Computers in Biology and Medicine 136 (2021) 104664

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier

Anju Prabha *, Jyoti Yadav, Asha Rani, Vijander Singh

Division of Instrumentation and Control Engineering, Netaji Subhas Institute of Technology, University of Delhi, New Delhi, India

ARTICLE INFO

Keywords:
Diabetes detection
PPG
MFCC
Feature selection
XGBoost

ABSTRACT

In this work, a non-invasive diabetes mellitus detection system is proposed based on the wristband photoplethysmography (PPG) signal and basic physiological parameters (PhyP) to enable easy detection of diabetes mellitus (DM). A dataset of 217 participants with diabetes, prediabetes and normal conditions is used to develop the system. The Mel frequency cepstral coefficients (MFCC) extracted from 5s PPG signal segments and the PhyP are used as input for the machine learning algorithms. The K-nearest neighbors, support vector machine, random forest and extreme gradient boost (XGBoost) classifiers are used for classification. In addition, a hybrid feature selection method (Hybrid FS) is proposed to reduce the size of the input data. The Hybrid FS-based XGBoost system achieves a high accuracy of 99.93 % for non-invasive diabetes detection with fewer features and less computational effort. The analysis suggests that the PPG signal from a wearable sensor is a good alternative for simple non-invasive blood glucose measurements in routine applications.

1. Introduction

Diabetes mellitus (DM) has become an epidemic in this century, hence advances in technology for easy and affordable screening are inevitable. The disease needs to be diagnosed at an early stage in order to alleviate the burden on the public health system. It causes difficulty in regulating blood sugar levels due to muscle insulin resistance, insufficient insulin secretion or both [1]. Over time, hyperglycaemia caused by diabetes may damage various organs such as the kidneys, heart, eyes, nerves, and blood vessels. The associated complications include diabetic retinopathy, neuropathy, cardiovascular disease (CVD) etc. and may even lead to stupor and coma [2]. There is no cure for DM and the treatments are intended to control its effects.

The fasting plasma glucose test (FPG) is a general method of diagnosing DM. FPG below 100 mg/dL is considered normal, while 100 mg/dL to 125 mg/dL indicate prediabetes and 126 mg/dL or higher indicate diabetes [3]. A condition in which the blood sugar level is between normal and diabetes is called prediabetes. Prediabetes cases tend to develop diabetes and related complications. According to a survey by the International Diabetes Federation (IDF), around 463 million people were affected by DM in 2019, and it will be 700 million in 2045 [4]. Polypharmacy is also a serious concern in diabetes mellitus (DM), especially in the elderly. Findings from Dobrica et al. [5] show that DM patients receive more medication than their non-DM counterparts and are therefore exposed to more drug-drug and food-drug interactions. Timely diagnosis and regular monitoring of blood sugar levels are essential to contain the repercussions of DM. This can reduce the effects of polypharmacy and keep blood glucose levels within limits. In the current scenario, however, 1 in 2 people with DM are not diagnosed. Family history of DM, lack of symptoms, and obese BMI are the main traits of people with undiagnosed diabetes [6]. The blood test for FPG, the haemoglobin A1C test, the oral glucose tolerance test and minimally invasive portable glucose meters are the predominant diagnostic and monitoring tools for DM. In general, these methods are either invasive or semi-invasive and expensive, which prevents regular monitoring. This necessitates the development of a non-invasive and economical DM detection system.

The rest of this paper is structured as follows: Section 2 describes the related work. The detailed discussion of the materials and methods used to design of the proposed system can be found in Section 3. The results of this work are discussed in Section 4 and finally Section 5 concludes the work.

2. Related work

The application of soft computing techniques in healthcare has

* Corresponding author. Division of Instrumentation and Control Engineering, Netaji Subhas Institute of Technology, Dwarka Sector 3, Dwarka, New Delhi, India.
E-mail addresses: anjuprabha.mec@gmail.com (A. Prabha), bmjyoti@gmail.com (J. Yadav), asha.rani@nsit.ac.in (A. Rani), vijaydee@nsit.ac.in (V. Singh).

https://doi.org/10.1016/j.compbiomed.2021.104664
Received 30 March 2021; Received in revised form 18 July 2021; Accepted 18 July 2021
Available online 22 July 2021

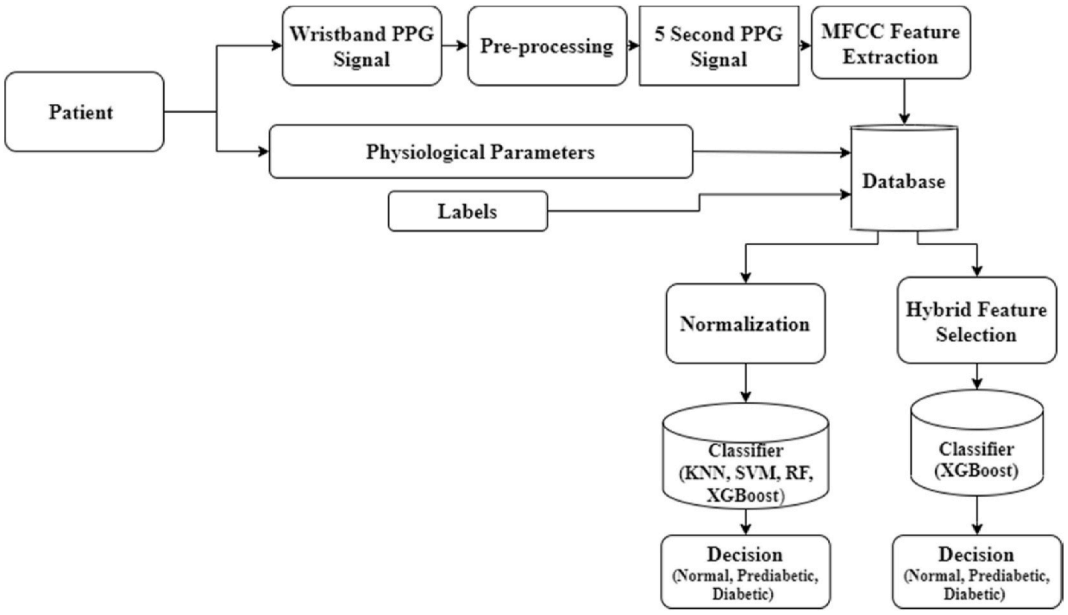

Fig. 1. Schematic diagram of non-invasive DM detection scheme.

