Prediction of Diabetic Obese Patients using Fuzzy KNN Classifier based on Expectation Maximization, PCA and SMOTE Algorithms

Ibrahim Eldesouky Fattoh[1]
Computer Science Dept., Faculty of Computers and Artificial Intelligence, Beni-Suef University
Beni-suef, Egypt

Soha Safwat[2]
Software Engineering and Information Technology Dept
The Egyptian Chinese University
Cairo, Egypt

***Abstract*—Diabetes is a long-term disease. Inappropriate blood sugar level control in diabetic patients can lead to serious issues like kidney and heart diseases. Obesity is widely regarded as a major risk factor for type 2 diabetes. In this research, a model proposed to predict diabetic obese patients based on Expectation Maximization, PCA, and SMOTE Algorithms in the preprocessing and feature extraction phases, and using Fuzzy KNN classifier in the prediction phase. The model applied on real dataset and the accuracy of prediction results reflects the positive effect of the preprocessing techniques. The accuracy of the proposed model is 95.97% and outperforms other model applied on the same dataset.**

Keywords*—KNN classifier; SMOTE; PCA; diabetic obese patients

I. Introduction

Using Data Mining (DM) and Machine Learning (ML) techniques in data mining research are a common way for making use of large amounts of available knowledge-based data. Machine Learning is extremely essential in the realm of medical diagnostics. Data mining is a great goal in science and medical research, which necessarily generates massive amounts of data owing to the special societal effect of the serious disease. As a result, Machine learning and data mining approaches are unquestionably of great importance for aspects of clinical administration, diagnosis, and treatment As part of this work, challenges were undertaken to examine the recent literature on ML and DM methodologies in many diseases especially in the diseases of the chronic diabetes. Diagnosis in the healthcare sector is an ideal subject for ML algorithms [1]. Many of these may be identified using pattern recognition on large amounts of data. An algorithm should be trained on a small number of tests to be useful in the field, medical diagnostics must be able to tolerate noisy and empty datasets. Many researches on machine learning in the sector of healthcare have been undertaken. Healthcare ML has emerged as a top goal for many academics. Different data mining approaches and procedures in hidden pattern recognition can be used to gain insights. The medical science primary roles are to prevent or help treat diseases. One of the chronic illnesses marked by hyperglycemia is Diabetes mellitus. It can lead to a slew of difficulties [2]. As a result of higher mortality rates in recent years, According to WHO (World Health Organization) forecasts by 2040, the world's population of diabetes is anticipated to reach 642 million [3], suggesting that one out of every ten people would suffer from diabetes in the future. There are three types of Diabetes [4], namely; Gestational Diabetes, Type-1 Diabetes Mellitus, Type-2 Diabetes Mellitus. Type-2 diabetes mellitus patients are frequently classified as having a fatty liver disease in which it could be either nonalcoholic or alcoholic fatty liver disease (NAFLD|AFLD) [5]. Type-2 diabetes mellitus has been postulated as a primary cause of NAFLD development, or nonalcoholic steatohepatitis, which likely reflects in Type-2 diabetes mellitus with rapid advancement of weight gain and resistance of insulin. Obesity and diabetes, both multifactorial, difficult illnesses, have become major public health issues across the world [6]. Many conditions, on the other hand, may be prevented. Obesity is a significant growing health concern; some refer to it as the New World Syndrome [7]. The occurrence of obesity and fatty liver in persons with diabetes of Type-2 has long been documented as they are strongly associated with each other [8]. It is often viewed as an accidental finding with small to no therapeutic value. Sedentary lifestyles or poor dietary habits result in weight gain. It may also increases the chances of facing a metabolic syndrome over time. Avoiding the significant consequences that result in massive issues in health, since early detection is the beginning point for a good life without the disease reflects the significance of using the recommended method for predicting patients suffer from diabetes and affected by obesity and NAFLD. Diabetes mellitus and its consequences, in particular, must be prevented and managed in poor and middle-income countries. The following is how this paper is arranged; Section II outlines the related work. Section III, details the suggested model as well as the dataset used. The Section IV offers the obtained results, followed by the conclusion and the future work in Section V.

