Sigma J Eng & Nat Sci 36 (4), 2018, 1265-1282

Sigma Journal of Engineering and Natural Sciences
Sigma Mühendislik ve Fen Bilimleri Dergisi

Review Article

DIAGNOSIS OF DIABETES MELLITUS USING STATISTICAL METHODS AND MACHINE LEARNING ALGORITHMS

Ebru PEKEL*[1], Tuncay ÖZCAN[2]

[1]*Ondokuz Mayis University, Dep. of Industrial Engineering, SAMSUN;* ORCID: 0000-0001-7717-6790
[2]*İstanbul University-Cerrahpasa, Dep. of Industrial Eng., Avcılar-İSTANBUL;* ORCID:0000-0002-9520-2494

Received: 22.03.2018 Revised: 26.09.2018 Accepted: 19.10.2018

ABSTRACT

The early diagnosis of the diabetes condition is crucial for cure process, because an early diagnosis provides the ease of treatment for the patient and the physician. At this point, statistical methods and data mining algorithms can provide important opportunities for early diagnosis of diabetes mellitus. In the literature, many studies have been published for solution of this problem. In this study, firstly, these studies are analyzed in detail and classified according to their methodologies and solution approaches. The main aim of this paper is to provide the comprehensive and detailed review of the diagnosis of diabetes by statistical methods and machine learning algorithms. Also, this paper presents a literature review on the diagnosis diabetes up to the end of 2017. It's identified over 425 papers, highly cited 100 ones are presented in detailed. This paper provides to guide future research and knowledge accumulation and creation of classification and prediction techniques in diagnosis of diabetes. This study shows it is clear that the combination of different machine learning algorithms and optimization models can lead to more meaningful and powerful results.
Keywords: Classification, diabetes mellitus, machine learning, prediction, statistical methods.

1. INTRODUCTION

Diabetes mellitus is a group of metabolic disorders with one common manifestation: elevated blood sugar or hyperglycemia [1]. The detection and the diagnosis of the diabetes is the most crucial point due to chronic hyperglycemia causes damage to the eye, kidney, nerves, heart, and blood vessels which causes the permanent damages. This contingency makes the diagnosis of the diabetes indeed important. The traditional diagnosis methods may be more painful and slower such as blood analysis [2]. A physician commonly determines decisions by evaluating the current blood analysis results of a patient. Therefore, diagnose of diabetes for the physicians and the patient is very difficult matter. For this reason, a lot of the intelligent diagnosis system for diabetes has been evolved by inspiring human-being biological constructers [3].

These evolved methods predict whether the probable patient suffers diabetes mellitus or not without including any surgeon progress [4]. In this context, this study aims to present a comprehensive literature review for the diagnosis of diabetes mellitus. The remainder of this study is organized as follows: the next section provides a comprehensive review of relevant

* Corresponding Author: e-mail: pkl.ebru@gmail.com, tel: (362) 312 19 19 / 1054

literature. In the third part, the mostly used techniques are described in detail. In section 4, the performance criteria are described which are the most widely used in the evaluation of classification algorithms in the literature. Chapter 5 gives general and comprehensive information about the studies in the literature. In the final part, the results of these study and literature review are discussed.

2. LITERATURE REVIEW

Many studies can be found for the early diagnosis of diabetes mellitus in the literature. These studies can be summarized as follows: Boyle et al. [5] managed a duality analysis to predict the diabetes population in 2050. Another study evaluated the average cost value caused by diabetes mellitus [6]. They estimated diabetes-related costs by using Cardiff Diabetes Cost-Benefit Model which takes into consideration the probabilistic processes. Some authors constructed a micro simulation model to evaluate the various scenarios for diabetes population [7]. Upadhyay and Patel [8] proposed a fuzzy classifier models to classify the diabetes condition. Their model allowed a splendid classification performance with 98.88% accuracy rate.

The regression based forecasting models have been commonly used by now and the regression models are the one of the oldest prediction models. A various statistical models were derived on diagnosis of diabetes as seen Table 1.

Additionally, the definition of the most suitable hybrid methods is essential due to acquire the best results. Moreover, the artificial intelligent technique provides to increase diagnostic accuracy and reduces costs and human resources [28]. Temurtas et al. [29] predicted on the same diabetes data by using Levenberg Marquardt learning algorithm and Probabilistic Neural Network (PNN) with 50 neurons for each hidden layer. Their model gives 82.37% accuracy rate. Polat and Güneş [30] presented a new hybrid model which occurs two stages. At first stage, principal component analysis is applied for reducing the number of features. At second stage, they predict diabetic condition. They acquire 89.47% accuracy rate. Doğantekin et al. [31] applied a hybrid method which is integrated Linear Discriminant Analysis (LDA) and ANFIS. Their prediction accuracy is 84.61%. Besides, Kala et al. [32] compared to three different neural network methods which are ANFIS and Evolutionary Artificial Neural Networks (EANN). According to their results, the best result is reach by EANN with 77.38 % accuracy rate. Drezet and Harrison [33] and Georga et al. [34] predicted by using support vector regression integrated with other technique. According to Karahoca et al. [35], the ANFIS provides the better results in comparison to Multinomial Logistic Regression under condition that dependent variables has more than two values such fuzzy numbers. They had a different database and considered glucose rate as input variable. Their error rate (RMSE) is 0.17%. Sharifi et al. [36] argued a hierarchical takagi-sugeno type fuzzy system for diabetes mellitus forecasting. Their accuracy rate is 78.73% which is the best result comparing to some conventional methods. Smith et al. [37] applied a neural network method with ADAP learning algorithm and their sensitivity rate was calculated as 76% while the accuracy rate was not calculated. Former intelligent methods are combined with other algorithms because the developed hybrid method's efficiency allows reaching the better results than former intelligent methods.

Table 1. The classification of the studies based on statistical methods

Ref.	Year	Methodology	Independent Variables
[9]	1993	Static and Dynamic Regression Model	Age, gender, frequency of diabetes
[10]	2003	Regression	Age, gender, weight, educational status, body mass index, waist circumference, fasting blood sugar.
[11]	2004	Multiple Regression Analysis	Age, usage of alcohol, usage of cigarette, physical activity, usage contraceptive, chronic pancreas history,

