International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Jun-24)

http://dx.doi.org/10.12785/ijcds/1501121

Hybrid K-means and Principal Component Analysis (PCA) for Diabetes Prediction:

Ahmed Abed Mohammed[1,2], Putra Sumari[3] and kassem Attabi[1,2]

[1]*Department of Technical Engineering, Islamic University, Najaf, Iraq*
[2]*Department of Technical Engineering, Islamic University, Dewania, Iraq*
[3]*School of Computer Science, Universiti Sains Malaysia, 11800 Penang, Malaysia*

Received 14 Sep. 2023, Revised 24 Mar. 2024, Accepted 7 Apr. 2024, Published 1 Jun. 2024

Abstract: Diabetes is the ”silent killer,” stealing the lives of millions of people worldwide. There are many reasons for diabetes, such as increasing glucose, Cholesterol, systolic BP, and Age. These are considered to be the four primary causes of diabetes. The challenge in diabetes is predicting the human illness early to start treatment immediately after discovering diabetes; this can be the most challenging thing in diabetes discovery because tens of features may cause diabetes. This study proposes a model consisting of data mining and Machine Learning (ML) algorithms to predict if humans can have diabetes or not in the future. The prediction is made up of compensating two datasets; one dataset is used to reconfirm the other dataset in order to make a more accurate prediction. This can be performed using the k-means-PCA hybrid model and the highest weight selection of features that widely cause diabetes. The selected features help the ML algorithm predict the model's accuracy, which indicates the prediction model's accuracy. Simulation results show that the number of predict-diabetic patients increased from 53 from the original datasets to 142 after applying the proposed model. Simulation outcomes also prove that the Random Forest ML model gives the highest accuracy of other ML models, reaching 95.2%.

Keywords: Machine Learning, Data Mining, Feature Extraction, Hybrid Model, Diabetes

1. INTRODUCTION

Diabetes is a common disease that impacts millions of individuals. There are 1.6 million people who die around the world because of diabetes, as the World Health Organization (WHO) reports [1] 2017; as an example of diabetes statistics, there were 451 million diabetes patients around the world, and expected to reach around 693 million over the next decade [2], diabetes is simply a flaw in turning food into energy. In ordinary food consumption, the food is broken down into glucose and released into the bloodstream. Insulin is released into the body to absorb the untreated glucose and transfer it to energy. That means that the leading cause of diabetes is the high amount of sugar in the body that is not treated or absorbed [3]; early diagnosis is essential in controlling the patient's condition and not worsening his health. Data mining is applied in various science fields, including health care and medicine. It is the recognition of patterns, prediction of diseases, and their classification using data mining techniques. Building models and prediction models for the early detection of illnesses have used various strategies [4]; after the technological boom that the world witnessed at the beginning of the twentieth century, technology entered several fields, including health. Artificial intelligence (AI), especially machine learning, is now at the heart of medical and healthcare applications of information and communication technology (ICT) [5], [6]; medical diagnosis with the assistance of ML is a dynamic and ever-expanding field of study, while this approach can be from a medical perspective to help patients avoid significant cost burdens. the results proved that machine learning helped diagnose many diseases with higher accuracy, more speed, and lower cost [7], [8], Nowadays, extensive data analysis is rising fast, especially in the health care sector. As the data increases, the quality of extracting meaningful results increases. Problems dealing with big data include hidden patterns and many features that increase the difficulty of recognizing the data. Since big data is essential for the healthcare industry, Machine Learning (ML) has been applied due to its efficient algorithms and ability to find insights from data and predict diseases [9]. ML is a part of artificial intelligence (AI) that enables computers to automatically learn from data and previous