II. Related Work

In [9], Kumar purposed various data mining approaches in medical sector to highlight data mining applications based on the nature of the information; In order to predict Parkinson's illness, Support Vector Machines and Artificial Neural Networks were used and resulted in 95 percent accuracy. In addition, it improved detection rate by employing an ANN to diagnose cancer of breast to 98.8 percent, and employed Artificial Neural Networks. Basma Boukenze et al. in [10] assessed the DM techniques performance in medical health

sector using multiple learning techniques. The result simulation indicated that the decision tree (DT) performed better than other learning techniques in forecasting kidney failure chronic disease. Furthermore, M. Abdullah and S. Al-Asmari in [11], clarified the same DM approaches to designate the type of anemia patients suffer from anemia. DT executed with an accuracy result of 93.75 percent. While only support vector machine algorithm was used in categorizing diabetes disease, while in [12] Kumari and Chitra used the Matlab tool version 2010a in order to identify the diabetic patients by 78 percent accuracy. Developing DT and DM classification approaches assists medical practitioners in gaining better medical judgments to detect diseases timely [13]. El-Halees and Shurrab in [14] developed a model that can discriminate between individuals with blood tumors and normal blood illness utilizing multiple association rules and ANN, results with 79.45 percent accuracy. In addition, in order to predict diabetes in many circumstances various researches have been conducted in which the authors of [15] used a regression-based approach of DM to introduce diabetes therapy predictive analysis. The Oracle Data Miner was used as a mining software to forecast diabetic treatment methods. For the experimental investigation, the support vector machine technique was applied. They conclude that pharmacological therapy for patients under the age of 18 can be postponed to minimize negative effects. The authors used four classifiers in [16] to categorize the diabetes mellitus risk. First, four categorization models were investigated: DT, Logistic Regression, ANN and Naive Bayes. Then, to improve the resilience of such models, Bagging and Boosting strategies were examined. According to the findings, the Random Forest (RF) algorithm performs the best in illness risk categorization. They suggested an early diabetes prediction model in [17], and they discovered a high correlation between diabetes, glucose level and body mass index (BMI), that was retrieved using the Apriori technique. Diabetes was predicted using RF, ANN, and K-means approaches. The ANN approach achieved the highest accuracy of 75.7 percent. For the prediction of diabetes, the authors of [18] employed KNN and the Naïve Bayes approach. Their method was implemented as a program of expert software, in which users submit input in the form of patient data and the determination of whether or not the patient is diabetic. The authors of [19] propose an attribute selection technique of firefly and cuckoo search-based for the PIMA Indian diabetes database from University of California Irvine (UCI), with the goal of greater accuracy and lesser training overhead. They also said that the proposed structure promises to be more accurate than the usual technique. The authors of [20] applied a ML model to forecast the occurrence of Type-2 Diabetes mellitus, using information from the present year (Y). From 2013 to 2018, electronic health records were collected at a private medical facility for this investigation. Key characteristics were initially picked for the prediction model using chi-squared tests, ANOVA tests and recursive variable reduction approaches. Based on these variables, they used random forest, logistic regression, XGBoost, SVM and ensemble ML methods in order to foresee the result as diabetic, non-diabetic or pre-diabetic. The model performed pretty well in anticipating the occurrence of Type-2 diabetes in the Korean population. The authors of [21] applied two machine-learning techniques for two-phase classification; SVM and ANN to predict diabetes mellitus. They used a real dataset from Al-Kasr Al-Aini Hospital in Egypt. In the first phase, they used SVM to predict patients with fatty liver disease with accuracy of 95%. Then in the second phase they used ANN for prediction of diabetic patients based on phase 1 and another 8 different attributes.

III. Proposed Solution and Dataset

As the dataset of this problem was collected manually as will be described in next section, it had many issues like missing values, and the data was unbalanced, so we applied a preprocessing phase for the dataset. The algorithms used in the proposed model are described in this section.