Ref.	Year	Methodology	Independent Variables
			hypertension history, education level, monthly income, weight, standing and sitting height lengths, waist and hip circumference.
[12]	2004	PearsonKi-Kare & ANOVA	Age, gender, body mass index, normal or overweight status, obesity status, type 2 diabetes status, health insurance.
[13]	2009	Cox Regression Model	Age, body mass index, diabetes history, social status, ethnicity.
[14]	2010	Lineer Regression	Ethnicity, age, poverty status.
[15]	2010	Pearson Partial Correlation	Age, race, waist circumference, hypertension status, cholesterol status, physical activity, smoking, alcohol use, diabetes in the family.
[16]	2011	Kohen's Kappa Regression	Blood glucose levels, cholesterol, triglyceride, body mass index, blood pressures, age, monthly income.
[17]	2011	Regression	Age, Diabetes predigree, height, systolic blood pressure, hip circumference measure, body mass index, cholesterol, non-HDL blood pressure, triglyceride, fasting blood sugar, physical activity, c-reactive protein, family income, smoking, alcohol use, use of lıquid-decreasing drugs .
[18]	2011	Regression	Age, gender.
[19]	2011	Statistical T Test	Age, gender, educational status, income status, the birth, smoking, alcohol use, physical activity, body mass index.
[20]	2012	Regression	Weight, height, waist and hip measures, HbA1c, glucose, uric acid, AST, ALT, GGT.
[21]	2013	C Statistical Analysis	Age, gestational status, body mass index, diabetes status in family tree, blood pressure, fasting sugar, fasting insulin concentration.
[22]	2013	Multi-Variable Adaptive Regression Curves	The number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps subcutaneous thickness, 2 hour serum insulin, diabetes pedigree function, age.
[23]	2014	Cox Regression Model	Having a parent with diabetes, having a parent with 2 diabetes, having a sibling with at least 1 diabetes, age, height, waist circumference, hypertension predisposition, physical activity, smoking, whole grain consumption, coffee consumption, red meat consumption.
[24]	2015	Multiple Regression Analysis	Waist circumference measure, body mass index, smoking, use of hypertension drugs, blood pressure values, plasma glucose ratio, HbA1c, cholesterol values, triglyceride ratio.
[25]	2015	Pennsylvania Clinics	Chronic liver diseases, high alanine aminotransferase, reflux state, hypertension, hA1c ratio.
[26]	2015	Wilcoxon Signed-Rank Regression Model	Age, body mass index, overwight, obesity, hypertension, diabetic state.
[27]	2016	General Regression Networks.	The number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps subcutaneous thickness, 2 hour serum insulin, diabetes pedigree function, age.

The classification of major studies that use machine learning algorithms in literature is presented in Table 2.

Table 2. The classification of the studies based on machine learning algorithms

Ref.	Year	Methodology	Independent Variables
[38]	2005	Artificial Neural Network	The number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps subcutaneous thickness, 2 hour serum insulin, diabetes pedigree function, age.
[39]	2007	K Nearest Neighbour Algorithm	Age, diagnosis time, HbA1c, blood sugar, triglyceride, cholesterol, body mass index, systolic blood pressure, diastolic blood pressure.
[40]	2010	Support Vector Machines	Age, gender, family history of diabetes, body mass index, waist and hip measurements, systolic blood pressure, diastolic blood pressure, cholesterol, fasting blood glucose, 2-hour glucose.
[41]	2011	Multi-Layer Perceptron	The number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps subcutaneous thickness, 2 hour serum insulin, diabetes pedigree function, age.
[42]	2011	Artificial Neural Network with Genetic Algorithm	The number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps subcutaneous thickness, 2 hour serum insulin, diabetes pedigree function, age.
[43]	2012	Back-Propagation Neural Networks	Smoking, usage of alcohol, body mass index, waist measure, family history, blood pressure.
[44]	2012	Machine Learning Algoritms	A1c1, A1c2, Sys-BP1, Sys-BP2, Dias-BP1, Dias-BP2, Serum-GLU1, Serum-GLU2,body mass index, keratin, HDL, MDRD, triglesirid, race, gender, age, diabetes condition.
[45]	2013	Feed-Forward Multi-Layer Neural Network	Heart attack history, cholesterol level, length.
[46]	2013	ROC Curve	Age, height, waist circumference, hypertension predisposition, physical activity, smoking, full-grain consumption, coffee consumption, red meat consumption, alcohol consumption.
[47]	2013	Artificial Neural Network with Levenberg Marquardt	The number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps subcutaneous thickness, 2 hour serum insulin, diabetes pedigree function, age.
[48]	2013	Support Vector Machines	The number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps subcutaneous thickness, 2 hour serum insulin, diabetes pedigree function, age.
[49]	2013	Regression + Genetic Programming + K Nearest Neighbour Alg.	The number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps subcutaneous thickness, 2 hour serum insulin, diabetes pedigree function, age.
[50]	2014	K Nearest Neighbour Algorithm	Age, gender, body mass index, blood pressure, blood pressure, plasma glucose ratio, triceps skin fold thickness, 2-hour serum insulin, diabetes pedigree, cholesterol, weight.
[51]	2014	Support Vectors with Basic Component Analysis	The number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps subcutaneous thickness, 2 hour serum insulin, diabetes pedigree function, age.
[52]	2014	Artificial Neural Network	Age, diabetes predegree, weight, gender, usage of alcohol and cigarette, frequency of thirst, urinary frequency, height, feeling of fatigue easily.
[53]	2014	K-Means Clustering Algorithm	The number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps subcutaneous thickness, 2 hour serum insulin, diabetes pedigree function, age.
[54]	2014	Naive Bayes Classifier	The number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps subcutaneous thickness, 2 hour serum insulin, diabetes pedigree function, age.
[55]	2014	Lasso Statistical Analysis with Bayesian Approach	Age, occupational status, nutritional status.
[56]	2014	Support Vector Machine with Principal Component Analysis	Hair and urine values(Li, Cr, Fe, Zn, Cu, Mg, Ni,V.)
[57]	2014	Artificial Neural Network	Brain Cancer Indications.

Ref.	Year	Methodology	Independent Variables
		Algorithms	
[58]	2015	ROC &Hosmer-Lemeshow	Blood pressure values, anthropometric measures, fasting blood sugar.
[59]	2015	Recursive Neural Network	Total bilirubin, BUN, Keratinine, Glucose AC, Glucose PC, Thyroxine, Uric Acid, Cholesterol, Triglyceride, HDL, Glucose, Gene, Age, Vital Capacity, Estimated vital capacity, FEV1, PFR, Albumin, Total Protein, SGOT, SGPT, ELDL, LDL.
[60]	2016	Artificial Neural Networks	Gender, age, height, weight, body mass index, diabetes history, pregnancy history, gestational diabetes history, abortion history, high blood pressure history, use of blood pressure drugs and history, systolic and diastolic blood pressure.
[61]	2016	Cognitive Development Optimization Algorithm based Support Vector	The number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps subcutaneous thickness, 2 hour serum insulin, diabetes pedigree function, age.

On the other hand, there have been probabilistic approaches to the diagnosis like Markov Models [62, 63]. A classification of markov models based studies in literature on the diagnosis of diabetes mellitus is given in Table 3.