E-mail address: a.alsherbe@gmail.com, putras@usm.my, msckassem@gmail.com

experiences by identifying patterns to generate predictions with minimal human intervention. This concept is essential in the case of diabetes. It can be used for diabetes prediction, so it helps in starting treatment early. The concept of ML is quickly becoming attractive to healthcare. The main goal of this research is to use data mining and machine learning approaches to create a reliable model for early diabetes onset prediction. Our study attempts explicitly to Investigate different data mining methods that may be used to uncover pertinent characteristics and trends from datasets related to healthcare and Explore ML techniques, including supervised learning classifiers, which may be used to forecast the start of diabetes based on traits that have been detected. Novel methodology combines data mining and machine learning methods to provide a unique methodology for predicting diabetes at an early stage. This method utilizes extensive healthcare data to enhance its effectiveness. Predictions and analyses of medical data sets are important because they help design appropriate treatments and disease prevention and prediction precautions. ML algorithms are applied to aid dataset analysis, decision-making, and predictions [10]; this work proposes a model of diabetes based on data mining and ML. The proposed work increases the accuracy of diabetes classification by allowing data mining for a generated dataset. one contains three classes representing diabetes, no diabetes, and unknown. The second one contains two classes (Diabetes and no diabetes) to train the model to enhance a classification's precision. The concept is to classify whether or not no diabetes patients will have diabetes based on knowing the best features that affect humans.

2. LITERTURE REVIEW

Jammuna.et al. suppose model k-means clustering and a machine learning model (decision tree, support vector machine, naive Bayes, random forest, k-nears- neighbour) to predict diabetes at an early period. The Pima Indian diabetes dataset was created using this dataset. Outcomes show that SVM outperforms with a high accuracy of 82% compared to other algorithms [11]; in 2020, Nora and Mohammad suggested that model k-means and SVM predict diabetes using PIMA Indian Dataset; they use 10=fold cross-validation for classification. The accuracy reached is 82% [12], Kadhm et al. suggest a model that predicts diabetes that uses Naive Bayes for prediction and K-means for clustering. The suggested system employed 768 instances with 8 characteristics for each PID dataset. The unnecessary data is removed during the pre-processing of the usable data. The characteristics analysis and classification components were the main emphases of the suggested system. The ideas in these sections provide the best results. Experiment findings showed an accuracy of 79.56% [13], Singh Rajesh proposed a model consisting of a k-means cluster and logistic regression to predict heart disease in 2019. They used the heart dataset, and both performance and accuracy were increased. Therefore, the Normalisation pre-processing step strengthens the Euclidean distance by calculating more nearby centers. This results in decreased repetitions, decreasing the computing time compared to k-means clustering. Finally, using the data recovered by K-Means Clustering, the classifiers are created using logistic regression. With these strategies, they achieved 90% accuracy [14]; in 2021, Ali et al. suggested Utilising both PAM and K-Means, a hybrid technique combining K-Means and Partitioning Around Medoids (PAM) known as K-MP, which may effectively build a model for forecasting patient status. The model recommended for the dataset was gathered using a questionnaire from 400 patients at several clinics in Iraq. When we compared the accuracy of the offered methods, we discovered that K-MP is more effective at determining a patient's state than K-Means and PAM, with an accuracy of 87.5% [15]; in 2019 Saru Subashree proposed a Decision Tree algorithm to detect a diabetic person with an accuracy of 94.44%. They employed three classification models: Decision Tree, K-nearest neighbors, and Naive Bayes. They focused on increasing the accuracy by using resampling techniques. They used the Pima Indian diabetes dataset, then pre-processed it, and did the feature selection phase by the PCA [16]; this article presents the computationally effective strategy Deka and Sarma devised for localizing the various characteristics in a fundus retinal picture. The study suggests several methods for extracting characteristics related to diabetic retinopathy using Singular value decomposition and Principal component analysis, then training an ANN using the composite form. To identify diabetic retinopathy and stop vision loss, hemorrhages and exudates must be found. According to experimental findings, ANN-SVD+PNN is an accurate, trustworthy, and computationally simple method of detecting diabetes; the accuracy of NN with PCA reaches 94% [17], Osareh Shadgar identified benign and malignant breast tumors using principle component analysis (PCA) feature extraction, sequential forward selection-based feature selection, and k-nearest-neighbor with signal-to-noise ratio feature ranking. Using k-nearest-neighbour classifier models on commonly used datasets for breast cancer, the most outstanding overall accuracy for diagnosis is 88%. [18]; the goal of the article suggested by Mushtaq et al. is to distinguish between tumorous (malignant) and non-tumorous (benign) cells in the dataset. The UCI machine learning repository's Wisconsin Breast Cancer Data (WBCD) was utilized—a supervised learning classifier using dimensionality rebate methods based on PCA. PCAs based on decision trees had an accuracy of 94.3% [19], Zou et al. suggested research; we used random forests in this work to forecast diabetes mellitus. The hospital physical examination statistics in Luzhou, China, comprise the dataset. There are 14 qualities in it. Five-fold cross-validation was utilized in this work to evaluate the models. We selected a few ways that perform better than others to carry out independent test trials to verify the methods' general applicability. As a training set, we chose data from 68994 healthy persons and 68994 diabetes patients at random. We randomly extracted data five times as a result of the data imbalance. The outcome is the mean of these five trials. To minimize the dimensionality in this investigation, we employed (PCA) and minimal

