



# A decision support system for diagnosis of diabetes using Logistic regression model

G.U Chukwuemeka<sup>1</sup>, V.I.E Anireh<sup>2</sup>, E.O Bennett<sup>3</sup>

Department of Computer Science, Rivers State University, port Harcourt, Nigeria<sup>1,2,3</sup>

**Abstract:** The Healthcare industry contains very large and sensitive data and needs to be handled very carefully. Diabetes is one of the growing, fatal diseases all over the world. Medical professionals want a reliable prediction system to diagnose Diabetes. Different machine learning techniques are useful for examining the data from diverse perspectives and synopsisizing it into valuable information. The accessibility and availability of huge amounts of data will be able to provide us with useful knowledge if certain data mining techniques are applied to it. The main goal is developing a decision support system for the diagnosis of diabetes using a machine learning model. Diabetes contributes to heart disease, kidney disease, nerve damage and blindness. So, efficiently mining the diabetes data is a crucial concern. The data mining techniques and methods will be discovered to find the appropriate approaches and techniques for efficient classification of the Diabetes dataset and in extracting valuable patterns. The RStudio software was employed for diagnosing diabetes. The Rivers State University Teaching Hospital dataset was acquired from the Health Database of the Hospital used for analysis. The dataset was studied and analyzed to build an effective model that predicts and diagnoses diabetes disease. In this study, we aim to apply the bootstrapping resampling technique to enhance the accuracy and the applying Logistic regression model.

**Keywords:** Logistic Regression Classification, Healthcare, Diabetes.

## 1.INTRODUCTION

Emerging realities in the world today have necessitated an intensified research in technology and computing, seeking ways to explore the usefulness of computers and information technology devices concerning the deluge of data/information being exponentially generated daily. It has found its way into academics, industry, and other social activities. Again, there is an increasing transformation in the medical field attributed to the recent advances in emerging computing technologies (machine learning and soft computing) and communications which have evolved into a profitable endeavor, leading towards a more computer-enabled medical industry. The major factor responsible for transforming theory to reality in this regard is the continuous development of basic hardware and software (armed with soft computing) components enabling the computer and associated computing devices to make use of high powered processors to sense the physical environment, exploit new forms of wireless communications and even assist medical personnel in making medical decisions. Computers have been used widely in the medical sector in recent times, from a local and global patient, medicine databases to emergency networks, and digital archives. Meanwhile, in the case of medical diagnosis, due to the complexity of the task, it has not been realistic to expect a fully automatic, computer-based, medical diagnosis system. However, recent advances in the field of intelligent systems are materializing into a wider usage of computers, armed with Artificial Intelligence (AI) and soft computing techniques. It is therefore imperative to have decision support to assist in diagnostic decision-making. A decision support system in this context is a computer-based information system that supports medical staff in diagnostic decision-making. A properly designed medical decision support systems are interactive software whose intent is to help medical practitioners semantically sieve through a deluge of raw data to identify and solve medical problems. The major task of medical science is to prevent and diagnose diseases.

Here our focus is the second task, which as mentioned before, is not a direct and simple task at all. Brause highlighted that almost all physicians are confronted during their formative years by the task of learning to diagnose. Central to good diagnosis is the ability of an experienced physician to know what symptoms or vitals to throw away and what to keep in the diagnostic process as stated in a report by [1]. The death of medical experts in the developing world and to be specific Nigeria has made a large percentage of its populace die due to preventable ailments since the few medical experts always opt to practice in the urban cities and as such preventable ailments are not timely diagnosed in rural areas thus leading to other complications and consequent deaths. Confusability of disease diagnosis has constituted a major threat not only to the existence of human but also the effectiveness and productivity in most developing countries due to inaccurate and untimely diagnosis procedures used by physicians in the region. Poor diagnosis in this regard has continuously affected the management of these diseases whose symptoms are in most cases confusing and overlapping. Diagnosis is the first stage towards a set of therapeutic action; - a mistake at this stage is disastrous and may hinder the effective control of these diseases. Each method has its merit and weakness, a review given by shows that a predictive model constructed by a probabilistic neural network had a significantly high classification accuracy for breast cancer early detection. However,