received considerable interest in recent years as it is fast, precise, economical, reliable and requires less human expertise. Significant research is being carried out to develop DM detection systems based on artificial neural networks (ANN), support vector machine (SVM), K-nearest neighbor (KNN), deep learning (DL), hybrid techniques [7] and so on. The publicly accessible and self-created datasets are usually used for the development of automated DM recognition systems. The Pima Indian diabetic database (PIDD), which contains information such as pregnancy, BMI, insulin level, age, etc., is also used by many researchers for this purpose [8,9]. A recent study [10] compared the performance of ANN, Naive Bayes (NB), decision tree (DT) and DL with PIDD dataset for diabetic detection. The results show that the classifiers' performances are in the range of 90–98 %, with DL being the best classifier to indicate the onset of diabetes. In a comprehensive review of machine learning (ML) and artificial intelligence based DM detection [11] it is observed that DL and SVM generally deliver better classification results, while random forest and ensemble classifiers are the second choice. However, a study of recent advances in non-invasive DM detection [12] reveals that traditional classification techniques such as ANN, SVM, decision tree, etc., pose challenges when applied to real-time systems, while deep learning algorithms face overfitting problems. Barakat et al. [13] developed an intelligible SVM to diagnose DM with 94 % accuracy using the dataset with questionnaire data and clinical details, including FPG. Soliman et al. [14] proposed a least-squares SVM-based system for DM classification using PIDD data. This achieves an average accuracy of 97.83 %. Daghistani et al. [15] used data mining algorithms, i.e. self-organizing map, C4.5 and random forest, on a huge dataset from the Ministry of National Guard Health Affairs in Saudi Arabia that contains 18 attributes including 11 laboratory tests and demographics. While all of these studies provide significant accuracy in diagnosing DM, they are developed based on laboratory test data and are invasive.

Much research has been carried out on non-invasive DM detection technologies in the past few decades. Ling et al. [16] developed a non-invasive hypoglycaemia monitoring system based on the electrocardiography (ECG) signal from children aged 14.6 ± 1.5years. Furthermore, Swapna et al. [17] proposed a deep learning based DM detection system that uses heart rate variability signals. The scheme achieves an accuracy of 95.7 % with the long short-term memory and the convolutional neural network. However, ECG recording is practically inconvenient for routine analysis, thus recent research has focused on the alternative data obtained with wearable sensors or smartphones. Various features extracted from the photoplethysmography (PPG) signal are also used with ML classifiers for DM detection [18–20]. Recently, DM detection systems with smartphone-captured PPG signal have also been built [21,22], but the accuracy of these systems is quite low.

The PPG signal analysis is an emerging electro-optical technique for measuring microvascular blood volume changes in tissue [23]. The PPG device consists of a light emitting diode and a photodetector to detect the intensity of the light reflected from the blood. Low price, simple and convenient handling are the attributes that lead researchers to use the PPG signal in various applications. The literature shows that the PPG signal is a very good alternative to non-invasive DM detection [24], but the method has not been validated on wearable PPG sensors. The use of the PPG signal to extract cardiovascular and respiratory parameters is already established [25]. It is shown that the information about arterial stiffness derived from the hemodynamic properties of the PPG signal serves as an early indicator for DM [26,27]. It has also been observed that the age-related increase in total body fat and visceral adiposity up to the age of 65 often leads to DM [28]. Therefore, the physiological parameters such as height, weight and age also play an important role in DM detection.

In the previous work [29], Mel frequency cepstral coefficients (MFCC) of the PPG signal and the physiological parameters (PhyP) of the subjects are used to detect DM with the help of SVM and KNN classifiers. The SVM classifier achieved the highest classification accuracy of 84.49 % with 107 features. However, for disease diagnosis, the classification accuracy needs to be improved with a smaller number of input features in order to develop a more efficient and reliable system with less computational effort. Therefore, more sophisticated machine learning algorithms and feature optimization techniques are analyzed, and the current work is an improvement on the conceptual framework outlined in the previous paper [29]. The improvement of the classification performance is achieved through the use of advanced classifiers and the selection of the most important features for the classification. Thus, different feature selection algorithms are applied to the feature set and a hybrid feature selection (Hybrid FS) algorithm based on majority voting is also proposed. Application of the Hybrid FS reduces the computational complexity of the system. In addition, the classification performance of more advanced classifiers such as random forest (RF) and extreme gradient boosting (XGBoost) is analyzed. In the present research work, an intelligent diabetes mellitus detection system based on hybrid feature selection and the XGBoost classifier is being developed and the

limitations of the study are also analyzed.

3. Materials and methods

3.1. Database

The database used in this work [30] includes blood sugar values determined with the Accu-chek performa glucometer and the clinical-chemical and immunological analyzer (Cobas 6000), physiological parameters (weight (kg), height (m) and age), PPG signal from the handle Empatica E4 wristband and class labels for 217 patients in a hospital in Cuenca, Ecuador. There are three classes based on blood sugar levels: normal, prediabetes, and diabetes. The PPG signal is recorded while the patient is resting for an average of 2.5 min. The sampling frequency of the signal is 64 Hz. The database includes 59.9 % normal, 32.7 % diabetic and 7.37 % prediabetic patients.

3.2. System description

The schematic diagram of the proposed scheme is shown in Fig. 1. In this work, the wristband PPG signal and the PhyP collected from the patients are considered. After preprocessing, Mel frequency cepstral coefficients (MFCC) features are extracted from 5s PPG signal segments. The MFCC features, PhyP and the corresponding class labels are used for the development of the diabetes mellitus detection system (DMDS). The dataset under consideration is divided into training and test datasets. The designed DM detection system is simulated in MATLAB 2017 and Python 3.9.0 on an Intel (R) Core (TM) i5-6200U CPU @ 2.30 GHZ 2.40 GHZ and 8 GB RAM PC.

3.2.1. Pre-processing

The PPG signal of each subject is segmented into 5s segments to increase the number of samples. The thresholding method is used to remove the segments with high amplitude disturbances. In this way 7263 samples are obtained. Furthermore, signal segments are filtered with a Butterworth bandpass filter with a pass band of 0.5–15 Hz.

3.2.2. Feature extraction

The MFCC features play a vital role in speech and speaker recognition applications. MFCC allows a compact representation of the signal amplitude spectrum [31]. The feature extraction process is inspired by human auditory perception. The literature shows the use of MFCC features of physiological signals such as ECG [32], electroencephalogram (EEG) [33], phonocardiogram [34] etc. In this work, MFCC features of the PPG signal are used to represent the hemodynamic characteristics. The approximation of the Mel frequency unit results from Eq. (1).

$$f_{Mel} = 2595 \ log_{10}\left(1+\frac{f}{700}\right) \tag{1}$$

where f_{Mel} is the Mel frequency and f is the normal frequency in Hz.

The steps for MFCC feature extraction are explained as follows:

1. *Windowing*: The analysis of quasi-stationary signals is carried out in short segments [35] in order to obtain stable characteristics. Therefore, the signal is framed into short segments (Eq. (2)) with 50 % overlap between the frames.

$$y(n) = x(n) * w(n); \tag{2}$$

where x(n) is the input signal and w(n) is the window function. The Hamming window function [36] is used to obtain smooth edges and to reduce the edge effect.

$$w(n) = 0.54 - 0.46 * cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \tag{3}$$

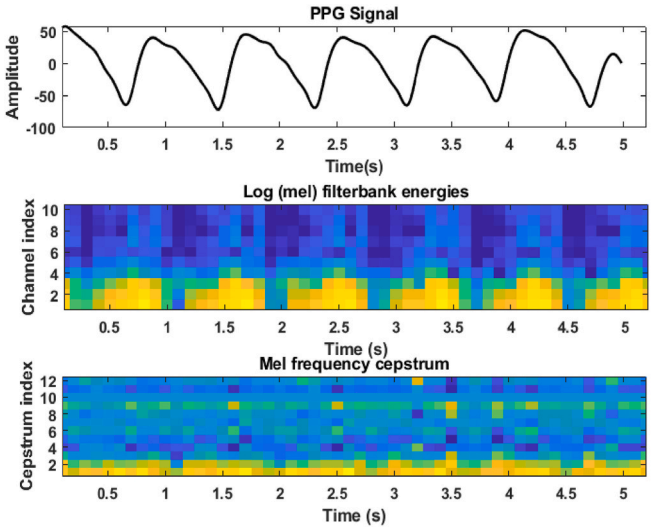

Fig. 2. Mel frequency cepstrum of 5s PPG signal.

where N is the number of samples in each frame.