redundancy maximum relevance (mRMR). The findings showed that prediction accuracy using random forest and PCA could reach 73.95% [20]. This research paper introduces a novel methodology for an intrusion detection system (IDS) that employs a clustering technique based on Principal Component Analysis (PCA) with an enhanced K-Means algorithm. This technique aims to enhance the efficiency and accuracy of hostile activity detection in cloud computing, social networks, and mobile cloud computing, therefore addressing the growing significance of Intrusion Detection Systems (IDS). It attains improved precision and effectiveness in comparison to current techniques.The empirical findings confirm the effectiveness of the suggested methodology and emphasize its potential for practical use in safeguarding network security [21]. This study proposed a hybrid framework to address the challenges of type-2 diabetes prediction, (SMOTE) is employed to balance the imbalanced dataset, PCA to dimensionality reduction, and classification techniques to address the challenges of type-2 diabetes prediction, including logistic regression (LR), naïve Bayes, support vector machine (SVM), and k-nearest neighbors (KNN). Among these techniques, logistic regression achieved the highest accuracy of 98.96%, demonstrating the effectiveness of the proposed hybrid model [22]. The paper introduces a unique strategy that combines optimization algorithms for feature selection, showcasing innovation in the area. Using the K-nearest neighbor algorithm for classification demonstrates encouraging outcomes, with an accuracy rate of 91.65%. The significance of preventive healthcare measures is emphasized in this strategy to reduce the risks linked with diabetes and infectious illnesses such as COVID-19. In general, this literature review emphasizes the notable advancements in using artificial intelligence (AI) for illness diagnosis and treatment, specifically focusing on improving early detection and patient outcomes [23].

3. Methodology

A. General Methodology Description

A general description of the methodology is discussed here to provide an overview of the work. The main contribution of this work is to detect whether humans will have diabetes or not based on data mining and ML algorithms. To achieve this, we suggested a methodology that consists of three phases, as shown in Figure 1

In phase 1, two datasets are used. The two datasets are of different sizes and different numbers of classes. The two datasets are merged to increase the size of the unsupervised dataset. The combined dataset has some data preparation, such as normalization and Filtering. In phase 2, The best features that affect human health and can cause diabetes are selected using data mining techniques. This data mining depends on the k-mean clustering and PCA to create a hybrid model, "K-means-PCA," to detect the highest weight feature for each class. In phase 3, the resulting supervised dataset enters the ML algorithm detection using 10-fold cross-validation and a different number of ML algorithms. The accuracy is obtained for each ML algorithm. Finally,

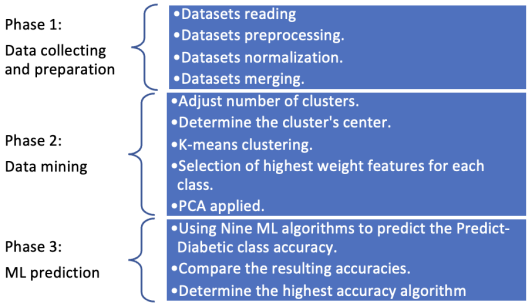

Figure 1. Methodology Description

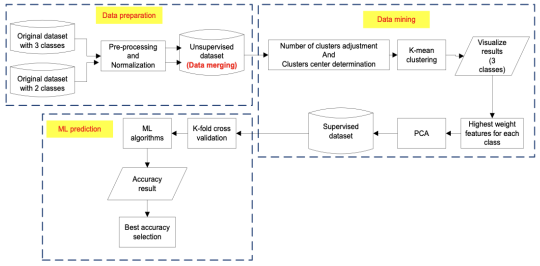

Figure 2. The proposed model block diagram

the resulting model is evaluated by combining data mining techniques with one of the machine learning algorithms. The best algorithm is selected according to the highest accuracy.

