International Journal of Scientific Engineering and Science
Volume 5, Issue 11, pp. 19-32, 2021. ISSN (Online): 2456-7361

An Improved Method for Predicting Diabetes Mellitus Using Adaptive Neuro-Fuzzy Inference System

Kenneth A. Akpado[1], Ugochukwu J. Njonu[2*], P. C. Obioma[3], Anthony N. Isizoh[4]
[1, 2, 3, 4]Department of Electronic and Computer Engineering, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria
Email address: njonuugochukwu(at)gmail.com[2*]

***Abstract*— *Chronic diseases are considered the major cause of death and disability worldwide. Diabetes is a chronic disease that occurs when the pancreas cannot produce enough insulin or when the body does not use the insulin effectively. According to World Health Organization (WHO), in the year 2019 alone, diabetes was the direct cause of about 1.5 million deaths. Since diabetes mellitus type 2 has become one of the major causes of premature diseases such as heart disease and kidney disease leading to death in many countries, it is important that an expert system be implemented and used in the diagnosis of this condition. Although several systems have been proposed and designed to diagnose diabetes mellitus type 2, the accuracy of different data mining and machine learning techniques is still not very high. Also, in cases where the accuracy of the prediction was high, it was discovered that very few input metrics were considered, which can actually be affected in a real life scenario. In this paper, a fuzzy logic based expert system model for diagnosing diabetes mellitus type 2 was developed. The developed model was evaluated alongside a similar fuzzy expert system. Several experiments were carried out to analyze the performance of the two models. Results showed that the model needed about 25 iterations to attain the global minimum, while the developed model needed 15 iterations, thus consuming less computation resources. Results also showed that the developed model outperformed the other model since it employed an augmented dataset. When tested with test dataset from the locally generated dataset, the developed fuzzy expert gave a prediction accuracy of 97%, with a specificity of 95%, a sensitivity of 94%, and a precision of 93% when compared to the other system that had a corresponding accuracy of 89%, specificity of 86%, sensitivity of 87% and a precision of 80%. This helps to establish the fact that there is a need to incorporate datasets that are local or unique to a group of persons or region so as to improve the accuracy of the developed model.*

***Keywords*— *Adaptive Neural Fuzzy Inference System (ANFIS), Diabetes Mellitus, Membership function, MATLAB, Expert System, Datasets*

I. INTRODUCTION

Today, the world is fast-paced, and with this trend came the surge of eating less healthy foods, fast-food products, reduced physical activities, and reduced resting periods. This change in lifestyle and habits opened the door to chronic and deadly disease developments.

Chronic diseases are considered the major cause of death and disability worldwide. Diabetes is a chronic disease that occurs when the pancreas cannot produce enough insulin or when the body does not use the insulin effectively. In the medical dictionary, diabetes mellitus is defined as a chronic disease associated with abnormally high levels of the sugar glucose in the blood (Shiel, 2017), (Gizem Koca, 2020). It is a major cause of heart attacks, kidney failure, blindness, lower limb amputation and strokes. In 2014, the number of people diagnosed with diabetes rose from 108 million in 1980to 422million. Between year 2000 and 2016, there was a 5% increase in premature mortality from diabetes. Research showed that in 2019, about 463 million people, who are between 20 years old and 79 years old, had diabetes mellitus (International Diabetes Federation, 2019), (Gizem Koca, 2020). According to World Health Organization (WHO), in the year 2019 alone, diabetes was the direct cause of about 1.5 million deaths (World Health Organization, 2021). The World Health Organization estimates that diabetes will be the 7th leading cause of death in 2030 (C. D. Mathers and D. Loncar, 2006). In addition to this, more than 80% of diabetes-related deaths occur in low and middle-income countries (World Health Organization, 2016), (Gizem Koca, 2020). There are three types of diabetes, namely type 1 diabetes, type 2 diabetes, and gestational diabetes. Type 2 diabetes which is the most common form of diabetes, accounts for about 90% of diabetes cases. It is a long-term metabolic disorder that is characterized by high blood glucose and insulin resistance. In addition, it results from the body's ineffective use of insulin. There are two main causes of type 2 diabetes, namely an increase in body weight and a lack of physical activity (World Health Organization, 1999), (National Institute of Diabetes and Digestive and Kidney Diseases, 2014). The Rates of this type of diabetes have increased considerably since 1960 in conjunction with increasing rates of obesity (M. Truglio-Londrigan and S. B. Lewenson, 2012). The number of type 2 diabetic patients increased from approximately 30 million in 1985 to around 368 million in 2013 (S. Smyth and A. Heron, 2006), (T. Vos et al, 2015). Until recently, type 2 diabetes was seen only in adults, but is now becoming increasingly common in young people (World Health Organization, 1999).

Conventionally, experts depend on blood glucose level and several other factors to diagnose and detect different types of diabetes mellitus. The success of this method often depends on the medical practitioners' or doctors' expertise level, experience, and perception. Recently, some partial automatic systems, which can be replaced by the conventional methodologies for time and cost point of view, have been developed to diagnose diabetes mellitus (Gizem Koca, 2020).

The shortfall of this type of system is that it still needs the input of the medical practitioners so as to verify results. Advances in research has seen the development of fully automated expert systems that makes use of Machine Learning Techniques, such as Artificial Neural Network, K-Means Clustering Algorithm, Support Vector Machine (SVM) Technique, Data Mining Methods Decision Trees, Decision Support Systems, and Adaptive Neuro-Fuzzy Inference Systems (ANFIS). In this thesis, an artificial intelligence based prediction system for diabetes mellitus is designed.

II. Review of Related Works

Several research works have been done in this research area, in this section, a review of these works is done.

Mitushi Soni, (2020) employed machine learning classification techniques, such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting (GB) and Random Forest (RF) for the prediction of diabetes mellitus. The aim was to determine which classification method gives a better result. From the experiment, it was discovered that RF achieved a higher accuracy compared to the other techniques.

Adeli and Neshat, (2015) proposed an expert system for heart disease diagnosis using fuzzy logic. The Mamdani inference technique was used to build this system. In addition, the Heart Disease Data Set of the V.A. Medical Centre, Long Beach, and Cleveland Clinic Foundation database were used to implement the system. This data set included 13 attributes and 303 instances. However, this study used 11 out of 13 attributes of the original data set. Input attributes included age, sex, chest pain type, cholesterol level, resting electrocardiography, blood sugar, blood pressure, maximum heart rate, old peak, exercise, and thallium scan. The output relates to the presence of heart disease in the patient. There were five fuzzy sets of the output that indicate the exact stage of the heart disease development process: healthy, mild, moderate, severe, and very severe.

M. Kalpana and A. Kumar, (2012) developed a fuzzy expert system for the diagnosis of diabetes using a fuzzy determination mechanism was implemented by Kalpana and Kumar. Their system diagnosed youths (from 25 to 30 years of age). The Mamdani fuzzy inference method was applied and the Pima Indian Diabetes Dataset (PIDD) was also used. The PIDD consisted of 9 attributes and 768 records. Some of the instances that relate to young patients were used. Moreover, six of the nine attributes of the original dataset were used to build this system.

The system developed by Thakur, Raw and Sharma, (2016) was used for the diagnoses of the Thalassemia disease by using the Fuzzy Logic. The proposed system consisted of 3 inputs, output and 15 fuzzy conditional statements. The system's performance evaluation completed by using 15 patients' information, and the accuracy of the proposed techniques was roughly 80%. However, it is understandable that the proposed system should test and train more patient information for reaching efficient results.

Shankar and Manikandan, (2019) explored the diagnosis of diabetes mellitus diseases using an optimized fuzzy rule set by grey wolf optimization. In this work, the dataset comprised of only female patients. A total of 17 fuzzy rules were produced by using eight features and two classes from the Pima Indians Data Set. The optimal rules for output were provided by using the grey wolf optimization algorithm. The algorithm took the fuzzy rules and created the optimal rules. Accuracy, precision and recall metrics used for performance evaluation of the model. "The base model worked on the concept of Ant Colony Optimization and fuzzy rule, which does not provide sufficient accuracy because the algorithm optimizes the local features only and gives 71% accuracy". The drawback of the study is to use the same data set as other studies. However, the proposed model based on grey wolf optimization gave higher accuracy.

The work by Chakraborty, *et al.* (2016) used the Fuzzy C-Means Clustering algorithm for developing Sugeno-Takagi Fuzzy Inference System for detecting Parkinson disease. The performance evaluation values for the systems were up to 96.4% for Fuzzy C-Means based Fuzzy Inference System results and 85.71% for subtractive clustering-based Fuzzy Inference System. The proposed methodologies' performance was better than previous studies in the literature, and C-Means Clustering gave an outstanding result than subtractive clustering.

El-Sappagh *et al.* (2018) proposed a semantically intelligent hierarchical FRBS for diabetes mellitus diagnosis. The study helped for Clinical Decision Support systems (CDSSs) for diabetes mellitus. Ontology and Fuzzy Logic in a novel manner combined in the proposed study. The system had two different layers; the patient's risk level determination and the Mamdani min-max inference mechanism. There were 39 inputs, and types of membership functions were triangular and trapezoidal. The output variable depended on the patient's health situation: diabetic and non-diabetic. The system tested 60 patients, which distribute as 53% diabetic and 47% non-diabetic. The proposed system difference was to use a complete list of diabetes mellitus' attributes, including never used features in a similar type of system. The proposed system used real cases for producing accurate results different from other studies.

Srinivasa R, *et al.* (2020) used three (3) machine learning classification methods namely – Decision tree, Support Vector Machine (SVM), and Naïve Bayes in their experiment to detect diabetes. Their aim was to design a model that could prognosticate the likelihood of diabetes. Results showed that the Naïve Bayes method outperformed the others with an accuracy of 76.30%. Though relatively, this result is still considered poor when compared to other available algorithms already developed for diabetes prognosis.

