Improving Diabetes Diagnosis in Smart Health Using Genetic-based Ensemble learning algorithm Approach to IoT Infrastructure

Jafar Abdollahi [1]. Babak Nouri Moghaddam [2]. Mehdi Effat Parvar [3]

1- Department of Computer Engineering, Islamic Azad University, Ardabil Branch, Ardabil, Iran
Email: ja.abdollahi77@gmail.com(Corresponding author)
2- Department of Computer Engineering, Ardabil Branch, Islamic Azad University, Ardabil, Iran
Email: Babak_nouri67@ live.com
2- Department of Computer Engineering, Ardabil Branch, Islamic Azad University, Ardabil, Iran
Email: Me.effatparvar@gmail.com

Received 15 May 2019 ***Revised*** 08 September 2019 ***Accepted*** 19 September 2019

ABSTRACT:
Chronic diabetes mellitus is one of the leading causes of mortality around the world. One of the main causes of this disease is the presence of high metabolites such as glucose. In 2014, there were about 378 million diabetics worldwide, with an estimated burden of $ 13,700 per year. This will more than double by 2030, according to the World Health Organization (WHO) report. Therefore, if diabetes can be predicted based on some variables, the cost of treatment can be significantly reduced by using machine learning and feature selection techniques to help diagnose diabetes early diagnosis of diabetes - that is to say it prevent diabetes progression and its many complications. It gets. In this paper, ensemble learning algorithms combined with hybrid feature selection are used to more accurately diagnose and predict diabetes, through educational data from actual data on Indian diabetes patients published on the University of California website. The results show that the proposed method performs better than the basic methods and accuracy reaches 93%.

KEYWORDS: smart health, machine learning, IoT, ensemble learning, hybrid feature selection.

1. Introduction

Continuous monitoring of patients or children will the many costs for the government and parents that ICT (information and communication technology) with artificial intelligence to reduce the costs of treatment. Diabetes is one of the leading causes of death worldwide; with the availability of vast amounts of medical information demonstrate the need for powerful learning tools to help medical professionals in the diagnosis of diabetes. Machine learning techniques are very helpful in diagnosing diabetes and increasing its efficiency. According to this research, diabetes is a vital issue and at the same time the most common disease in the world [1]. For this reason, diabetes has been considered as a chronic disease and is accounted the one of the most important health difficulties and

also the fifth leading cause of death [2]. Therefore, diabetes is a chronic endocrine disorder that affects the body's metabolism and causes structural changes, so diabetes is one of the most common diseases in the human body that causes the disease. Since 2014; Next, the prevalence of disease has risen from 100 million to 422 million patients [3].The disease is usually divided into type 1 and type 2 diabetes, a type of diabetes that is a worldwide influential disease that is still on the rise and is one of the leading causes of mortality leading to the progression of the disease. It also helps the heart. Therefor; type 2 diabetes is a worldwide influential disease that is still on the rise and the one of the leading causes of mortality and at the same time in the progression of heart disease [4]. Other side effects can mention cataracts, glucose, Retinopathy (ophthalmic vascular disease). Health care in the simplest form of diagnosis and prevention or treatment of any medical damages plays an important role in providing a useful life for the community [5]. Therefore, efforts have been made to reduce the number of chronic disease screening tests to reduce overall costs. A probable solution is to use machine learning techniques in healthcare data that are used to find frequent patterns in a large database to obtain useful information. In addition, the prevalence of type 2 diabetes in adolescents and youngsters is significantly increased. In people with type 2 diabetes, a history of type 2 diabetes, the main risk factors for obesity, family history and lifestyle are examined.

The remainder of the paper is as follows: The second part is the related works that has been done in this area, the third section deals with the organization of IoT in the field of intelligent health, fourth section is for proposed method and finally fifth section is conclusion and corresponding's.