Table 3. The classification of the studies based on Markov models

Ref.	Year	Methodology	Markov Status
[64]	2003	Dynamic Markov Model	Age, race, ethnicity, gender.
[65]	2006	Markov Model	Undiagnosed diabetes status, diagnosed diabetes status, death status.
[66]	2009	Markov Model	Gender, ethnicity, blood pressure, cholesterol level, GHb level, duration of diabetes.
[67]	2010	Markov Model	Diabetes status, Obesity status, Smoking.
[68]	2011	Markov Model	Diabetes status, Obesity status, Smoking.
[69]	2012	Markov Model	Diabetes status, Obesity status, Smoking.
[70]	2013	Markov Model	Diabetes status, Obesity status, Smoking.
[71]	2013	Discrete Markov Model	Demographic changes, disease dynamics, age and gender.
[72]	2014	Monte Carlo with Markov Model	Death, fertility, migration, body mass index, genotype, participation in work.
[73]	2014	Markov Model	Age, ethnicity, marital status, level of education, occupation, family income, relatives status, body mass index, physical activities, smoking, sleep duration, family history of diabetes.
[74]	2015	Markov Model	Diabetes status, Obesity status, Smoking.
[75]	2017	Dynamic Markov Model	Undiagnosis diabetic state, Type 2 Diabetes status, Type 1 Diabetes status, death.

In order to improve the performance measures of machine learning algorithms, hybridization approach with optimization algorithms has been used in recent years [76]. The idea that machine learning algorithms can be hybridized with optimization algorithms is first proposed by Davis [77]. Later on, this work was first conducted by Kelly and Davis [78]. The authors showed that the K-nearest neighbor algorithm was hybridized with the genetic algorithms and increased the performance values. Although the introduction of a new idea by Kelly and Davis in the literature began in the 1990s, the full dissemination of the idea became possible from the 2000s. Some highly cited studies are summarized in Table 4.

Table 4. Hybrid Models

Ref.	Year	Method	Hybrid
[79]	1990	Machine Learning Algorithms	Genetic Algorithm
[80]	1996	Artifcial Neural Networks	Multi-Variable Discriminant Analysis
[81]	2001	Support Vector Machine	Independent Component Analysis
[82]	2003	Artifcial Neural Networks	Decision Tree
[83]	2004	Decision Tree	Genetic Algorithm
[84]	2005	Support Vector Machine	Genetic Algorithm
[85]	2006	Support Vector Machine	Genetic Algorithm
[86]	2006	Support Vector Machine	Genetic Algorithm
[87]	2006	NONMEM	Genetic Algorithm
[88]	2007	Support Vector Machine	Genetic Algorithm
[89]	2007	K-Nearest Neighbors	Fuzzy Artificial Immune Recognition System
[90]	2007	K-Nearest Neighbors	Tabu Search Algorithm
[91]	2007	Artifcial Neural Networks	Genetic Algorithm
[92]	2007	Artifcial Neural Networks	Ant Colony Optimization
[93]	2008	Support Vector Machine	Particle Swarm Optimization
[94]	2008	Support Vector Machine	Genetic Algorithm
[95]	2009	K Harmonic Means	Particle Swarm Optimization
[96]	2009	Artifcial Neural Networks	Decision Tree
[97]	2010	Support Vector Machine	Genetic Algorithm
[98]	2010	Support Vector Machine	Independent Component Analysis
[99]	2010	K-Nearest Neighbors	Genetic Algorithm
[100]	2010	K-means Algorithm	Particle Swarm Optimization
[101]	2012	Support Vector Machine	Simulated Annealing
[102]	2013	Support Vector Machine	Particle Swarm Optimization
[103]	2013	Artifcial Neural Networks	Genetic Algorithm
[104]	2014	Support Vector Machine	K-Means Algorithm

3. METHODOLOGIES

The machine learning algorithms are frequently used in the Diabetes Mellitus prediction and classification problems as can be figured out from the previous sections [106, 107, 108, 109]. In this section, the most 4 popular machine learning algorithms are introduced and explained in their general form.

3.1. Decision Tree

Decision Tree (DT) is the one of the supervised learning algorithm that is mostly used in classification problems and works on both categorical and continuous input and output variables [110]. It is one of the most widely used and practical methods for inductive inference. Decision trees learn and train themselves from given examples and predict for unseen situations.

Each branch node represents a choice between a number of alternatives and each leaf node represents a decision. In DT, there have been some measures that can help us in selecting the best choice such entropy, gained information. In data mining, entropy is a measure of the uncertainty about a source of messages or a degree of disorganization in the data set. Given a collection S containing positive and negative examples of some target concept, the entropy of S relative to this boolean classification is calculated as in Equation (1).

$$Entropy\ (S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{1}$$

More precisely, the information gain Gain (S, A) of an attribute A relative to a collection of examples S is defined as in Eq(2).

$$Gain\ (S, A) = Entropy\ (S) - \sum_{v \in values\ (A)} \frac{S_v}{S} . Entropy\ (S_v) \tag{2}$$

S = Each value v of all possible values of attribute A
S_v = Subset of S for which attribute A has value v
$|S_v|$ = Number of elements in S_v
$|S|$ = Number of elements in S

Decision trees, while providing easy to view illustrations, can also be unwieldy. Even data that is perfectly divided into classes and uses only simple threshold tests may require a large decision tree. Large trees are not intelligible, and pose presentation difficulties.

3.2. Naive Bayes

Naive Bayes classifier is a useful algorithm for the classification problem and is based on Bayes' theorem with independence assumptions between predictors [111].

Bayes theorem provides a way of calculating the posterior probability for each class, *P(c|x)*, from *P(c)*, *P(x)*, and *P(x|c)*. Naive Bayes classifier assume that the effect of the value of a predictor *(x)* on a given class *(c)* is independent of the values of other predictors. This assumption is called class conditional independence (Eq. 3)

$$P(c\ |x\ ) = \frac{P(\ x|c)P(c)}{p(x)} \tag{3}$$

$P(c|x)$ is the posterior probability of class (target) given predictor (attribute).
$P(c)$ is the prior probability of class.
$P(x|c)$ is the likelihood which is the probability of predictor given class.
$P(x)$ is the prior probability of predictor.

Naïve bayes is easy to implement and fast to solve problems. It scales linearly with the number of predictors and data points and can be used for both binary and multiclass classification problems.

3.3. Support Vector Machine

In machine learning, Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and also regression analysis [112, 113]. SVM is composed in the framework of statistical learning, which has been developed by Vapnik and Chervonenkis.

SVR maps the input data x into a higher dimensional feature space through a nonlinear mapping Φ and then a linear regression problem is obtained and solved in this feature space.

With the given training data $\{(x_1,y_1),\ldots, (x_i,y_i),\ldots, (x_n,y_n)\}$, the mapping function can be formulates as in Eq (4).

$$f(x) = \sum_{i=1}^{n} w_i \Phi_i x_i + b \tag{4}$$

Where ω_i and b are the parameters that need to be defined. SVR is to find a function *f(x)* that has at most ε deviation from the actually obtained targets y_i for all the training data and at the same time is as flat as possible. Flatness in this case means to reduce the model complexity by minimizing $||\omega|^{|2}$, so that this problem can be written as an optimization problem as seen in Eq(5) and Eq (6).

$$Min\ \frac{1}{2}||w||^2 \tag{5}$$

$$s.t.\begin{cases} y_i - \Phi(w, x_i) - b \le \varepsilon \\ \Phi(w, x_i) + b_i - y \le \varepsilon \end{cases} \tag{6}$$

Equation (7) defines a constrained optimization problem. Equation (8) shows the solution of this problem.

$$Max\, W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^{n} \alpha_i \alpha_i\, y_i y_i x_i^T x_j \tag{7}$$

$$s.t. C \ge \alpha_i \ge 0, \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{8}$$

While SVM has a regularization parameter, which makes the user think about avoiding over-fitting, it does not give class probabilities and being rather cumbersome for multiclass problems.