Nazari, Fallah, *et al.* (2018) developed the clinical decision support system for heart disease. The system aims to calculate the likelihood of developing heart disease. The developed system was based on the Fuzzy Analytic Hierarchy Process and Fuzzy Inference System. The proposed system tested by the attendance of 100 real patients and seven medical doctors. According to doctors' observation, 81 patients needed to do further investigation and test, and 20 patients out of 81 patients suffered from heart disease. However, the proposed system found that 26 patients out of 81 patients, including the 20 patients, who suffered from heart disease, need further investigations. The findings show that by using the proposed support system, the cost and resources can save.

Omisore, *et al*. (2017) conducted a study related to tuberculosis diagnostic. The proposed methodology used Fuzzy Logic, neural network, and genetic algorithm together and developed a Genetic-Neuro-Fuzzy Inference System. Twenty-four input variables and an output variable existed; the proposed system tested by ten patient information. The performance evaluation of the proposed system completed with sensitivity and accuracy, which are 60% and 70%, respectively. The drawback of the proposed methodology is less amount of data used for performance evaluation. The system should be evaluated with more patient information.

Mansourypoor and Asadi (2017) developed a diagnosis system for diabetes mellitus using a Reinforcement Learning-Based Evolutionary Fuzzy Rule-Based System. The proposed model evaluated by using two different databases, which are the Pima Indian Dataset and BioSat Diabetes Dataset. The proposed model comprised a two-step process; “(1) reducing the number of rules and conditions and (2) using the Genetic Algorithm (GA) and Reinforcement Learning (RL) to increase the consistency among the rules” (Mansourypoor & Asadi, 2017). The first step presented rule learning, rule pruning, pruning rule antecedents and evolutionary rule selection. The rule base building process’ numbers of rules were 7593 for Pima Indian Diabetes and 10530 for BioSat Diabetes Dataset in the rule learning step, selecting 200 rules from each dataset in the rule pruning step, 140 rules for Pima Indian Dataset and 136 rules for BioSat Diabetes Dataset in the pruning rule antecedents step, and 19 rules for Pima Indian Dataset and six rules for BioSat Diabetes Dataset in the evolutionary rule selection step. Also, the second step displayed evolutionary rule tuning, adjusting weights and rule stretching. After the implementation of the steps, the accuracy of the proposed system increased. The proposed model gave higher accuracy than other methodologies for both datasets, and the accuracy was 84% for Pima Indian Diabetes Dataset and 99% for BioSat Diabetes Dataset.

Neha Prerna *et al.* (2020) developed a machine learning classification method for type 2 diabetes. The system was aimed at accessing the risk of diabetes among individuals based on their lifestyle and family background. In order to conduct the experiment, 952 instances of diabetes were collected through an online and offline based questionnaire. The system suffered from subjective, as the data could be influenced by random answers from patients or users who filled the questionnaire.

Gizem Koca (2020) designed an intelligent system for type 2 diabetes mellitus diagnosis. The proposed system consists of five different Fuzzy Inference Systems, two central and three subsystems. The main systems diagnose the patient using the subsystems. The Primary diagnostic system uses personal features, biological features, and lifestyle habits; while the Secondary diagnostic system considers personal features and morphological features. Furthermore, the proposed system performance has been evaluated with accuracy, sensitivity, and specificity using three different diabetes datasets: Pima Indian Diabetes Dataset (PIDD), Biostatistics Diabetes Dataset (BSDD) and Randomized produced dataset (RPD). The number of attributes is different for each dataset. There are six attributes for PIDD, eight attributes for BSDD and 14 attributes for RPD. The specificity of the proposed system is 69.2% for PIDD, 92.5% for BSDD, and 95.89% for RPD; while the sensitivity of the proposed method is 93.75% for PIDD, 98.33% for BSDD, and 100% for RPD. Also, the performance evaluation results show that when the number of attributes is lower than system needs, the performance of the proposed system reduces drastically. The shortfall of this system is with its complexity and amount of computational power required.

Abdelgader and Hagras, (2018) presented a diabetes mellitus diet recommendation system in their paper. In the study, Abdelgader and Hagras aim to generate a white box artificial intelligence model which should generate from data models which could be easily analyzed and interpreted by diabetes patients and dietitian. The proposed system needed more than one parameter (age, gender, weight, height, and activity level) for creating the best diet. The reason behind using Type-2 Fuzzy Logic Systems is that it can deal with uncertainty, noise, and imprecision. The personalized diet results were one of the biggest challenges in the research.

Benamina, Atmani and Benbelkacem, (2018) developed a Diabetes diagnosis system by Case-Based Reasoning using Fuzzy Logic. The primary aim of the proposed methodology is to improve the accuracy of diabetes mellitus classification. Also, the paper aims to show the importance of a Fuzzy Inference System guided by data mining in case-based reasoning modelling. The proposed methodology was comprised of two parts; “the modeling part fuzzy realized by Fispro and the reasoning part realized by the platform. The reason behind using Fuzzy Logic is to reduce the complication of the degree of similarity calculation that can exist between individuals who require different monitoring plans. The comparison between proposed methodology and other techniques, which are k-nearest neighbors, decision tree and proposed decision tree, shows that the proposed methodology’s accuracy was higher than other techniques on the same cases. Accuracies were 66% for J Colibri k-nearest neighbors, 73% for Weka decision tree and 81% for Fispro fuzzy decision tree. One of the drawbacks of the proposed methodology is the complicated diagnosis domain for diabetes mellitus.

Mahata *et al.* (2017) developed a mathematical model for Glucose-Insulin Regulatory System on Diabetes Mellitus in Fuzzy and Crisp Environment. The model solved with numerical results for both cases. Hukuhara derivative concept was used to explain the fuzzy solution in the model. The model variables were the plasma glucose concentration at time t, the generalized insulin variable for the remote compartment at time t, the plasma insulin concentration at time t, the basal pre-injection value of plasma-glucose, insulin-independent rate constant of glucose rate uptake in muscles, liver and adipose tissue, the rate of decrease in tissue glucose uptake ability, the insulin-independent increase in glucose uptake ability in tissue per unit of insulin concentration, the rate of the pancreatic beta cells release of insulin after the glucose injection and with glucose concentration, the threshold value of glucose above which the pancreatic beta-cells release insulin, and the first-order decay rate for insulin in plasma pancreatic beta-cells release insulin.

Ambilwade and Manza (2016) addressed the prognosis of diabetes mellitus by using the Fuzzy Inference System and Multilayer Perception. In the proposed system, the Fuzzy Inference System used for predicting the initial risk of prediabetes and type 2 diabetes mellitus using blood tests to measure the sugar/glucose levels in different situations. In the proposed system, three hundred eighty-five patients' information has used, and the dataset collected from Diabetes Care and Research Center, Pune. The performance evaluation criteria of the system were accuracy, sensitivity and specificity, and the results were 91.16% for accuracy, 91.3% for sensitivity and 94.6% for specificity. One of the drawbacks of the proposed system is the lack of patients' data. If more patients' data exists, the performance evaluation results can change, and the system will give a better solution in future implementation.

Abdullah *et al.* (2018) developed a Fuzzy Expert System for the diagnosis of diabetes mellitus. In this study, the expert system was used for risk estimation of diabetes mellitus. In the Fuzzy Expert System, there were 17 inputs and six outputs variables. The inputs were age, body-mass index (BMI), systolic blood pressure, diastolic blood pressure, waist circumference (WC) for male and female, waist-to-hip ratio (WHR) for male and female, cigarettes intake, exercise, alcohol, glycated hemoglobin (HBA1C), triglycerides (TG), high-density lipid (HDL), low-density lipid (LDL), glucose level, education level. The type of membership function for each input and output has been triangular. The center of gravity has used for defuzzification methodology. The six risk categories were very low (0%-10%), low (9%-20%), medium (19%-40%), high (39%-60%), very high (59%-80%), very-very high (79%-100%). If the patient's risk category level is between 79% and 100%, the result shows the confirmation of diabetes mellitus in the patient. However, the calculation of accuracy has not done in the study. The lack of performance evaluation creates a barrier to implement the developed system in real-life cases.

Lukmanto and Irwansyah, (2015) used the Fuzzy Hierarchical Model for the development of the early detection of diabetes mellitus. The proposed model was the computational intelligence application by the usage of the Fuzzy Hierarchical Model that can detect diabetes mellitus early. The designed model architecture was based on how the doctors' decision-making system works against potential disease risk, and the proposed model has justified with the real patient data token from the laboratory. The proposed model used three symptoms, polyuria, polydipsia and polyphagia, fasting blood glucose, 2-hour postprandial blood glucose and age as an input, and the system output can be related to the potential risk of diabetes mellitus. Five Fuzzy Inference Systems used nine rules in each system. The accuracy of the proposed model was 87.46%. According to data, if the patient has diabetes mellitus, but the age of the patient is not in the potential risk groups, the result of the proposed system will be the only potential against diabetes mellitus.

Kumar, Vijav and Devaraj (2013) developed a Hybrid Colony Fuzzy System for Analyzing Diabetes Microarray Data. Ant colony optimization and artificial bee colony algorithms combined for analysis of the datasets, and it was the strength of the proposed system. The optimal rule set created by using ant colony optimization and a total of 4 optimal rules existed. The accuracy of the optimal rules was 98.5%. The utilization of the artificial bee colony algorithm was membership functions' points in the hybrid system. The proposed system gave more compact, accurate, and interpretable results than other studies, such as genetic swarm algorithm.

The robustness of data mining algorithms is an essential factor in choosing the best methodology for the development of a suitable intelligent system. Visalatchi, Gnanasoundhari and Balamurugan (2014) surveyed to select the better data mining techniques for diabetes mellitus. In the paper, five data mining algorithms performance evaluated. The chosen data mining techniques were the C4.5 algorithm, the k-nearest neighbor algorithm, naïve Bayes algorithm, support vector machines and the apriori algorithm. The data source of the study was the Pima Indians Diabetes Database, and there were nine different attributes (pregnancy, plasma, pres, skin, insulin, mass, pedi, age and class). The accuracy of each algorithm helped to evaluate the system's performance. The analysis results were 86%, 78%, 75%, 75% and 74.8%, C4.5 algorithm, k-nearest neighbor algorithm, naïve Bayes algorithm, apriori algorithm and support vector machine algorithm, respectively. The analysis showed that the C4.5 algorithm classifies diabetes mellitus better than other algorithms.