2. *Discrete Fourier Transform (DFT) spectrum:* The DFT of each frame is calculated to obtain the magnitude spectrum.

$$Y(k) = \sum_{n=0}^{N-1} y(n) e^{-j\frac{2\pi nk}{N}} \quad 0 \le k \le N-1 \tag{4}$$

3. *Mel spectrum:* The Mel spectrum is obtained by passing the power spectrum of the magnitude spectrum Y(k) through a set of Mel triangle weighted filters (5). The product is calculated at each frequency.

$$s(m) = \sum_{k=0}^{M-1} |Y(k)|^2 H_m(k) \quad 0 \le m \le N-1 \tag{5}$$

where $H_m(k)$ is the weight of the k^{th} energy spectrum bin that provides the m^{th} output band, M is the number of triangular Mel weighing filters. $H_m(k)$ is calculated with Eq. (6). In the warped axis, Eq. (1) is used to simulate the perception of the human ear.

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \le k \le f(m) \\ \dfrac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{6}$$
$$0 \le m \le M-1$$

4. *Discrete Cosine Transform (DCT) of the Logarithmic Spectrum:* The logarithm of the Mel spectrum is computed before the DCT is computed. The Mel cepstral coefficients are computed as follows:

$$c\left(n\right) = \sum_{m=0}^{M-1} log_{10}\left(s\left(m\right)\right) cos\left(\frac{\pi n(m-0.5)}{M}\right) \tag{7}$$
$$n = 0, 1, 2..., C-1$$

where C is the number of MFCCs. In this work 13 MFCCs are extracted,

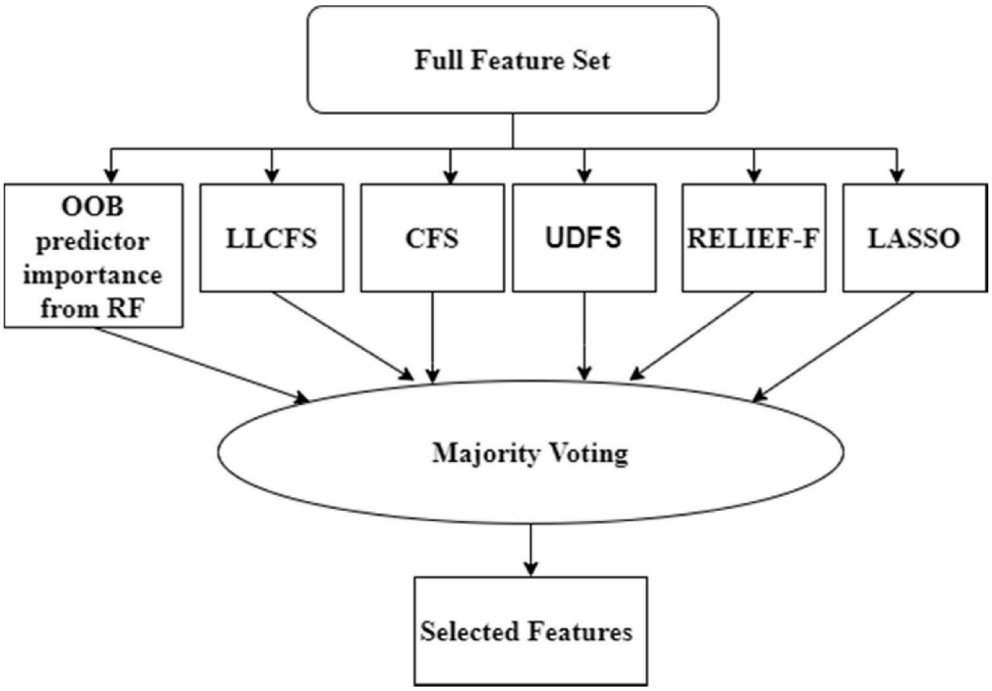

Fig. 3. Hybrid feature selection technique.

resulting in a 13 X n MFCC matrix. The matrix is converted into a one-dimensional feature array by taking the mean and the upper diagonal of the variance from the MFCC matrix. Thus 104 MFCC features are calculated for each 5s PPG signal segment (Fig. 2).

3.2.3. Normalization

The data points are organized by converting them to a common scale using normalization. In this work, min-max and Z-score normalization techniques [37] are used. The data are scaled to the range of [0, 1] with min-max normalization. In Z-score normalization, the Z-score of each feature vector is calculated using Eq. (8), where $\bar{X}$ and σ denote mean and standard deviation, respectively.

$$Z = \frac{x - \bar{X}}{\sigma} \tag{8}$$

3.2.4. Classification

To classify the data into normal, diabetic and prediabetic classes, various supervised learning classifiers KNN, SVM, RF and XGBoost are taken into account. The training is done with 80 % of the data and the remaining 20 % of the data is used for testing.

1) *KNN:* Because of its simplicity, it is a widely used classification technique. It works on the principle that objects of the same class are placed in close proximity [38]. The majority vote among the K nearest neighbors is the criterion for the classification in KNN. It is a lazy learning method with no computations in the training phase. All available data (training data and class labels) are stored in the training phase. In the classification phase, voting is done by the K nearest neighbors, which are found by the Euclidean distance between the data points in space. A suitable value of K should be chosen in order to achieve the minimum error [38].
2) *SVM:* It is a binary classifier that finds a hyperplane to separate the classes with maximum margin and minimum error [39]. Using various kernel tricks, the data is projected into a higher-dimensional space in which the hyperplane can be found with better classification efficiency [40]. SVM is used in a wide variety of applications because of its low computational cost and the ability to generalize performance. In this work, the one vs. one technique is used for multi-class SVM [41].
3) *RF:* It is a popular classifier for multi-class problems [42,43]. It is a collection of tree-based classifiers [44] in which each tree assigns a class label to the input and finally the class with the most votes among the trees is assigned. Each tree is built on an independently sampled random vector. RF is fast, resistant to noise, and can handle highly non-linear data.
4) *XGBoost:* Due to the high computing speed and improved performance, it is a highly efficient ML classifier [45]. The model is created by boosting tree models i.e. the existing models are adjusted by recursively adding new models. Gradient boosting uses the gradient descent algorithm for minimization when a new model is added. XGBoost uses multi-threading of the CPU, which in turn reduces the computing time.

To improve efficiency and reduce computational complexity, a hybrid feature selection technique is also applied to the extracted features.