3.4. Artificial Neural Network

The purpose of ANNs is to present a development to mimic the basic biological neural systems including the human brain [114]. ANNs have a number of interconnected simple processing points. If an input signal is picked by each node and operated through an activation or transfer function and a transformed output signal is generated. Though each function is implemented by each individual neuron quite slowly, a network can execute an amazing number of tasks efficiently.

As an advantage, ANN is the flexibility in changing the encoding of the data to fit different statements of the problem and is capable to conform to the real world.

4. MACHINE LEARNING PERFORMANCE MEASURES

The confusion matrix is a two by two table that contains four outcomes produced by a binary classifier. Various measures, such as error-rate, accuracy, specificity, sensitivity, and precision, are derived from the confusion matrix (Table 5).

Table 5. Confusion Matrix [115]

	Prediction outcome			
Actual value		**p**	**N**	**Total**
	p'	True Positive	False Negative	P'
	n'	False Positive	True Negative	N'
	Total	P	N	

There are basic performance metrics that can be used to evaluate the methods applied in machine learning. The most popular and basic method used to measure model performance is the accuracy of the model. The accuracy rate is the ratio of the number of true classified samples *(TP + TN)* to the total number of samples *(TP + TN + FP + FN)* (Equation 9).

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \tag{9}$$

Precision is the ratio of the number of True Positive *(TP)* samples predicted as class 1 to the total number of samples *(TP + FP)* predicted as class 1 (Equation 10).

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

Sensitivity is the ratio of the number of correctly classified positive samples *(TP)* to the total number of positive samples *(TP + FN)* (Equation 11).

$$Sensitivity = \frac{TP}{TP+FN} \tag{11}$$

Precision and sensitivity metrics alone are not enough to make a meaningful comparison in the application comprised from a few machine learning algorithms. It can be obtained better comparison results by evaluating both criteria together. Therefore, F-measure *(F)* is used for comparing the algorithms. The F-criterion is the harmonic mean of the precision *(P)* and the sensitivity *(S)* (Equation 12).

$$F - Measure = \frac{2*P*S}{P+S} \tag{12}$$

5. DESCRIPTIVE ANALYSIS

In the present, machine learning algorithms are mostly popular among the other techniques due to yield outstanding classification performance according to former techniques. Nevertheless, it can be constructed much more effective machine learning algorithms by hybridizing some optimization techniques.

The distribution of articles by year of publication is shown in Figure 1. It is obvious that publications which are related to application of machine learning techniques in diagnosis diabetes mellitus have increased significantly from 2007 to 2017. In 2006, the largest increase has taken place between 2015-2016 years with 75%.

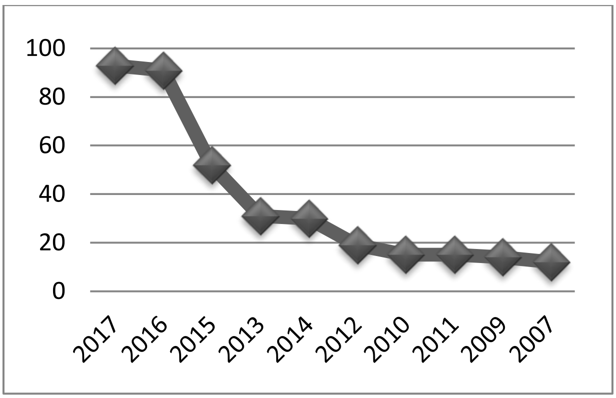

Figure 1. Machine learning in diabetes mellitus

Table 6 shows the distribution of articles by countries. Articles related to application of machine learning techniques in diabetes diagnosis are distributed across 10 countries. Of these, "USA", which focuses on the knowledge of the application of expert and intelligent systems in diabetes more than 25% (110 of 425 articles) of the total number of articles published.

Table 7 shows the distribution of science category by classification and prediction on diabetes from Web Of Science. Among 425 papers which have been applied in diabetes diagnosis, artificial intelligence field is the most commonly used in literature. It has been described in 96 (22.5%) out of 425 articles in total. Following are computer science interdisciplinary applications and engineering electrical electronic which have been described in 74 (17.41%) and in 74 (17.41%) fields respectively.

Table 6. Studies by countries

Country	Amount	Ratio
USA	110	0.25882
India	63	0.14824
China	44	0.10353
England	30	0.07059
Australia	21	0.04941
Turkey	19	0.04471
Canada	15	0.03529
Japan	14	0.03294
Malaysia	14	0.03294
South Korea	14	0.03294

Table 7. Studies by work categories

Work Categories	Amount	Ratio
Computer Science Artificial Intelligence	96	0.22588
Computer Science Interdisciplinary Applications	74	0.17412
Engineering Electrical Electronic	74	0.17412
Computer Science Theory Methods	66	0.15529
Medical Informatics	64	0.15059
Engineering Biomedical	55	0.12941
Computer Science Information Systems	50	0.11765
Mathematical Computational Biology	43	0.10118
Health Care Sciences Services	27	0.06353
Endocrinology Metabolism	23	0.05412

Table 8 shows the top 10 of articles by journal. Articles related to application of machine learning techniques in diabetes diagnosis are distributed across 55 journals. Of these, "Expert Systems with Applications", which focuses on the knowledge of the application of expert and intelligent systems in diabetes diagnosis, contains more than 3% (14 of 425 articles) of the total number of articles published.

Table 8. Studies by journals

Journals	Amount	Ratio
Expert systems with applications	14	0.03294
Lecture notes in computer science	10	0.02353
Artificial intelligence in medicine	9	0.02118
Plos one	9	0.02118
Journal of biomedical informatics	7	0.01647
Ieee engineering in medicine and biology society conference proceedings	6	0.01412
Journal of medical systems	6	0.01412
Lecture notes in artificial intelligence	6	0.01412
Computer methods and programs in biomedicine	5	0.01176
Diabetes	5	0.01176

The distribution of articles by machine learning techniques is shown in Figure 2. It is obvious that support vector machines were significantly used in application of machine learning techniques in diagnosis diabetes mellitus compared to other techniques. It can be said that the most 4 popular machine learning algorithm are support vector machines, naïve bayes, neural network and decision consecutively.

