Egyptian Informatics Journal 21 (2020) 183–190

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Full length article

GDD: Geometrical driven diagnosis based on biomedical data

Ahmed E. Mohamed *, Mona Farouk

Cairo University, Egypt

ARTICLE INFO

Article history:
Received 27 June 2019
Revised 21 September 2019
Accepted 23 April 2020
Available online 13 May 2020

Keywords:
GDD
Medical diagnosis
Classification
Machine learning
Diabetes
Big data
Bioinformatics

ABSTRACT

Modern medical diagnosis heavily rely on bio-medical and clinical data. Machine learning algorithms have proven effectiveness in mining this data to provide an aid to the physicians in supporting their decisions. In response, machine learning based approaches were developed to address this problem. These approaches vary in terms of effectiveness and computational cost. Attention has been paid towards non-communicable diseases as they are very common and have life threatening risk factors. The diagnosis of diabetes or breast cancer can be considered a binary classification problem. This paper proposes a new machine learning based algorithm, Geometrical Driven Diagnosis (GDD), to diagnose diabetes and breast cancer with accuracy up to 99.96% and 95.8% respectively.

1. Introduction

Diabetes has two categories [1]. Type 1 diabetes [2] is a chronic condition where the $\beta - cells$ of the pancreas fail to generate a sufficient amount of the insulin hormone for controlling the blood glucose level. In contrast, type 2 diabetes [3] (insulin resistance) is an endocrine disease where the body does not consume the insulin efficiently, thus starts by pressuring the pancreas to greater amounts of more insulin.

Type 2 diabetes [4] is the most common type of diabetes. The total number of type 2 patients can be estimated by 415 million humans worldwide. The estimated number of undiagnosed diabetes patients is 193 millions. Early diagnosis of diabetes along with the proper treatment, can effectively limit the catastrophic effects of diabetes on health.

Type 2 diabetes is a leading cause of death [5] especially in the developing countries. It can have multiple complications including but not limited to: coronaropathy, stroke, heart failure, blindness, etc...

* Corresponding author.
E-mail addresses: ahmed.e.mohamed@eng1.cu.edu.eg (A.E. Mohamed), mona_farouk@eng.cu.edu.eg (M. Farouk).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

Although diabetes can be accurately detected through regular screening [6] or DNA sequencing [7,8], their costs are relatively high for the developing countries population. In response, machine learning algorithms have been developed to address this problem [9,10].

Breast cancer is a condition where the breast cells reproduction activities grows above the normal rate forming a malignant tumours [11].

A study is curried out about the growth of 36 cancers in 185 countries [12]. Then they estimated the number of new breast cancer cases in 2018 to be 2,088,849 worldwide, and the total number of deaths to be 626,679. Breast cancer is rated as the most frequent in homo sapiens' females worldwide. Although the tumour can be visually detected and examined, the tumour classification (benign or malignant) is not a simple task, thus women are advised to perform genetic testing [13]. However, these kind of tests are not available or expensive in the developing countries.

To the best of our knowledge, most of the proposed algorithms rely on separating the two classes of the binary classification problem, represented in the diabetes or breast cancer diagnosis, by a hyper-dimensional plane, or a polygon. Whereas in some cases the two classes cannot be simply separated by the aforementioned method. Alternatively different models have to be created for each class to identify the density centres for each class. In order to address this problem, we used k-means clustering [14] algorithm to find out the density centroids' positions of each class.

The rest of the paper is organised as follows: Section 2 reviews the most recent advances in the field of machine learning for diagnosis. Section 3 explains the proposed approach. Section 4 presents

https://doi.org/10.1016/j.eij.2020.04.002
1110-8665/

the experiments and the results and proves the excellence of the proposed approach over other recent algorithms in the field. Section 5 provides the conclusion for the proposed approach and the experiments carried. Section 6 discusses the limitations of this work and future work possibilities.

2. Related work

While most machine learning algorithms that target medical diagnosis are different in the part concerning data cleaning and dimension reduction, they are always based on one of two main approaches. These approaches are: Artificial deep neural networks such as multi-layer perceptrons [15] and Classical machine learning algorithms such as Support Vector Machine (SVM) [16].

[17] proposed a method for diagnosing diabetes by using Association Rule Mining (ARM) [18] techniques to analyze and detect patterns in the patients' data. Then using a multi-layer perceptrons Artificial Neural Network (ANN) to develop a diabetes classifier based on clinical data. They performed data pre-processing to eliminate any unnecessary, redundant, missing, or improperly formatted values from the dataset. Then the Apriori algorithm [19] is used to perform ARM. Singh et al. used the association rules generated with confidence higher than 0.75 to eliminate redundant features. The total number of generated rules were 17 rules. Based on these rules, a set of 18 features were selected as independent features for the classification phase. The reduced dataset was split into two parts; a training set and a test set containing 70% and 30% respectively. This dataset was used for training and testing of an ANN model. This model contains 18-inputs, 2-hidden layers of 4-perceptrons each and an output layer. This model achieved a total accuracy of 88.5%.