3.2.5. Hybrid feature selection

Feature selection is an important step in removing redundant, noisy, and unimportant features when building the ML system. This increases the system efficiency by reducing the computing time. In this work a Hybrid FS method based on different feature selection techniques (Fig. 3) is used to select a minimal set of the best features. All techniques are used independently for feature selection and the final selection is based on majority voting. The feature selection techniques considered for this purpose are listed below:

1) *OOB predictor importance from RF*: In RF, the percentage importance of each feature is computed using an out-of-bag (OOB) dataset of each tree. The OOB data are the part of the dataset that is left out when building a tree [44]. The RF performance for each feature with and without a randomly permuted feature value in OOB is compared. Performance is assessed using the misclassification rate (MCR); An increase in the MCR indicates a greater importance of this particular feature. Hence, the predictor importance is obtained for each feature.

Table 1
Performance of DMDS using different classifiers.

Classifier		System	Accuracy (%)		
			Without Normalization	Min-Max Normalization	Z-score Normalization
KNN		**DMDS1**	**83.87 (K = 1)**	71.47 (K = 5)	67.95 (K = 21)
SVM	Linear	**DMDS2**	68.64	65.27	68.99
	Polynomial		71.26	58.51	67.54
	RBF		**84.49**	64.44	72.16
RF		**DMDS3**	98.76 (N t = 94)	**99.24 (N.t = 79)**	98.55 (N.t = 36)
XGBoost		**DMDS4**	**99.79**	99.65	98.62

*N.t – number of trees for random forest.

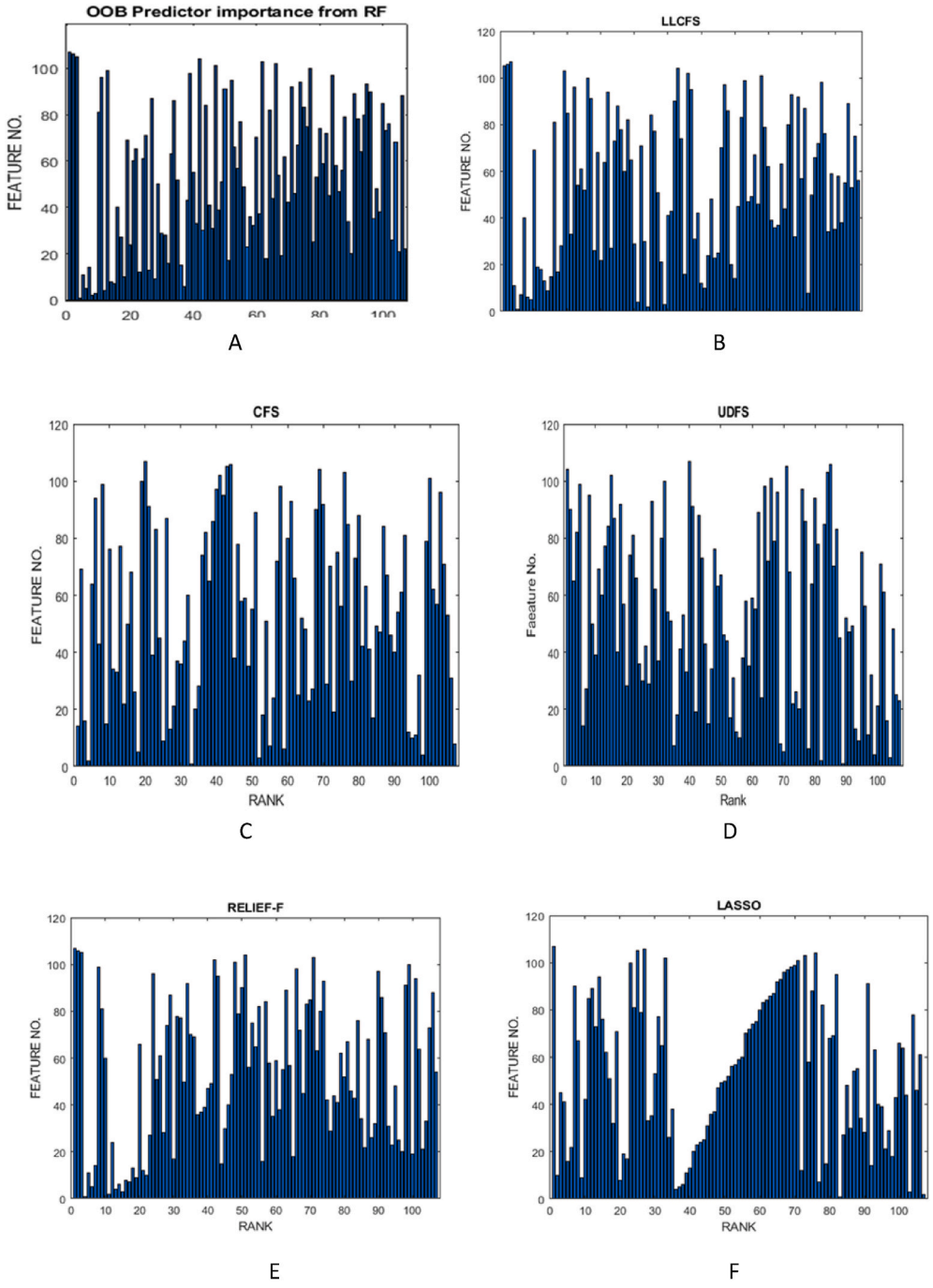

Fig. 4. Results of feature selection by A- OOB predictor importance from RF, B - LLCFS, C–CFS, D – UDFS, E – RELIEF-F, F - LASSO.

2) *Local Learning-based Clustering Feature Selection (LLCFS)*: The selection of the features is made taking into account the feature relevance in the regularization of local learning on the basis of clustering. The Laplacian graph is iteratively updated in the scheme [46].
3) *Correlation-based Feature Selection (CFS)*: In this scheme, features are selected by correlation in pairs [47]. This method assumes that the irrelevant features have a low correlation with class labels.
4) *Unsupervised Discriminative Feature Selection (UDFS)*: The distinguishing features are selected in batch mode. A joint framework consisting of discriminant analysis and $L_{2,1}$ norm normalization is implemented [48].
5) *RELIEF-F*: It is an extension of the RELIEF algorithm to handle multi-class problems [49]. It is a supervised feature weighting method based on the nearest neighbor concept. The feature score decreases if a feature value difference is found in neighbors of the same class, while the feature score increases if the feature value is the same in neighbors of another class.
6) *Least Absolute Shrinkage and Selection Operator (LASSO):* A regularization process is applied to the model that penalizes the regression coefficients to shrink some of them to zero. The variables with non-zero coefficients are selected during the feature selection [50].