Lee and Wang (2011) developed a diabetes mellitus decision support application by using a Fuzzy Expert System. The focus of the paper was a novel Fuzzy Expert System. The system included a novel five-layer fuzzy ontology (a fuzzy knowledge layer, fuzzy group relation layer, fuzzy group domain layer, fuzzy personal relation layer, and fuzzy personal domain layer), fuzzy concepts and fuzzy relations for diabetes mellitus application. The proposed system examined with the Pima Indians Diabetes Database. C++ Builder 2007 programming language has used for the development of the Fuzzy Expert System. The proposed methodology gave the best results for an age group slightly old and slightly young, 91.2% and 90.3%, respectively. However, the accuracy rate for very very young, very young and more or less young classes was higher than 75%. The uncertainties of the proposed techniques depended on the dataset, dataset domain changes effect, and fuzzification methods were testing. One of the drawbacks of the system is that when the dataset changes happen, the fuzzy rules redesign can be necessary.

Lalka and Jain (2015), developed a Fuzzy Based Expert System for Diabetes Diagnosis and Insulin Dosage Control. The method dealt with uncertainty and vagueness about type 1 diabetes mellitus diagnosis. The inputs were body mass index (BMI), plasma glucose level, minimum blood pressure and serum insulin level for diagnostic of type 1 diabetes mellitus, and plasma glucose level and body mass index also used for insulin dosage control. The membership function type is trapezoidal, and the defuzzification method of the system is the centroid area. Also, there were 60 rules in the system. JAVA programming language used for expert system design. The system verified with real-time patient results. It proves that the effective and efficient diagnosis of type 1 diabetes mellitus is possible with the usage of the proposed system.

Bashir, *et al*. (2014) presented a study, called 'An Efficient Rule-Based Classification of Diabetes Using ID3, C4.5 & CART Ensembles. The primary aim of the proposed methodology was to find the best ensemble techniques for decision trees. The used ensembles were Majority Voting, Adaboost, Bayesian Boosting, Stacking and Bagging. The techniques evaluated by using methodologies such as accuracy, sensitivity, specificity, and f-measure with taking advantage of the Pima Indian Diabetes Dataset and BioStat Diabetes Dataset. Model building, learning, and testing completed using RapidMiner5 Machine Learning Toolbox. The training set was 90% of data, while the testing set covered 10% of data. The bagging approach gave better performance results than the other approaches. The accuracy of the bagging approach was the highest for both datasets.

III. Materials and Methods

The method that would be adopted in this research would be based on simulations and experimentations carried out using MATLAB derived functions in the Neuro-fuzzy toolbox. The research flow would follow the step-wise approach illustrated below:

3.1 Training Data Acquisition

The goal of data acquisition is to find datasets that can be used to train machine learning models. There are largely three approaches to this, which are: data discovery, data augmentation, and data generation. Data discovery is necessary when one wants to share or search for new datasets and has become important as more datasets are available on the Web and corporate data lakes. Data augmentation complements data discovery where existing datasets are enhanced by adding more external data. Data generation can be used when there is no available external dataset, but it is possible to generate crowd sourced or synthetic datasets instead. In the machine learning community, adding pre-trained embedding's is a common way to increase the features to train on. In the data management community, entity augmentation techniques have been proposed to further enrich existing entity information. In many cases, datasets are incomplete and need to be filled in by gathering more information. The missing information can either be values or entire features.

For this work, data augmentation approach was used. The work uses an augmentation of data from seven (7) different databases, shown in appendix B. The databases include – The Pima Indian Diabetes Dataset, Bio-Statistics Diabetes Dataset, Nnamdi Azikiwe University Teaching Hospital (NAUTH), Enugu State University Teaching Hospital (ESUTH) Parkline Enugu state, University of Nigeria Teaching Hospital (UNTH), Federal Teaching Hospital Abakaliki, Ebonyi State. (FETHA) and Federal Medical Centre Owerri, Imo state, Nigeria.

The Pima Indian Diabetes Dataset comprise of data from Pima Indian people in the North America and Caribbean region and the Native American community which is known with the highest prevalence rate of type 2 diabetes mellitus and other diseases. The dataset has 768 patients with eight different attributes, which are – pregnancy situation, plasma glucose concentration, diastolic blood pressure, skin thickness level, insulin, body mass index, diabetes pedigree function and age.

The second dataset is the Bio-Statistics Diabetes Dataset. This dataset covers 403 patients information with 19 attributes namely - patient id, total cholesterol, stabilized glucose, high-density lipoprotein, ratio of cholesterol and HDL levels, glycosylated hemoglobin, location, age, gender, height, weight, frame, first systolic blood pressure, first diastolic blood pressure, second systolic blood pressure, second diastolic blood pressure, waist, hip and postprandial time when labs were drawn. This data was drawn from Buckingham and Louisiana (Department of Biostatistics, Vanderbilt University, 2019). The differences between the Pima Indian Diabetes Dataset and Bio-Statistics Diabetes Dataset are the number of attributes and the location of the people. These two databases have been widely used in most research that has to do with prediction of diabetes mellitus using machine learning approach. What influenced the choice of these two databases is the number of attributes considered and the number of patients considered.

The data from indigenous databases used were collected manually from some of the federal teaching hospitals in eastern Nigeria. The dataset from Nnamdi Azikiwe University Teaching Hospital has 48 patients with four different attributes, which are – plasma glucose concentration, diastolic blood pressure, Gender, and age. In the dataset, 28 patients out of 48 patients indicated as a type 2 diabetes mellitus patient.

The dataset from FETHA has records from 2163 patients with ten different attributes, which are –sex, cholesterol, pregnancy situation, alcohol consumption, Physical activity, plasma glucose concentration, diastolic blood pressure, Gender, and age.

The dataset from ESUTH has records from 2045 patients with ten different attributes, which are – physical activity, gender, sex, cholesterol, pregnancy situation, alcohol consumption, plasma glucose concentration, diastolic blood pressure, and age.

The dataset from UNTH has records from2062 patients with ten different attributes, which are – sex, cholesterol, pregnancy situation, alcohol consumption, physical activity, plasma glucose concentration, diastolic blood pressure, gender, and age.

The dataset from Federal Medical Centre Owerri, Imo state has 52 patients with five (5) different attributes, which are – insulin, glucose concentration, diastolic blood pressure, Gender, and body mass index. In the dataset, 32 patients out of 52 patients indicated as a type 2 diabetes mellitus patient.

3.2 Data Pre–processing

This step is one of the most important phases in the training process. It prepares and transforms the initial dataset. Raw data is generally incomplete, inconsistent, and noisy. Analyzing data that has such problems can produce misleading results. Thus, data preprocess in methods can be applied to raw data before running an analysis. Data pre-processing methods involve replacing missing values, normalization, data discretization, data transformation, data integration, feature extraction, etc.

In this paper, the Diabetes Dataset obtained from the various repositories has missing values for some of the attributes. In this paper, all attributes and instances of the original dataset were used.

The data from FETHA, UNTH, and ESUTH did not have any missing attribute.

Improving accuracy or reducing computational cost are the main approaches of machine learning techniques, but it depends heavily on the test data used. Even more so when it comes from real-world data that contain a high level of missing values. It is very important to select a method to that is capable of replacing these missing values with plausible values.

In this paper, the Multiple Imputation method was used to handle the missing values in the original dataset. The multiple imputation technique by D. Rubin, (2004) was selected based on the percentage and pattern of the missing values. The Multiple Imputation is an approach that replaces each deficient or missing value with more than one acceptable value representing a distribution of possibilities. It looks at the pattern of the available data, and based on probability judgment, attempts to find the best matches, replacing the missing values with imputed values. Replacement is performed repeatedly in order to find the perfect fit. In this thesis, IBM SPSS version 22 was used to perform the multiple imputation process.

3.3 Fuzzification of the input and output variables

The fuzzification process comprises the process of transforming crisp values into grades of membership for linguistic terms of fuzzy sets. As stated in the literature review, 10 inputs were used in this paper, and they consists of Blood pressure (systolic/diastolic), Cholesterol, Blood Glucose Level, Pregnancy Situation, Body Mass Index (BMI), Gender, Age, Family History of Patients and Physical Activity Level.