only high accuracy is not enough, using neural networks along with logistic regression for breast cancer diagnosis guarantees the sensitivity and specificity of predictive models. For cancer, just a simple diagnosis is not enough. More importantly, it is worth taking care of the prognostic prediction after the malignant lump has been surgically excised. To be more specific, predicting the outcome of recovery by a particular treatment plan is helpful for further clinical research. Furthermore, the prediction of whether a patient will recur at a specific time is also important. Prognosis helps in determining the case for whether a patient is recurring or not as well as the case for the time to recur. The study summarizes that breast cancer prognosis is mainly analysed under ANNs, this method provides an efficient way to classify patients and predict the survival time with high accuracy. Last but not least, there is a beneficial step in the data mining procedure called data pre-processing. In the cancer research area, the preparation of data is particularly important, as a suitable dataset after cleaning up and transforming irrelevant or invalid data helps in modelling a more sensitive predictor of a cancer diagnosis.

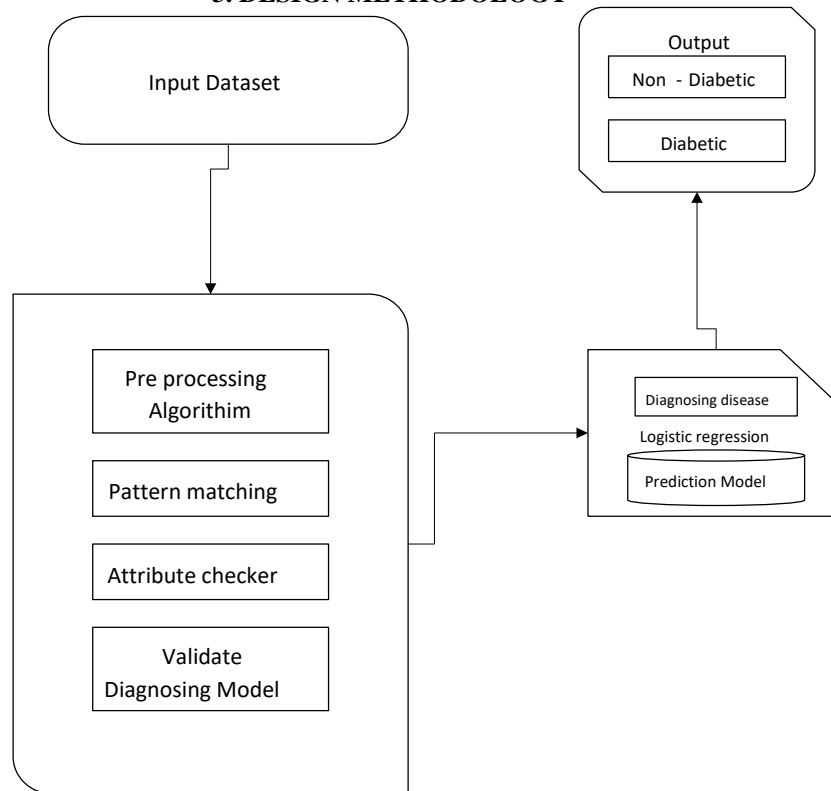
## 2. RELATED WORKS

The paper “Survey of Machine Learning Algorithms for Disease Diagnostics by [2] stated that in medical imaging, Computer-Aided Diagnosis (CAD) is a rapidly growing dynamic area of research. In recent years, significant attempts are made for the enhancement of computer-aided diagnosis applications because errors in medical diagnostic systems can result in seriously misleading medical treatments. Machine learning is important in Computer-Aided Diagnosis. After using an easy equation, objects such as organs may not be indicated accurately. So, pattern recognition fundamentally involves learning from examples. In the field of bio-medical, pattern recognition and machine learning promise the improved accuracy of perception and diagnosis of disease. They also promote the objectivity of the decision-making process. For the analysis of high-dimensional and multimodal biomedical data, machine learning offers a worthy approach for making classy and automatic algorithms.