3.3. Performance analysis

The receiver operating characteristic (ROC) curve and the entropy triangle are considered to evaluate the performance of ML classifiers. ROC is often used in clinical diagnostic test models [51]. For different classification thresholds, the true positive rate (TPR) (Eq. (9)) is plotted against the false positive rate (FPR) (Eq. (10)). TPR stands for the sensitivity and FPR is a measure for the specificity ($\boldsymbol{FPR} = 1 - \boldsymbol{specificity}$). The area under the ROC curve (AUC) indicates the classifier's ability to distinguish between classes. The value of the AUC is in the range [0, 1]. AUC is 1 for a perfect classifier and is represented by a line parallel to the x-axis. The diagonal line represents a classifier with an AUC value of 0.5. This corresponds to a random classifier without the ability to differentiate (50 % sensitivity, 50 % specificity). In this work, the ROC curve is plotted for each class, as this is a multi-class problem. The micro-average and macro average are also computed. The micro-average is calculated by summing the individual values for true positive (TP), true negative (TN), false positive (FP) and false negative (FN), while the macro-average is calculated as the average performance for all classes.

$$TPR = \frac{TP}{TP + FN} \tag{9}$$

$$FPR = \frac{FP}{FP + TN} \tag{10}$$

In addition, the entropy triangle (ET) helps to visualize the performance of multi-class classifiers [52]. The position of different classifiers on the ET provides information about the classification performance. The classifier with the highest accuracy is placed at the apex of ET. The classifiers that cannot give results for difficult problems are placed at the left vertex, they are the worst classifiers. The classifiers at the right vertex process simple data and therefore offer high accuracy.

4. Results and discussion

The PPG signal and PhyP from the database [30] are used to develop the non-invasive DMDS. The MFCC features are extracted from the preprocessed PPG signal segments. 104 MFCC features together with the PhyP (weight (kg), height (m), age) form the input feature set for the proposed DMDS. All features are considered for the classification without normalization, with min-max normalization and z-score normalization. Table 1 shows the percent accuracy of various classifiers. It turns out that KNN and radial basis function (RBF) kernel SVM classifiers perform better for input data without any normalization (DMDS1 and DMDS2, respectively). However, the best performance of RF-based DMDS is achieved with min-max normalized input data (DMDS3). XGBoost achieves the highest performance among the other classifiers with input data without any normalization (DMDS4).

Table 2
Selected features by different FS techniques in Hybrid FS.

Feature Selection Method	Features in the First 5 Ranks
OOB predictor importance from RF	F107, F106, F105, F1, F11
LLCFS	F105, F106, F107, F11, F1
CFS	F14, F69, F16, F2, F64
UDFS	F104, F90, F65, F82, F99
RELIEF-F	F107, F106, F105, F1, F11
LASSO	F107, F10, F45, F41, F16
Selected features based on majority voting	**F1, F11, F105, F106, F107**

Furthermore, Hybrid FS is applied to the training dataset without normalization to select the 5 most discriminatory features. The results of the various feature selection algorithms in Hybrid FS are shown in Fig. 4. The features in the first 5 ranks of each FS method are listed in Table 2. The feature set contains 107 features, of which 104 features are MFCC of the PPG signal and the remaining three are weight, height and age. The final selection is made by majority vote and the five selected features are F1, F11, F105, F106 and F107. Therefore, these features are used with the XGBoost in the Hybrid FS-based XGBoost DMDS (DMDS5).

Table 3 shows the statistical analysis of selected features in the three classes including analysis of variance (ANOVA). The performance indexes include mean, standard deviation, degrees of freedom (Df), p-value, and F-value. ANOVA uses the F-distribution to compare the means of more than two groups [53]. The test statistic, F-value is calculated according to Eq. (11), where MS_B and MS_W are the mean square between the samples and the mean square within the samples, respectively.

$$F = \frac{MS_B}{MS_W} \tag{11}$$

where,

$$MS_B = \frac{Sum\ of\ square\ between\ sample}{Df_B}$$

$$MS_W = \frac{Sum\ of\ square\ within\ sample}{Df_W}$$

Df_B represents the degrees of freedom for the variation between groups, given by k-1, k is the number of levels (number of classes), in this work it is 3. Df_w is the degrees of freedom for the variation within the group and is equal to N-k, where N is the total sample size. The p-value is the probability of observing an F-statistic in the F-distribution that is greater than or equal to the obtained F-value. A p-value <0.05 indicates a significant difference between the classes and there is an inverse relationship between the F-value and the p-value, i.e., a higher F-value corresponds to a significantly lower p-value. The statistical analysis shows that there is a significant difference in the features between the classes and this scheme is able to correctly distinguish the classes.

The performance of XGBoost DMDS with various feature selection techniques is shown in Table 4. The highest accuracy of 99.93 % is achieved with the Hybrid FS-based system. The system also offers high sensitivity and specificity, i.e., 99.93 % and 99.94 %, respectively. A 10-fold cross-validation is also done to review the overfitting problems. The 10-fold cross-validation accuracy is $99.95 \pm (6.3 \times 10^{-4})$, the small value of the standard deviation (6.3×10^{-4}) implies that the proposed system has no overfitting.

The confusion matrix of the DMDS5 is shown in Fig. 5, with classes 1, 2 and 3 representing normal, prediabetic and diabetic patients, respectively. The diagonal elements represent the TP cases for each class, while the prediction error is represented by the off-diagonal elements. It is also

Table 3
Statistical analysis of selected features.

Feature	Mean			Standard Deviation			ANOVA TEST			
	Normal	Pre-diabetic	Diabetic	Normal	Prediabetic	Diabetic	Df_B	Df_W	F-value	p-value
F1	−5.11	−4.50	−4.84	2.42	2.39	2.89	2.0	7260	1.37×10^{-8}	18.16
F11	1.30×10^{-13}	1.23×10^{-13}	1.27×10^{-13}	2.91×10^{-14}	2.94×10^{-14}	3.49×10^{-13}	2.0	7260	7.15×10^{-9}	18.80
F105	74.67	78.18	78.49	16.55	13.94	16.40	2.0	7260	5.13×10^{-21}	47.02
F106	1.60	1.59	1.61	0.10	1.59	1.61	2.0	7260	3.75×10^{-9}	19.45
F107	45.57	49.58	55.52	11.90	49.58	55.52	2.0	7260	≈0	728.90

Table 4
Performance of XGBoost DMDS with different feature selection methods.

Feature selection method	Features	Sensitivity (%)	Specificity (%)	Accuracy (%)
CFS	F14, F69, F16, F2, F64	38.14	70.20	57.82
UDFS	F104, F90, F65, F82, F99	36.13	68.79	56.72
LASSO	F107, F10, F45, F41, F16	66.84	85.44	77.19
Hybrid FS	**F1, F11, F105, F106, F107**	99.93	99.94	**99.93**

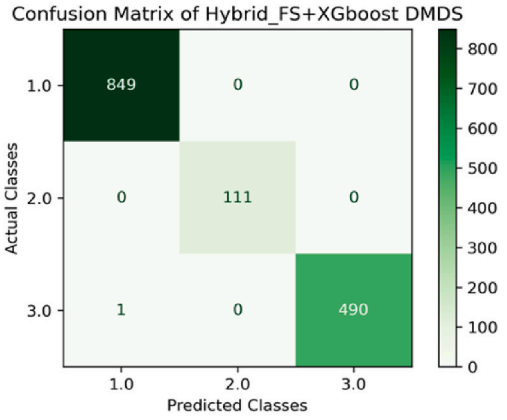

Fig. 5. Confusion matrix of Hybrid-FS + XGBoost DMDS.