A. Expectation Maximization Algorithm for Estimating the Missing Values

Dempster et al. 1977 in [22], demonstrated that the Expectation Maximization (EM) algorithm can be applied when $XMIS$ (the missing data joint distribution) and $XOBS$ (the observed data) is candid. For all ($\theta \epsilon Rd$), let the density function probability of (;) be X=($XOBS$,$XMIS$). The estimation of θ get the most out of the observed data log eventuality in which the expectation maximization algorithm aims to find.

$$(\theta; X_{OBS})=\log(X_{OBS}; \theta)=\log\int f(X_{OBS}, X_{MIS}; \theta) dX_{MIS} \quad (1)$$

In general, because this number cannot be estimated explicitly, the EM method calculates the MLE by iteratively maximizing the anticipated log-likelihood of complete-data in (2)

$$l\,(X; \theta)=\log f\,(X_{OBS}\,, X_{\,MIS}; \theta \quad (2)$$

Begin with a value of $\theta^0$and let θt be the estimate of at the t^{th} iteration, then below is two steps of the next EM iteration:

E step: Calculate the expectation of log-likelihood of complete-data in relation to the missing covariate conditional distribution parameterized by (θ^t):

$$Q(\theta,\theta^{(t)}=E\,[l(X; \theta) \mid X_{OBS}; \theta^{(t)}]=\int l(X; \theta) f(X_{MIS}| X_{OBS}; \theta^{(t)}\, d\, X_{MIS} \quad (3)$$

M step: Define ((t+1)) by maximizing the Q function:

$$^{(t+1)}\in argmax_{\theta}\,(\theta, ^{(t)}) \quad (4)$$

B. Feature Reduction using Principal Component Analysis Algorithm

Principle Component Analysis (PCA) is an extracting features statistical approach that employs to turn a set of possibly associated annotations to a set of variables uncorrelated transformed linearly known as principle components. PCA may be used to reduce the feature dimensions [23]. Because the eigenvectors number exceeds the columns number, the dimension of the projected output data is smaller than the dimension of the input data. The method of PCA utilized in the feature reduction step is as follows.

Algorithm 1 " PCA"

Assume: samples of N $(x_1, \dots . x_N)$ each $x_n \in R^D$
Aim: D dimensions data project to K dimensions (K<D)
Then captures the projected data maximum possible variance

Consider
$(u_1, \dots, u_D)$ *be the principle components, assumed to be orthognal such that:* $u_i^T u_j = 0$ *if* $i \neq j$ *and should be orthonormal such that* $u_i^T u_i = 1$
Each peincipal component is a vector of size $D \times 1$
We will select the frirst K principal components
1: *Compute the mean of the data*

$$\bar{X} = \frac{1}{N}\sum_{n=1}^{N} X_n \quad (5)$$

2: *Compute the Covariance matrix*

$$s = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T \quad (6)$$

3: *Find the eign values* 4
4: *Take the top k eign vectors according* (*corresponding to the top k eign values*)
5: *Call these vectors as* $u_1, u_2, \dots u_k$
(*such that* $\lambda_1 \geq \lambda_2 \dots \dots \geq \lambda_{k-1} \geq \lambda_k$
6: $Z_n = U^T \times X_n \quad (7)$

C. Handling the Un-Balanced Data using SMOTE Algorithm

Chawla presented the Synthetic Minority Oversampling Technique (SMOTE) in 2002 [24]. In contrast to random oversampling, in the SMOTE method the minority class is oversampled by producing samples of synthetic rather than oversampling with replacement. The SMOTE method generates fake instances based on similarities between existing minority cases in feature space rather than data space [24, 25]. These synthetic instances are constructed by connecting a portion or all of the minority class's K-Nearest Neighbor (KNN). Neighbors from the KNN are picked at random depending on the quantity of oversampling necessary. Algorithm 2 represents the used SMOTE algorithm for handling the un-balanced dataset.