The paper “A Method for Classification Using Machine Learning Technique for Diabetes” by [3] diseases occur when the production of insulin is insufficient or there is improper use of insulin. The data set used in this work is the Pima Indian diabetes data set. Various tests were performed using the WEKA data mining tool. In this data-set percentage split (70:30) predict better than cross-validation. J48 shows 74.87% and 76.96% accuracy by using CrossValidation and Percentage Split Respectively. Naive Bayes presents 79.57% correctness by using PS. Algorithms show the highest accuracy by utilizing the percentage split test. Metalearning algorithms for diabetes disease diagnosis was first introduced by Sen and Dash (2015). The employed data set is Pima Indians diabetes which is received from the UCI Machine Learning laboratory.

The paper “Application of fuzzy logic and genetic algorithm in heart disease risk level prediction” by [4] have suggested the work on Naive Bayes to predict diabetes Type-2. Diabetes disease has 3 types. The first type is Type-1 diabetes, Type-2 diabetes is the second type and the third type is gestational diabetes. Type-2 diabetes comes from the growth of Insulin resistance. The data set consists of 415 cases and for purpose of variety; data are gathered from dissimilar sectors of society in India. MATLAB with SQL server is used for the development of the model. 95% correct prediction is achieved by Naive Bayes. Yasodha et al (2018) made use of classification techniques on diverse types of datasets that can be accomplished to decide if a person is diabetic or not. The diabetic patient’s datasets are established by gathering data from the hospital warehouse which contains two hundred and forty-nine instances with seven attributes. These instances of this dataset are referring to two groups i.e. blood tests and urine tests. In this study, the implementation can be done by using WEKA to classify the data and the data is assessed utilizing a 10-fold cross-validation approach, as it performs very well on small datasets, and the outcomes are compared. The naïve Bayes, J48, REP Tree and Random Tree are used. It was concluded that J48 works best showing an accuracy of 60.2% among others. [3] work was aimed at how to discover solutions to detect diabetes by investigating and examining the patterns originate in the data via classification analysis by using Decision Tree and Naïve Bayes algorithms. The research hopes to propose a faster and more efficient method of identifying the disease that will help in a well-timed cure for the patients. Using the PIMA dataset and cross-validation approach the study concluded that the J48 algorithm gives an accuracy rate of 74.8% while the naïve Bayes gives an accuracy of 79.5% by using a 70:30 split.

## 3. DESIGN METHODOLOGY



**Figure 1: Architecture of the Proposed System design**

In principle, there are steps/components for diagnosing any disease using machine learning Algorithms. This proposed system is divided into three components which are as follow: i. Data Collection

- ii. Data Pre-processing
- iii. Diagnosing of disease

In this work, we concentrate on both the pre-processing and classification part as a proof-of concept methodology for a diabetes diagnosis. Therefore, to classify the diabetic or nondiabetic subjects, R is used as a tool in building and diagnosing our model.

For the diagnosis of diabetes, we have used a non-parametric method as it does not estimate the parameters in an assumed manner such as the linear form encountered in linear regression. Logistic regression algorithm, sometimes called the logit model, is a common model for dichotomous output variables and was extended for disease classification prediction<sup>12</sup>. Suppose that there are  $p$  input variables where their values are indicated by  $x_1, x_2, \dots, x_n$ . Let  $z$  be a probability that an event will occur and  $1-z$  be a probability that the event will not occur. The logistic regression model is given by  $\text{logit}(Y) = \text{natural log(odds)} = \ln(\pi) = \alpha + \beta X$  (1.0)  $1-\pi$