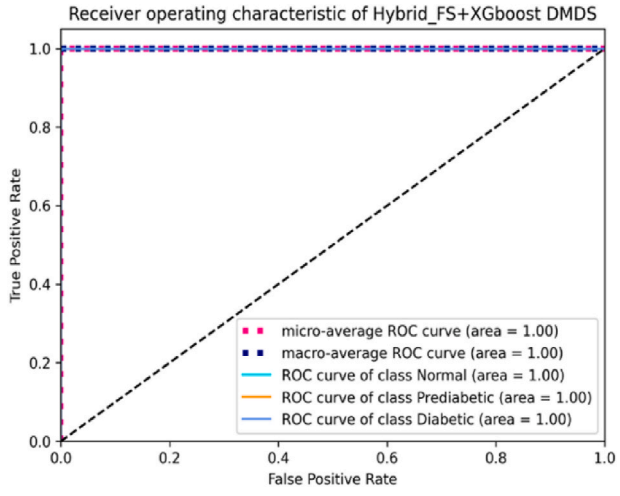

Fig. 6. Receiver operating characteristic curve of Hybrid-FS + XGBoost DMDS.

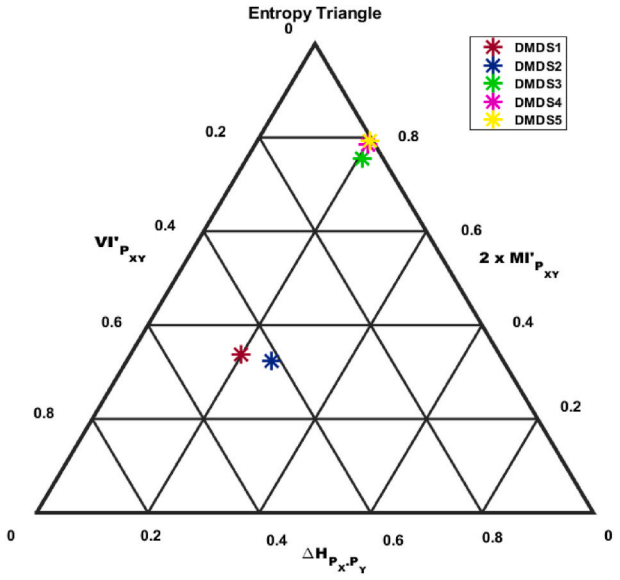

Fig. 7. Entropy triangle of different DMDS.

observed that most of the off-diagonal elements in the confusion matrix are zero, which means that the miss-classification is quite less. The ROC curve of this system is plotted (Fig. 6) and an AUC value of 1 is obtained, indicating a perfect classifier.

4.1. Entropy triangle analysis of DMDS

The placement of various DMDS on the ET is shown in Fig. 7. It is observed that all designed systems are positioned near the left side of the ET, which means that the application under consideration presents a difficult classification problem. Also, DMDS3, DMDS4, and DMDS5 are closer to the apex. The proposed DMDS5 with 5 selected features shows the highest performance among the other examined systems and thus proves that it can work efficiently with real data.

4.2. Comparison with state-of -the–art techniques

The proposed DMDS is compared with the state-of-the-art technologies for DM detection based on the PPG signal. Keikhosravi et al. [18] used a modified model of the human upper vasculature that uses bilateral PPG signals to differentiate between diabetic and healthy subjects. Singular value decomposition is used to reduce the number of features and the classification is done using the Naïve Bayes classifier. Further research [19] uses features related to heart rate variability (HRV) and the signal pattern of the PPG signal for DM classification, and SVM with weighted fusion of features is the classifier. Another study [20] used for comparison uses 10 features of the toe PPG signal with SVM classifier. In addition, the PPG signal recorded by the smartphone is used in a 34-layer CNN-based model for diabetes detection [21]. Recently, features of the PPG signal extracted from fingertip videos [22] are classified using Gaussian SVM for diabetes detection. The PPG signal is formed

Table 5
Comparison with state-of-the-art techniques.

Author (year)	Input data	Classifier	Classes	Accuracy (%)
Keikhosravi et al. (2013) [18]	Bilateral finger PPG parameters	Naïve Bayes	Normal and Diabetic	93.5
V. Reddy et al. (2017) [19]	Features related to heart rate variability (HRV) and signal pattern of PPG signal	SVM + weight based fusion of features	Normal and Diabetic	89
Nirala et al. (2019) [20]	Features of toe PPG signal	SVM	Normal and Diabetic	97.87
R. Avram et al. (2019) [21]	Smartphone-based PPG signal	Deep Learning	Normal and Diabetic	77.2
G. Zhang et al. (2020) [22]	Features of PPG signal retrieved from smartphone - fingertip video	GSVM	Normal and Diabetic	81.49
Proposed DMDS	MFCC features of wrist band PPG signal and Physiological parameters (5 features)	**XGBoost**	Normal, Prediabetic and Diabetic	**99.93**

Table 6
Between class and within class mean square error of selected features.

Features	Between class mean square error (MSE_B)	Within class mean square error (MSE_W)
F1	1.21×10^{2}	6.70
F11	1.83×10^{-26}	9.74×10^{-28}
F105	1.25×10^{4}	2.66×10^{2}
F106	1.94×10^{-1}	9.97×10^{-3}
F107	7.7×10^{4}	1.05×10^{2}

using the changes in the pixel intensity in each frame, and the features are extracted from the Gaussian characteristic parameters of the PPG signal. However, Table 5 shows that the Hybrid FS + XGBoost DMDS proposed in this work turns out to be the perfect classification method compared to the existing schemes. Since there is a three-class classification, it helps to identify even the prediabetic cases. Thus, the application of the novel MFCC features from the PPG signal with hybrid feature selection can lead to an effective non-invasive DM detection system.

The limits of this research are also analyzed. This work is based on data collected from 217 participants in a hospital in Cuenca, Ecuador [30]. The sample size is increased by sampling the data for reliable performance while applying ML algorithms. This leads to input data with less variance. An ANOVA test of selected features (Table 6) shows that the mean square error within the class (MSE_W) is significantly less than the mean square error between classes (MSE_B), e.g. for the feature set F11, $MSE_W = 9.74 \times 10^{-28}$ and $MSE_B = 1.83 \times 10^{-26}$. The low value for MSE_W indicates that there is less variance within each class. This problem can be effectively solved by considering more participants for each case to increase the data size. In addition, data from different population groups in different geographic locations can be considered for more reliable testing of the DM screening system. Also, this research is intended to develop an intelligent non-invasive screening system for DM that classifies normal, diabetic, and prediabetic cases. Therefore, continuous glucose levels cannot be measured with this system, which leaves room for further study to improve the system. Over time, dynamic physiological factors such as temperature, sweat, etc., can affect the system. It has been shown that the results can be improved by taking these parameters into account during calibration. The Monte Carlo (MC) methodology [54] is used to assess the effectiveness of calibration for non-invasive continuous glucose monitoring systems (NI-CGM). This shows that the estimation of the glucose level by means of multi-sensor signals and a mathematical model, in which the influence of other physical parameters is taken into account, does not require any further calibration after sweat events. Therefore, the development of a mathematical model based on external parameters can be considered for future work. Another limitation of the proposed XGBoost classifier is its sensitivity to outliers. In the future, studies on the outliers in the database can also be carried out. In the present work, however, the highest performance is achieved and a 10-fold cross-validation results in an accuracy of 99.95 % with a standard deviation of 6.3×10^{-4}. This ensures that overfitting does not occur. The results of this research are promising and form the basis for the development of intelligent non-invasive DM screening systems that can play a major role in the control and management of DM.