The Membership functions' type used for both the input and output of the system is the triangular type. The choice of this type of membership function is because of its computational efficiency and ease in dealing with real-life implementations. Also, the triangular membership function gives more accurate output results. The membership functions of each of the following features are shown in this section below as depicted in Table 1−10 and Figure 1−10:

A. Blood pressure (systolic/diastolic):

TABLE 1: Blood Pressure values

Membership Function	**Lower Limits**	**Upper Limits**
SYSTOLIC BLOOD PRESSURE		
Low	0 mm Hg	120 mm Hg
Medium	119 mm Hg	130 mm Hg
High	129 mm Hg	140 mm Hg
Very High	139 mm Hg	300 mm Hg
DIASTOLIC BLOOD PRESSURE		
Low	0 mm Hg	80 mm Hg
Medium	79 mm Hg	90 mm Hg
High	89 mm Hg	120 mm Hg
Very High	119 mm Hg	200 mm Hg

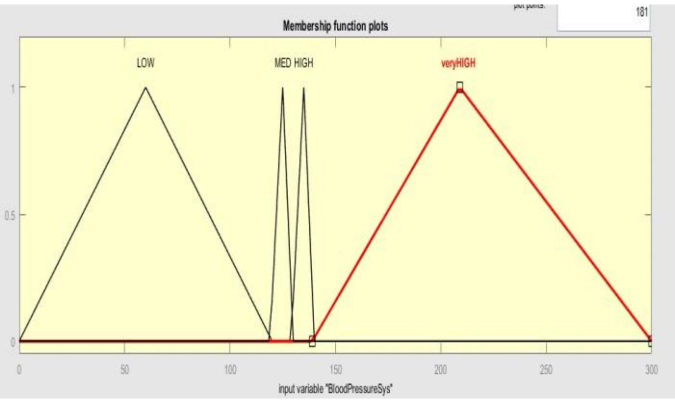

Figure 1: Membership functions for Systolic Blood Pressure

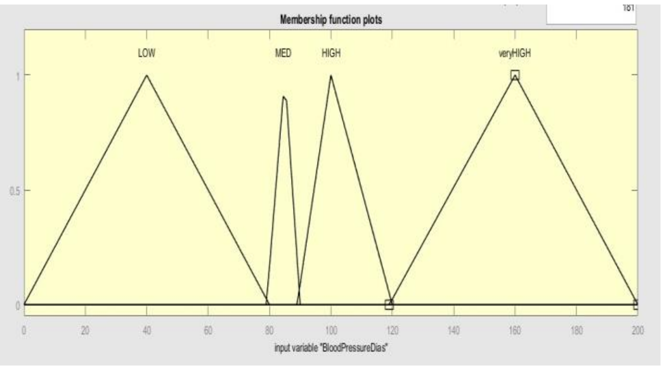

Figure 2: Membership functions for Diastolic Blood Pressure

B. Cholesterol

TABLE 2: Cholesterol Membership function values

Membership Functions Level	**Lower Limits**	**Upper Limits**
Low	0mg/dL	200mg/dL
Medium	99mg/dL	240mg/dL
High	239mg/dL	500mg/dL

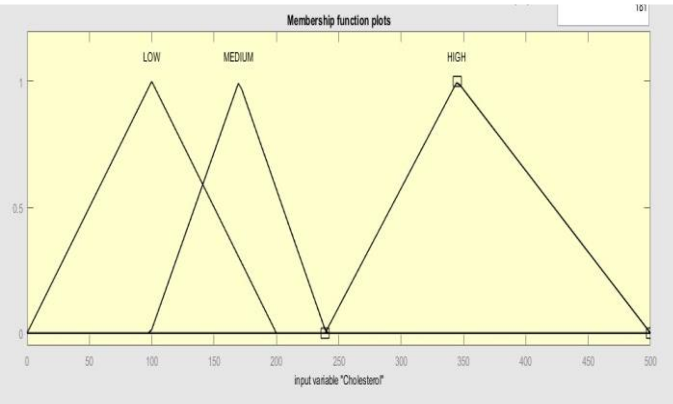

Figure 3: Cholesterol Membership function values

C. Blood Glucose Level

TABLE 3: Blood Glucose Level values

Membership Functions Level	**Lower Limits**	**Upper Limits**
Low	0mg/dL	100mg/dL
Medium	99mg/dL	126mg/dL
High	125mg/dL	800mg/dL

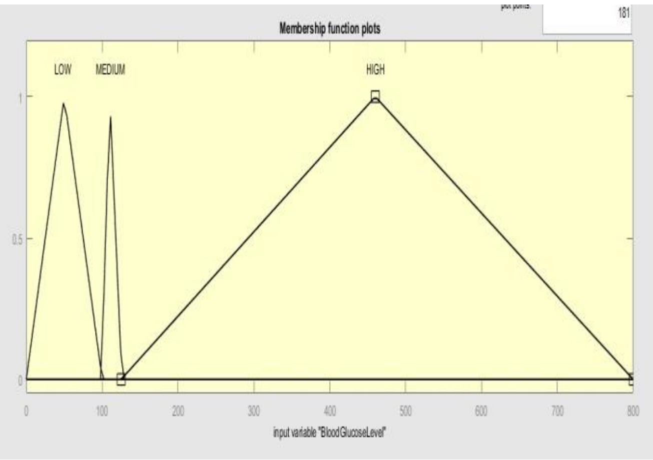

Figure 4: Blood Glucose Level Membership function values

D. Pregnancy Situation

TABLE 4: Pregnancy situation Membership function values

Category	Risk Level
Positive	High Risk
Negative	Low Risk

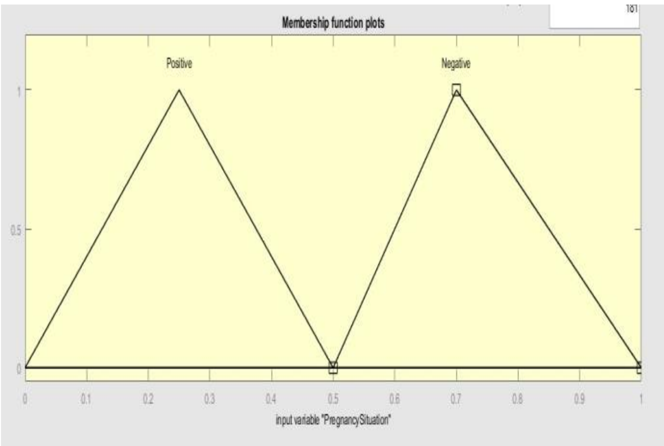

Figure 5: Pregnancy situation Membership function values

E. Body Mass Index (BMI)

TABLE 5: BMI Membership function values

Membership Functions Level	Lower Limits	Upper Limits
Normal	18.5	24.09
Under Weight	0	> 18.5
Over Weight	25	29.9
Obesity	30	39.9
Morbidity obesity	40	< 40

F. Gender

TABLE 6: Gender Membership function values

Gender	Risk Level
Male	High Risk
Female	Low Risk

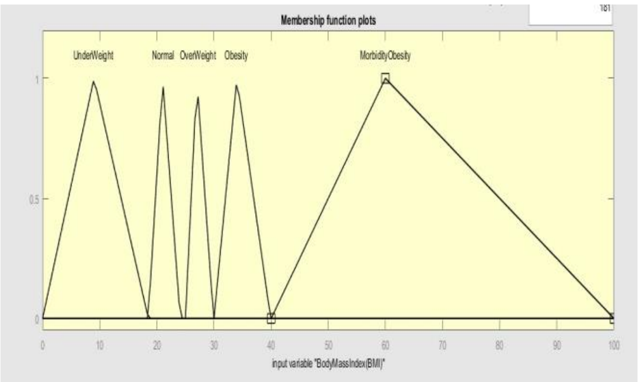

Figure 6: BMI Membership function value

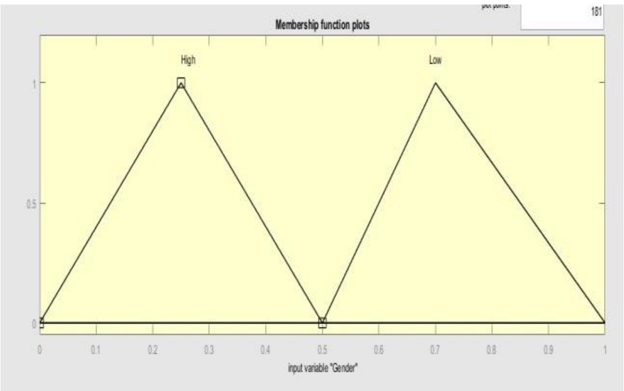

Figure 7: Gender Membership function values

G. Age:

TABLE 7: Age Membership Functions Level

Category	Low Limit	High Limit
20 – 29 Years Old	20	30
30 – 39 Years Old	29	40
40 – 49 Years Old	39	50
50 – 59 Years Old	49	60
60 – Older Years Old	59	100

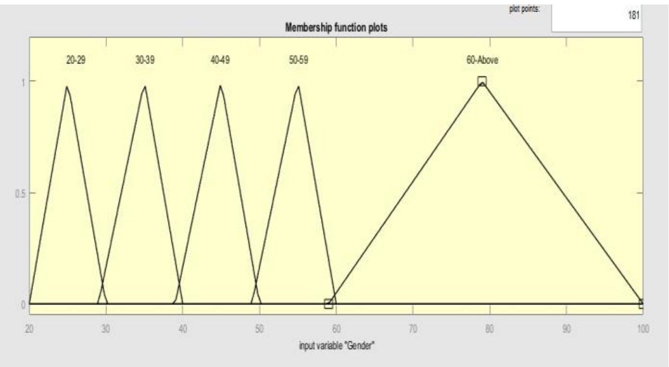

Figure 8: Age Membership Function values

H. Physical Activity Level

TABLE 8: Family History Membership Function Values

Physical Activity Categories	Activeness Level of Patient
Sedentary	Less than 150 minutes moderate or 75 minutes vigorous
Low Activity	150 minutes moderate or 75 minutes vigorous
Active	300 minutes moderate-insensitivity
Very Active	More than 300 minutes moderate-insensitivity

I. *Family History of Patients*

TABLE 9: Family history Membership function values

Category	Risk Level
No History/Unknown History	Low Risk
Just Mother Has Diabetes	Low Risk
Just Father Has Diabetes	Medium Risk
Both Parents Have Diabetes	High Risk
Sibling(s) Has Diabetes	Low Risk

J. *Alcohol consumption*

TABLE 10: Alcohol Membership function values

Membership Functions	**Alcohol Consumption Level**	**Risk Level**
No Usage	Never used alcohol	Low Risk
Low Usage	Women average ≤ I drink/day Men average ≤ 2 drinks/day	Low Risk
Regular Usage	Women average < 4 drinks/in occasion. average < 7 drinks/week Men average < 5 drinks/one occasion. Average < 14 drinks/week	High Risk High Risk