D. Fuzzy KNN Classifier

Keller et al introduced the fuzzy KNN classifier[26], which assign to each sample a class memberships, as a function from of its KNN training samples of each sample's distance. Because it is easy and provides information on the certainty of the classification result, the fuzzy KNN classifier is a popular choice for applications. According to Keller et al, the major benefit of utilizing the FKNN model may not be the reduction in error rate. More crucially, the model provides a level of assurance that may be combined with the "refuse-to-decide" option. Objects with overlapping classes can thus be discovered and treated independently as in Algorithm 3.

Algorithm 2 " SMOTE"

The input:
X: original set of training sample
N: percentage of oversampling
K: nearest neighbors value
The output: the oversampled training set
n ← # observations
m ← # attributes
nmin ← # min observations
if N < 100 then
  Stop: warning "N should be greater than 100"
end if
N ← int(N/100)
$S_{(n*N)\times m}$ ← empty array for synesthetic samples
$for\ i \leftarrow 1 \rightarrow\ n_{min}$
Do
  Discover the KNN for each *i* then save the indexes in the *nn* newindex ← 1
    while N $\neq$ 0 do
      K_c ← number between (1& K) randomly

        $for\ j \leftarrow 1 \rightarrow\ m$
        do
          dif f ← X[nn[K_c]][j] − X[i][j] (8)
          gap ← uniform(0, 1)
          synthetic[newindex][j] ← X[i][j] + gap × dif f (9)
        end for
        newindex+ = 1
        N− = 1
    end while
end for
Return (X & synthetic)

Algorithm 3 " Fuzzy K-Nearest Neighbor

From sample $x, get\ the\ KNN.$

Soft labels, input x, *a membership vector*

$(\mu(x) = [\mu_{c1}(x), \dots, \mu_{ci}(x), \dots . \mu_{cl}(x)])$

$$\mu_{ih}(x) = \mu_i(x_i) = \begin{Bmatrix} 0.51 + \left(\frac{n_i}{k}\right) * 0.49, if\ c(x_j) = i \\ \left(\frac{n_i}{k}\right) * 0.49, if\ c(x_j) \neq i \end{Bmatrix} \quad (10)$$

- *Membership function calculation*

$$\mu_i(x) = \frac{\sum_{j=1}^{k} \mu_{ij}(1/(\|x - x_j\|^{2/(m-1)})}{\sum_{j=1}^{k}(1/(\|x - x_j\|^{2/(m-1)})} \quad (11)$$

- *The target class max* $\mu_i(x)$

E. Dataset Description

The dataset used in this study was obtained from Cairo University, Faculty of Medicine, Al-Kasr Al-Aini Hospital [21]. The dataset contains 30 variables; Gender, Age, Alcohol consumption, Smoking, Schistosomiasis, steroids, History of hypertension, Oral contraceptive pill, Waist circumference, Body Mass Index, Hemoglobin test (HGB), Liver disease, Primed lymphocyte test, Basic Insulation Level, Aspartate Aminotransferase (AST), Alanine Aminotransferase (ALT), White blood cells (WBCs), Albumin level in blood (ALB), Protein C test, Alkaline phosphatase (ALP), Gamma-Glutamyl Transferase (GGT), Total cholesterol, Triglycerides test (TGs), High-density lipoprotein (HDL), Low-density lipoprotein (LDL), International Normalized Ratio (INR), Spleen size, Fasting blood sugar, History of diabetes, and Hemoglobin A1c (HBA1C). This was preprocessed as will be shown in the proposed model section through different phases. The algorithms used in the data-preprocessing phase are expectation maximization algorithm, which estimate missing values. PCA algorithm is used in feature reduction phase, while SMOTE algorithm used to generate new sample in the minority class to overcome the unbalanced data issue that affects the measures.

F. Proposed Model

Fig. 1 shows the basic steps used in the proposed model for the prediction of diabetic obese patients. At the first, we read the dataset and apply a data preprocessing phase on it. The first step in the data preprocessing is estimating the missing values by the EM algorithm. The next step is applying the PCA algorithm to reduce the features in the dataset. The basic steps of the PCA are calculating the covariance matrix, then calculating the Eigen values, then sorting the attributes in descending order, then normalizing the values, and calculating the weight value for each attribute. The third step is solving the unbalanced data using SMOTE algorithm described in last section above. The SMOTE algorithm used for generating new sample in the minority class. The last step in the proposed model is classifying the new samples using fuzzy KNN classifier.