Taking the antilog of Equation 1 on both sides, one derives an equation to predict the probability of the occurrence of the outcome of interest as follows:

$$\pi = \text{Probability}(Y = \text{outcome of interest} | X = ea + \beta x$$

$x$ , a specific value of  $X$ ) =  $1 + e^{\alpha + \beta x}$  (1.1) where  $\pi$  is the probability of the outcome of interest or “event,” such as a child’s referral for remedial reading classes,  $\alpha$  is the Y-intercept,  $\beta$  is the regression coefficient, and  $e = 2.71828$  is the base of the system of natural logarithms.  $X$  can be categorical or continuous, but  $Y$  is always categorical. According to Equation 1, the relationship between  $\text{logit}(Y)$  and  $X$  is linear. Yet, according to Equation 2, the relationship between the probability of  $Y$  and  $X$  is nonlinear. For this reason, the natural log transformation of the odds in Equation 1 is necessary to make the relationship between a categorical outcome variable and its predictor(s) linear. The value of the coefficient  $\beta$  determines the direction of the relationship between  $X$  and the logit of  $Y$ . When  $\beta$  is greater than zero, larger (or smaller)  $X$  values are associated with larger (or smaller) logits of  $Y$ . Conversely, if  $\beta$  is less than zero, larger (or smaller)  $X$  values are associated with smaller (or larger) logits of  $Y$ . Within the framework of inferential statistics, the null hypothesis states that  $\beta$  equals zero, or there is no linear relationship in the population. Rejecting such a null hypothesis implies that a linear relationship exists between  $X$  and the logit of  $Y$ . If



a predictor is binary, then the odds ratio is equal to  $e$ , the natural logarithm base, raised to the exponent of the slope  $\beta$  ( $e^\beta$ ). Extending the logic of the simple logistic regression to multiple predictors as follows:

$$\text{logit}(Y) = \ln \left( \frac{\pi}{1-\pi} \right) = \alpha + \beta_1 X_1 + \beta_2 X_2. \text{ Therefore, } \pi =$$

$$\text{Probability} ( Y = \text{outcome of interest} | X_1 = x_1, X_2 = x_2 =$$

$$\frac{e^{\alpha + \beta_1 X_1 + \beta_2 X_2}}{1 + e^{\alpha + \beta_1 X_1 + \beta_2 X_2}} \quad (1.2)$$

#### 4. EXPERIMENT

The system uses R Studio framework in the implementation of the Logistic Regression Model Algorithm in classifying diabetes patient's data sets. The method applied in this dissertation is the Object-Oriented System Analysis and Design method (OOAD) where an existing system is studied from the perspective of objects and similar objects are grouped as classes and their properties are handled as fields while their behaviours are treated as the actions or methods within the same bundle of an object. The choice of this methodology is clear since it is a method developed in Software Engineering during the last decades to develop computational models of reality and it is a type of tool needed when one deal with the development of complex computational applications like biological systems. This software development methodology is made up of three aspects, which are Object-Oriented Analysis (OOA) that deals with the design requirements and overall architecture of a system which is focused on describing what the system should do as regards to key objects in the problem domain; the second one is the object-oriented design (OOD) that translates the system architecture interfaces and classes, and the third one is the object-oriented programming (OOP) that implements these programming constructs. First, the use of this methodology helps us to exploit the expressive power of object-based and object-oriented programming languages. As [5] points out, "It is not always clear how best to take advantage of languages such as C++, Java, Python, and R". Significant improvements in productivity and code quality have consistently been achieved using JAVA, R and Python with a bit of data abstraction thrown in where it is useful. However, further and noticeably larger improvements have been achieved by taking advantage of class hierarchies in the design process. This is often called object-oriented design and this is where the greatest benefits of using Java, Python and R have been found. The experience has been that, without the application of the elements of the object model, the more powerful features of languages such as Smalltalk, C++, Java, Python and so forth are either ignored or greatly miss used. Second, the use of the object model encourages the reuse not only of software but of entire designs, leading to the creation of reusable application frameworks. We have found that object-oriented systems are often smaller than equivalent non-object-oriented implementations. Not only does this mean less code to write and maintain, but greater reuse of software also translates into cost and schedule benefits. However, reuse does not just happen. If reuse is not the primary goal of your project, it is unlikely that it will be achieved. Plus, designing for reuse may cost you more when initially implementing the reusable component. The good news is that the initial cost will be recovered in the subsequent uses of that component. Third, the use of this methodology produces systems that are built on stable intermediate forms, which are more resilient to change. This also means that such systems can be allowed to evolve, rather than be abandoned or completely redesigned in response to the first major change in requirements.

