International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12– No.39, August 2022 – www.ijais.org

A Diabetic Prediction Model using Firefly Algorithm with K-Nearest Neighbor Classifier

Sulaiman Olaniyi Abdulsalam
Department of Computer Science,
Kwara State University,
Malete, Nigeria

ABSTRACT

Diabetes is one of the illnesses that lasts for a long time, it has led to a lot of mortality yearly. If it is not treated, it can affect how well other human organs functions. So, early detection is an important part of living a healthy life. According to the World Health Organization, about 104 million people had diabetes in 1980. By 2014, that number had risen to 422 million, and it is expected to double by 2030. Machine learning is an area of Artificial Intelligence that focuses on making tools that can learn, or automatically pull out the knowledge hidden in data. Along with statistics, it is the most important part of a smart analysis of data. Both machine learning and data mining are based on the same idea: the machine learns from examples and then uses that model to solve the problem. The results of the finding obtained an accuracy of 91% compared to existing related works. Hence, this paper suggests a firefly-based attribute selection algorithm with K-nearest neighbor (KNN) classifier for the PIMA Indian diabetic database from University of California, Irvine (UCI).

General Terms

Machine learning, Bioinformatics, Data Science

Keywords

Diabetes; Firefly; K-Nearest Neighbor; Classification; Prediction

1. INTRODUCTION

Diabetes Mellitus is a disease that stops our bodies from using the energy gotten from food in the right way. When human eat carbohydrates, our bodies turn the energy from the food into glucose and send it to the bloodstream. Diabetes can show up in these ways; either the pancreas isn't making enough insulin, or it is making enough insulin, but it isnot working the way it should, this is called insulin resistance [1]. When human eat food, our bodies break it down into glucose, which gives us the energy needed to go about our daily lives. From the blood vessel, sugar moves to the liver cells. The pancreas sends out insulin, which helps break down the glucose. Without insulin, glucose can't get to the cells of the body to be used as energy. Instead, it gets stored in the stomach, which raises the amount of glucose in the blood [2].

One of the type of diabetes that causes damage to the beta cells and makes insulin is known as Type 1 diabetes. Insulin from outside the body is injected into the body to control this type of diabetes. This helps the body break down food into glucose, which lowers the amount of glucose in the body. Type 2 diabetes is another form of diabetes that occurs when the body doesn't make enough insulin or when the insulin doesn't work right [3]. This type of diabetes usually affects people over the age of 30, and it can be managed with a healthy diet, regular exercise, and oral medications. Gestational diabetes happens when the baby in the womb needs more glucose, which happens when a woman is pregnant. Gestational diabetes goes away after giving birth, but the woman is more likely to get type 2 diabetes in the future. Because there are so many things that can make you more likely to get diabetes, it is important to spot it early [4].

There's no doubt that machine learning and data mining are very important for clinical administration, diagnosis, and treatment. Machine learning algorithms are a great fit for the field of health care diagnosis [5]. Many of these can be found by analyzing large amounts of data for patterns. To be useful in the field, medical diagnostics must be able to handle noisy and empty datasets. An algorithm should be trained on a small number of tests. A lot of research has been done on machine learning in the field of healthcare. Machine learning in health care has become one of the top goals of many academics. Different data mining methods and procedures can be used to find out hidden patterns. The main jobs of medical science are to help stop or treat diseases [6].

This study uses a firefly feature selection algorithm to choose the right features to use with a K-nearest neighbor (KNN) classifier to predict diabetes in human.

2. RELATED WORKS

The role of KNN in finding Diabetes Mellitus was investigated [7]. Diabetes is one of the illnesses that lasts for a long time. According to the World Health Organization, about 104 million people had diabetes in 1980. By 2014, that number had risen to 422 million, and it is expected to double by 2030. In this paper, the PIMA Indians diabetes dataset was used to test the supervised K-nearest neighbor machine learning algorithm. The K-nearest neighbor algorithm looks at how similar new data is to data that has already been stored. After the proposed algorithm was used, the accuracy went up from 70.1% to 78.58%, which is an increase of 8.48%.