B. The Proposed Model: Block Diagram

There are three phases of this work, as described in Figure 1 Each one of these stages has several operations that are performed to get the final accurate result. Figure 2 shows the proposed model's block diagram and the workflow between each block. The following sections describe each block that appeared in the block diagram.

4. Phase 1: Data Collection and Preparation

The first phase of this work starts by reading two different datasets, each dataset of size M×N. M is the number of samples in the datasets, and N is the number of diabetes features. The two datasets used in this work are imbalanced due to the low number of samples in the class Predict-Diabetic. Thus, data pre-processing has been applied to the datasets, such as renaming and reindexing columns, removing unnecessary columns, data transformation and normalization, and merging the two datasets to obtain one unsupervised dataset by removing all labels available.

A. Dataset Description

1- First Dataset with 3 Classes : This first dataset was acquired from the laboratory of Medical City Hospital and the Specializes Centre for Endocrinology and Diabetes-Al-Kindy Teaching Hospital in Iraq. It is chosen because it contains the three classes (diabetic, non-diabetic, and

Figure 3. First dataset with 3 classes

Figure 4. Second dataset with 2 classes

predict-diabetic) from which the ML algorithm learns to predict diabetes. This dataset includes 1000 patients and 14 columns of their medical information (features), which are ID, the Number of Patients, Gender, Age, Urea, Creatinine ratio (Cr), HBA1C, Cholesterol (Chol), Triglycerides (TG), HDL Cholesterol, LDL, VLDL, Body Mass Index (BMI). The last column represents the dataset's classes. After reading the dataset, I found that it contains 844 patients represented as 1 class. It refers to diabetes, with 103 patients represented in the 0 class. It refers to non-diabetic, and 53 patients are represented as two. It refers to pre-diabetic patients, as shown in Figure 3

2- Second Dataset with 2 Classes: The second dataset was acquired from Vanderbilt University Medical Center in Nashville, the capital of the U.S. state of Tennessee, and contains 390 patients and 18 columns of their medical information (features), which are Patient number, Cholesterol, Glucose, HDL Chol, Chol/HDL ratio, Age, Gender, Height, Weight, BMI, Systolic BP, Diastolic BP, waist, hip, Waist/hip ratio, Unnamed: 16, Unnamed: 17 and Diabetes column represents the dataset's classes, There are 330 patients classed as 0, which refers to non-diabetic, and 60 patients classed as 1, which refers to diabetic, as shown in Figure 4

B. dataset Preparation

The two datasets used have imbalanced data, meaning that the predicted diabetes class has a low number of samples; to fix that, the two datasets must be merged and unlabelled to apply data mining techniques to the resulting merged dataset. The data pre-processing and normalization are crucial to getting good results and extracting insights from the dataset because the two used datasets have different column names and are not in the same order. This pre-processing has:

1- Removing Unnecessary Columns.

Figure 5. The merged unsupervised dataset

2- Renaming Columns .
3- Reindexing Columns.
4- Data Transformation.
5- Data Normalisation.

C. Datasets Merging (unsupervised dataset generation)

In this step, the two pre-processed datasets must be merged to apply the data mining technique. That is because it increases the number of classes of predicted diabetes. The resulting merged dataset, as shown in Figure 5, has 403 patients represented as 0, 904 patients represented as 1, and 53 patients represented as 2. After merging, classes of the merged dataset are removed to obtain an unsupervised dataset. It is essential to allow the data mining technique to classify the unsupervised dataset into new classes based on k-means and PCA and according to the highest weight feature extraction.

5. Phase 2: Clustering and Features Selection

This is the second phase of this work, the clustering phase, where its input is the unsupervised merged datasets, and its output is the supervised datasets. This supervised dataset contains only the highest weight features for each class obtained from clustering. The two objectives of this

stage are to balance the merged dataset and determine the new classes of this merged dataset by increasing the number of samples in the class Predict-Diabetic. Data mining techniques have been applied to satisfy these objectives, such as using the K-means algorithm to divide the unsupervised dataset into three classes. This clustering extracts the dataset's highest weight feature, obtaining a supervised dataset with three balanced classes. At the end of this stage, the PCA was applied to visualize the distribution of classes and features and their correlation.