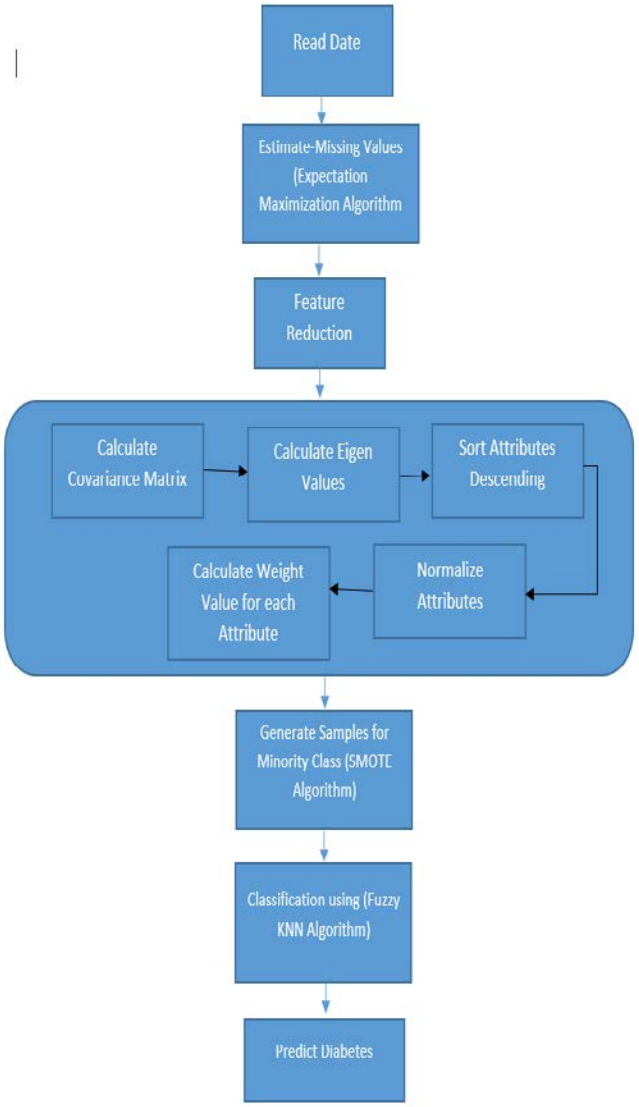

Fig. 1. Proposed Model Steps for Prediction of Diabetic Obese Patients.

IV. Evaluation Methods and Results

A. Evaluation Method

Evaluation of the performance of algorithms using the precision and recall criteria is very valuable. When making a choice, precision is the proportion of the time that the model properly predicts a good outcome. Precision is defined as the accurately identified or predicted positive examples divided by all the positive examples given. The proportion of properly recognized positives out of all existing positives is referred to as recall; it is calculated by dividing by the truly categorized positive cases by all the number of genuine examples in the set of positive testing. An optimal model must have both high recall and great accuracy. The F-measure is the consistent measure of accuracy and recall. The F-measure runs from zero to one, in which one indicating a classifier that properly captures accuracy and recall.

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

$$Sensitivity = \frac{TP}{TP+FN}, \quad (13)$$

$$F-measure = \frac{2(Precision)(Sensitivity)}{(Precision)+Sensitivity} \quad (14)$$

In which, the true positive (TP) represent the positive cases that predicted positive, the false negative (FN) represents the cases that were positive. However, it predicted negative and the false positive (FP) are the negative cases that were positively predicted.

B. Results

In this part, we report the findings obtained when the fuzzy KNN classifier used with the proposed model on the dataset described, and applying the fuzzy KNN classifier on the raw data of the dataset. Table I shows the proposed model output applied on the dataset after preprocessing compared to the same classifier but without data preprocessing.