Firefly and Cuckoo Search Algorithms were used to come up with a hybrid method for predicting Type-1 and Type-2 diabetes [8]. Machine learning is an area of Artificial Intelligence that focuses on making tools that can learn, or automatically pull out the knowledge hidden in data. Along with statistics, it is the most important part of the smart analysis of data. Both machine learning and data mining are based on the same idea: the machine learns from examples and then uses that model to solve the problem. This paper suggests a firefly and cuckoo search-based attribute selection algorithm for the PIMA Indian diabetic database from UCI.

The goal is to improve accuracy and reduce training time. Using the UCI dataset and the KNN classifier, the experimental setup has been made. As an evaluation criterion, the accuracy, precision, and recall have been calculated. When compared with the optimized structures of the Cuckoo search and Firefly algorithm, the proposed structure claims to be more accurate than the traditional method.

A study was proposed using Machine Learning Algorithms for predicting diabetes [9]. Diabetes Mellitus is one of the most serious diseases, and a lot of people have it. Diabetes Mellitus can be caused by age, being overweight, not getting enough exercise, having a family history of diabetes, a bad diet, high blood pressure, and other things. People with diabetes are more likely to get heart disease, kidney disease, stroke, eye problems, nerve damage, and other diseases. Different tests are used in hospitals to get the information needed to diagnose diabetes, and the right treatment is given based on the diagnosis. In the healthcare industry, Big Data Analytics plays a big role. There are a lot of databases in the healthcare industry. Using "big data analytics," you can look at very large datasets to find hidden information and patterns. This lets you learn from the data and predict what will happen based on what you find. With the methods existing, it's not so easy to classify things and make predictions. In this paper, a way to predict diabetes based on factors like glucose, body mass index (BMI), age, insulin, etc., as well as a few outside factors that can cause diabetes is suggested. With a new dataset, the accuracy of classification is better than with an existing dataset. Also, a pipeline model for predicting diabetes was put in place to improve the accuracy of classification.

An investigation using Machine Learning Classification Methods was proposed to predict Type 2 Diabetes [10]. Diabetes affects more than 30 million people in India, and a lot of other people are at risk. So, early diagnosis and treatment are needed to stop diabetes and the health problems it can cause. The goal of this study is to figure out how likely people are to get diabetes based on how they live and what their family history is like. Using different machine learning algorithms, the risk of Type 2 diabetes was estimated. These algorithms are very accurate, which is very important in the health field. Once the model is trained to be accurate, people will be able to figure out how likely they are to get diabetes on their own. In order to do the experiment, 952 responses to an online and paper questionnaire with 18 questions about health, lifestyle, and family history have been collected. The Pima Indian Diabetes database was also run through the same algorithms. For both sets of data, the Random Forest Classifier gives the best results.

Machine learning techniques were used to make predictions about diabetes mellitus [11]. Diabetes Mellitus is a chronic disease that is becoming more common. It is caused by the body's inability to process glucose. The goal of this study was to make a good predictive model with high sensitivity and selectivity that could better identify Canadian patients who might have Diabetes Mellitus based on demographic information about the patients and the lab results from their visits to medical facilities. Recent records of 13,309 Canadian patients between the ages of 18 and 90, as well as their laboratory information (age, sex, fasting blood glucose, body mass index, high-density lipoprotein, triglycerides, blood pressure, and low-density lipoprotein), to make predictive models. The ability of these models to tell different things apart was judged by the area under the receiver operating characteristic curve (AROC). The adjusted threshold method and the class weight method to improve sensitivity, which is the number of Diabetes Mellitus patients that the model correctly predicted. They compared these models to other ways for machines to learn, like Decision Tree and Random Forest. The proposed GBM model has an AROC of 84.7% and a sensitivity of 71.6%, while the proposed Logistic Regression model has an AROC of 84.0% and a sensitivity of 73.46%. The Random Forest and Decision Tree models don't work as well as the GBM and Logistic Regression models. Using some commonly used lab results, our model can predict with high accuracy which patients have Diabetes. These models can be put into an online computer program to help doctors predict which patients will get diabetes in the future and take the right steps to prevent it. The model was made and tested on the Canadian population, which makes it more accurate and useful for Canadian patients than models made for the US or other populations. In these models, the most important predictors were fasting blood glucose, body mass index, high-density lipoprotein, and triglycerides.