**Table 1: Diabetes dataset sample**

patient id	Gender	Age	Pedigree function	Diagnostics criteria	Pregnant	Glucose	Blood pressure	Skin thickness	Insulin	BMI	Out come
1	Male	79	0.127	FBG	0	172	73	17	59	33.6	1
2	female	67	0.233	FBG	3	61	77	15	267	38.2	1

This dataset is created using the RSUTH data collected and modelled after PIMA Indian Dataset parameters. The data set which was gotten from RSUTH consists of two thousand (2000) rows of observations from respective patients and fourteen (14) variables. The variables consist of thirteen independent variables and one dependent variable.

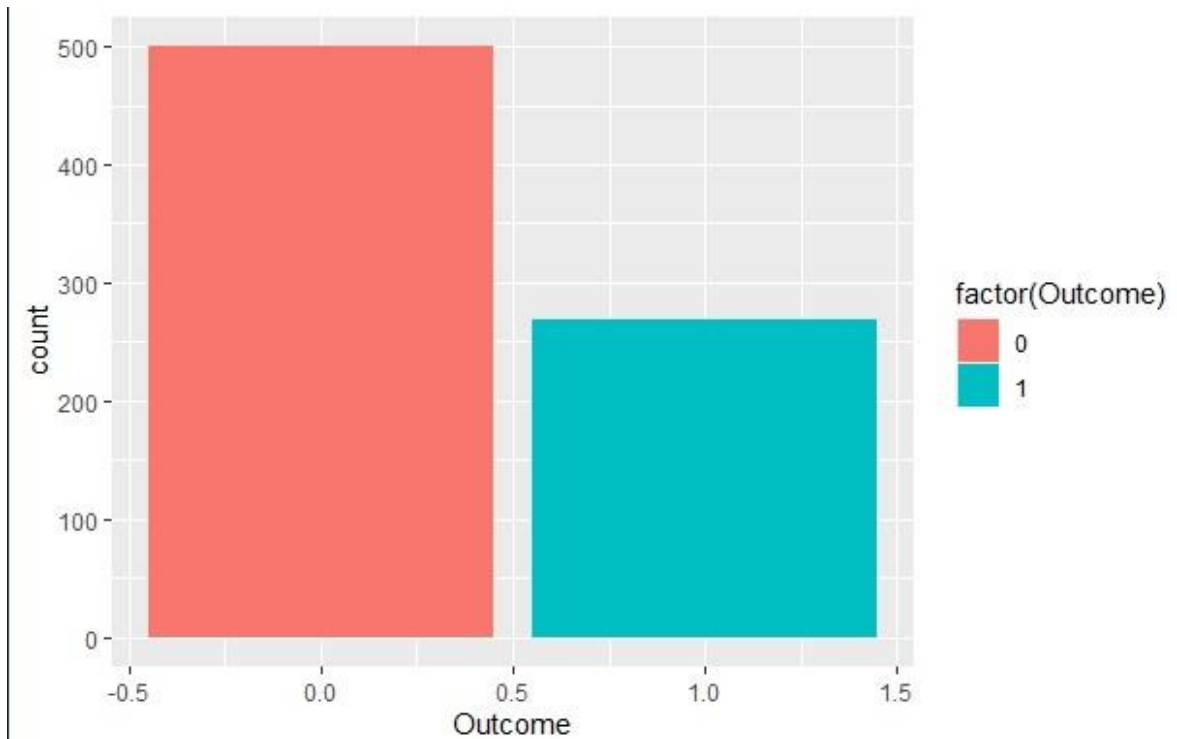