TABLE I. Fuzzy KNN (Proposed) with Data Preprocessing Compared to Fuzzy KNN without Data Preprocessing

	FKNN Proposed Model Fuzzy Parameter = 0.5 & K= 5	**F-KNN Raw data Fuzzy Parameter = 0.5 & K= 5**
Sensitivity	1.0000	0.7156
Precision	0.9197	0.5735
Accuracy	0.9579	0.6577
F1 Score	0.9582	0.6367

Table I and Fig. 2 shows that the data preprocessing steps, estimating the missing values, feature reduction and solving the problem of unbalanced data enhanced the all measurement values resulted from the classifier.

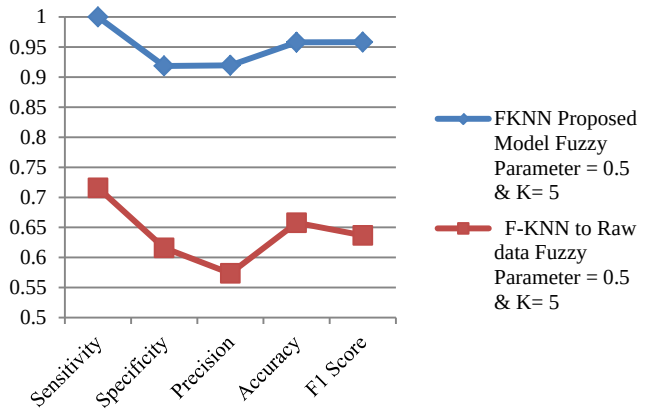

Fig. 2. Proposed Model with Data Preprocessing Vs Fuzzy KNN without Data Preprocessing.

The obtained results were compared to the results obtained in [21], as they used the same dataset. They proposed a two-phase classifier for predicting the potential diabetic obese patient as mentioned in related work section. Table II shows the basic differences in the proposed model and the model in [21].

TABLE II. The Difference in Methodologies used in the Proposed Model with the Model in [21]

	Proposed model	**Model in [21]**
Estimating missing values	Expectation Maximization Algorithm	weighted sum of linear interpolations from the closest accessible points.
Feature reduction	PCA algorithm	Correlation Feature Selection
Handling unbalanced data	SMOTE algorithm	Using K-fold cross validation
Classification	Fuzzy KNN classifier	SVM for phase1 and ANN for phase 2

Table III shows the comparison between results of the proposed model and results in [21].

TABLE III. Accuracy Comparison between Proposed Model and Model in [21]

	Proposed model	**Model in [21]**
Accuracy	**95.97%**	**86.56%**

From Tables II and III, we can observe that the algorithms and techniques used in the proposed model to prepare the data before training and testing were affected positively the data especially the steps of estimating missing values and handling unbalanced data, also the proposed classifier introduces a promising classification accuracy compared to the results introduced in [21].

V. Conclusion and Future Work

In this research a model for prediction of Diabetic Obese Patients was proposed, the model was based on Expectation Maximization, PCA, and SMOTE Algorithms in data preparation and preprocessing phase, and the fuzzy KNN classifier was used in prediction phase. The dataset used in this research was obtained from Cairo University, Faculty of Medicine, Al-Kasr Al-Aini Hospital. The algorithms used in the preprocessing enriched the clearness and effectiveness of the dataset which reflected in the prediction phase as shown in the results. The prediction accuracy reached to 95.97% in the proposed model and this result outperforms a corresponding model applied on the same dataset mentioned in the related work. We can suggest some improvements in the preprocessing phase afterwards like adopting another feature selection algorithm and other algorithms for handling imbalanced data, and estimating the missing values. In addition, an ensemble model can be provided on more than one classifier in order to enhance the precision value.

References

[1] Nilashi M, bin Ibrahim O, Ahmadi H, Shahmoradi L. An analytical method for diseases prediction using machine learning techniques. Computers & Chemical Engineering 2017;106:212-23.

[2] Cichosz SL. Predictive models in diabetes: Early prediction and detecting of type 2 diabetes and related complications: Aalborg Universitetsforlag; 2016.