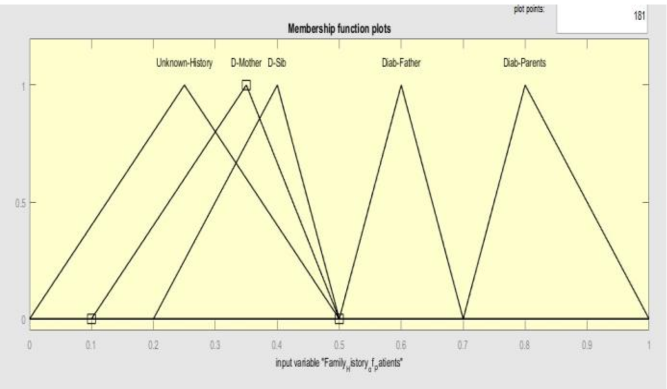

Figure 9: Age Membership function values

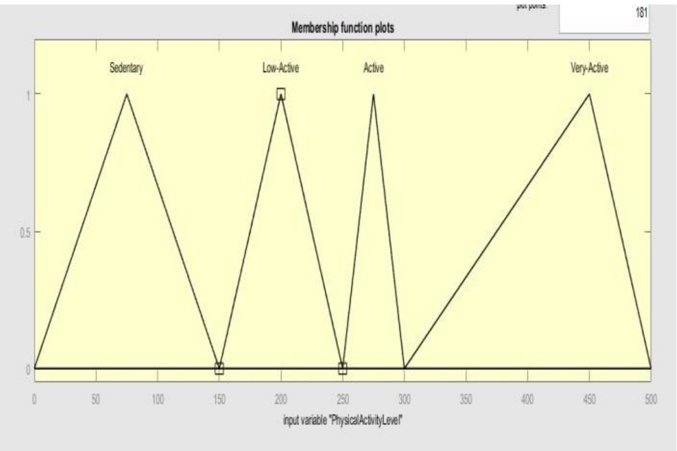

Figure 10: Family history Membership function values

3.4 *Training the Model Using ANFIS function*

To train the model, a knowledge base is created using a rule base which comprises the selection of fuzzy conditional statements, and defines the membership functions in the fuzzy conditional statements. The fuzzy inference engine completes the generation of fuzzy conclusions from the knowledge base. The fuzzy inference engine takes the fuzzified input variables and describes the conclusions by evaluating the fuzzy conditional statements. Finally, all conclusions of each fuzzy conditional statement produce the fuzzy output distribution.

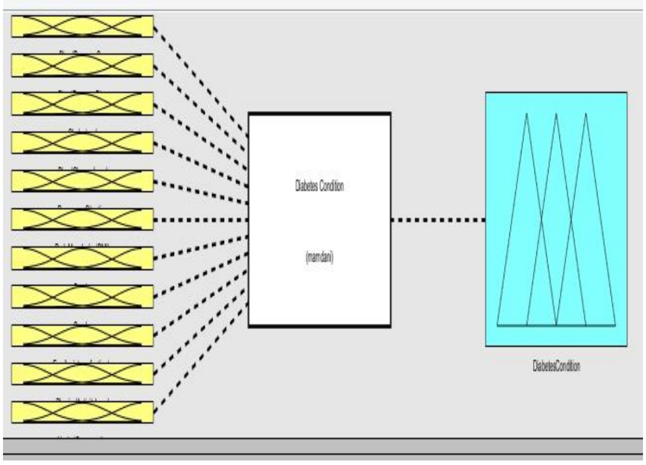

Figure 11: Image of the complete Membership function

The designed Fuzzy Inference System in Figure 11 is evaluated by using ‘evalfis’ function in the MATLAB software. The function helps to train the designed system. The function takes the inputs values and calculates the output value by using the Fuzzy Inference System.

Each layer contains several nodes, which are expressed by that node’s function. Adaptive nodes, shown as squares, display the set of parameters that are adjustable in the respective node. Also, fixed nodes, shown as circles, display parameters that are constant in the model. These layers include:

i. *First Layer:* The first layer contains adjustable nodes, whose membership functions are Gaussian or Bell-shaped with a maximum of 1, and a minimum of 0. Membership function parameters, which are the same as parameters of fuzzy rules, are adjusted based on lingual expression of variables and fuzzy subspaces and based on hybrid methods.

$$O_{1,i} = \mu_{A_i}(x), for\ i = 1,2 \qquad (1)$$

$$O_{1,i} = \mu_{B_{i-2}}(y), for\ i = 3,4 \qquad (2)$$

Where x and y are the input node, and A and B are the linguistic labels associated with this node. μ(x) and μ(y) are the membership functions. The triangular shaped function is adapted.

ii. *Second Layer:* Nodes of the second layer are considered to be constant. These nodes multiply two input signals and deliver the result to the network as their output. The input signals to these nodes are the rate of input adaptability with each of the membership functions and their output is the weight associated with each of the rules.

$$O_{2,i} = W_i = \mu_{A_i}(x) \,.\, \mu_{B_{i-2}}(y), for\ i = 1,2 \quad (3)$$

Where $O_{2,\,i}$ is the output of the second layer.

iii. *Third layer:* The nodes of the third layer are also fixed and their functions is to calculate the normalized weight of each of the rules. The results are represented by equation (3.4).

$$O_{3,i} = \overline{W} = \frac{W_i}{W_1 + W_2} \qquad (4)$$

Where $O_{3,\,i}$ is the output of the third layer.

iv. *Fourth layer:* The nodes of the fourth layer multiply the normalized weight of each fuzzy rule by the latter part of that rule. Every node *i* is an adaptive node. With a node function as shown in equation (5)

$$O_{4,i} = \bar{W}_i \,.\, f_i\, , for\ i = 1,2 \qquad (5)$$

Where f_1 and f_2 are if-then fuzzy rules as described below:

I. Rule 1: if x is the same as A_1 and y is the same as B_1, then $f_1 = p_1x + q_1y + r_1$

II. Rule 2: if x is the same as A_2 and y is the same as B_2, then $f_2 = p_2x + q_2y + r_2$

Where p_i q_i and r_i are specified parameters, which are known as the consequent parameters.

v. Fifth Layer: The fifth layer node collects all output signals from the fourth layer nodes and delivers them to the network as shown in (6):

$$O_{5,i} = \sum_i \bar{W}_i f_i = \frac{\sum_i W_i f_i}{W_i} = f_{out} = overal\ output \qquad (6)$$

During the training, the ANFIS function looks for the minimum value of the error function in weight space using a technique called the gradient descent. The weights that minimize the error function is then considered to be a solution to the learning problem. The flow chart of the training and predicting process is as shown in figure 12.

From figure 12, the system starts by calculating the error of the model by checking the amount of deviation of the model output from the actual output. If y, y' be vectors in R^n, the error function $E(y, y')$ measuring the difference between the target output$y$ and real output $y'$is the square of the Euclidean distance between the vectors $y\ and\ y'$. Mathematically, this function is given as:

$$E(y, y') = \frac{1}{2}||y - y'||^2 \qquad (7)$$

The error function over n training examples is written as average of losses over individual examples as:

$$E = \frac{1}{2n}\sum_x ||(y(x) - y'(x))||^2 \qquad (8)$$

Once this is determined, the system checks if the error is minimized or not. If the error is huge, the system will update the parameters (i.e. weights), either by increasing or decreasing the weight of descent. After that process, the system checks the error again to see if it is minimized. This process is repeated until the error becomes minimum. Once the error is greatly minimized to the global minimum, then the model is ready to make a prediction. At this point, one can feed some inputs to the model and it will produce the output. The following steps are followed to Model the system using MATLAB.

1. Create local workspace variables for test and train data in the MATLAB command window.
2. Open the variables by double clicking the variable name in the workspace.
3. Insert the training data and testing data in the cells of the variables.
4. Open the ANFIS editor from application tab or giving a command “anfisedit” in command window.

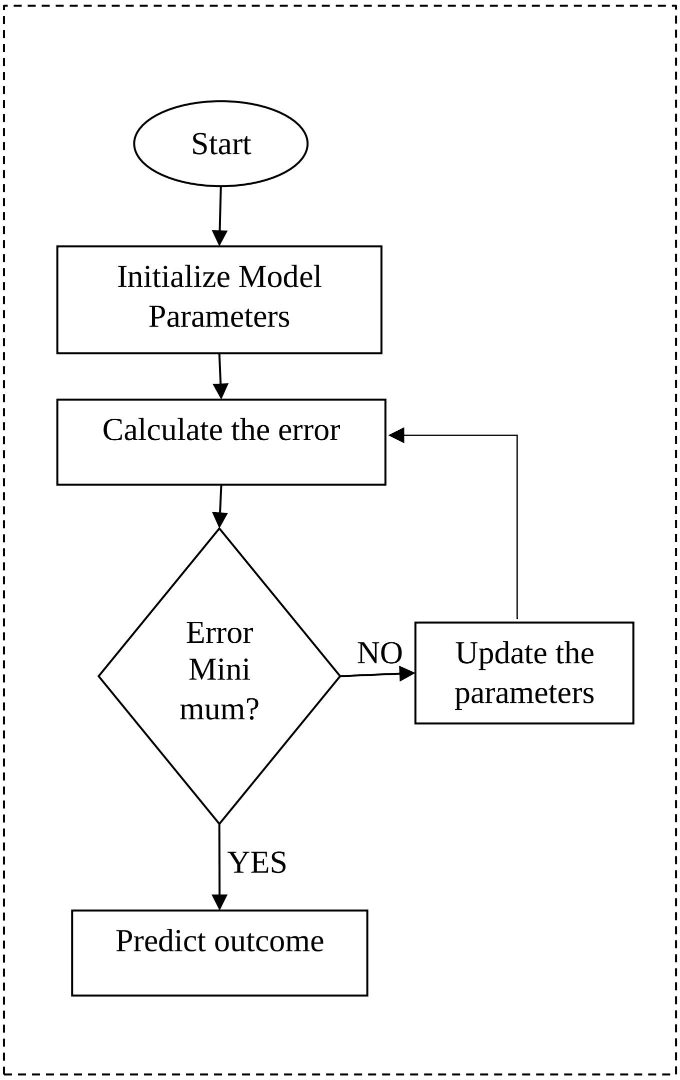

Figure 12: ANFIS Training procedure

5. Load the training data first and generate the FIS by selecting grid partitioning on data.
6. The FIS model structure will be generated and can be viewed in structure.
7. Now the structure needs to be trained.
8. Clicking on train now button trains the FIS. Training involves adjusting the membership function parameters.
9. Insert the testing data in variable using aforementioned method and test the system.