Prediction using Machine learning techniques was proposed for diabetes mellitus [12]. Hyperglycemia is a symptom of diabetes mellitus, which is a long-term disease. It could lead to a lot of trouble. Based on the rising death rate in recent years, there will be 642 million diabetic patients in the world by 2040. This means that one in ten adults will have diabetes in the future. Without a doubt, this scary number needs a lot of attention. With the fast growth of machine learning, it has been used in many areas of medicine and health care. In this study, they tried to figure out who would get diabetes mellitus by using a decision tree, a random forest, and a neural network. The data set is made up of information about hospital physical exams in Luzhou, China. It has 14 characteristics. In this study, the models were looked at with five-fold cross validation. So that they could be sure that the methods could be used anywhere, they chose some of the best methods and did tests on them independently. As a training set, they chose at random the data of 68994 healthy people and 68994 diabetic patients. Due to the unevenness of the data, they randomly pulled out 5 times the data. And the result is the average of these five tests. In this study, they reduced the number of dimensions by using principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR). The results showed that when all the attributes were used, random forest predictions were most accurate (ACC = 0.8084).

Making a better model for predicting diabetes by using improved firefly feature selection and a hybrid random forest algorithm was carried out [13]. Diabetes is a long-term illness that kills a lot of people every year. If diabetes isn't treated, it can affect how well other organs in the body work. So, early detection is an important part of living a healthy life. Most of the time, the performance of classification is hurt by the high dimensionality of medical data. In this study, a system model is proposed for the Pima dataset to improve the accuracy of classification by getting rid of features that don't matter. Because of this, it is important to choose a feature selection method that can predict diseases more accurately than previous studies. So, Improved Firefly and a hybrid Random Forest algorithm were proposed as new ways to choose and classify features. With 96.3 percent accuracy, this study gives a better answer. The effectiveness of this study is compared to the effectiveness of the classification methods used before.

Diabetes was predicted using machine learning, and a smart web app was made for people with diabetes [14]. Diabetes is a very common illness that affects people all over the world. Diabetes makes you more likely to have long-term problems like heart disease and kidney failure, among other things. If this disease is caught early, people might live longer and better lives. Diabetes can be diagnosed early on with the help of different supervised machine learning models that have been trained with the right datasets. The goal of this work is to find good machine-learning-based classifier models that can use clinical data to find people who have diabetes. In this article, you can train Decision tree (DT), Naive Bayes (NB), k-nearest neighbor (KNN), Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR), and Support Vector Machine (SVM) with several datasets (SVM). They used good pre-processing techniques, such as label-encoding and normalization, to make the models more accurate. Also, they have found and ranked a number of risk factors by using different ways to choose features. Using two different sets of data, a lot of tests have been done to see how well the model works. When our model is put up against some recent research, the results show that the proposed model can be more accurate by 2.71 to 13.13 percent, depending on the dataset and the ML algorithm used. Finally, the most accurate machine learning algorithm is chosen for further development. They used the Python Flask Web Development Framework to add this model to a web application. Based on the results of this study, a good preprocessing pipeline for clinical data and ML-based classification may be able to accurately and quickly predict diabetes.

3. MATERIALS AND METHODS

3.1 Data Description

Seven predicting features and one target feature of PIMA Indians Diabetes dataset was used in this investigation. Diastolic blood pressure, the number of pregnancies a woman has had, plasma glucose concentration, the two-hour serum insulin, age of a patient, triceps skinfold thickness, and body mass index are seven of the most important factors. The UCI's repository contains 800 records with the source; https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes.

3.2 Firefly

Soft computing's most popular algorithm is the firefly algorithm[8. Two critical components go into the Firefly algorithm's construction: the varying light intensity and the specifics of how the attraction is structured. Lightness is used as a metric for determining whether or not fireflies find a particular thing attractive. As a result of a bioluminescence method, each of the 2000 present living firefly species has its own unique flashing pattern. These displays have two primary functions: to attract prey and to mate with other predators. FA is used in this work to solve the problem of feature selection. The feature selection procedure is made more efficient by adding the weight option to the basic firefly technique. This approach is applied on our Cleveland dataset to study the firefly's natural behavior. It is common practice to compute an intensity value based on light attraction and the fitness function in order to select the best features. To improve the outcome of this investigation, however, the intensity value is weighted.