[20] studied the variation in the performance of the different classification algorithms while treating different data types (alphabetical, numeric and alphanumeric). The study was performed on three different types of algorithms: Random Forest [21], K-Nearest Neighbor (KNN) [22] and Naïve Bayes Classifier (NBC) [23]. To evaluate the performance of the three algorithms in terms of accuracy and computational cost, the study was performed on 8 different datasets of different number of instances, attributes and data types. KNN and Random Forest algorithms required parameter tuning. For the KNN the k-parameter is required. This parameter was chosen to be an odd value ranging from 1 to 50 i.e. $k \in [1:50]$. Different number of trees was chosen for the Random Forest algorithm. These numbers belong to the list {1, 5, 10, 15, 25}. Results show that both Random Forest and KNN had identical performance, while the NBC performance was inferior in terms of accuracy. Random Forest required more computational time and space than KNN. NBC was the lightest. For example considering the Diabetes 130-US hospitals for years 1999–2008 dataset which is an alphanumeric dataset with 52 attributes, the accuracy of the Random Forest, KNN and NBC algorithms was: 79.19%, 78.9% and 31.43% respectively, while the computational time in seconds was 854.97, 114.815 and 2.23 respectively.

[24] studied the computational cost of different data clustering algorithms with biomedical data. These algorithms are K-Means Clustering, density-based clustering [25] and hierarchical clustering [26]. Thangaraju et al. used WEKA [27] to perform the clustering analysis. This analysis was applied to the Diabetes 130-US hospitals for years 1999–2008 dataset [28]. Results show that the hierarchical clustering was the most expensive to be built where the computing time was 3.91 s, the density based clustering took 0.06 s and the least expensive was the K-Means clustering algorithm with a time of only 0.03 s.

[29] studied the efficiency of multi-label classification algorithms for diagnosis and classification of cervical cancer. These algorithms are: NBC, J48 decision tree [30], sequential minimal optimization [31] and Random Forest. These algorithms were evaluated in terms of: accuracy, exact match, hamming loss and rank loss. Results show that the classification accuracy of NBC, sequential minimal optimization and Random Forest were around 80%, in contrast with the J48 decision tree which achieved lower accuracy. The Random Forest achieved the highest exact match ratio of 0.890, while the J48 decision tree and sequential minimal optimization scored 0.887, 0.866 respectively. The NBC came last with 0.839. The hamming distances and rank loss of all of the discussed algorithms were nearly the same.

[32] proposed two SVM-based algorithms to diagnose cervical cancer risk factors. These algorithms are: SVM-based recursive feature elimination and SVM-based Principal Component Analysis [33] (SVM-PCA). In the SVM-based recursive feature elimination method the features were ranked by relevance through training and applying the SVM classifier to each individual feature. The feature that resulted in higher accuracy got a higher rank. Similarly, The SVM-PCA uses the PCA to decompose the feature space and eliminate the redundant and irrelevant features before training and applying the SVM classifier to the data. Wu et al. used the Cervical Cancer (Risk Factors) dataset [28]. Results show that the SVM-PCA accuracy was equal to or slightly higher than the SVM-based recursive feature elimination approach. Both approaches achieved higher accuracies than the naïve SVM.

3. Methods

In this section, a classification algorithm for medical diagnosis is proposed. This section is divided into 4 sub-sections: Pre-processing, Feature selection, Classification algorithm and Classification metrics.

3.1. Pre-processing

The pre-processing phase aims to transform the raw data into a more suitable form for machine learning algorithms. In this stage the obviously irrelevant features for the classification process are eliminated such as: Patient ID from the Breast Cancer dataset, and, Hospital name, Room number, Encounter ID, Length of stay, Payer code, medical specialty of admitting physician, number of outpatients and inpatient from the diabetes dataset. Then features with low variance are removed. After that all the categorial features are converted into numeric features using one hot encoding technique [34] to enhance the quality of data representation. Then features which contain missing values above 50% are removed (for example "weight" in the diabetes dataset has 97% missing values, the breast cancer dataset does not have any missing values). Records as a whole are filtered as well; records with most of the fields missing are removed. After that, the rest of the missing values are filled by the average. Finally, all the remaining features are normalized.

3.2. Feature selection

For the diabetes dataset, the extra trees classifier algorithm [35] is used from the python Scikit-learn machine learning library [36] to ascertain the feature importance. Extra trees classifier is an ensemble algorithm similar to the random forest algorithm, however, it selects the cut point at random instead of figuring out the optimal cut point. Afterwards, the PCA is applied to reduce the data dimensionality to 5 axes (the number of axes is chosen empirically).