3.5 Validate trained model using test data

To validate the model, the testing dataset is obtained from the training dataset by randomly picking 20% of the data in each of the database used. During the test, the result of the test can either be positive (classifying the person as having diabetes mellitus) or negative (classifying the person as not having diabetes mellitus). The result of the test for each person may or may not match the person's actual status. To accommodate these scenarios, the following instances after the simulation are postulated:

1. True positive (TP): Diabetic people correctly identified as diabetic
2. False positive (FP): Non-diabetic people incorrectly identified as diabetic
3. True negative (TN): Non-diabetic people correctly identified as non-diabetic
4. False negative (FN): Diabetic people incorrectly identified as non-diabetic

The performance metrics used to evaluate the performance of the fuzzy expert system for the incidence of diabetes are – accuracy metric, specificity metric, sensitivity metric, and precision metric.

The accuracy of a classifier on a given test is the percentage of test set tuples that are correctly classified by the classifier layer (Isizoh *et al.*, 2021) as shown in equation (9). The specificity metric (also called true negative rate) refers to the test's ability to correctly detect patients who do not have diabetes, whereas the sensitivity metric (also called recall, or true positive rate) relates to the test's ability to correctly detect patients who do have diabetes. In other words, sensitivity is the proportion of correct positive classifications (TP) from cases that are actually positive. On the other hand, precision is the proportion of correct positive classifications (TP) from cases that are predicted to be positive. The equations of the performance metrics are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \qquad (9)$$

$$Specificity = \frac{TN}{FP + TN} \times 100 \qquad (10)$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \qquad (11)$$

$$Precision = \frac{TP}{TP + FP} \times 100 \qquad (12)$$

IV. Results and Discussion

The experiments were carried out by means of MATLAB derived functions in the Neuro-fuzzy toolbox. The simulation parameters are as shown in table 11.

TABLE 11: Simulation parameters

Parameter	Value
Number of nodes	555
Number of linear parameters	2304
Number of nonlinear parameters	48
Total number of parameters	2352
Number of training data pairs	538
Number of checking data pairs	0
Number of fuzzy rules	256

To properly analyze and validate the performance of the developed fuzzy expert system, the results obtained from the system was compared with those obtained by Gizem Koca (2020). The reason for comparing the result with those of Gizem Koca (2020), is because the work employed as much parameters and dataset as the model developed in this thesis.

Several performance metrics were used to evaluate the performance of the fuzzy expert system and that of Gizem Koca (2020) for the incidence of diabetes, which are accuracy, specificity, sensitivity and precision.

It is important to note that during the medical diagnosis, specificity (also called true negative rate) refers to the test's ability to correctly detect patients who do not have diabetes, whereas sensitivity (also called recall, or true positive rate) relates to the test's ability to correctly detect patients who do have diabetes. Hence sensitivity is the proportion of correct positive classifications (TP) from cases that are actually positive. On the other hand, precision is the proportion of correct positive classifications (TP) from cases that are predicted to be positive.

In the following section, different experiments were carried out and analyzed, and the model with the highest accuracy, specificity, sensitivity, and precision was considered the best predictive model.

4.1 Experiment 1

In the first experiment, the two models were analyzed to check which among them took less time and system resource in minimizing the error to the global minimum during training, before getting the model ready to make a prediction.

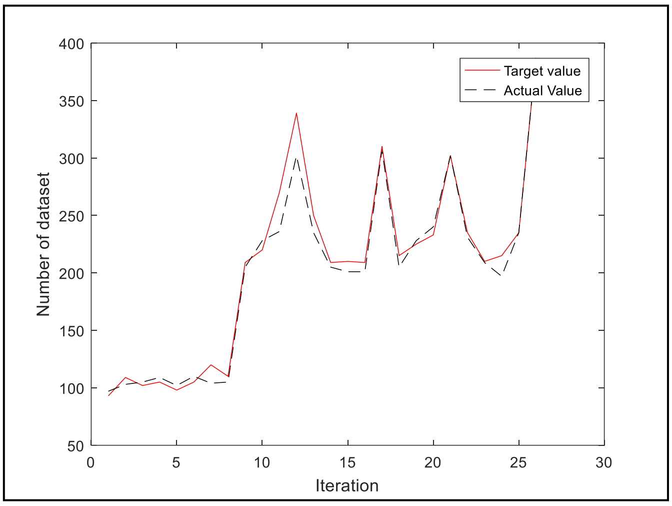

Figure 13: Experiment 1 results for Gizem Koca (2020)

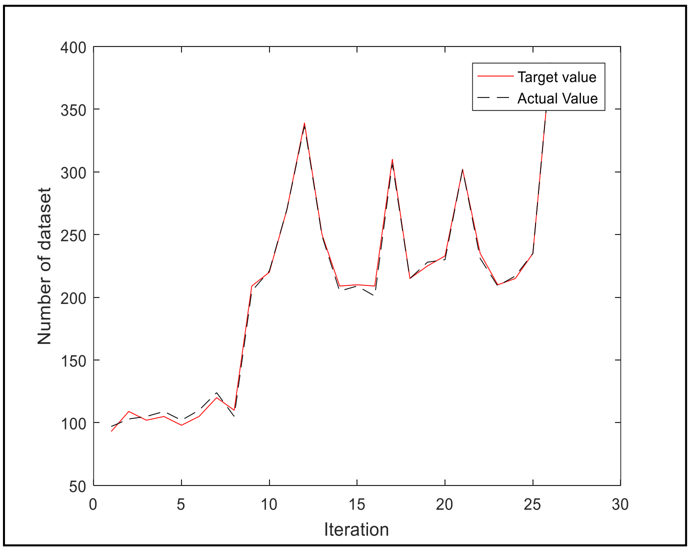

Figure 14: Experiment 1 result for the developed fuzzy expert system

From figure 13 and figure 14, it can be seen that the developed fuzzy expert system takes less iterations to minimize the error to the global minimum. The model by Gizem Koca (2020) needed about 25 iterations to attain the global minimum, while the developed model needed 15 iterations, thus also reducing the amount of system resource consumed.

4.2 Experiment 2

After pre-processing the dataset it was split into training dataset and testing dataset. The test data was obtained from the dataset by splitting the dataset to a ratio of 20:80. Thus 20% of the data was used for testing the model.

The dataset used in this thesis augmented the one used by Gizem Koca (2020) by including data that are indigenous and unique to the African context. Comparison is made to check the performance of the two models.

The first evaluation metric calculated was the accuracy metrics, which is the fraction of true results (both true positives and true negatives) among the total number of cases examined. After this, specificity, sensitivity, and precision metric were calculated. The experiment involved predicting the cases for diabetes when only the dataset considered by Gizem Koca (2020) was used. To carry out this comparison, test data was extracted from both the dataset available online, and that obtained locally in Nigeria. Two set of test data was created for the test – the first one was fetched from just the dataset obtained online while the second tests data was fetched from the dataset obtained locally. These two datasets were used to evaluate the performance of the two expert systems.

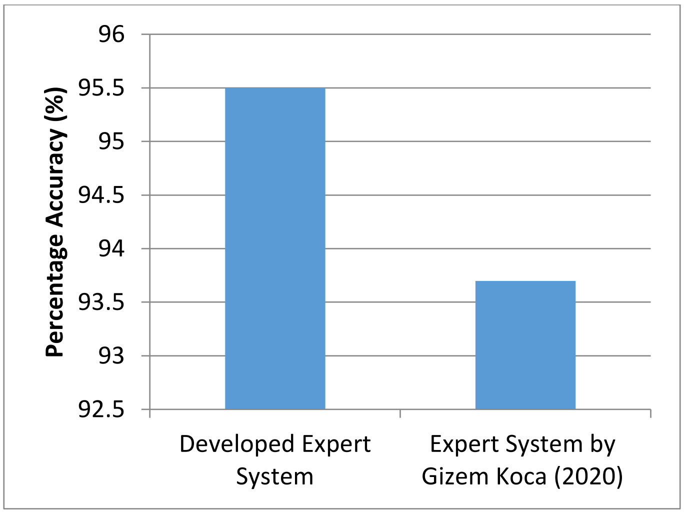

Figure 15: Accuracy of Each Classifier using the test data from only the online database

From figure 15, the accuracy of the developed expert system is 95.5%, while the accuracy of the system by Gizem Koca (2020) is 93.7%. As can be clearly seen, the fuzzy expert system has the highest accuracy with the lowest number of false positives and negatives. The test data set obtained from the local dataset was also used for prediction and the accuracy is as shown in figure 16.

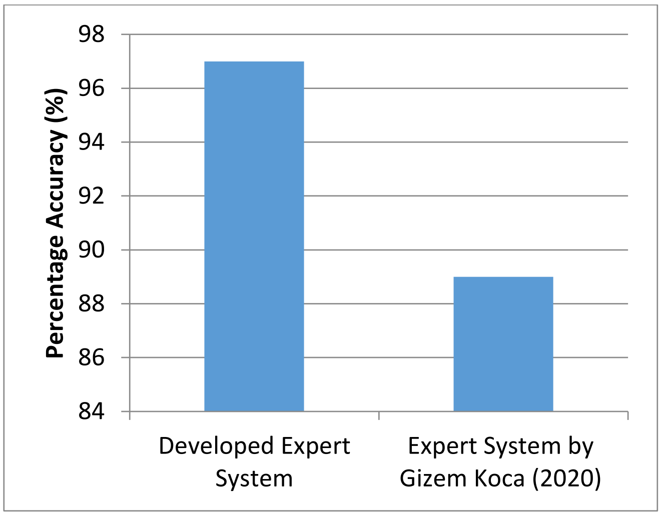

Figure 16: Accuracy of Each Classifier using the test data from the local dataset

From figure 24, it can be seen that the accuracy of the system by Gizem Koca (2020) reduced to 89% as against the developed system that had an accuracy of 97%. The reason for this reduction in accuracy is as a result of the test data used. The system by Gizem Koca (2020) did not incorporate the data from the local content, and thus the prediction accuracy greatly reduced. The developed system augmented the data available online with the locally sourced data, thus enhancing the prediction accuracy. This makes the expert system also suitable for use even for local users.