3.3 K-Nearest Neighbor

While KNN is a machine learning algorithm that may be used for classification and regression, it is also capable of learning unsupervised [10][15]. As a lazy learner, it uses all of the data for classification and training. New data is compared to previously available data to see if there are any commonalities. Based on the similarity of the new data point, it is assigned a new classification. Rather than learning from a set of pre-trained data, the algorithm acts directly on the dataset itself. When new data is input into the algorithm, it is assigned to the category that is the most comparable to the previously stored data. Classification difficulties can be solved using the K-Nearest Neighbor (KNN) method, which is most commonly used in business to tackle regression problems. Simple translation and low processing time are its main advantages. Data points in a KNN are distanced using Euclidean distance functions.

3.4 Evaluation

The performance evaluation measures used in this study are; classification accuracy, sensitivity, specificity, precision, and F1-score, are used to examine the prediction findings.

There are a certain number of right predictions for each sample, and this is called the classification accuracy ratio. The number of correct guesses divided by the total number of forecasts is reported in this format:

As an output, it provides a matrix that characterizes the model's performance in detail. A True Positive; a True Negative; a False Positive; and so on.

A matrix's accuracy is easily figured out by averaging the values along the major diagonal line. the TP+TN/TP+TN+FP+FN formula is used. in which N is the total number of samples

Accuracy is measured by the F1-score. When comparing two variables, the Harmonic Mean (or F1-Score) is used. There is a zero-to-one range for F1 Score. It informs you both how exact and robust your classifier is. To get the best of both worlds, F1 Score aims to strike the right balance between accuracy and recall.

The number of correct positive findings divided by the number of projected positive results is called precision. It is the number of true positive outcomes divided by the total number of relevant samples.

4. RESULTS AND DISCUSSIONS

In this study, a firefly with KNN model is developed on a diabetic data, the firefly selects relevant features from the dataset. A Confusion Matrix was created as a result of the proposed strategy.

Table 1 shows the Firefly with K-nearest neighbor Confusion Matrix, which includes a TP, TN, FP, and FN for the Diabetic and Non-diabetic characteristics.

In addition, K-nearest neighbor machine learning classifier and a 10-fold cross validation technique was usedin analyzing the dataset of diabetic patients. There are a number of parameters that can be used to help classify signals. These include: f-measure, ROC, and AUC. Figure 1 provides an overview of the findings.

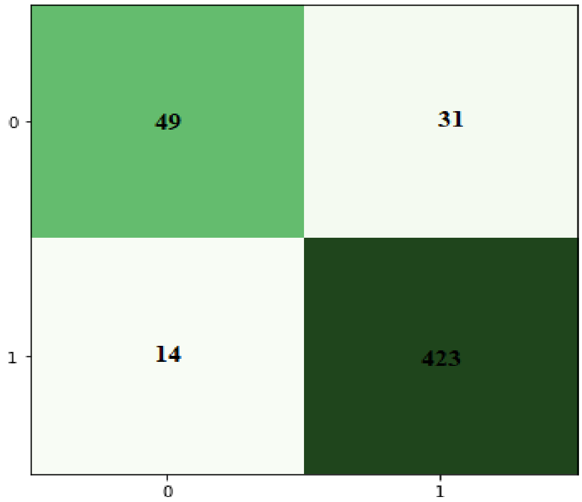

Figure 1: Confusion Matrix for the Developed Model

Table 1: Evaluation Performance of the Developed Model

Measure	**Value (%)**	**Derivations**
Sensitivity	77.78	TPR = TP / (TP + FN)
Specificity	93.17	SPC = TN / (FP + TN)
Precision	61.25	PPV = TP / (TP + FP)
Negative Predictive Value	96.80	NPV = TN / (TN + FN)
False Positive Rate	6.83	FPR = FP / (FP + TN)
False Discovery Rate	38.75	FDR = FP / (FP + TP)
False Negative Rate	22.22	FNR = FN / (FN + TP)
Accuracy	91.30	ACC = (TP + TN) / (P + N)
F1 Score	68.53	F1 = 2TP / (2TP + FP + FN)

Overweight, an unhealthy lifestyle, a heavy workload, and high levels of stress are all risk factors for developing diabetes mellitus.

This study tested firefly attribute selection method based on diabetesdataset obtainedfrom UCI database, and the findings are presented in this study. The KNN classifier was used to test the suggested method's classification performance. With the help of preprocessing approaches and optimization of key parameters of the lazy classifier, in this study a developed method for predicting diabetes utilizing firefly's k-nearest neighbor, a lazy algorithm is carried out. The proposed methodology yielded an accuracy of 91 percent, which is a significant improvement.