For the breast cancer dataset, only the 10 features with variance higher than 1 are selected.

3.3. Classification algorithm

In the classification stage, the algorithm begins with separating the dataset into two subsets: one containing only the positive class

and the other the negative class. Then the positive class data is duplicated one or more times to balance the positive and negative classes. After that, each subset is split into training and testing sets 70% and 30% respectively. The training datasets are clustered using

Table 1
Feature Variance for the Breast Cancer dataset.

Feature	Variance
Fractal dimension se	0.000007
Smoothness se	0.000009
Concave points se	0.000038
Fractal dimension mean	0.000050
Symmetry se	0.000068
Smoothness mean	0.000198
Compactness se	0.000321
Fractal dimension worst	0.000326
Smoothness worst	0.000521
Symmetry mean	0.000752
Concavity se	0.000911
Concave points mean	0.001506
Compactness mean	0.002789
Symmetry worst	0.003828
Concave points worst	0.004321
Concavity mean	0.006355
Compactness worst	0.024755
Concavity worst	0.043524
Radius se	0.076902
Diagnosis	0.234177
Texture se	0.304316
Perimeter se	4.087896
Radius mean	12.418920
Texture mean	18.498909
Radius worst	23.360224
Texture worst	37.776483
Perimeter mean	590.440480
Perimeter worst	1129.130847
Area se	2069.431583
Area mean	123843.554318
Area worst	324167.385102

Table 2
Diabetes dataset classification results comparison.

Algorithm	GDD	[17]	[20]
Score	3.994	Not available	Not available
Accuracy	**0.999**	0.885	0.7919
Sensitivity	1.0	Not available	Not available
Specificity	0.999	Not available	Not available
Positive Predictive Accuracy	1.0	Not available	Not available
Negative Predictive Accuracy	0.994	Not available	Not available

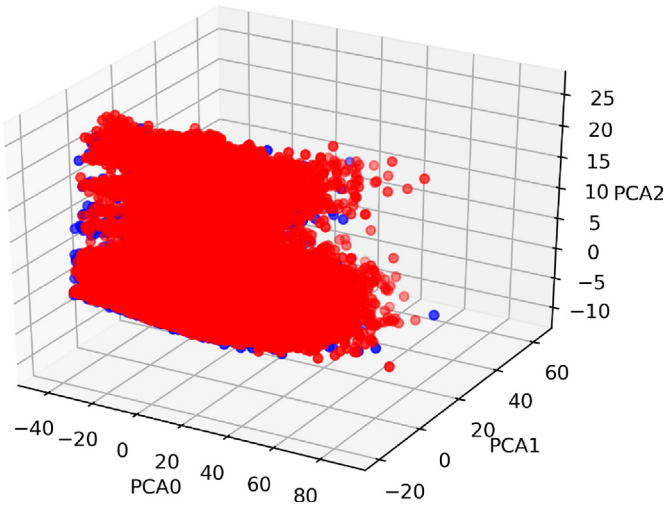

Fig. 2. Diabetes dataset after dimension reduction in 3D.

k-means clustering algorithms into k clusters where $k \in [1{:}20]$ (where k can be different for positive and negative subsets). The center of each cluster is saved for each combination of clusters. Consequently, the performance of each group of clusters is