A. K-means

K-means is a clustering technique belonging to the partitioning-based methods of grouping based on the iterative relocation of data across clusters. It is divided into non-overlapping clusters based on extracted characteristics, either the cases or variables of datasets. The algorithm is applied to the cases or variables of datasets depending on which dimension a dimensionality reduction is required with a primary goal of producing clusters of cases or variables with a high degree of similarity within the cluster and a low degree of similarity between clusters. K-means can also be described as a centroid model since one vector representing the mean is used to describe each cluster. It is beneficial in exploratory data analysis and data mining. Considering the exponential growth in computer power and the occurrence of large data sets, the ease of implementation, computational efficiency, and low memory consumption allowed the k-means algorithm to stay on top of the list on matters of popularity even when compared with other techniques. A secondary goal of K-means is the reduction of the complexity of data, which can be used as an initialization step to prepare for more computationally expensive algorithms. The formula of k-means calculates the distance between each cluster (class).

$$distance(C_j, p) = \sqrt{\sum_{i=1}^{d} (C_{ji} - p_i)^2} \quad (1)$$

Where C is the cluster's center, and p represents the values of the features.

Feature selection plays a crucial role in improving the clustering distribution in this phase due to its impact on the performance of the clustering model.To choose the right features, k means feature-importance, which gives the critical features of each class. K-means was selected for its simplicity, effectiveness, and widespread use in various applications, including medical data analysis. It is a versatile algorithm that can handle large datasets efficiently. It identifies clusters of related data points based on their features. It is beneficial in a medical context where identifying subgroups within the data can inform diagnosis and treatment strategies. K-means depends on the distance in splitting the dataset into clusters, which is useful when we want to predict the healthy people "not have diabetes" in the second dataset based on the first dataset. Also, it works well when we use PCA because it focuses on some features; also, merging with PCA will take more time, but the k-means is comparatively simple to implement and Easily adjusts to new instances [24]

B. Principal Components Analysis (PCA)

Principle Components Analysis (PCA) is one technique that allows the representation of high-dimensional data into a more tractable lower-dimension form using the dependencies between the variables without losing too much information. The central idea of principal component analysis is dimensionality reduction. It aims to reduce the dimensionality of a data set consisting of a large number of interrelated variables while still retaining as much of the variation present in the data set as possible. That can be achieved by transforming to a new set of variables, also called principal components (PCs), which are uncorrelated and ordered so that the first few retain most of the variation in all of the original variables. In other words, PCA can be defined as identifying patterns in data and expressing said data in a way that highlights the patterns' similarities and differences since patterns in data can be complex to find in data of high dimensions where the luxury of graphical representation is unavailable. PCA formula is:

$$cor(\ ,\ j) = \sum_{i=1}^{n} p_i \left(\frac{x_{ij}}{s_j}\right) \left(\frac{p_i}{\ }\right) \quad (2)$$

There are many features; some are unimportant, and some are not defined. To get an efficient model, we must choose the valuable features only; it saves time and space for storage, and it is possible to use the PCA effectively and efficiently, especially when combined with K-means; it also helps to enhance the work of K-means [25].

6. Phase 3: Diabetes Prediction Model

This is the third and last phase of this work. It is simply using an ML algorithm to determine the prediction accuracy of the Predict-Diabetic class. This means that this stage determines whether the human will have diabetes or not in the future based on the highest weighted features obtained from stage two. Nine ML algorithms are used: Random Forest, Gradient Boosting, XGboost, Multinomial Naive Bayes, Logistic Regression, KNN, Decision Tree, SVM(RBF), and SVM(linear).

It contains two steps to get the predict-diabetic accuracy after data mining. The two steps used here are:

1- Dataset splitting: the dataset was divided into two sets. A train set contains the data used to train the ML algorithm, and a test set contains the data used to test the ML algorithm. 70% of the dataset for the training set contains 1112 samples, and 30% for the testing set contains 278 samples.

2- ML algorithm uses: 9 ML algorithms have been trained by the three classes, which are 0 as non-diabetic, 1 as diabetic, and 2 as predict-diabetic.

Figure 6. The proposed model description

7. The Proposed Model: Description and Flowchart

Figure 6, shows the proposed model description. It starts by describing the Inputs of the model, which are the two datasets used. The pre-processing procedures are summarised in this Figure, and the normalization process is mentioned. The k-means data mining for clustering was used. The highest weight features have been determined and selected to improve the Ml algorithm accuracy results. All processes have been repeated for a distinct number of ML algorithms.