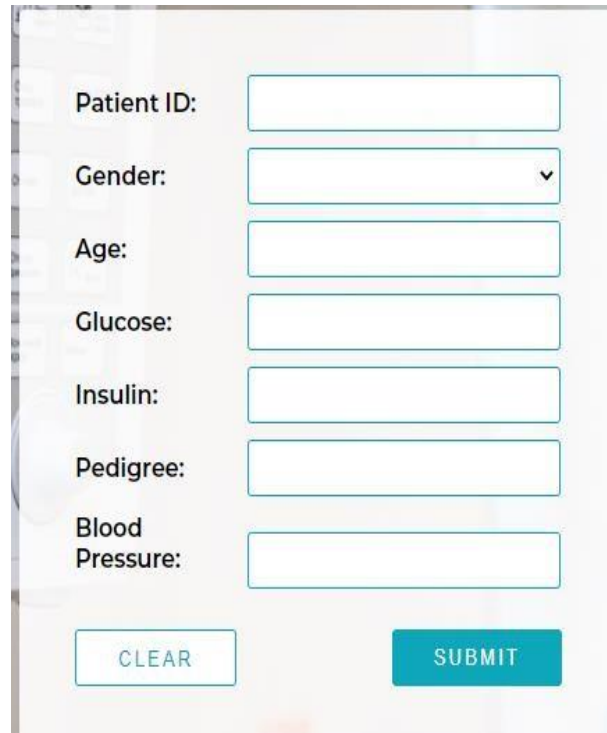
Fig. 2: Graph showing the number of people with diabetes (0) and those without (1)

5. RESULT

The diagram below shows an accuracy of 86.7%, After a successful training and confusion matrix test and Random Forest show an accuracy of 84.5%, while the Decision Tree show an accuracy of 78.3%. it clearly shows that the Logistic Regression which is also a classifier is the right way to go when looking for a concise result in diagnosis. With little parameters, you can get an accurate prediction.

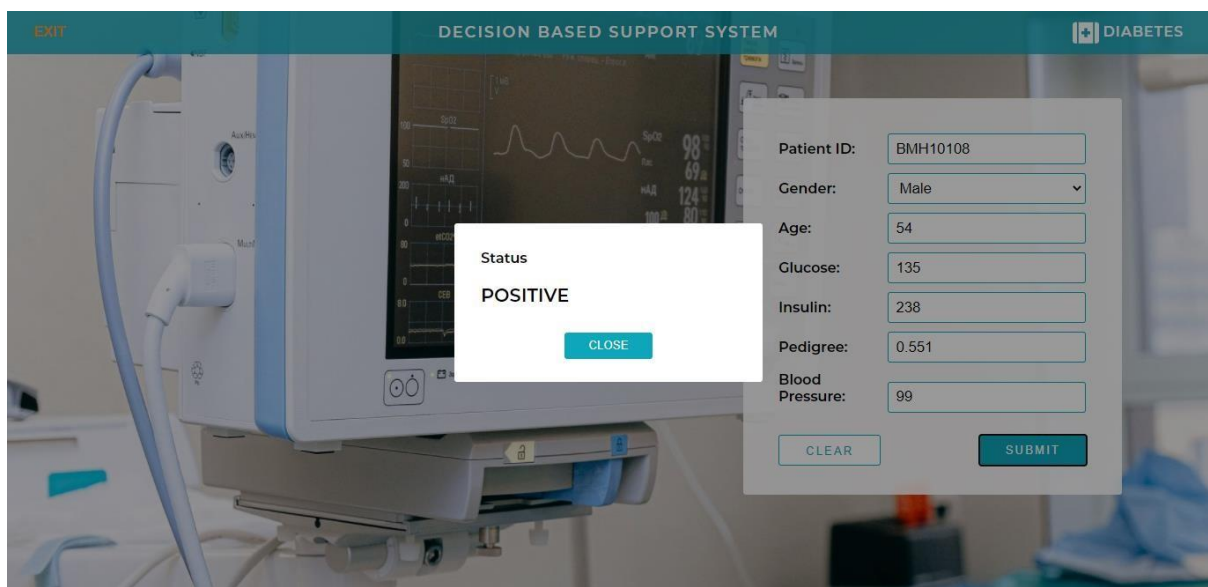
Table 4.4: Comprehensive view of machine model precision on predicting medical diseases