5. Conclusion

This work focuses on the design of a non-invasive diabetes mellitus detection system (DMDS) using novel Mel frequency cepstral coefficients (MFCC) features of the wristband photoplethysmogram (PPG) signal and physiological parameters (PhyP). Various classifiers are used to develop DMDS that distinguish between normal, diabetic and prediabetic cases. The XGBoost DMDS (DMDS4) achieves an accuracy of 99.79 % for the input feature set without normalization. Furthermore, a hybrid feature selection technique is used, which selects the 5 best features from the feature set and thus reduces the complexity of the system. The Hybrid FS based XGBoost DMDS (DMDS5) offers a significant improvement in accuracy (99.93 %) compared to the existing techniques. The performance of the proposed system is also assessed by a 10-fold cross-validation. It is concluded that by using appropriate features of the PPG signal, an effective, inexpensive, and computationally less complex non-invasive DM screening tool can be built. In the future, DMDS can be modified to continuously monitor blood sugar levels and a hardware implementation of the system can also be made.

References

[1] Z. Punthakee, R. Goldenberg, P. Katz, Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome, Can. J. Diabetes 42 (2018) S10–S15, https://doi.org/10.1016/j.jcjd.2017.10.003. Available:.
[2] A. Kharroubi, Diabetes mellitus: the epidemic of the century, World J. Diabetes 6 (6) (2015) 850, https://doi.org/10.4239/wjd.v6.i6.850. Available:.
[3] R. Khan, Z. Chua, J. Tan, Y. Yang, Z. Liao, Y. Zhao, From pre-diabetes to diabetes: diagnosis, treatments and translational research, Medicina 55 (9) (2019) 546, https://doi.org/10.3390/medicina55090546. Available:.
[4] P. Saeedi, et al., "Global and Regional Diabetes Prevalence Estimates for 2019 and Projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th Edition", Diabetes Research and Clinical Practice, 157, 2019, p. 107843, https://doi.org/10.1016/j.diabres.2019.107843. Available:.
[5] E. Dobrică, M. Găman, M. Cozma, O. Bratu, A. Pantea Stoian, C. Diaconu, Polypharmacy in type 2 diabetes mellitus: insights from an internal medicine department, Medicina 55 (8) (2019) 436, https://doi.org/10.3390/medicina55080436. Available:.
[6] G. Bantie, et al., Prevalence of undiagnosed diabetes mellitus and associated factors among adult residents of Bahir Dar city, northwest Ethiopia: a community-based cross-sectional study", BMJ Open 9 (10) (2019), e030158 https://doi.org/10.1136/bmjopen-2019-030158. Available:.
[7] D. Shankaracharya, S. Odedra, Samanta, A. Vidyarthi, Computational intelligence in early diabetes diagnosis: a review, Rev. Diabet. Stud. 7 (4) (2010) 252–262, https://doi.org/10.1900/rds.2010.7.252. Available:.
[8] S. El-Sappagh, M. Elmogy, A. Riad, A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis, Artif. Intell. Med. 65 (3) (2015) 179–208, https://doi.org/10.1016/j.artmed.2015.08.003. Available:.
[9] A. Choudhury, D. Gupta, A survey on medical diagnosis of diabetes using machine learning techniques, Adv. Intell. Syst.Comput. (2018) 67–78, https://doi.org/10.1007/978-981-13-1280-9_6. Available:.
[10] H. Naz, S. Ahuja, Deep learning approach for diabetes prediction using PIMA Indian dataset, J. Diabetes Metab. Disord. 19 (1) (2020) 391–403, https://doi.org/10.1007/s40200-020-00520-5. Available:.
[11] J. Chaki, S. Thillai Ganesh, S. Cidham, S. Ananda Theertan, Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: a systematic review, J. King Saud Univ.Comput. Inf.Sci. (2020), https://doi.org/10.1016/j.jksuci.2020.06.013. Available:.

[12] S. Lekha, M. S, Recent advancements and future prospects on E-nose sensors technology and machine learning approaches for non-invasive diabetes diagnosis: a review, IEEE.Rev. Biomed. Eng. 14 (2021) 127–138, https://doi.org/10.1109/rbme.2020.2993591. Available:.

[13] N. Barakat, A. Bradley, M. Barakat, Intelligible support vector machines for diagnosis of diabetes mellitus, IEEE Trans. Inf. Technol. Biomed. 14 (4) (2010) 1114–1120, https://doi.org/10.1109/titb.2009.2039485. Available:.

[14] O.S. Soliman, E. AboElhamd, Classification of diabetes mellitus using modified particle swarm optimization and least squares support vector machine, Int. J. Comput. Trends Technol. 8 (1) (2014) 38–44, https://doi.org/10.14445/22312803/ijctt-v8p108. Available:.

[15] T. Daghistani, R. Alshammari, Diagnosis of diabetes by applying data mining classification techniques, Int. J. Adv. Comput. Sci. Appl. 7 (7) (2016), 070747, https://doi.org/10.14569/ijacsa.2016. Available:.

[16] S. Ling, P. San, H. Nguyen, Non-invasive hypoglycemia monitoring system using extreme learning machine for Type 1 diabetes, ISA (Instrum. Soc. Am.) Trans. 64 (2016) 440–446, https://doi.org/10.1016/j.isatra.2016.05.008. Available:.

[17] G. S, R. V, K.P. S, Diabetes detection using deep learning algorithms, ICT Express 4 (4) (2018) 243–246, https://doi.org/10.1016/j.icte.2018.10.005. Available:.

[18] A. Keikhosravi, H. Aghajani, E. Zahedi, Discrimination of bilateral finger photoplethysmogram responses to reactive hyperemia in diabetic and healthy subjects using a differential vascular model framework, Physiol. Meas. 34 (5) (2013) 513–525, https://doi.org/10.1088/0967-3334/34/5/513. Available:.

[19] V. Reddy, A. Dutta Choudhury, S. Jayaraman, N. Kumar Thokala, P. Deshpande, V. Kaliaperumal, PerDMCS: weighted fusion of PPG signal features for robust and efficient diabetes mellitus classification", Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (2017) 553–560, https://doi.org/10.5220/0006297205530560. Available:.

[20] N. Nirala, R. Periyasamy, B.K. Singh, A. Kumar, Detection of type-2 diabetes using characteristics of toe photoplethysmogram by applying support vector machine, Biocybern. Biomed. Eng. 39 (1) (2019) 38–51, https://doi.org/10.1016/j.bbe.2018.09.007. Available:.