2. RELATED WORKS

In this section, health care is the simplest form of diagnosis and prevention or treatment of any medical damage that has an important role to play in providing life to the community. One of concerns is that how to provide better services with lower costs; machine learning techniques help us to reach this goal. Significant advances in biotechnology and medical sciences has been lead to large datasets, such as genetic data and clinical information which generated from electronic health records (EHRs). To this end, the use of machine learning and data in the biologic sciences now is more than ever necessary and vital in the quest to intelligently convert all available information into valuable knowledge. Researchers in [6] have used machine learning techniques to provide a hybrid intelligent method using principal component analysis (PCA) and Gaussian mixture modeling, classification and regression of Cart trees to predict chronic diseases. Researchers in [7] have provided a new hybrid intelligent system for classifying diabetes disease using machine learning techniques whereas utilizes clustering algorithms and support vector machine (SVM) to cluster patient datasets and classify different types of diseases, respectively.

The PCA approach for dimensionality reduction has been used and the results show that this method of clustering consisting of PCA, SVM, has achieved good classification accuracy. In addition, researchers have compared the accuracy of classification of different data sets using machine learning techniques to predict chronic heart disease using algorithms such as: RBF, SVM, SCRI, NB, MLP, KNN, and decision tree. They also used reinforcement algorithms such as Bagging, Boosting, and Stacking. Also in [9], researchers have developed a semi-autonomous machine learning-based framework for the identification of type 2 diabetes through an electronic health record in which to evaluate and compare the efficacy of techniques to identify many models of machine learning such as: SVM, RF, DT, KNN, NB, and logical regression were used. The experimental results show that the proposed framework can detect people with type 2 diabetes by an average of 98%. Finally, in the [10] researchers have examined a variety of different machine learning techniques and algorithms and tools were used for disease analysis and decision-making processes to analyze machine learning algorithms for the detection of various diseases. Such as heart disease, diabetes, liver, and hepatitis. As a result, machine learning and data mining play an important role in discovering a database that has a iterative process of data cleaning; data integration, data selection and pattern recognition and data recognition. It also has significant presence in the medical field and plays an important role for the detection of chronic diseases such as diabetes, lung cancer, heart disease, kidney failure, kidney stones and liver disorders used.

3. ORGANIZING IOT IN SART HEALTH

The Internet of Things (IoT) is regarded as a global infrastructure for the information society and provides advanced services by connecting (physical and virtual) based on existing and evolving information and communication, as well as providing many different technologies, services and standards, it is believed to be the cornerstone of the IT market for the next ten years. Logically, an IoT system can be used as a set of collaborative, interoperable smart devices to demonstrate a common goal of integrating the IoT with the information world and integrating common services and applications to improve human's lifestyle. Undoubtedly, the communication of creatures (objects or things) in the field of IoT plays an active role in human activities, devices and its processes [11, 12, and 13].

3.1. SMART HEALTH

The promising potential of the emerging IoT for connected medical devices and sensors plays an important role in the next generation of healthcare industry for the high quality patient care. With the increasing number of elderly and disabled people, there is an urgent need for real-time healthcare infrastructure to analyze patient health care data and prevent predictable mortality [14], as health care providers continuously monitor elderly's behavior and health. The expected system should perform tasks such as identifying and preventing accidents and transmitting body parameters such as: heart rate, body temperature, blood pressure, blood glucose level, etc. to the workplace (hospital) using intelligent health. Wearable devices usage has expanded its capabilities in the past decades [15]. The benefits of IoT-based healthcare in patient care can be:

- Ability to monitor patient's health (partially and accurately) at any time
- Monitoring the status of patients in hospitals and nursing homes
- Avoid unnecessary medical expenses and provide appropriate medical support at the right time
- Improving the quality of life for people in need of constant support or supervision

3.2. SMART HEALTH CHALLENGE

While IoT technology has potential advantages in the personal care industry, IoT technologies are still widely used, but along with many challenges in childhood. [16]. IoT or smart health advancements are not without challenges, main IoT challenges in smart health are:

- Security of people's health data using energy
- Lack of standardization in different sensors and protocols
- Accurate diagnosis and prediction of chronic disease outcomes

4. SUGGESTED METHOD

The use of machine learning technology has become increasingly popular in recent years, where computers try to learn or discover a hidden pattern from a training data. In this paper, we use a group learning approach that proposes a reinforcement algorithm to improve the prediction accuracy of diabetes mellitus. Before using this method, we use a hybrid feature selection approach to find features that have more variance in the rapid diagnosis of diabetes, to obtain better prediction accuracy using hybrid feature selection results.