Author(s)	Years	Disease	Machine Learning Techniques	Tools	Precision
Yasodha et al.,	2014	Cancer	KNN	WEKA	0.3625
Lees et al.,	2015	Diabetes	Decision Tree	R	0.7837
Aishwarya et al.,	2013	Diabetes	Random Forest	R	0.8452
Proposed System	2020	Diabetes	Logistic Regression	R	0.8676



A screenshot of a patient information form. The form contains the following fields: Patient ID (text input), Gender (dropdown menu), Age (text input), Glucose (text input), Insulin (text input), Pedigree (text input), and Blood Pressure (text input). At the bottom, there are two buttons: 'CLEAR' and 'SUBMIT'.

**Fig. 3 Patient Form**



A screenshot of a 'DECISION BASED SUPPORT SYSTEM' interface. The background shows a medical monitor with vital signs. A modal window is open, displaying a 'Status' prediction of 'POSITIVE'. To the right, a form is filled with patient data: Patient ID: BMH10108, Gender: Male, Age: 54, Glucose: 135, Insulin: 238, Pedigree: 0.551, and Blood Pressure: 99. The form has 'CLEAR' and 'SUBMIT' buttons.

**Fig. 4: Prediction of a patient status**

### 6. CONCLUSION & RECOMMENDATIONS

A developing country like Nigeria that is battling with developments needs information technology to expand both in length and breadth in every sector. This paper aims to illuminate the trends in the computing industry that suggest that they have the potential to impact the administrative activities and service delivery in the medical industry which would ultimately have a positive impact on both patients and medical personnel. The future is more computer enabled, more connected and more reliant on computing devices and technologies which is expected as a norm. It is expected that medical diagnosis will move more and more outside the hospitals into the patient environment for real and virtual mediated by computing devices and technologies. We recommend that every medical personnel should endeavour to have the system in place of diagnosing diabetes which would help in boosting the quality of their diagnosis or medical



decisions. It is our sincerest recommendation that other diseases such as Cancer, Aids etc. can borrow from the methodology of our proposed system. The system is not replacing medical personnel rather it complements them

### REFERENCES

- [1] O. A. P. Ejiata, "An improved framework for diagnosing confusable diseases using neutrosophic based neural network.," *Neutrosophy Sets and Systems*, Port Harcourt, June 1, 2017.
- [2] M. P. Meherwar Fatima, "Survey of Machine Learning Algorithms for Disease Diagnostics," *Journal of Intelligent Learning Systems and Applications*, p. 16, 24 January 2017.
- [3] G. P. J. Aishwarya. R, "A Method for Classification Using Machine Learning Technique fro Diabetes," *International Journal of Engineering and Technology (IJET)*, p. 6, Jun/Jul 2013.
- [4] P.Sharma and K. Saxena, "Application of fuzzy logic andgenetic algorithm in heart disease risk level prediction," *International Journal of System Assurance Engineering and Management*, 2017.
- [5] B. Stroustrup, in *Programming Priciples and Practices Using C++*, Indiana, Pearson, 2008, p. 1227.
- [6] S. J. R. S. Aiswarya Iyer, "Diagnosis of Diabetes Using Classification Mining Techniques," *International Journal of Data Mining & Knowledge Management Process (IJDMP)*, vol. 5, p. 15, January 2015.
- [7] H. P. a. J. M. R. Miller, "an Experimental Computer-Based Diagnostic Consultant for General Internal Medicine," *New England Journal of Medicine*, vol. 307, pp. 468-476, 1982.
- [8] W. v. Melle, "MYCIN: a knowledge-based consultation program for infectious disease diagnosis," *International Journal of Man-Machine Studies*, vol. 10, pp. 313-322, 1978.
- [9] F. J. F. W. J. Z. W. L. a. S. C. X. Wei, "An Ensemble Model for Diabetes Diagnosis in Largescale and Imbalanced Dataset," in *CF17*, 2017.
- [10] P. K. B. K. D. & K. S. V. (. Groves, "The 'big data' revolution in healthcare – Accelerating value and innovation (Rep.)," *Pharmaceutical Research and Manufacturers of America*, 2015.