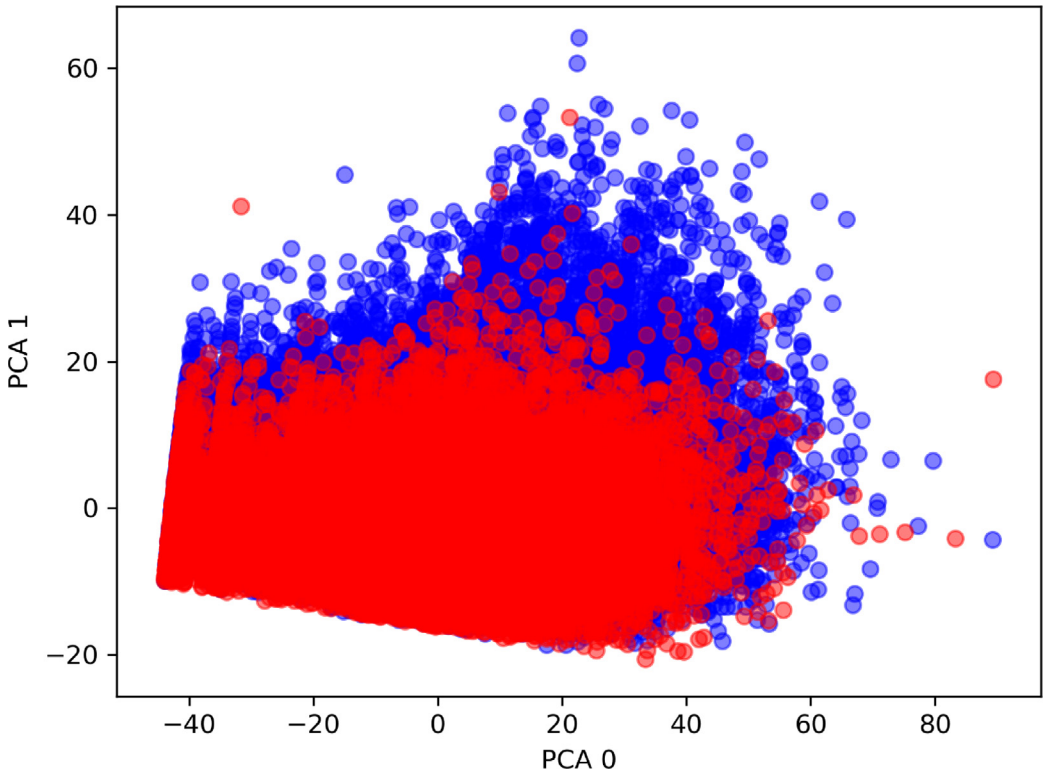

Fig. 1. Diabetes dataset after dimension reduction in 2D.

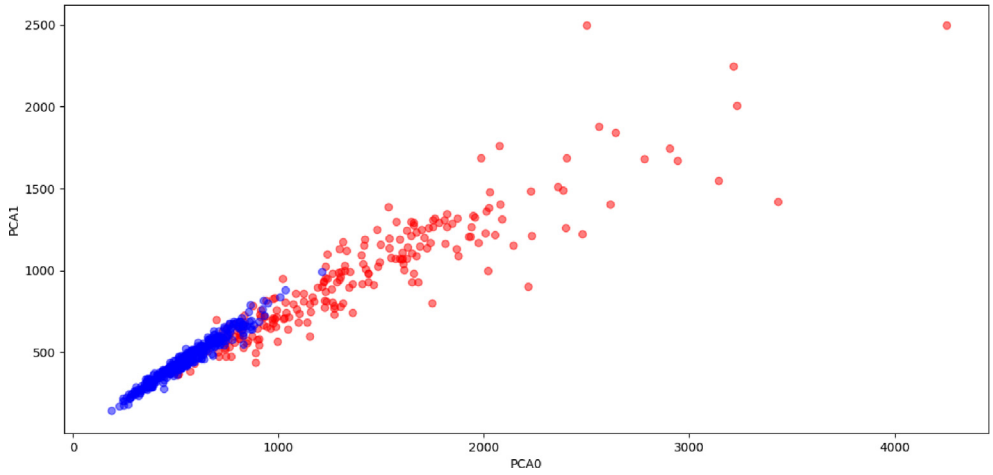

Fig. 3. Breast Cancer dataset after dimension reduction in 2D.

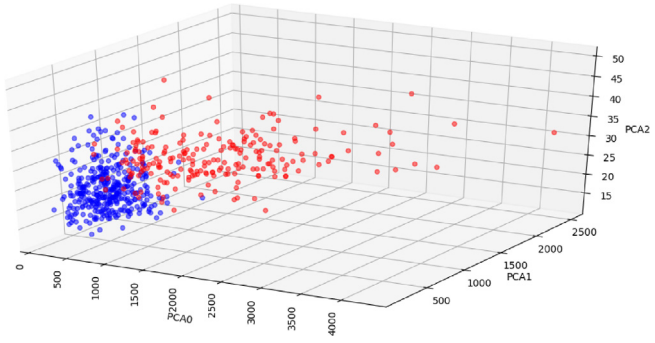

Fig. 4. Breast Cancer dataset after dimension reduction in 3D.

evaluated by the test set to get the clusters combination with the best classification metrics.

The classifier classifies each new point by measuring the Euclidean (geometrical) distance and assigns that point to the nearest centroid. The pseudo code for this algorithm is presented in 1.

3.4. Classification metrics

While dealing with medical data the total classification accuracy is not the only measure used as an indicator of right diagnosis. Thus different performance measures have to be taken into consideration. These measures include:

- Sensitivity: Is the ratio of positive cases that are correctly classified.
- Specificity: Is the ratio of negative cases that are correctly classified.
- Positive predictive accuracy: Is the accuracy of the positive classification.
- Negative predictive accuracy: Is the accuracy of the negative classification.
- Classification score: Is the total algorithm ability to produce quality results.

$$total\ accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Positive\ Predictive\ Accuracy = \frac{TP}{TP + FP} \tag{4}$$

$$Negative\ Predictive\ Accuracy = \frac{TN}{TN + FN} \tag{5}$$

$$Classification\ score = Sensitivity + Specificity + Positive\ Predictive\ Accuracy + Negative\ Predictive\ Accuracy \tag{6}$$

Where the terms True Positive (TP) is the number of positive cases that is classified correctly, False Positive (FP) is the number of negative cases that is misclassified, True Negative (TN) is the number of negative cases that is classified correctly and False Negative (FN) is the number of positive cases that is misclassified.