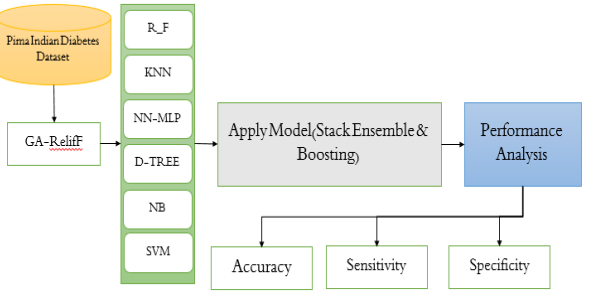

Fig.1.shows our proposed flowchart

Ensemble learning is one of the methods of machine learning that uses meta-learning algorithms to obtain better predictive results in order to achieve better accuracy of each learning algorithm. Many popular machine learning algorithms are actually Ensemble. In this paper, we use Stack Ensemble algorithms consisting of various algorithms that will be described below.In fact a class of algorithms that includes "strong learners" training to find the optimal combination of basic learners. The details of these algorithm are described in Algorithm 1:

Algorithm 1: Stacking Algorithm

Input: training data $D = \{Xi, yi\}^{m}{}_{i=1}$
Ouput: ensemble classifier **H**
Step 1: learn base-level classifiers
for t=1 to T **do**
    learn h_t based on D
end for
Step 2: construct new data set of predictions
for i=1 to m **do**
    D_h = {Xi, yi}, where Xi } = {h1 (Xi), ..., h_T(Xi)}
end for
Step 3: learn a meta-classifier
learn **H** based on D_h
return **H**

Ensemble learning is the process by which several models, such as classifiers, are strategically generated and combined to solve a particular computational information problem. Ensemble learning is mainly used to improve (classify, predict, approximate performance, etc.) the performance of a model, which is why ensemble methods are included in many valid

educational contests, such as Netflix, KDD 2009 and Kaggle.

Hybrid Feature Selection: Feature selection is the process of finding relevant variables for a prediction model. These techniques can be used to identify and remove unnecessary and nonsensical features that reduce the accuracy of the prediction model. Feature selection and sample selection are two important steps in data preprocessing.The former is intended to remove some of the inappropriate or irrelevant features from a dedicated database, and the second is to remove faulty data. Genetic algorithms have been widely used for these tasks in related studies. However, these two tasks have been considered separately in the literature prior to information processing. Therefore, the purpose of this study was to select samples based on genetic algorithms using different priorities to evaluate the classification performance of different datasets.

In this paper we use Genetic Algorithm (GA) to select the optimal features. This method minimizes the potential subset space to obtain a set of features that maximum predicted accuracy and minimizes inappropriate attributes and a multiple correlation in a fitness function used by GA to evaluate the suitability of each attribute domain relative to the domain, we introduce it. Experimental results show that the performance of the proposed method is more effective than the traditional method in all the cases studied. In this work we use Reliff algorithm to rank and use genetic algorithm based on decision tree algorithm to find optimal features. Fig 2 shows the proposed flowchart of the hybrid method based on the ensemblelearning algorithm.