The ow chart of our paper consists of 5 main steps, as shown in Figure 7:

1. Data Preparation: Two datasets are read at the start of the procedure, and columns are renamed, reindexed, and super uous columns are eliminated; data transformation procedures are used; after merging the datasets, labels are eliminated.
2. Clustering or Feature Weight Calculation: - Clustering is the next stage; samples are allocated to clusters. If successful, in case clustering is not successful, feature weights are computed directly.
3. Feature Weight Calculation (Alternative Path): - Feature weights are determined based on cluster assignment. If clustering is booming, "Is weight the best?" is up for debate.
If so, the ow moves on to choose the best characteristics. If not, the process returns to grouping samples into clusters.
4. Machine Learning Prediction: - The ow permits machine learning prediction after selecting the characteristics with the most signi cant scores. The forecast turns out to be accurate.
5. Comparison of Accuracy: - Multiple algorithms go through this procedure again, and the accuracy is compared at the end.

Figure 7. Flow chart of the simulation

8. Result

A. K-means-PCA

This section provides the k-means clustering results, the highest weight features selection, and the PCA results. At the end of this section, the supervised data is obtained. Figure 8: shows the clustered dataset after the k-means clustering of the unsupervised dataset. The data was partitioned into 3 clusters depending on the dataset's features. Figure 8: shows that there are 476 clustered for non-diabetics, represented as 0, 603 for people with diabetes, represented as 1, and 311 clustered for predicted diabetics, represented as 2.

Feature selection is crucial in improving the dataset performance and the clustering distribution shown in Figure 8. Figure 9: shows the highest weight of the 0 clusters using the k-means-feature-importance function. It shows that the four critical features for the 0 cluster are Cholesterol, systolic BP, Glucose, and Age. Figure 9: shows that the highest weight feature comes from Cholesterol with 630 weights. In comparison, the second highest feature is Systolic BP with 360 weights. Glucose. Age and all other used features are less than 100 in weight. Glucose

Figure 8. Comparison between the (a) old merged dataset and (b) the new clustered dataset

has 90 weights, and Age is around 50 in weight.

Cholesterol, Systolic BP, Glucose, and Age are the four highest weight features for the 1 cluster, as shown in Figure 10. 460, 210, 85, and 52 are weights of the four highest weight features for 1 cluster, respectively.

Figure 9. Highest weight features in 0 cluster

Figure 9 and Figure 10 show that the two best features used in diabetes tests to indicate whether a human has diabetes are Cholesterol and Systolic BP. They give the highest weight, meaning the highest probability of accurately determining the diabetes test. It is clear from the two gures that the Chol/HDL ratio does not aect the diabetes test result because it gives no weight in both clusters.

Figure 10. Highest weight features in 1 cluster

Figure 11 shows the highest weight features of the 2 clusters. The four highest weights are Cholesterol, Systolic BP, Glucose, and Diastolic BP, with 230, 140, 100, and 65 weight values, respectively. It is clear from the Figure that the Age feature is not among the four highest ones. It can be concluded that the age feature contributes to diabetes prediction with less ect than the other highest weight features. Another notice from Figure 11 is that the Chol/HDL ratio weights so it can contribute to diabetes prediction. In general, simulation results shown in Figure 9, Figure 10, and Figure 11.

After clustering and feature selection, PCA enhances the clustering distribution. This leads to using the K-mean PCA clustering technique. Figure 12 shows the primary K-mean clustering. Figure 12 shows three clustering labels with purple, yellow, and blue coloured circles. As mentioned, clusters are distributed logically based on two features: Cholesterol and systolic BP values. The distance separating the di erent clusters is noticeable, and the distance between each point of the dierent clusters is close, de ning the borders of each cluster. Figure 13 shows the k-means PCA clustering distribution. As shown in the Figure, there are 3 clustering labels. Black refers to non-diabetic, green refers to diabetic, and red refers to predict-diabetic. A noticeable intersection between the red and green clusters is visible in the graph due to the diabetic class and the predict-diabetic class sharing the same values of certain features.

K-means results depend on the k-value chosen. Figure 14 shows that the best k-value with the highest accuracy obtained is 3.

