ISBN: 978-1-118-87613-8.

32. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995; 20: 273-97. doi: 10.1007/BF00994018.

33. Hossin M, Sulaiman M. A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manag Process. 2015; 5.

34. Strack B, Deshazo JP, Gennings C, *et al.* Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. Biomed Res Int. 2014; 11. doi: 10.1155/2014/781670.

35. Types of diabetes. [cited 2021 Mar 23]. Available at: https://www.idf.org/aboutdiabetes/what-is-diabetes.html.

36. Perry IJ, Wannamethee SG, Walker MK, *et al.* Prospective study of risk factors for development ofnon-insulin dependent diabetes in middle aged British men. BMJ. 1995; 310: 560-64.

37. Asuncion A, Newman D (2007) UCI machine learning repository. [cited 2020 Aug 10]. Available at: https://archive.ics.uci.edu/ml/citation_policy.html.

38. MIMIC critical care database. [cited 2020 Aug 10]. Available at: https://mimic.physionet.org/.

39. Dataset files. [cited 2020 Oct 18]. Available at: https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2013.

40. International classification of Diseases, Ninth revision, clinical modification (ICD-9-CM). [cited 2020 Aug 10]. Available at: https://www.cdc.gov/nchs/icd/icd9cm.htm#:~:text=ICD-9-CM is the,10 for mortality coding started.

41. Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng. 2014; 40: 16-28.

42. Larose DT. Data mining methods and models. Wiley Online Library; 2006: Print ISBN: 9780471666561, Online ISBN: 9780471756484. doi: 10.1002/0471756482.

43. Abe N, Zadrozny B, Langford J. Outlier detection by active learning. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006. pp 504-9.

44. Jerez JM, Molina I, Garcia-Laencina PJ, *et al.* Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med. 2010; 50: 105-15.

45. Fernández A, García S, Galar M, *et al.* Learning from imbalanced data sets. 1st ed. Springer International Publishing; 2018: XVIII, 377, eBook ISBN: 978-3-319-98074-4. doi: 10.1007/978-3-319-98074-4.

46. Fushiki T. Estimation of prediction error by using K-fold cross-validation. Stat Comput. 2011; 21: 137-46. doi: 10.1007/s11222-009-9153-8.

47. Hall M, Frank E, Holmes G, *et al.* The WEKA data mining software: an update. ACM SIGKDD Explor Newsl. 2009; 11: 10-18. doi: 10.1145/1656274.1656278.

48. Python language. [cited 2021 Mar 3]. Available at: https://devdocs.io/python~3.8/.

Corresponding author
Leila Ismail can be contacted at: leila@uaeu.ac.ae