Table 3
Breast Cancer dataset classification results comparison.

Algorithm	GDD	[38]	[39]
Score	3.819	Not available	Not available
Accuracy	**0.958**	0.81	0.920
Sensitivity	0.924	Not available	Not available
Specificity	0.977	Not available	Not available
Positive Predictive Accuracy	0.960	0.60	Not available
Negative Predictive Accuracy	0.956	0.89	Not available

Table 4
Feature Importance for the Diabetes dataset.

Feature	Importance factor
Tolazamide Up	0.299
A1Cresult None	0.100
1ange	0.090
Insulin Down	0.065
Acetohexamide Steady	0.059

Table 5
Feature Importance for the Breast Cancer dataset.

Feature	Importance factor
Radius mean	0.287
Area worst	0.266
Area mean	0.209
Perimeter se	0.119
Texture worst	0.116

Algorithm 1: GDD Algorithm

```
Input: Dataset
Result: Classification model
initialization;
dataset ← readCSV();
foreach record ∈ dataset do
    if record.class = positive then
        positiveClass.append(record);
    else
        negativeClass.append(record);
    end
end
while positiveClass.size() < NagativeClass.size() do
    replicate positiveClass;
end
postiveClass training set ←
 randomly selected 70%of the PostiveClass recods;
postiveClass test set ←
 randomly selected 30%of the PostiveClass recods;
negativeClass training set ←
 randomly selected 70%of the negativeClass recods;
negativeClass test set ←
 randomly selected 30%of the negativeClass recods;
testset ← Concatenate(PostiveClass test set, negativeClass test set);
classification score ← 0;
```

3.5. Datasets

In this study two datasets were used. The first one is the Diabetes 130-US hospitals for years 1999–2008 dataset, which is downloaded from the UCI machine learning library [28]. This dataset was made publicly available in 2014 [37]. The dataset contains 100,000 instances and 55 attributes. Five of these attributes contain missing values. The attributes types include: boolean, categorical and numeric attributes.

The second dataset is the Breast Cancer Wisconsin (Diagnostic) dataset, which is downloaded from the UCI machine learning library [28]. This dataset was made donated in 1995-11-01. The dataset contains 569 instances and 32 attributes. These attributes do not have any missing values. All the attributes are numeric. Only ten attributes were selected because of their relatively high variance (higher than 1) compared with other attributes. This is illustrated in Table 1. These attributes are: 'area worst','area mean', 'area se', 'perimeter worst', 'perimeter mean', 'texture worst', 'radius worst', 'texture mean', 'radius mean' and 'perimeter se'.

4. Results and discussion

This section discusses the classification results obtained by running the GDD algorithm on the datasets illustrated in Section 3.5.

To visualize the diabetes dataset, PCA dimension reduction is applied to reduce the data dimensionality into 2D, 3D as shown in Fig. 1 and Fig. 2 respectively. In both figures, the red points represent the sample (positive class), while the blue data points represent the healthy controls (negative samples). The results show that the data points are interleaved and relatively hard to be separated.

To visualize the breast cancer dataset, the same process was applied. PCA dimension reduction is applied to reduce the data dimentionality into 2D, 3D as shown in Fig. 3 and Fig. 4 respectively. In both figures, the red points represent the sample (positive class), while the blue data points represent the healthy controls (negative samples). The results again show that the data points are interleaved and relatively hard to be separated.

The results of the GDD algorithm compared with the state of the art algorithms discussed in Section 2 are represented in Table 2

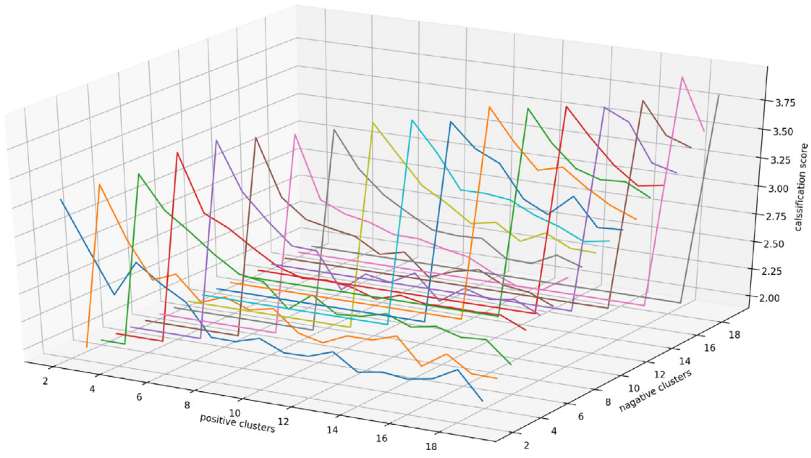

Fig. 5. Score Progress in Diabetes dataset with changing the number of positive and negative classes clusters.