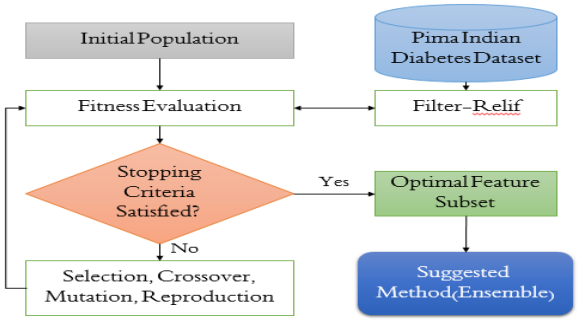

Fig. 2. Proposed flowchart of the hybrid method

- K Nearest Neighbor: This algorithm is used to estimate the data label of the training data and to classify the test data based on the training algorithms [17].
- Decision tree: A type of supervised learning algorithm that is mostly used for classification problems which is used for both independent and dependent variables. In addition, the decision tree is one of the classification methods and classification is also one of the supervised learning methods [18].
- Artificial Neural Network: An artificial neural network, in fact, creates a structure similar to the biological brain of the human brain, so that it can like the human brain have the power to generalize and make decisions that are made up of simple components called neurons that work in parallel. Divided into monolayer and multilayer perceptron neural network.
- Naive Bayesian: A Bayesian theorem-based classification technique assuming independence among predictors is a probability-based classification that predicts the probability of class membership [17].
- AdaBoost: One of the most well-known ensemble algorithms used to solve classification problems is a meta-algorithm that is computed as an average method.
- Random forest: Consists of a set of decision trees that classify a new object based on the characteristics of each tree, which is used together to predict and classify the algorithm, which is suitable for dealing with a very large dimension dataset [19].

As a result, we aim to present a new approach using the hybrid feature selection approach with a combination of decision tree based genetic algorithm and the use of statistical and probability algorithms and improve the basic algorithms using AdaBoost and Boosting methods in the proposed (ensemble learning) method. In this way, we increase the accuracy of diagnosis and prediction of diabetes compared to the basic algorithms.

5. EVALUATION METHODS

In order to evaluate the methods used and to determine how one of the available models can be considered, a model that has the most predictive accuracy compared to the other methods is compared to the categorization methods and to find the appropriate one. And efficiently in this article, we have used cost-benefit analysis (disruption matrix), ROC curve and other model selection issues such as accuracy and so on.

Performance measurement is used to determine the effectiveness of the classification algorithm so that, in the case of two-dimensional classification problems, one can show the cost of classification with a cost matrix for two types of false positive (FP) and false negative (FN) errors and two types of classification into positive true (TN) and negative true (TN) that give different costs and benefits. As shown in Table 1.

Table 1. Confusion Matrix

Confusion Matrix		Classified As:	
		Negative	Positive
Actual Class	Negative	TN	FP
	Positive	FN	TP

Classification accuracy means the percentage of correctly predicted categories and is calculated from the following formula:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Sensitivity: True Positive Rate: If the result is positive for the individual, in some cases the model will also be positive, calculated from the following formula.

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (2)$$

Specificity: True Negative Rate: If the result is negative for the individual, in a few percent of the cases the model will also have a negative result calculated from the following formula.

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \quad (3)$$

TPR and FPR Characteristics (ROC): which draws a graph for evaluating models, each of which has a higher level diagram. Used graphically.

$$TPR = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (4)$$

$$FPR = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \quad (5)$$

PPV: If the model yields a positive result, what is the probability that a person will develop diabetes?

$$PPV = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (6)$$

NPV: If the model is positive, what is the probability that a person will develop diabetes?

$$NPV = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} \quad (7)$$

5.1. SIMULATION

In this section we simulate and evaluate our article. Our simulation environment is implemented in the Jupyter_Notebook module based on Python version 3.6 programming. Also the performance of different classifiers can be compared using one example to simulate classifiers from the Diabetic Patients dataset. Using the UCI site, this dataset contains 9 variables and 768 records. These variables and abbreviations are listed in Table 2.