5. CONCLUSION

Diabetes is a chronic disease that is increasingly affecting the world's population and early detection of diabetes is very essential for its successful treatment. In this study, a diabetic prediction model which employed machine learning technique was used for predicting diabetes disease in patients. Experiment on various existing works on diabetic prediction have shown the need for improvement of classification accuracy of classifiers and selection of right feature subsets. This research experimented Firefly feature selection algorithm and KNN classifier for the development of a diabetic prediction model which yielded classification accuracy of 91% based on PIMA Indian diabetes dataset. For future work, this research intend to use other types of diabetes dataset.

6. REFERENCES

[1] G. D. Kalyankar, S. R. Poojara, and N. V. Dharwadkar, "Predictive analysis of diabetic patient data using machine learning and Hadoop," Proc. Int. Conf. IoT Soc. Mobile, Anal. Cloud, I-SMAC 2017, pp. 619–624, 2017, doi: 10.1109/I-SMAC.2017.8058253.

[2] Y. Jian, M. Pasquier, A. Sagahyroon, and F. Aloul, "A Machine Learning Approach to Predicting Diabetes Complications," Healthcare, vol. 9, no. 12, p. 1712, Dec. 2021, doi: 10.3390/healthcare9121712.

[3] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," IEEE Access, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[4] K. Kantawong, S. Tongphet, P. Bhrommalee, N. Rachata, and S. Pravesjit, "The Methodology for Diabetes Complications Prediction Model," 2020 Jt. Int. Conf. Digit. Arts, Media Technol. with ECTI North. Sect. Conf. Electr. Electron. Comput. Telecommun. Eng. ECTI DAMT NCON 2020, pp. 110–113, 2020, doi: 10.1109/ECTIDAMTNCON48261.2020.9090700.

[5] S. M. D. A. C. Jayatilake and G. U. Ganegoda, "Involvement of Machine Learning Tools in Healthcare Decision Making," J. Healthc. Eng., vol. 2021, pp. 1–20, Jan. 2021, doi: 10.1155/2021/6679512.

[6] C. Toh and J. P. Brody, "Applications of Machine Learning in Healthcare," in Smart Manufacturing - When Artificial Intelligence Meets the Internet of Things, IntechOpen, 2021.

[7] R. Saxena and S. Kumar Sharma Manali Gupta, "Role of K-nearest neighbour in detection of Diabetes Mellitus," Turkish J. Comput. Math. Educ., vol. 12, no. 10, pp. 373–376, 2021.

[8] R. Haritha, D. S. Babu, and P. Sammulal, "A hybrid approach for prediction of type-1 and type-2 diabetes using firefly and cuckoo search algorithms," Int. J. Appl. Eng. Res., vol. 13, no. 2, pp. 896–907, 2018.

[9] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," ProcediaComput. Sci., vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.

[10] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods,"

ProcediaComput. Sci., vol. 167, pp. 706–716, 2020, doi: 10.1016/j.procs.2020.03.336.

[11] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," BMC Endocr. Disord., vol. 19, no. 1, p. 101, Dec. 2019, doi: 10.1186/s12902-019-0436-6.

[12] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," Front. Genet., vol. 9, Nov. 2018, doi: 10.3389/fgene.2018.00515.

[13] B. Senthil Kumar and R. Gunavathi, "An enhanced model for diabetes prediction using improved firefly feature selection and hybrid random forest algorithm," Int. J. Eng. Adv. Technol., vol. 9, no. 1, pp. 3765–3769, 2019, doi: 10.35940/ijeat.A9818.109119.

[14] N. Ahmed et al., "Machine learning based diabetes prediction and development of smart web application," Int. J. Cogn. Comput. Eng., vol. 2, pp. 229–241, Jun. 2021, doi: 10.1016/j.ijcce.2021.12.001.

[15] Arowolo, M. O., Adebiyi, M. O., Adebiyi, A. A., &Olugbara, O. (2021). Optimized hybrid investigative based dimensionality reduction methods for malaria vector using KNN classifier. Journal of Big Data, 8(1). https://doi.org/10.1186/s40537-021-00415-z.