The specificity metric was also analyzed using the two different test data, and the result is as presented in figure 17 and 18.

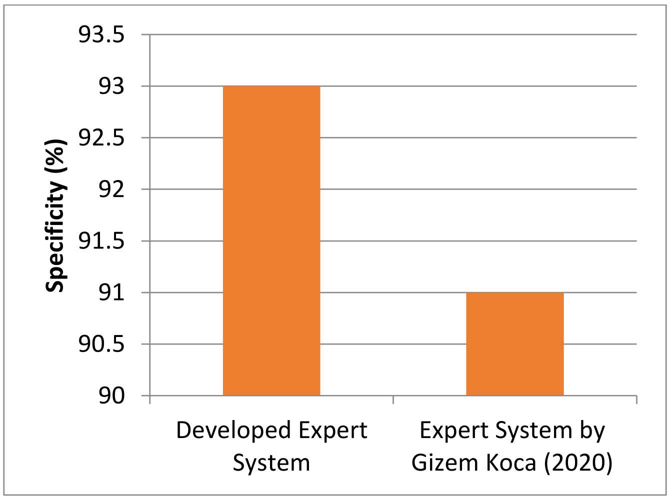

Figure 17: Specificity of Each Classifier using test data from online database

As we can see in the above figure, the fuzzy expert system has the highest number of true negative cases. The specificity for the developed fuzzy based expert system was 93%, while that of Gizem Koca (2020) is 91%. This shows that the developed system has a higher ability to correctly detect patients who do not have diabetes more than that of Gizem

Koca (2020). The results obtained when the test data from the local data set was used are as shown in figure 18.

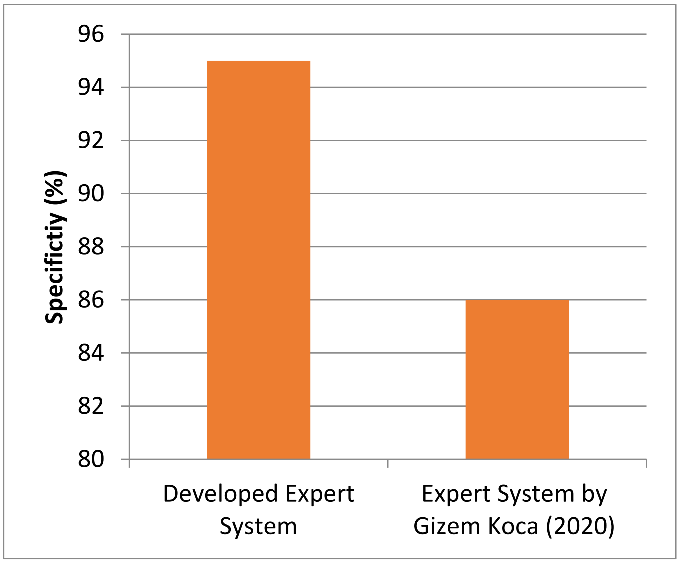

Figure 18: Specificity of Each Classifier using test data from local database

From figure 18, it can be seen that the specificity of the system by Gizem Koca (2020) reduced to 86% as against the developed system that had a specificity of 95%. The reason for this reduction in accuracy is as a result of the test data used. The system by Gizem Koca (2020) did not incorporate the data from the local content, and thus the prediction accuracy greatly reduced.

The sensitivity metric was also analyzed using the two different test data, and the result is as presented in figure 19 and 20.

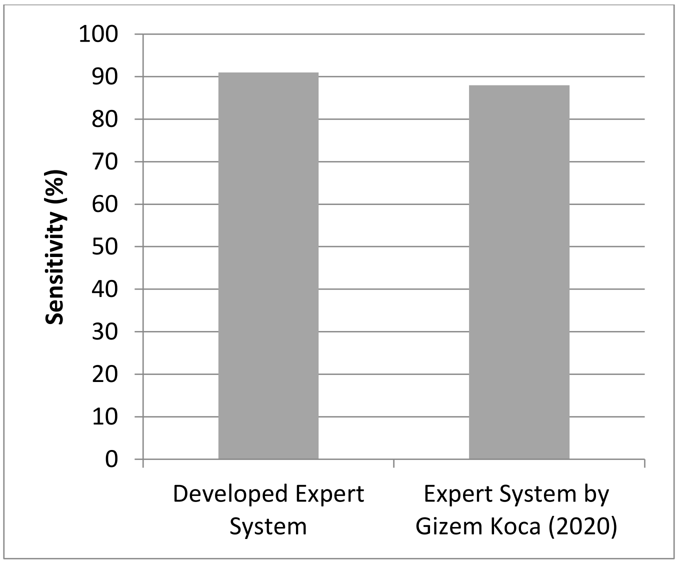

Figure 19: Sensitivity of Each Classifier using test data from online database

As we can see in figure 19, the fuzzy expert system has the highest number of true positive cases. The sensitivity for the developed fuzzy based expert system was 91%, while that of Gizem Koca (2020) is 88%. This shows that the developed system has a higher ability to correctly detect patients who do have diabetes. In other words, the proportion of correct positive classifications (TP) from cases that are actually positive is higher when compared to Gizem Koca (2020).

The results obtained when the test data from the local data set was used are as shown in figure 20.

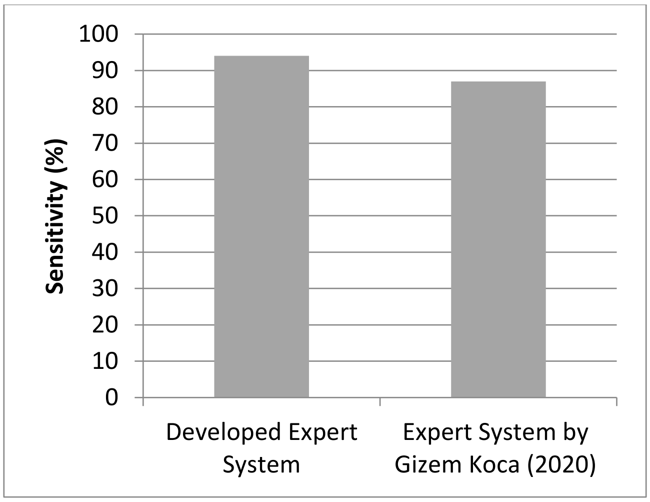

Figure 20: Sensitivity of Each Classifier using test data from local database

From figure 20, it can be seen that the sensitivity of the system by Gizem Koca (2020) reduced to 87% as against the developed system that had a specificity of 94%. The reason for this reduction in accuracy is as a result of the test data used. The system by Gizem Koca (2020) did not incorporate the data from the local content, and thus the prediction sensitivity reduced.

The precision metric was also analyzed using the two different test data, and the result is as presented in figure 21 and 22.

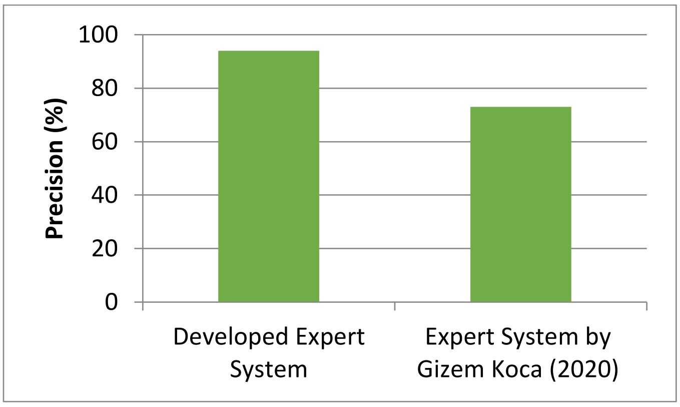

Figure 21: Precision of Each Classifier Using test data from online database

The bar graph of figure 21 illustrates that the fuzzy expert system has the highest precision value. This implies that the lowest number of false positive errors was committed by this classifier. By comparison, the expert system by Gizem Koca (2020) had a lower precision value, and thus a higher number of false positive cases compared to the developed system.

The results obtained when the test data from the local data set was used are as shown in figure 22.

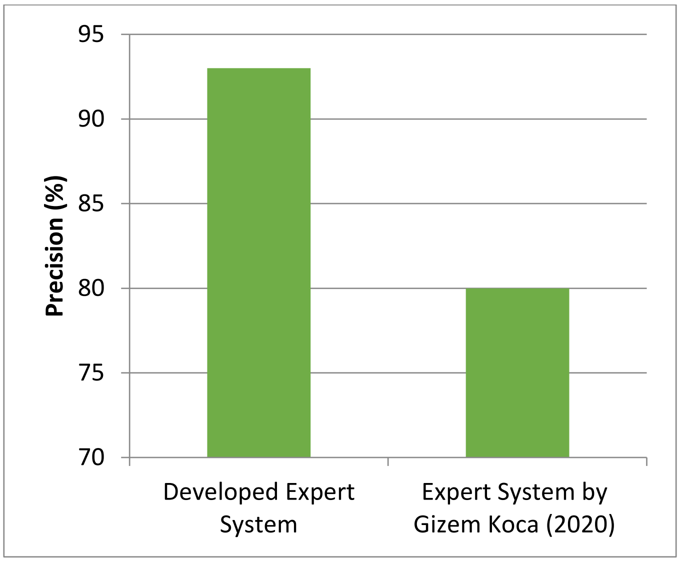

Figure 22: Precision of Each Classifier using test data from local database

Figure 22 also shows that the fuzzy expert system has the highest precision value. The developed expert system had a precision value of 93%, while the expert system by Gizem Koca (2020) had a precision value of 80%. This implies that the lowest number of false positive errors was committed by this classifier. By comparison, the expert system by Gizem Koca (2020) had a lower precision value, and thus a higher number of false positive cases compared to the developed system which is also attributed to the choice of dataset used.