Table 2. Problem Variables and Abbreviations

P.NO	Number of times Pregnant
PG	Plasma Glucose Concentration
DBP	Diastolic Blood Pressure
TSFT	Triceps Skin Fold Thickness
SI	Two Hour Serum Insulin
BMI	Body Mass Index
DPF	Diabetes Pedigree Function
AGE	Age
C	Class Variable

As a result, the following figures show the simulation results for each of the different algorithms used. We have used a genetic algorithm to extract features that are effective in the rapid diagnosis of diabetes. The simulation results show that among the data set variables, 1-frequency pregnancy, 2-hour glucose concentration, 3-hypertension, 4-insulin level, and 5-age were selected as effective fitters with 86% accuracy. We classify the results based on these same features. The following table shows the parameters selected for the genetic algorithm.

Table 3. Selected Parameters for Genetic Algorithm

Population number	Gen	Mutation rate	Rate of intercourse	Average
40	20	0.1	0.2	86

The following will be described below. We first used the Holdout estimation method to evaluate the basic algorithms, as shown in the simulation results below.

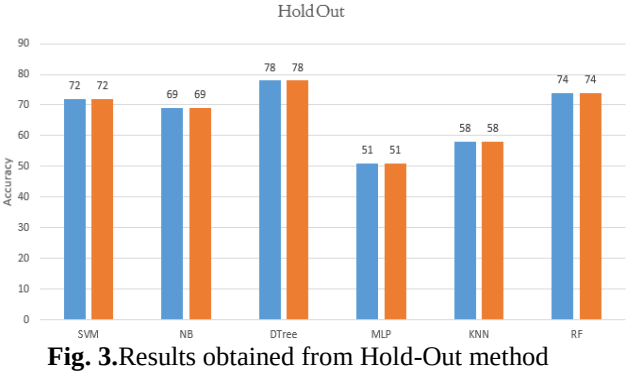

Fig. 3. Results obtained from Hold-Out method

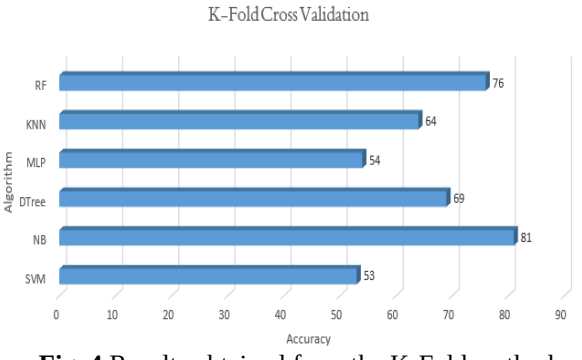

Fig. 4. Results obtained from the K-Fold method

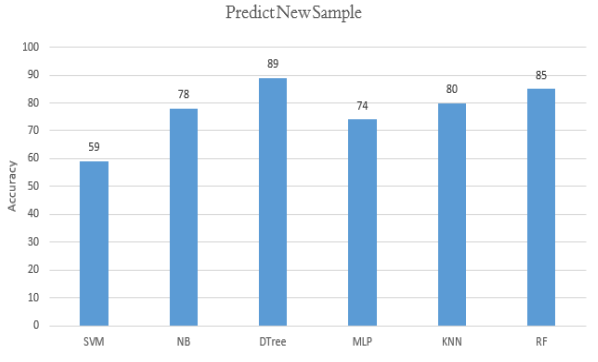

Fig. 5. Results obtained from the prediction accuracy

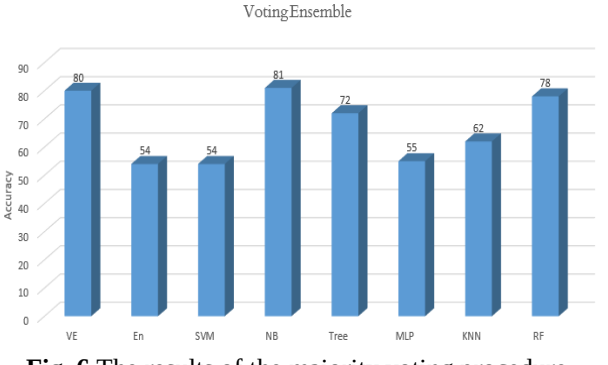

Fig. 6.The results of the majority voting procedure

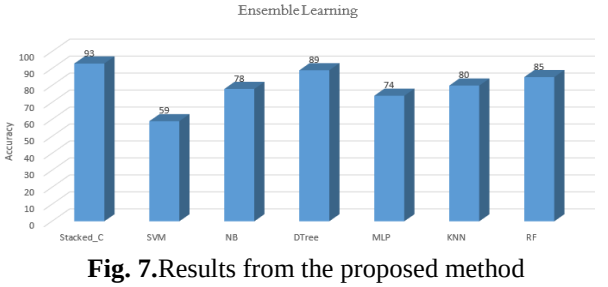

Fig. 7. Results from the proposed method

Finally, performance criteria (clutter matrix) for different methods on diabetic patients are presented in Table 4. The performance criteria of the proposed algorithm are:

Table 4. Performance Criteria for Different Methods

	KNN	SVM	GNB	ANN	RFC	TREE
Accuracy	0.73	0.73	0.74	0.70	85	89
Sensitivity	0.65	0.70	0.83	0.76	85	89
Specificity	0.55	0.75	0.56	0.55	85	89
PPV	0.65	0.87	0.78	0.8	85	89
NPV	0.65	0.55	0.64	1.5	85	89
TPR	0.76	0.70	0.83	0.76	0.99	89
FPR	0.44	0.29	0.43	0.44	0. 25	0.42

6. CONCLUSION AND FUTURE WORK

In this study, we have analyzed open health issues with various solutions and we have seen in many cases that smart health can help improve health care by continuously monitoring the elderly and paving the way for smart health. Finally, our proposed approach, after evaluating and comparing with the basic algorithms, it has been shown to be advanced compared to the above methods and the accuracy achieved with this model is 93%. Table 5 shows the best modeling results with different classification techniques.