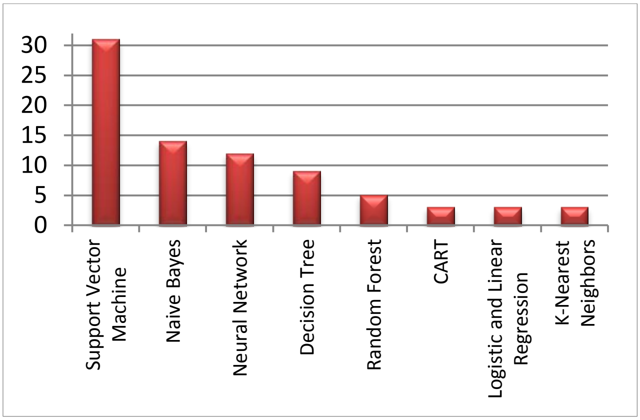

Figure 2. Machine learning algorithm in general

6. DISCUSSION AND CONCLUSIONS

In this study, firstly a detailed classification of studies in literature about the diagnosis of diabetes mellitus.

The diagnosis problem of the diabetes mellitus is the oldest research topics. This problem is the crucial problem due to expect that the amount of the diabetes patient may increase until 2025. While some papers perform to predict the spread of DM, some papers perform to analyze the diabetic condition of people. The regression models were used extensively to forecast the diabetic condition of persons in earlier years due to the statistical methods is the former techniques among the forecasting techniques. The Markov models emerged to predict diabetic condition people by time. These models have been widely placed in the literature due to has an advantage which is taking into consideration probabilistic factors.

Among the 97 articles, 31 described support vector machine in the diabetes classification problems. Support vector machine can be applied easily in classification due to allow get only the binary outputs. Thus, it is not surprising that support vector machine were used in a wide range of diabetes classification. Naïve bayes and neural networks techniques rank after support vector machine in popularity of application diagnosis of in diabetes.

The majority of diabetes studies in the literature have been conducted on the PIMA Indian data set with 0.67% [116]. The performance measures on this dataset show a change from 73.83% to 96.00%. Performance measures are higher in diabetes prediction and classification problems that are conducted in data sets collected from different sources.

This study might have some limitations such that only surveyed articles published between 2000 and 2017, which were extracted based on combination of keywords search of “diabetes” and “classification” “prediction “or “machine learning”.

In the literature, traditional machine learning algorithms have been replaced by models that hybridize with optimization algorithms over time. As the optimization algorithm, it was seemed that the most commonly used and best-resultant method is genetic algorithm. In addition, nearly half of the hybrid models proposed in the literature are hybridized with genetic algorithms (13 of 27). Following is particle swarm optimization algorithm with 14.81% (4 of 27).

It is obvious that the combination of different machine learning algorithms and optimization models can lead to more meaningful and powerful results. Among machine learning algorithms which are employed in hybridization, neural networks and support vector machines are widely used in diagnosis of diabetes. It might be obtained more powerful results by hybridizing new prediction and classification methods such as Extreme Learning Machines. On the other hand, to be used genetic algorithm in hybridization yields good results and increases the performance of traditional machine learning algorithms in general.

REFERENCES

[1] Gavin, J. R. (1998). New classification and diagnostic criteria for diabetes mellitus. *Clinical cornerstone*, *1*(3), 1-12.

[2] Gray, L. J., & Khunti, K., (2013). Type 2 diabetes risk prediction—Do biomarkers increase detection?. *Diabetes research and clinical practice*, *101*(3), 245-247.

[3] Kayaer, K., & Yıldırım, T., (2003), Medical diagnosis on Pima Indian diabetes using general regression neural networks. In *Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP)* (pp. 181-184).

[4] Magliano, D. J., Peeters, A., Vos, T., Sicree, R., Shaw, J., Sindall, C., Zimmet, P. Z., (2009), Projecting the burden of diabetes in Australia–what is the size of the matter?. *Australian and New Zealand journal of public health*, *33*(6), 540-543.

[5] Boyle, J. P., Honeycutt, A. A., Narayan, K. V., Hoerger, T. J., Geiss, L. S., Chen, H., & Thompson, T. J. (2001). Projection of diabetes burden through 2050. *Diabetes care*, *24*(11), 1936-1940.

[6] McEwan, P., Peters, J. R., Bergenheim, K., & Currie, C. J., (2006), Evaluation of the costs and outcomes from changes in risk factors in type 2 diabetes using the Cardiff stochastic simulation cost-utility model (DiabForecaster). *Current medical research and opinion*, *22*(1), 121-129.

[7] Shi, L., van Meijgaard, J., & Fielding, J., (2011), Forecasting diabetes prevalence in California: a microsimulation. *Prev Chronic Dis*, *8*(4), A80.

[8] Upadhyay A., Patel V. R., (2016), Comparative Study - Prediction of Diabetes and Heart Disease using Data Mining Approaches, International Journal of Engineering Technology, Management and Applied Sciences, 4(1), 70-76

[9] Ruwaard, D., Hoogenveen, R. T., Verkleij, H., Kromhout, D., Casparie, A. F., & Van der Veen, E. A. (1993). Forecasting the number of diabetic patients in The Netherlands in 2005. *American journal of public health*, *83*(7), 989-995.

[10] Gu, D., Reynolds, K., Duan, X., Xin, X., Chen, J., Wu, X., (2003), Prevalence of diabetes and impaired fasting glucose in the Chinese adult population: International Collaborative Study of Cardiovascular Disease in Asia (InterASIA). *Diabetologia*, *46*(9), 1190-1198.

[11] Rosenthal A. D. Et al., (2004), Body fat distribution and risk of diabetes among Chinese women, International Journal of Obesity, 28, 594-599.

[12] McNeely, M. J., & Boyko, E. J., (2004), Type 2 diabetes prevalence in Asian Americans. *Diabetes Care*, *27*(1), 66-69.

[13] Cox, J., Coupland, C., Robson, J., Sheikh, A., & Brindle, P. (2009). Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *Bmj*, *338*, b880.

[14] Holman, N., Forouhi, N. G., Goyder, E., & Wild, S. H., (2011), The Association of Public Health Observatories (APHO) diabetes prevalence model: estimates of total diabetes prevalence for England, 2010–2030. Diabetic Medicine, 28(5), 575-582

[15] Chao, C., Song, Y., Cook, N., Tseng, C. H., Manson, J. E., Eaton, C., Liu, S. (2010). The lack of utility of circulating biomarkers of inflammation and endothelial dysfunction for

type 2 diabetes risk prediction among postmenopausal women: the Women's Health Initiative Observational Study. *Archives of internal medicine*, *170*(17), 1557-1565.

[16] Ahasan N., Islam, Z., Alam, B., Miah, T., Nur, Z.,(2011), Prevalence and Risk Factors of Type 2 Diabetes Mellitus Among Secretariat Employees of Bangladesh, Bangladesh Journals Online, 12(2) : 125-130

[17] Onat, A., Can, G., Yüksel, H., Ayhan, E., Dogan, Y., & Hergenç, G., (2011), An algorithm to predict risk of type 2 diabetes in Turkish adults: contribution of C-reactive protein. Journal of endocrinological investigation, 34(8), 580-586.