V. Conclusion and Recommendation

In this paper, a fuzzy logic based expert system model for diagnosing diabetes mellitus type 2 was developed. The developed expert system included the four steps of the Mamdani fuzzy inference system, namely fuzzification, rule evaluation, aggregation of the outputs, and difuzzification. The developed model was trained using real data obtained online, and locally collected data from federal medical institutions in eastern Nigeria. Data pre-processing was also carried out to handle some of the missing variables in the dataset. Two different test dataset was obtained from both the dataset obtained online and that collated locally. The developed model was evaluated alongside a similar fuzzy expert system developed by Gizem Koca (2020). Four performance metrics namely accuracy, specificity, sensitivity, and precision were used to evaluate the performance of the two models. Several experiments were carried out to analyze the performance of the two models. Results showed that the model by Gizem Koca (2020) needed about 25 iterations to attain the global minimum, while the developed model needed 15 iterations, thus consuming less computation resources. We found that the performance of the fuzzy expert system. Results also showed that the developed model outperformed the other model since it employed an augmented dataset. When tested with test dataset from the locally generated dataset, the developed fuzzy expert gave a prediction accuracy of 97%, with a specificity of 95%, a sensitivity of 94%, and a precision of 93% when compared to the other system that had a corresponding accuracy of 89%, specificity of 86%, sensitivity of 87% and a precision of 80%. This helps to establish the fact that there is a need to incorporate datasets that are local or unique to a group of persons or region so as to improve the accuracy of the developed model.

a. Recommendation

It is recommended that an interface for the fuzzy expert system be developed in order to enhance its usability. In this thesis, the Mamdani fuzzy inference system was used to develop the fuzzy expert system. Another possible research avenue could be the use of the Sugeno fuzzy inference system, which is the other type of fuzzy inference system.

Contribution to Knowledge

Development of a fuzzy based expert system with improved accuracy in predicting Diabetes mellitus type 2.

References

1. Shiel, W. J. (2017, January 26), "Medical Definition of Diabetes Mellitus", *Medicine Net*. Retrieved November 22, 2019, from Medicine Net: https://www.medicinenet.com/script/main/art.asp?articlekey=2974
2. Gizem Koca (2020) "An Intelligent System for Type Ii Diabetes Mellitus Diagnostic", A Thesis Submitted to the Faculty of Graduate Studies and Research in Partial Fulfillment of the Requirements for the Degree of Master of Applied Science in Industrial Systems Engineering, University of Regina. Glucose-insulin regulation in vivo *(Master's thesis, University of Stavanger).*
3. C. D. Mathers and D. Loncar, "Projections of Global Mortality and Burden of Disease from 2002 to 2030," *PLOS Med*, vol. 3, no. 11, pp. 442-462, 2006.
4. M. Truglio-Londrigan and S. B. Lewenson (2012), "Public Health Nursing*"*, 2nd ed. Jones & Bartlett Publishers.
5. S. Smyth and A. Heron (2006) "Diabetes and obesity: the twin epidemics," *Nature Medicine*, vol.12, no. 1, pp. 75–80.
6. T. Vos, R. Barber, B. Bell, S. Biryukov, A. Bertozzi-Villa, I. Bolliger, and L. Duan (2015), "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries" *The Lancet*, vol. 386, no. 9995, pp. 743– 800.
7. Mitushi Soni (2020), "Diabetes prediction using machine learning techniques", *International Journal of Engineering Research and Technology*, Volume 9, ISSN 2278 – 0181
8. Adeli, A., & Neshat, M. (2010), "A Fuzzy Expert System for Heart Disease Diagnosis", *Proceedings of the International Multi Conference of Engineers and Computer Scientists Vol I - IMECS 2010.* Hong Kong.
9. Thakur, S., Raw, S. H., & Sharma, R. (2016). Design of a Fuzzy Model for Thalassemia Disease Diagnosis: Using Mamdani Type Fuzzy Inference System (FIS). *International Journal of Pharmacy and Pharmaceutical Sciences, 8*(4), 356-361.
10. Shankar, S. G., & Manikandan, K. (2019). Diagnosis of Diabetes Diseases Using Optimized Fuzzy Rule Set by Grey Wolf Optimization. *Pattern Recognition Letters*, 432-438. doi:10.1016/j.patrec.2019.06.005
11. Chakraborty, A., Chakraborty, A., & Mukherjee, B. (2016). Detection of Parkinson's disease Using Fuzzy Inference System. In S. Beretti, S. Thampi, & P. Srivasta, *Intelligent Systems Technologies and Applications, Advances in Intelligent Systems and Computing* (Vol. 384, pp. 79-90). Springer Cham. doi:10.1007/978-3-319-23036-8_7
12. El-Sappagh, S., Alonso, J. M., Ali, F., Ali, A., Jang, J., & Kwak, K. (2018, July 04). An Ontology-Based Interpretable Fuzzy Decision Support System for Diabetes Diagnosis. *IEEE Access, 6*, 37371-37394. doi:10.1109/ACCESS.2018.2852004
13. Srinivasa R, Yashaswini J, Venkatesh K. B, Yaswanth S. P (2020), "Prediction of diabetes using machine learning", *International Journal of Advanced Science and Technology*, Vol. 29, No. 6, pp 7593 – 7601
14. Nazari, S., Fallah, M., Kazemipoor, H., & Salehipour, A. (2018). A Fuzzy Inference-Fuzzy Analytic Hierarchy Process-Based Clinical Decision Support System for Diagnosis of Heart Disease. *Expert Systems with Applications, 95*, 261-271. doi:10.1016/j.eswa.2017.11.001

15. Omisore, M. O., Samuel, O. W., & Atajeromavwo. (2017). A Genetic-Neuro-Fuzzy Inferential Model for Diagnosis of Tuberculosis. *Applied Computing and Informatics, 13*, 27-37. doi:10.1016/j.aci.2015.06.001
16. Mansourypoor, F., & Asadi, S. (2017). Development of A Reinforcement Learning-Based Evolutionary Fuzzy Rule-Based System for Diabetes Diagnosis. *Computers in Biology and Medicine, 91*, 337-352. doi:10.1016/j.compbiomed.2017.10.024
17. Neha Prena Tigga, Shruti Garg (2020), Prediction of Type 2 Diabetes using machine learning classification methods", *International conference on computational Intelligence and data science*, Volume 167, pp 706 – 716
18. Abdelgader, H., & Hagras, H. (2018). Towards Developing Type 2 Fuzzy Logic Diet Recommendation System for Diabetes. *10th Computer Science and Electronic Engineering (CEEC)* (pp. 56-59). Colchester, United Kingdom: IEEE. doi:10.1109/CEEC.2018.8674186
19. Benamina, M., Atmani, B., & Benbelkacem, S. (2018, December). Diabetes Diagnosis by Case-Based Reasoning and Fuzzy Logic. *International Journal of Interactive Multimedia and Artificial Intelligence*, 72-80. doi:10.9781/ijimai.2018.02.001
20. Mahata, A., Mondal, S. P., Alam, S., & Roy, B. (2017). Mathematical Model of Glucose-Insulin Regulatory System on Diabetes Mellitus in Fuzzy and Crisp Environment. *Ecological Genetics and Genomics*, 25-34. doi:10.1016/j.egg.2016.10.002
21. Ambilwade, R. P., & Manza, R. R. (2016). Prognosis of Diabetes Using Fuzzy Inference System and Multilayer Perception. *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)* (pp. 248-252). Noida, India: IEEE. doi:10.1109/IC3I.2016.79177969
22. Lukmanto, R. B., & Irwansyah, E. (2015). The Early Detection of Diabetes Mellitus (DM) Using Fuzzy Hierarchical Model. *Procedia Computer Science, 59*, 312-319. doi:10.1016/j.procs.2015.07.571
23. Kumar, P. G., Vijav, S. A., & Devaraj, D. (2013). A Hybrid Colony Fuzzy System for Analyzing Diabetes Microarray Data. *2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (pp. 104-111). Singapore, Singapore: IEEE. doi:10.1109/CIBCB.2013.6595395
24. Visalatchi, G., Gnanasoundhari, S. J., & Balamurugan, M. (2014, February 2). A Survey on Data Mining Methods and Techniques for Diabetes Mellitus. *International Journal of Computer Science and Mobile Applications, 2*(2), 100-105.
25. Lee, C., Wang, M., & Hagras, H. (2010, April). A Type-2 Fuzzy Ontology and Its Application to Personal Diabetic-Diet Recommendation. *IEEE Transactions on Fuzzy Systems, 18*(2), 374-395. doi:10.1109/TFUZZ.2010.2042454
26. Lalka, N., & Jain, S. (2015). Fuzzy Based Expert System for Diabetes Diagnosis and Insulin Dosage Control. (pp. 262-267). Noida, India: IEEE. doi:10.1109/CCAA.2015.7148385
27. Bashir, S., Qamar, U., Khan, F. H., & Javed, M. Y. (2014). An Efficient Rule-Based Classification of Diabetes Using ID3, C4.5 & CART Ensembles. *2014 12th International Conference on Frontiers of Information Technology* (pp. 226-231). Islamabad, Pakistan: IEEE. doi:10.1109/FIT.2014.50
28. D. Rubin, 2004"Multiple imputation for nonresponse in surveys," *John Wiley & Sons*, vol. 81, p. 253.
29. Isizoh A.N *et al* (2021), "Development of convolution Neural Network-Based Diagnostic System for Detection of Coronavirus Using Magnetic Resonant Imaging (MRI)", *International Journal of Advances in Engineering and Management (IJAEM)*, Volume 3, Issue 7, pp:1865−1886. www.ijaem.net.