Table 5. Comparison of different techniques for diagnosis of chronic disease.

The method used	Target	accuracy and validity	The results
ANFIS and ANFIS synthetic inference system	Prediction and classification of diabetes	Best accreditation performance of %90.190 for fuzzy inference and %90.32 accuracy for neural network.	ANFIS shows greater accuracy with less error than ANN.
Development of a semi-autonomous machine learning based framework such as LR, SVM, RF, DT, NB, KNN	Identify people with T2DM based on HER	The average AUC is %0.98, which is significantly higher than the AUC of .0.71	The proposed approach has an accurate and efficient way of identifying T2DM sufferers using EHR.

Using machine learning techniques such as RBF, SVM, SCRI, NB, MLP, KNN as well as using reinforcement algorithms such as Stacking, Boosting, Bagging	Comparison of classification accuracy of heart disease data	SVM with an accuracy of %84.15 and SCRL with an accuracy of %69.96also DT with an accuracy of %78.54in addition to combining SVM with MLP with an accuracy of .%84.15	SVM approach using Boosting technique is more advanced than other methods with accuracy of .%76.8

Therefore, based on the results of this simulation, it can be decided that our proposed approach is better than other methods in terms of evaluating the results. And since the basic algorithms alone were not effective in detecting and predicting diseases, this paper compared Stack Ensemble algorithms with a combination of hybrid feature selection techniques to improve prediction accuracy and finally compared with basic algorithms. Our proposed algorithm is less error-free than the basic methods and has reached 93% accuracy. It is also recommended due to the high performance of the proposed method in more accurate diagnosis of diabetes outcomes as well as the high acceptability of noise data for future work. The proposed method is used to diagnose various chronic diseases and, considering the importance of diagnosing the disease, we intend to implement Real-time chronic disease diagnosis system and recommendation systems and short message systems in addition to the proposed method.

REFERENCES

[1] S. Sendra, L. Parra, J, Lloret, and J. Tomás,**"Smart system for children's chronic illness monitoring."** *Information Fusion, 40*, pp.76-86, 2018.