The resulting supervised dataset after the data mining process is shown in Figure 15 the supervised dataset that enters the ML prediction has three classes. Three 344 patients presented as 0 are non-diabetics, 904 patients presented as one are diabetic, and 142 patients presented as 2 are predict-diabetics. Compared to Figure 8 (a), the predicted number of diabetic patients increased from 53 to 142.

B. ML Prediction Results

Figure 16 and Table I show the accuracy results for all ML algorithms. The lowest accuracy comes from the SVM

Figure 11. Highest weight features in 2 cluster

Figure 13. K-Means PCA clustering

Figure 14. k-means value chosen

Figure 12. K-Means clustering

(Linear) algorithm, while the highest accuracy comes from the RF with 95.2 %.

Because the RF gives the highest accuracy result, some changes to the hyperparameters of this algorithm are performed.The Grid Search approach is applied to tune and improve the performance of the RF algorithm. The process consists of feeding the hyperparameters to the model and running a search of all possible combinations of values. The training of the RF algorithm is performed for each set of values; then, the grid search approach compares the resulting accuracy. The highest accuracy value obtained after the grid search approach is 95.2% using 500 estimators, the Gini approach for the criterion, 9 for the maximum depth,

Figure 15. Resulting supervised dataset

TABLE I. Accuracy of each ML algorithm

Algorithm	Accuracy
Random Forest	95.2%
Decision Tree	93.5%
XGBoost	93.2%
Gradient Boosting	93%
KNN	91.8%
SVM(RBF)	90.4%
Logistic Regression	76.7%
MultinomialNB	63.7%
SVM(linear)	27.3%

Figure 16. Accuracy of each ML algorithm

and auto-selection of the maximum features that should be used.

9. Conclusion and Future work

Diabetes is a disease affecting either the production or the use of insulin. It is the most widespread chronic illness, with more than 450 million patients around the globe in 2017. The problem with diabetes treatment is knowing the sickness early in the disease. This leads to thinking about predicting if the human will have diabetes or not; this is difficult to perform using human effort because tens of diabetes features can cause the illness. This is a great reason to propose a hybrid model K-means-PCA to predict diabetes, as this work presents. In this work, data mining for an unsupervised dataset of three classes uses k-means with PCA and the highest weight feature selection with nine machine learning algorithms to predict. The data mining technique converts the unsupervised dataset to a supervised one. Besides, the data mining technique increases the number of the predict-diabetic class from 53 to 142. The supervised dataset enters the ML prediction algorithms. The random Forest algorithm gives the highest accuracy, reaching 95.2%.

To ensure that the proposed approach works similarly with more evenly balanced training and testing datasets, it should be evaluated with more datasets in a future study. Additionally, the principle of the suggested model may be extended by using deep learning algorithms like CNN and LSTM to cope with the large dataset. The data mining method can be modified to assess the model's performance. It is also possible to use a proposed model that includes additional diseases like breast cancer and heart disease and to predict classes of people who will develop diabetes in the future using a combined dataset and another hybrid model.

References

[1] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," Ict Express, vol. 7, no. 4, pp. 432–439, 2021.

[2] A. Dagliati, S. Marini, L. Sacchi, G. Cogni, M. Teliti, V. Tibollo, P. De Cata, L. Chiovato, and R. Bellazzi, "Machine learning methods to predict diabetes complications," Journal of diabetes science and technology, vol. 12, no. 2, pp. 295–302, 2018.

[3] S. Sossi Alaoui, B. Aksasse, and Y. Farhaoui, "Data mining and machine learning approaches and technologies for diagnosing diabetes in women," in Big Data and Networks Technologies 3Springer, 2020, pp. 59–72.

[4] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," Scientific reports, vol. 10, no. 1, p. 11981, 2020.

[5] M. L. Giger, "Machine learning in medical imaging," Journal of the American College of Radiology, vol. 15, no. 3, pp. 512–520, 2018.

[6] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," Ieee Access, vol. 5, pp. 8869–8879, 2017.

[7] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," Cancer informatics, vol. 2, p. 117693510600200030, 2006.

[8] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," Jama, vol. 319, no. 13, pp. 1317–1318, 2018.

[9] R. Bhardwaj, A. R. Nambiar, and D. Dutta, "A study of machine learning in healthcare," in 2017 IEEE 41st annual computer software and applications conference (COMPSAC), vol. 2. IEEE, 2017, pp. 236–241.

[10] L. Chaves and G. Marques, "Data mining techniques for early