[3] Zou Q, Qu K, Ju Y, Tang H, Luo Y, Yin D. Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics 2018;9:515.

[4] Devi MR, Shyla JM. Analysis of various data mining techniques to predict diabetes mellitus. International Journal of Applied Engineering Research 2016;11:727-30.

[5] Mills EP, Brown KPD, Smith JD, Vang PW, Trotta K. Treating nonalcoholic fatty liver disease in patients with type 2 diabetes mellitus: a review of efficacy and safety. Therapeutic advances in endocrinology and metabolism 2018;9:15-28.

[6] Bhupathiraju SN, Hu FB. Epidemiology of obesity and diabetes and their cardiovascular complications. Circulation research 2016;118:1723-35.

[7] Ng M, Fleming T, Robinson M, Thomson B, Graetz N, Margono C, et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. The lancet 2014;384:766-81.

[8] Ballestri S, Zona S, Targher G, Romagnoli D, Baldelli E, Nascimbeni F, et al. Nonalcoholic fatty liver disease is associated with an almost twofold increased risk of incident type 2 diabetes and metabolic syndrome. Evidence from a systematic review and meta-analysis. Journal of gastroenterology and hepatology 2016;31:936-44.

[9] Kumar RN, Kumar MA. Medical Data Mining Techniques for Health Care Systems. International Journal of Engineering Science 2016;3498.

[10] Boukenze B, Mousannif H, Haqiq A. Performance of data mining techniques to predict in healthcare case study: chronic kidney failure disease. Int Journal of Database Managment systems 2016;8:1-9.

[11] Abdullah M, Al-Asmari S. Anemia types prediction based on data mining classification algorithms. Communication, Management and Information Technology–Sampaio de Alencar (Ed) 2017.

[12] Kumari VA, Chitra R. Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications 2013;3:1797-801.

[13] Daghistani T, Alshammari R. Diagnosis of diabetes by applying data mining classification techniques. International Journal of Advanced Computer Science and Applications (IJACSA) 2016;7:329-32.

[14] El-Halees, A. M., & Shurrab, A. H. (2017). Blood tumor prediction using data mining techniques. Health Informatics—An International Journal, 6.

[15] Aljumah, A. A., Ahamad, M. G., & Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University-Computer and Information Sciences, 25(2), 127-136.

[16] Nai-arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. Procedia Computer Science, 69, 132-142.

[17] Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., ... & Abbas, Z. (2019). A model for early prediction of diabetes. Informatics in Medicine Unlocked, 16, 100204.

[18] Shetty D, Rit K, Shaikh S, Patil N. Diabetes disease prediction using data mining.Innovations in information, embedded and communication systems (ICIIECS), 2017international conference on. 2017. p. 1–5.

[19] Haritha, R., Babu, D. S., & Sammulal, P. (2018). A hybrid approach for prediction of type-1 and type-2 diabetes using firefly and cuckoo search algorithms. International Journal of Applied Engineering Research, 13(2), 896-907.

[20] Deberneh, H.M.; Kim, I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. Int. J. Environ. Res. Public Health 2021, 18, 3317. https://doi.org/10.3390/ ijerph18063317.

[21] Ali, R. E., El-Kadi, H., Labib, S. S., & Saad, Y. I. (2019). Prediction of potential-diabetic obese-patients using machine learning techniques. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 8, 2019.

[22] Dempster, A.,Laird,N.,andRubin,D.(1977).Maximum likelihood from incom-plete data via the EM algorithm.Journal of the Royal Statistical Society, SeriesB,39(1):1–38.

[23] Lakhina, S., Joseph, S., & Verma, B. (2010). Feature reduction using principal component analysis for effective anomaly–based intrusion detection on NSL-KDD. International Journal of Engineering Science and Technology Vol. 2(6), 1790-1799.

[24] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16(1):321–357, 2002.

[25] Haibo He and Edwardo A Garcia. Learning from imbalanced data. Knowledge and Data Engineering, IEEE Transactions on, 21(9):1263–1284, 2009.

[26] Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. IEEE transactions on systems, man, and cybernetics, (4), 580-585.