[18] Lau, R. S., Ohinmaa, A., & Johnson, J. A., (2011), Predicting the Future Burden of Diabetes in Alberta from 2008 to 2035. *Canadian Journal of Diabetes*, *35*(3), 274-281.

[19] Lee, J. W. R., Brancati, F. L., & Yeh, H. C., (2011), Trends in the prevalence of type 2 diabetes in Asians versus Whites. *Diabetes care*, *34*(2), 353-357.

[20] Abbasi, A., Bakker, S. J., Corpeleijn, E., Gansevoort, R. T., Gans, R. O., Peelen, L. M. & Beulens, J. W. (2012). Liver function tests and risk prediction of incident type 2 diabetes: evaluation in two independent cohorts. *PloS one*, *7*(12), e51496.

[21] Kwak, S. H., Choi, S. H., Kim, K., Jung, H. S., Cho, Y. M., Lim, S., Jang, H. C., (2013), Prediction of type 2 diabetes in women with a history of gestational diabetes using a genetic risk score. *Diabetologia*, *56*(12), 2556-2563.

[22] Senthilkumar D., Paulraj S., (2013), Diabetes Disease Diagnosis Using Multivariate Adaptive Regression Splines, International Journal of Engineering and Technology, 5(5), 3922-3929.

[23] Mühlenbruch, K., Joost, H. G., Boeing, H., & Schulze, M. B., (2014), Risk prediction for type 2 diabetes in the German population with the updated German Diabetes Risk Score (GDRS). *Ernährungs Umsch*, *61*, 90-3.

[24] Nanri A. et al.,, (2015), Development of Risk Score for Predicting 3-Year Incidence of Type 2 Diabetes: Japan Epidemiology Collaboration on Occupational Health Study, PLOS ONE 10(11):e0142779.

[25] Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S., & Sontag, D., (2015), Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*, *3*(4), 277-287.

[26] Hu, H., Huff, C. D., Yamamura, Y., Wu, X., & Strom, S. S., (2015), The relationship between Native American ancestry, body mass index and diabetes risk among Mexican-Americans. *PloS one*, *10*(10).

[27] Alby S. and Shivakumar B. L., (2016), A Prediction Model for Type 2 Diabetes Risk Among Indian Women, ARPN Journal of Engineering and Applied Sciences, 11 (3), 2037-2043.

[28] Polat, K., Şahan, S., & Güneş, S., (2006), A new method to medical diagnosis: Artificial immune recognition system (AIRS) with fuzzy weighted pre-processing and application to ECG arrhythmia. *Expert Systems with Applications*, *31*(2), 264-269.

[29] Temurtas, H., Yumusak, N., & Temurtas, F., (2009), A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with applications*, *36*(4), 8610-8615.

[30] Polat, K., & Güneş, S., (2007), An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing*, *17*(4), 702-710.

[31] Dogantekin, E., Dogantekin, A., Avci, D., & Avci, L. (2010). An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS. *Digital Signal Processing*, *20*(4), 1248-1255.

[32] Kala, R., Shukla, A., & Tiwari, R., (2009), Comparative analysis of intelligent hybrid systems for detection of PIMA indian diabetes. In *Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on* (pp. 947-952). IEEE.

[33] Drezet, P. M., & Harrison, R. F. (2001). A new method for sparsity control in support vector classification and regression. *Pattern Recognition*, *34*(1), 111

[34] Georga, E. I., Protopappas, V. C., Ardigò, D., Marina, M., Zavaroni, I., Polyzos, D., & Fotiadis, D. I., (2013). Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. *IEEE journal of biomedical and health informatics*, *17*(1), 71-81.

[35] Karahoca, A., Karahoca, D., & Kara, A., (2009), Diagnosis of diabetes by using adaptive neuro fuzzy inference systems. In *Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control, 2009. ICSCCW 2009. Fifth International Conference on* (pp. 1-4). IEEE.

[36] Sharifi, A., Vosolipour, A., Sh, M. A., & Teshnehlab, M., (2008), Hierarchical Takagi-Sugeno type fuzzy system for diabetes mellitus forecasting. In *Machine Learning and Cybernetics, 2008 International Conference on* (Vol. 3, pp. 1265-1270). IEEE.

[37] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S., (1988), Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Annual Symposium on Computer Application in Medical Care (p. 261). American Medical Informatics Association.

[38] Farhanah, S., Jafan, B., & Ali, D. M. (2005). Diabetes Mellitus Forecast using Artificial Neural Networks (ANN*). In Asian Conference on sensors and the international conference on new techniques in pharamaceutical and medical research proceedings (IEEE)* (pp. 135-138).

[39] Huang, Y., McCullagh, P., Black, N., & Harper, R., (2007), Feature selection and classification model construction on type 2 diabetic patients' data. *Artificial intelligence in medicine*, *41*(3), 251.

[40] Barakat, N., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE transactions on information technology in biomedicine*, *14*(4), 1114-1120.

[41] Acar, E., Özerdem, M. S., & Akpolat, V. (2011). Diabetes Mellitus Forcast Using Various Types of Artificial Neural Networks. In *6th International Advanced Technologies Symposium* (pp. 196-201).

[42] Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A., (2011), Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes. *International Journal on Soft Computing*, *2*(2), 15-23.

[43] Luangruangrong, W., Rodtook, A., & Chimmanee, S., (2012), Study of Type 2 diabetes risk factors using neural network for Thai people and tuning neural network parameters. In *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on* (pp. 991-996). IEEE.

[44] Mani, S., Chen, Y., Elasy, T., Clayton, W., & Denny, J. (2012). Type 2 diabetes risk forecasting from EMR data using machine learning. In *AMIA annual symposium proceedings* (Vol. 2012, p. 606). American Medical Informatics Association.

[45] Mohamed, N., Ahmad, W. M. A. W., & Aleng, N. A., (2013), Modeling Multi Layer Feed-forward Neural Network Model on the Influence of Hypertension and Diabetes Mellitus on Family History of Heart Attack in Male Patients. *Applied Mathematical Sciences*, *7*(41), 2047-2053.

[46] Mühlenbruch, K., Jeppesen, C., Joost, H. G., Boeing, H., & Schulze, M. B., (2013), The value of genetic information for diabetes risk prediction–differences according to sex, age, family history and obesity. *PLoS One*, *8*(5), e64307.

[47] Khan, N., Gaurav, D., & Kandl, T., (2013), Performance evaluation of Levenberg-Marquardt technique in error reduction for diabetes condition classification. *Procedia Computer Science*, *18*, 2629-2637.

[48] Kumari, V. A., & Chitra, R., (2013), Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, *3*(2), 1797-1801.

[49] Ahmed, K. A. W. S. A. R., Jesmin, T. A. S. N. U. B. A., Fatima, U. S. H. I. N., Moniruzzaman, M., Emran, A. A., & Rahman, M. Z. (2012). Intelligent and effective diabetes risk prediction system using data mining. *Orient J Comput Sci Technol*, *5*, 215-221.