[21] R. Avram, et al., Predicting diabetes from photoplethysmography using deep learning", J. Am. Coll. Cardiol. 73 (9) (2019) 16, https://doi.org/10.1016/s0735-1097(19)33778-7. Available:.

[22] G. Zhang, et al., A noninvasive blood glucose monitoring system based on smartphone PPG signal processing and machine learning, IEEE Transactions on Industrial Informatics 16 (11) (2020) 7209–7218, https://doi.org/10.1109/tii.2020.2975222. Available:.

[23] Y. Sun, N. Thakor, Photoplethysmography revisited: from contact to noncontact, from point to imaging, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 63 (3) (2016) 463–477, https://doi.org/10.1109/tbme.2015.2476337. Available:.

[24] M. Ghamari, A review on wearable photoplethysmography sensors and their potential future applications in health care, International Journal of Biosensors & Bioelectronics 4 (4) (2018) 195, https://doi.org/10.15406/ijbsbe.2018.04.00125. Available:.

[25] A.L.A. J Dekker, "Monitoring Physiological Parameters Based on Variations in a Photoplethysmographic Signal", U.S. Patent 7,001,337, Feb. 21, 2006.

[26] K. Pilt, R. Ferenets, K. Meigas, L. Lindberg, K. Temitski, M. Viigimaa, New photoplethysmographic signal analysis algorithm for arterial stiffness estimation", Sci. World J. (2013) 1–9, https://doi.org/10.1155/2013/169035. Available:.

[27] I. Muhammad, et al., Arterial stiffness and incidence of diabetes: a population-based cohort study, Diabetes Care 40 (12) (2017) 1739–1745, https://doi.org/10.2337/dc17-1071. Available:.

[28] N. Gray, G. Picone, F. Sloan, A. Yashkin, Relation between BMI and diabetes mellitus and its complications among US older adults, South. Med. J. 108 (1) (2015) 29–36, https://doi.org/10.14423/smj.0000000000000214. Available:.

[29] A. Prabha, J. Yadav, A. Rani, V. Singh, Non-invasive diabetes mellitus detection system using machine learning techniques, in: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence),, 2021, pp. 948–953, https://doi.org/10.1109/confluence51648.2021.9377138. Available:.

[30] C. Salamea, E. Narvaez, M. Montalvo, Database proposal for correlation of glucose and photoplethysmography signals, in: Advances in Intelligent Systems and Computing, Springer, Cham, 2020, pp. 44–53, https://doi.org/10.1007/978-3-030-32033-1_5. Available:.

[31] B. Logan, Mel frequency cepstral coefficients for music modeling, In *Ismir* 270 (2000) 1–11.

[32] S. Zokaee, K. Faez, Human identification based on ECG and palmprint, Int. J. Electr. Comput. Eng. 2 (2) (2012) 261.

[33] M. Othman, A. Wahab, I. Karim, M. Dzulkifli, I. Alshaikli, EEG emotion recognition based on the dimensional models of emotions, Procedia - Social and Behavioral Sciences 97 (2013) 30–37, https://doi.org/10.1016/j.sbspro.2013.10.201. Available:.

[34] S. Ismail, I. Siddiqi, U. Akram, Localization and classification of heart beats in phonocardiography signals —a comprehensive review, EURASIP J. Appl. Signal Process. (1) (2018) 1–27, https://doi.org/10.1186/s13634-018-0545-9, 2018.

[35] J. Deller, J. Proakis, J. Hansen, Discrete-time Processing of Speech Signals, Institute of Electrical and Electronics Engineers, 2000.

[36] J. Picone, Signal modeling techniques in speech recognition, Proc. IEEE 81 (9) (1993) 1215–1247, https://doi.org/10.1109/5.237532. Available:.

[37] A. Jain, K. Nandakumar, A. Ross, Score normalization in multimodal biometric systems, Pattern Recogn. 38 (12) (2005) 2270–2285, https://doi.org/10.1016/j.patcog.2005.01.012. Available:.

[38] Z. Zhang, Introduction to machine learning: k-nearest neighbors", Ann. Transl. Med. 4 (11) (2016) https://doi.org/10.21037/atm.2016.03.37, 218-218.

[39] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[40] A. Prabha, A. Trivedi, A. Kumar, C. Kumar, Automated system for obstructive sleep apnea detection using heart rate variability and respiratory rate variability", in: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 1303–1307, https://doi.org/10.1109/icacci.2017.8126021. Available:.

[41] C.W. Hsu, C.J. Lin, A comparison of methods for multiclass support vector machines, IEEE Trans. Neural Network. 13 (2) (2002) 415–425, https://doi.org/10.1109/72.991427. Available:.

[42] X. Chen, M. Liu, Prediction of protein-protein interactions using random decision forest framework, Bioinformatics 21 (24) (2005) 4394–4400, https://doi.org/10.1093/bioinformatics/bti721. Available:.

[43] J. Ion Titapiccolo, et al., Artificial intelligence models to stratify cardiovascular risk in incident hemodialysis patients, Expert Syst. Appl. 40 (11) (2013) 4679–4686, https://doi.org/10.1016/j.eswa.2013.02.005. Available:.

[44] L. Breiman, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/a:1010933404324. Available:.

[45] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794 [Online]. Available: https://arxiv.org/abs/1603.02754.

[46] H. Zeng, Y. Cheung, Feature selection and kernel learning for local learning-based clustering, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1532–1547, https://doi.org/10.1109/tpami.2010.215. Available:.

[47] M.A. Hall, "Correlation-based Feature Selection for Machine learning." New Zealand, 1999.

[48] Y. Yang, H. Shen, Z. Ma, Z. Huang, X. Zhou, ℓ 2,1-Norm regularized discriminative feature selection for unsupervised learning, in: IJCAI International Joint Conference on Artificial Intelligence, 2011. Available: http://hdl.handle.net/10453/119490.

[49] I. Kononenko, Estimating attributes: analysis and extensions of RELIEF, Mach. Learn.: ECML- 94 (1994) 171–182, https://doi.org/10.1007/3-540-57868-4_57. Available:.

[50] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Stat. Soc. B 58 (1) (1996) 267–288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x. Available:.

[51] L. Lusted, Signal detectability and medical decision-making, Science 171 (3977) (1971) 1217–1219, https://doi.org/10.1126/science.171.3977.1217. Available:.

[52] F.J. Valverde-Albacete, C. Peláez-Moreno, "The Multivariate Entropy Triangle and Applications," *Lecture Notes In Computer Science*, 2016, pp. 647–658, https://doi.org/10.1007/978-3-319-32034-2_54. Available:.

[53] L. Sthle, S. Wold, Analysis of variance (ANOVA), Chemometr. Intell. Lab. Syst. 6 (4) (1989) 259–272, https://doi.org/10.1016/0169-7439(89)80095-4. Available:.

[54] M. Zanon, et al., Non-invasive continuous glucose monitoring with multi-sensor systems: a Monte Carlo-based methodology for assessing calibration robustness, Sensors 13 (6) (2013) 7279–7295, https://doi.org/10.3390/s130607279. Available:.