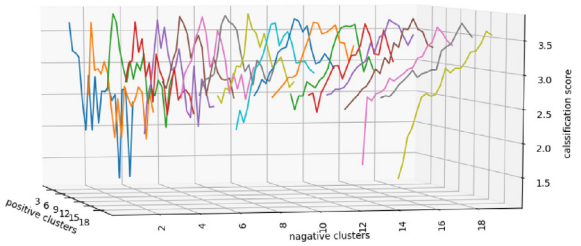

Fig. 6. Score Progress in Breast Cancer dataset with changing the number of positive and negative classes clusters.

and Table 3. While testing the datasets, the best results are obtained with k = 18, 9 for both the positive and negative classes for the diabetic and breast cancer datasets respectively. The features importance are measured for each feature during the feature selection phase using the extra trees classifier algorithm. The top 5 important features in the diabetic dataset are represented in Table 4 along with their importance factors. Table 4 and 5 show that the most dominant indicator is the 'tolazamide Up' in the diabetes dataset and 'radius mean' in the breast cancer dataset.

Figs. 5 and 6 relate the number of positive clusters, negative clusters and the total classification score achieved in the diabetic and the breast cancer datasets respectively. It is obvious that the best results are obtained with equal number of clusters for both the positive and negative classes. The score value keeps getting higher while increasing the number of clusters up to 18 clusters, and 9 clusters for the diabetic and the breast cancer datasets respectively. Then the score begins to decrease.

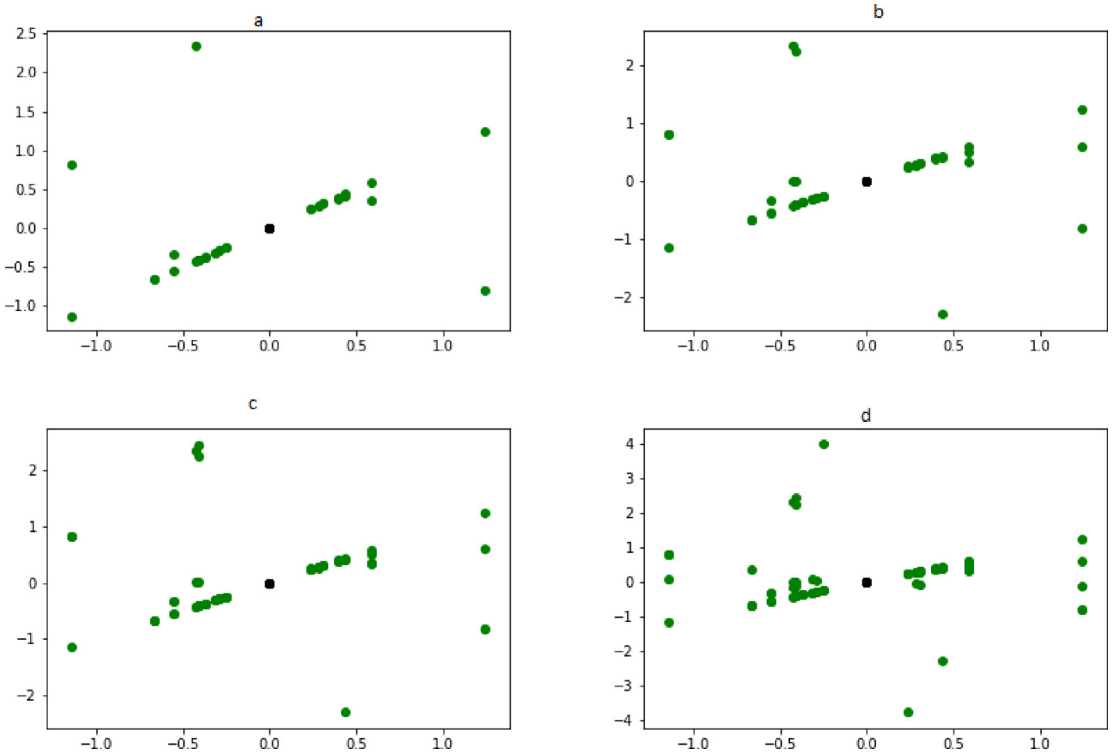

Fig. 7. Centroids for the best clusters case (k = 18) of the diabetes dataset projected in 2D.