[2] J. Amin., M. Sharif, M. Yasmin, H. Ali, and S. L. Fernandes, "**A method for the detection and classification of diabetic retinopathy using structural predictors of bright lesions**." *Journal of Computational Science*, 19, pp.153-164, 2017.

[3] T . Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You,... and Y. Chen, "**A machine learning-based framework to identify type 2 diabetes through electronic health records".** *International journal of medi*cal *informatics, 97*, pp.120-127, 2017.

[4] S. K. Somasundaram, and P. Alli,"**A Machine Learning Ensemble Classifier for Early Prediction of Diabetic Retinopathy."** *Journal of Medical Systems, 41(12)*, pp.201, 2017.

[5] M. Nilashi, O. bin Ibrahim, H. Ahmadi, and L. Shahmoradi,**"An Analytical Method for Diseases Prediction Using Machine Learning Techniques."** *Computers & Chemical Engineering, 116*, pp.212-23, 2017.

[6] M. Nilashi, O. Bin Ibrahim, A. Mardani, A. Ahani, and A. Jusoh,**"A soft computing approach for diabetes disease classification."** *Health Informatics Journal*, 24(4), pp. 379-393, 2018.

[7] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, **"A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease".** *In Computers and Communications (ISCC), IEEE Symposium on*, pp. 204-207, 2017.

[8] T. Zheng, W. Xie, L., Xu, X. He, Y. Zhang, M. You, ...and Y. Chen, **"A machine learning-based framework to identify type 2 diabetes through electronic health records"**. *International journal of medical informatics, 97*, pp. 120-127, 2017.

[9] M. Fatima, and M. Pasha, **"Survey of Machine Learning Algorithms for Disease Diagnostic"**. *Journal of Intelligent Learning Systems and Applications, 9(01)*, pp. 1, 2017.

[10] F. K. Shaikh, S. Zeadally, and E. Exposito,**"Enabling technologies for green internet of things".** *IEEE Systems Journal, 11(2)*,pp. 983-994, 2017.

[11] S. Sicari, A. Rizzardi, L. A. Grieco, G. Piro, and A. Coen-Porisini,**"A policy enforcement framework for Internet of Things applications in the smart health"**. *Smart Health, 3*, pp. 39-74, 2017.

[12] A. R. Sfar, E. Natalizio, Y. Challal, and Z. Chtourou, "**A roadmap for security challenges in the Internet of Things."** *Digital Communications and Networks 4(2)*, pp. 118-137, 2018.

[13] M. S. Hossain, and G. Muhammad**, "Cloud-assisted industrial internet of things (iiot)–enabled framework for health monitoring"**. *Computer Networks, 101*, pp. 192-202, 2016.

[14] A. Temko,**"Accurate wearable heart rate monitoring during physical exercises using PPG"**. *IEEE Transactions on Biomedical Engineering, 64(9)*, pp. 2016-2024, 2017.

[15] Y. Chen, W. Shen, H. Huo, and Y. Xu,**" A smart gateway for health care system using wireless sensor**

network". *In 2010 Fourth International Conference on Sensor Technologies and Applications*, pp. 545-550. IEEE, 2010.

[16] K. G. Nisha, and K. Sreekumar, **"A review and analysis of machine learning and statistical approaches for prediction".** *International Conference on In Inventive Communication and Computational Technologies (ICICCT)*, IEEE, pp. 135-139, 2017.

[17] C. Voyant, G. Notton, S. Kalogirou, M.L. Nivet, C. Paoli, F. Motte, and A. Fouilloy, **" Machine learning methods for solar radiation forecasting: A review"**. *Renewable Energy, 105*, pp.569-582, 2017.

[18] J. Abellán, C.J. Mantas, and J.G. Castellano, "**A Random *Forest approach using imprecise probabilities".*** *Knowledge-Based Systems, 134*, pp.72-84, 2017.