[50] Aslam, M. W., Zhu, Z., & Nandi, A. K. (2013). Feature generation using genetic programming with comparative partner selection for diabetes classification. *Expert Systems with Applications*, *40*(13), 5402-5412

[51] Saxena, K., Khan, Z., & Singh, S., (2014), Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm. International Journal of Computer Science Trends and Technology (IJCST).

[52] Jhaldiyal T. and Mishra P. K., (2014), Analysis and Prediction of Diabetes Mellitus Using PCA, REP and SVM, International Journal of Engineering and Technical Research (IJETR), 2 (8), 164-166.

[53] Sarwar A. and Sharma V., (2014), Comparative analysis of machine learning techniques in prognosis of type II diabetes, Al & Society, 29(1), 123-129.

[54] Thangaraju, P., Deepa, B., & Karthikeyan, T., (2014), Comparison of Data mining Techniques for Forecasting Diabetes Mellitus. International Journal of Advanced Research in Computer and Communication Engineering, 3(8).

[55] Diwani, S. A., & Sam, A. (2014). Diabetes Forecasting Using Supervised Learning Techniques. Advances in Computer Science: an International Journal, 3(5), 10-18.

[56] Shigemizu, D., Abe, T., Morizono, T., Johnson, T. A., Boroevich, K. A., Hirakawa, Y., & Maeda, S., (2014), The construction of risk prediction models using GWAS data and its application to a type 2 diabetes prospective cohort. *PLoS One*, *9*(3), e92549.

[57] Chen, H., Tan, C., Lin, Z., & Wu, T. (2014). The diagnostics of diabetes mellitus based on ensemble modeling and hair/urine element level analysis. *Computers in biology and medicine*, *50*, 70-75.

[58] Utomo, C. P., Kardiana, A., & Yuliwulandari, R., (2014), Breast cancer diagnosis using artificial neural networks with extreme learning techniques. *International Journal of Advanced Research in Artificial Intelligence*, *3*(7), 10-14.

[59] Tanamas, S. K., Magliano, D. J., Balkau, B., Tuomilehto, J., Kowlessur, S., Söderberg, S., Shaw, J. E., (2015), The performance of diabetes risk prediction models in new populations: the role of ethnicity of the development cohort. *Acta diabetologica*, *52*(1), 91-101.

[60] Chen, L. S., & Cai, S. J. (2015). Neural-Network-Based Resampling Method for Detecting Diabetes Mellitus. *Journal of Medical and Biological Engineering*, *35*(6), 824-832.

[61] Heydari, M., Teimouri, M., Heshmati, Z., & Alavinia, S. M., (2016), Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. *International Journal of Diabetes in Developing Countries*, *36*(2), 167-173.

[62] Kose, U., Guraksin, G. E., & Deperlioglu, O., (2016), Cognitive development optimization algorithm based support vector machines for determining diabetes. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *7*(1), 80-90.

[63] Amanda A. H., Boyle, JP., Broglio KR, Thompson TJ, Hoerger, TJ., Geiss, LS., Narayan, HM., 2003, A Dynamic Markov Model for Forecasting Diabetes Prevalence in the United States through 2050, Health Care Management Science, 6(3), 155-164.

[64] Boyle, J. P., Thompson, T. J., Gregg, E. W., Barker, L. E., & Williamson, D. F. (2010). Projection of the year 2050 burden of diabetes in the US adult population: dynamic

modeling of incidence, mortality, and prediabetes prevalence. *Population health metrics*, *8*(1), 29.

[65] Honeycutt, A. A., Boyle, J. P., Broglio, K. R., Thompson, T. J., Hoerger, T. J., Geiss, L. S., & Narayan, K. V., (2003), A dynamic Markov model for forecasting diabetes prevalence in the United States through 2050. Health *care management science*, 6(3), 155-164.

[66] Narayan, K. V., Boyle, J. P., Geiss, L. S., Saaddine, J. B., & Thompson, T. J., (2006), Impact of recent increase in incidence on future diabetes burden. *Diabetes care*, *29*(9), 2114-2116.

[67] Huang, E. S., Basu, A., O'grady, M., & Capretta, J. C., (2009), Projecting the future diabetes population size and related costs for the US. *Diabetes care*, *32*(12), 2225-2229.

[68] O'Flaherty M., Critchley J., Wild S., (2010), "018 Forecating Diabetes Prevalence Using a Simple Model: England and Wales 1993-2006", Journal of Eğidemiology & Community Health 2010, 64:A7

[69] O'flaherty, M., Critchley, J., Wild, S., Unwin, N., & Capewell, S., (2011), O1-5.6 Forecasting Diabetes Prevalence: validation of a simple model with few data requirements. *Journal of Epidemiology and Community Health*, *65*(Suppl 1), A17-A17.

[70] Abu-Rmeileh, N. M., Husseini, A., O'Flaherty, M., Shoaibi, A., & Capewell, S. (2012). Forecasting prevalence of type 2 diabetes mellitus in Palestinians to 2030: validation of a predictive model. *The Lancet*, *380*, S21.

[71] Al Ali, R., Mzayek, F., Rastam, S., Fouad, F. M., O’Flaherty, M., Capewell, S., & Maziak, W. (2013). Forecasting future prevalence of type 2 diabetes mellitus in Syria. *BMC Public Health*, *13*(1), 507.

[72] Waldeyer, R., Brinks, R., Rathmann, W., Giani, G., & Icks, A., (2013), Projection of the burden of type 2 diabetes mellitus in Germany: a demographic modelling approach to estimate the direct medical excess costs from 2010 to 2040. *Diabetic Medicine*, *30*(8), 999-1008.

[73] Phan, T. P., Alkema, L., Tai, E. S., Tan, K. H., Yang, Q., Lim, W. Y., Chia, K. S., (2014), Forecasting the burden of type 2 diabetes in Singapore using a demographic epidemiological model of Singapore. *BMJ open diabetes research & care*, *2*(1), e000012.

[74] Mutlu, F., Bener, A., Eliyan, A., Delghan, H., Nofal, E., Shalabi, L., & Wadi, N., (2014), Projection of diabetes burden through 2025 and contributing risk factors of changing disease prevalence: an emerging public health problem. *J Diabetes Metab*, *5*(2), 1000341.

[75] Saidi, O., O’Flaherty, M., Mansour, N. B., Aissi, W., Lassoued, O., Capewell, S., Romdhane, H. B., (2015), Forecasting Tunisian type 2 diabetes prevalence to 2027: validation of a simple model. *BMC public health*, *15*(1), 104.

[76] Utomo, C. P., Kardiana, A., Yuliwulandari, R. (2014). Breast cancer diagnosis using artificial neural networks with extreme learning techniques. International Journal of Advanced Research in Artificial Intelligence, 3(7), 10-14.