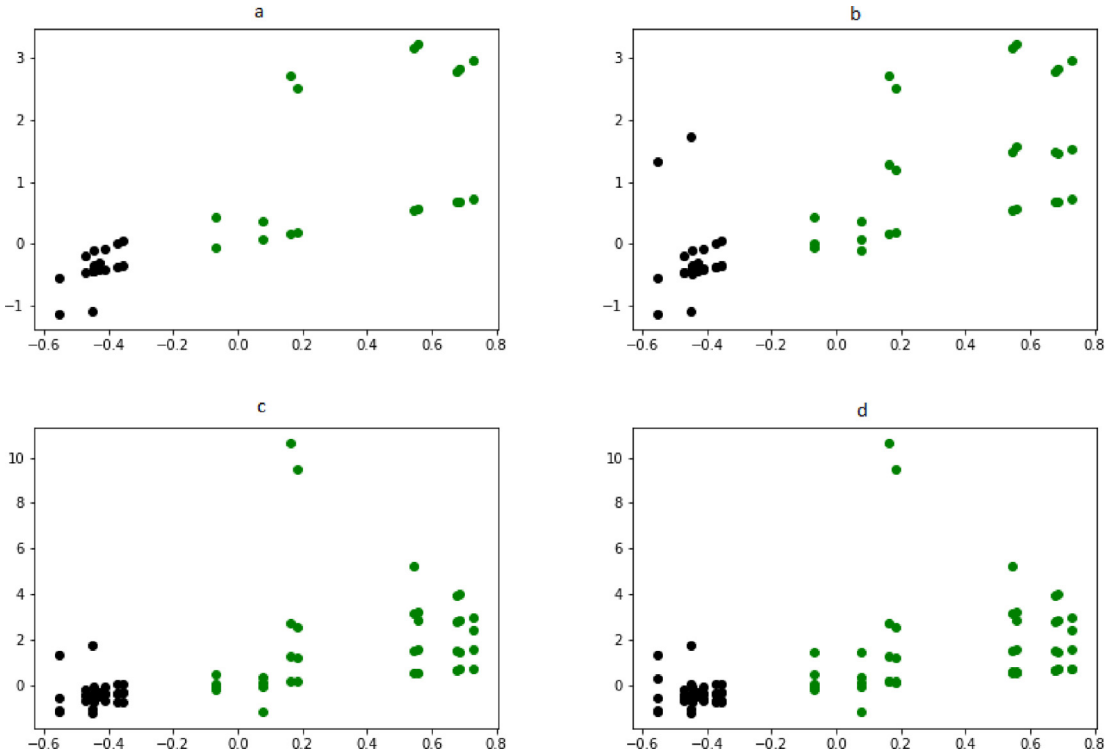

Fig. 8. Centroids for the best clusters case (k = 9) of the breast cancer dataset projected in 2D.

Figs. 7 and 8 represent the best clusters case centroids obtained by applying the GDD algorithm on the diabetes dataset (k = 18) and the breast cancer dataset (k = 9), respectively. In both figures the centroids are plotted after projecting them in 2D. For more illustration of the results, Fig. 7 shows the first PCA component plotted against the $2^{nd}, 3^{rd}, 4^{th}$ and 5^{th} PCA components in the sub-figures 7_a, 7_b, 7_c and 7_d respectively for the diabetes dataset. Also Fig. 8 shows the 'area worst' plotted against 'area mean', 'area se', 'perimeter worst' and 'perimeter mean' in the sub-figures 8_a, 8_b, 8_c and 8_d respectively for the breast cancer dataset. In both figures the green points represent the positive class centeroids, while the black points represent the negative class centeroids. The plotted results show that the negative class centroids lie in close proximity to each other, while the positive class centroids are spread across the figure.

5. Conclusion

This work proposes a new machine learning algorithm (GDD) for medical diagnosis. GDD algorithm starts by selecting the most relevant features for the problem using the extra trees classification algorithm for determining the features importance and then the PCA for data dimensionality reduction. The proposed approach outperformed the state of the art approaches as it built different models for each class. The analysis of each feature variance and importance as well as the conversion of categorical features into numeric showed which features are more relevant to be considered by the GDD classifier. The use of PCA technique to further reduce the problem dimensionality -in case of high dimensionality- was effective in making the algorithm faster and more reliable. GDD algorithm achieves an accuracy of 99.9% when experimented with the Diabetes 130-US Hospitals for years 1999–2008 dataset and 95.8% when experimented with the Breast Cancer Wisconsin (Diagnostic) Dataset.

6. Future work

It is recommended to find an optimized method for determining the clusters of each class that eliminates combinations where clusters from different classes intersect. This method might reduce the computational cost and result in a more accurate solution.

References

[1] Williams G, Pickup JC. Handbook of diabetes. Wiley-Blackwell; 2004.
[2] Association AD et al. 2. Classification and diagnosis of diabetes. Diabetes Care 2016;39(1):S13–22.
[3] DeFronzo RA, Ferrannini E, Groop L, Henry RR, Herman WH, Holst JJ, Hu FB, Kahn CR, Raz I, Shulman GI, et al. Type 2 diabetes mellitus. Nat. Rev. Disease Primers 2015;1:15019.
[4] Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. Lancet 2017;389 (10085):2239–51.
[5] Schlienger J-L. Type 2 diabetes complications, Presse medicale (Paris, France: 1983) 42(5) (2013) 839–848.
[6] Balkau B. Screening for diabetes; 2008.
[7] Shendure J, Ji H. Next-generation dna sequencing. Nat Biotechnol 2008;26 (10):1135.
[8] Ginsburg GS, Willard HF. Essentials of genomic and personalized medicine. Academic Press; 2009.
[9] Wei S, Zhao X, Miao C. A comprehensive exploration to the machine learning techniques for diabetes identification. In: Internet of Things (WF-IoT), 2018 IEEE 4th World Forum on. IEEE; 2018. p. 291–5.
[10] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Proc Comput Sci 2018;132:1578–85.
[11] A.C. Society, Cancer facts & figures, The Society; 2008.
[12] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer J Clinic 2018;68(6):394–424.
[13] Matsen CB, Lyons S, Goodman MS, Biesecker BB, Kaphingst KA. Decision role preferences for return of results from genome sequencing amongst young breast cancer patients. Patient Educ Counseling 2019;102(1):155–61.
[14] Aggarwal CC, Reddy CK. Data clustering: algorithms and applications. CRC Press; 2013.
[15] Schalkoff RJ. Artificial neural networks, vol. 1. New York: McGraw-Hill; 1997.
[16] Adankon MM, Cheriet M. Support vector machine. Encyclopedia of biometrics. Springer; 2009. pp. 1303–1308.

[17] Singh PP, Prasad S, Das B, Poddar U, Choudhury DR. Classification of diabetic patient data using machine learning techniques. In: Ambient Communications and Computer Systems. Springer; 2018. p. 427–36.
[18] Karthikeyan T, Ravikumar N. A survey on association rule mining. Int J Adv Res Comput Commun Eng 2014;3(1). 2278–1021.
[19] Rao S, Gupta P. Implementing improved algorithm over apriori data mining association rule Algorithm 1.
[20] Singh A, Halgamuge MN, Lakshmiganthan R. Impact of different data types on classifier performance of random forest, naïve bayes, and k-nearest neighbors algorithms. Int J Adv Comput Sci Appl 2017;8(12):1–10.
[21] Liaw A, Wiener M, et al. Classification and regression by randomforest. R news 2002;2(3):18–22.
[22] Peterson LE. K-nearest neighbor. Scholarpedia 2009;4(2):1883.
[23] Cichosz P. Naïve bayes classifier, Data Mining Algorithms: Explained Using R (2015) 118–133.
[24] Thangaraju P, Deepa B, Karthikeyan T. Comparison of data mining techniques for forecasting diabetes mellitus. Int J Adv Res Comput Commun Eng 2014;3(8).
[25] Kriegel H-P, Kröger P, Sander J, Zimek A. Density-based clustering. Wiley Interdisc Rev: Data Min Knowl Disc 2011;1(3):231–40.
[26] Everitt BS, Landau S, Leese M, Stahl D. Hierarchical clustering, Cluster Analysis. 5th ed, 2011. p. 71–110.
[27] Srivastava S. Weka: a tool for data preprocessing, classification, ensemble, clustering and association rule mining. Int J Comput Appl 2014;88(10).
[28] Lichman M, et al. Uci machine learning repository; 2013.
[29] Ceylan Z, Pekel E. Comparison of multi-label classification methods for prediagnosis of cervical cancer. Int J Intell Syst Appl Eng 2017;5(4):232–6.
[30] Bhargava N, Sharma G, Bhargava R, Mathuria M. Decision tree analysis on j48 algorithm for data mining. Proc Int J Adv Res Comput Sci Software Eng 2013;3(6).
[31] Platt J. Sequential minimal optimization: a fast algorithm for training support vector machines.
[32] Wu W, Zhou H. Data-driven diagnosis of cervical cancer with support vector machine-based approaches. IEEE Access 2017;5:25189–95.
[33] Bro R, Smilde AK. Principal component analysis. Anal Methods 2014;6(9):2812–31.
[34] Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining ACM. p. 785–94.
[35] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63(1):3–42.
[36] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12(Oct):2825–30.
[37] Strack B, DeShazo JP, Gennings C, Olmo JL, Ventura S, Cios KJ, Clore JN. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. BioMed Res Int 2014.
[38] Ferreira P, Dutra I, Salvini R, Burnside E. Interpretable models to predict breast cancer. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE. p. 1507–11.
[39] Kharya S, Soni S. Weighted naive bayes classifier: a predictive model for breast cancer detection. Int J Comput Appl 2016;133(9):32–7.