[77] Davis, L. (1990). Hybrid genetic algorithms for machine learning. In Machine Learning, IEE Colloquium on (pp. 9-1). IET.

[78] Kelly Jr, J. D., Davis, L. (1991). A Hybrid Genetic Algorithm for Classification. In IJCAI (Vol. 91, pp. 645-650).

[79] Carvalho, D. R., & Freitas, A. A. (2004). A hybrid decision tree/genetic algorithm method for data mining. Information Sciences, 163(1), 13-35.

[80] Lee, K. C., Han, I., & Kwon, Y. (1996). Hybrid neural network models for bankruptcy predictions. Decision Support Systems, 18(1), 63-72.

[81] Qi, Y., Doermann, D., & DeMenthon, D. (2001). Hybrid independent component analysis and support vector machine learning scheme for face detection. In Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on (Vol. 3, pp. 1481-1484). IEEE.

[82] Pan, Z. S., Chen, S. C., Hu, G. B., & Zhang, D. Q. (2003, November). Hybrid neural network and C4. 5 for misuse detection. In Machine Learning and Cybernetics, 2003 International Conference on (Vol. 4, pp. 2463-2467). IEEE.

[83] Niknam, T., & Amiri, B. (2010). An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. Applied Soft Computing, 10(1), 183-197.

[84] Li, L., Jiang, W., Li, X., Moser, K. L., Guo, Z., Du, L., ... & Rao, S. (2005). A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics, 85(1), 16-23.

[85] Min, S. H., Lee, J., & Han, I. (2006). Hybrid genetic algorithms and support vector machines for bankruptcy prediction. Expert systems with applications, 31(3), 652-660.

[86] Huerta, E. B., Duval, B., & Hao, J. K. (2006, April). A hybrid GA/SVM approach for gene selection and classification of microarray data. In Workshops on Applications of Evolutionary Computation (pp. 34-44). Springer, Berlin, Heidelberg.

[87] Bies, R. R., Muldoon, M. F., Pollock, B. G., Manuck, S., Smith, G., & Sale, M. E. (2006). A genetic algorithm-based, hybrid machine learning approach to model selection. Journal of Pharmacokinetics and Pharmacodynamics, 33(2), 195-221.

[88] Shon, T., & Moon, J. (2007). A hybrid machine learning approach to network anomaly detection. Information Sciences, 177(18), 3799-3821.

[89] Kelly Jr, J. D., & Davis, L. (1991, August). A Hybrid Genetic Algorithm for Classification. In IJCAI (Vol. 91, pp. 645-650).

[90] Şahan, S., Polat, K., Kodaz, H., & Güneş, S. (2007). A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Computers in Biology and Medicine, 37(3), 415-423.

[91] Kim, H. J., & Shin, K. S. (2007). A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets. Applied Soft Computing, 7(2), 569-576.

[92] Sivagaminathan, R. K., & Ramakrishnan, S. (2007). A hybrid approach for feature subset selection using neural networks and ant colony optimization. Expert systems with applications, 33(1), 49-60.

[93] Huang, C. L., & Dun, J. F. (2008). A distributed PSO–SVM hybrid system with feature selection and parameter optimization. Applied soft computing, 8(4), 1381-1391.

[94] Choudhry, R., & Garg, K. (2008). A hybrid machine learning system for stock market forecasting. World Academy of Science, Engineering and Technology, 39(3), 315-318.

[95] Aci, M., İnan, C., & Avci, M. (2010). A hybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm. Expert Systems with Applications, 37(7), 5061-5067.

[96] Tsai, C. F., & Wang, S. P. (2009, March). Stock price forecasting by hybrid machine learning techniques. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1, No. 755, p. 60).

[97] Kharrat, A., Gasmi, K., Messaoud, M. B., Benamrane, N., & Abid, M. (2010). A hybrid approach for automatic classification of brain MRI using genetic algorithm and support vector machine. Leonardo Journal of Sciences, 17(1), 71-82.

[98] Yang, H., Liu, J., Sui, J., Pearlson, G., & Calhoun, V. D. (2010). A hybrid machine learning method for fusing fMRI and genetic data: combining both improves classification of schizophrenia. Frontiers in human neuroscience, 4.

[99] Tahir, M. A., Bouridane, A., & Kurugollu, F. (2007). Simultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier. Pattern Recognition Letters, 28(4), 438-446.

[100] Yang, F., Sun, T., & Zhang, C. (2009). An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization. Expert Systems with Applications, 36(6), 9847-9852.

[101] Sartakhti, J. S., Zangooei, M. H., & Mozafari, K. (2012). Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA). Computer methods and programs in biomedicine, 108(2), 570-579.

[102] Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. Procedia Engineering, 53, 453-462.

[103] Aziz, A. S. A., Hassanien, A. E., Hanaf, S. E. O., & Tolba, M. F. (2013, December). Multi-layer hybrid machine learning techniques for anomalies detection and classification approach. In Hybrid Intelligent Systems (HIS), 2013 13th International Conference on (pp. 215-220). IEEE.

[104] Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), 1476-1482.

[105] Wong, L. Y., Toh, M. P. H. S., & Tham, L. W. C., (2017), Projection of prediabetes and diabetes population size in Singapore using a dynamic Markov model. *Journal of diabetes*, *9*(1), 65-75.

[106] Ahmed, K. A. W. S. A. R., Jesmin, T. A. S. N. U. B. A., Fatima, U. S. H. I. N., Moniruzzaman, M., Emran, A. A., & Rahman, M. Z. (2012). Intelligent and effective diabetes risk prediction system using data mining. *Orient J Comput Sci Technol*, *5*, 215-221.

[107] Jaafar, S. F. B., & Ali, D. M., (2005), Diabetes mellitus forecast using artificial neural network (ANN). In *Sensors and the International Conference on new Techniques in Pharmaceutical and Biomedical Research, 2005 Asian Conference on* (pp. 135-139). IEEE.

[108] Luo G., (2016), Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction, Health Information Sciences and Systems.

[109] Shi, L., van Meijgaard, J., & Fielding, J., (2011), Forecasting diabetes prevalence in California: a microsimulation. *Prev Chronic Dis*, *8*(4), A80.

[110] Sanidad, J. G., Bandala, A., Sanidad, B. G., & Marfil, S. S. Prediction of Diabetes on Women using Decision Tree Algorithm.

[111] Parthiban, G., Rajesh, A., & Srivatsa, S. K. (2011). Diagnosis of heart disease for diabetic patients using naive bayes method. *International Journal of Computer Applications*, *24*(3), 7-11.

[112] Polat, K., Güneş, S., & Arslan, A. (2008). A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. *Expert systems with applications*, *34*(1), 482-487.

[113] Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC medical informatics and decision making*, *10*(1), 16.

[114] Suhaimi, N., & Ismail, A. Comparing the Performance of Logistic Regression and Artificial Neural Networks Models: An Application to Type 2 Diabetes Mellitus.

[115] Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20-29.

[116] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.