INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

Methods of Machine Learning for Diabetes Prediction in the Early Stages

Shaik Mohammedjany[1], Dr.P.Dillep Kumar Reddy[2], Dr. G.Syam Prasad[3]
[1]Research scholar,Dept of CSE, Glocal University, Uttar Pradesh, India
[2]Professor, Dr.P.Dillep Kumar Reddy , Glocal University,Uttar Pradesh, India.
[3]Professor HOD,Dept of IT,Narasaraopeta Engineering College ,Guntur,A.P, India.

Abstract

Type 2 diabetes is rapidly becoming an epidemic in the United States. The lives of the millions of individuals who suffer with diabetes are profoundly impacted by the disease. In most cases, persons with diabetes go undiagnosed for years, despite the fact that early detection may lessen the disease's consequences and enhance patients' quality of life. Machine learning techniques used to existing patient data may aid in making accurate, timely diagnoses. Diagnosis may be made rapidly without the need for a blood test or glucose screening. If someone is diabetic or at risk for developing diabetes, they might find out by answering a short series of questions. Machine learning methods are used to diagnose diabetes in the proposed research. In this context, we build our models on top of a publically accessible diabetes dataset that has 16 characteristics gathered from 520 individuals. The dataset was run using eight different machine learning techniques. A 10-fold cross validation schema was used to verify the accuracy of the model outputs. Other performance indicators, including precision, recall, and f1 score, based on the confusion matrix were also presented. High levels of accuracy were achieved by every model that was developed. Using Naive Bayes, a simple machine learning method, we found that 88.85% was the lowest acceptable accuracy. Using a one-dimensional convolutional neural network model, we were able to achieve a 99.04% accuracy rate. The constructed Convolutional Neural Network model achieved the maximum performance across all measures, including a precision score of 100.00%, a recall score of 98.63%, and a f1 score of 99.31%. These results suggest that the developed 1D CNN model may be put to use in the identification of diabetes patients with little intrusion into patient privacy by use of a small number of screening questions.

Keywords: Diabetes, Machine Learning, K Nearest Neighbor, Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, XGBoost, Artificial Neural Network, Convolutional Neural Network.

1. Introduction

Diabetes is a chronic disease that causes irregularities of the blood sugar level. It damages many parts of the body such as heart, eyes, blood vessels and nerves. There are two types of diabetes. Type 1 diabetes is a condition in which the pancreas does not produce enough insulin. In type 2 diabetes, not only the lack of insulin, but also the problems about the insulin resistance are the major issues critically affecting human life. According to the reports of World Health Organization (WHO, n.d.), more than 400 million people are diabetic in the world population and 1.6 million deaths are recorded due to the diabetic issues each year.

ISSN: 2057-5688

Additionally, the number of diabetic patients is currently increasing at a growing rate every year when compared to earlier years. Therefore, the early stage detection for diabetics has crucial importance for human life.

It is proven that diabetes may be present for 4-12 years before diagnosis (Harris et al, 1992). When diagnosis is made, diabetesrelated damage occurs in half of the patients. Researchers prove that early detection of diabetes will help to prevent heart diseases, blindness, vascular complications (Ramachandran & Chamukuttan, 2008) and stroke, kidney failure and limb amputations (U.S. Department of Health & Human Services, 2004). So, early diagnosis is very important to reduce the effects of the disease and improve life quality of patients.

Machine learning is a sub-branch of computer science and artificial intelligence which focuses on the use of data to create predictive models for various problems (IBM Cloud Education, 2020). It can be used for analyzing data and retrieving critical information. Moreover, the estimation and prediction for the future according to the presented data is possible by using the machine learning techniques for many different fields. Similarly, these techniques can be also applied in the medical field for information retrieving and pattern understanding from existing data and predict the future trends of health problems. For instance, some symptoms like polyuria, polydipsia, obesity, itching, polyphagia and delayed healing are known as the early symptoms of diabetes. However, healthy people can also have some of these symptoms. By using machine learning techniques, predicting if a person is healthy or diabetic is possible. In the presented study, the prediction of diabetes stages is aimed to estimate with a high accuracy rate.

In this study, “early stage diabetes risk prediction dataset” obtained from the UCI Machine Learning Repository has been used in the evaluation of techniques. In the literature, several studies focused on this dataset have been found. Oladimeji et. al analyzed the data set to define the most essential attributes for classification step (Oladimeji et. al, 2021). Several attribute evaluator approaches are applied such as SymmetricalUncert Attribute Evaluator, GainRatio Attribute Evaluator, InfoGain Attribute Evaluator, Correlation Attribute Evaluator with the ranker search method. It is found that obesity, delayed healing and itching were redundant features so, these features are deleted from the dataset. As machine learning techniques, Random forest, Knearest neighbor, regular Decision Tree and Naive Bayes methods are applied by using a machine learning toolbox named ad WEKA. Results are evaluated and discussed by using a 10 fold cross validation schema. The highest accuracy achieved at this study is 98.31% by using random forest.

In another study on the same dataset, a feature selection algorithm is implemented to optimize the Multilayer Perceptron Classifier by reducing the number of input features (Le et. al, 2021). Grey Wolf Optimization and Adaptive Particle Swarm Optimization were utilized in the optimization step. The results are compared to several well-known machine learning methods such as decision tree, support vector machine,

ISSN: 2057-5688

k-nearest neighbor, naïve bayes, random forest and logistic regression. An 80:20 ratio training and testing split ratio is used for model evaluation and 97% accuracy is achieved by using the proposed optimization based Multilayer Perceptron Classifier.

Sadhu and Jadli compared the seven classical machine learning models by using the same dataset in another study (Sadhu and Jadli, 2021). Accuracy of 98.08% and ROC score of 99.79% are reported by using random forest method. Oleiwi et. al applied 10 fold cross-validation over the radial basis function network model which resulted in 98.80% accuracy and 100% sensitivity scores (Oleiwi et. al, 2020). In another research, a test accuracy of 98,08% is reported by implementing the K-nearest neighbor algorithm on the same dataset (Bilgin, 2021). Additionally, several more classical machine learning methods such as multilayer perceptron, decision tree, ensemble learning algorithms, support vector machine and linear discriminant analysis were also used and a diabetes early diagnosis kit was developed. Different from the classical machine learning application, a time domain specific deep learning model, LSTM, is also tested over the same data set by Özer (Özer, 2020). Özer reported an average F1 score of 98.9% by implementing LSTM networks and validated the result of the created model over 10fold cross validation. There are also other diabetic related datasets in literature. Nahzat and Yağanoğlu implemented random forest over different diabetic dataset. They reported an 88.31% accuracy score (Nahzat and Yağanoğlu, 2021).

In this paper, eight machine learning techniques are explained and used to create a predictive model for diabetes. Different from the previous studies, a 1D CNN model has been designed and tested over the dataset. The evaluated results are compared in terms of different metrics. Since early diagnosis helps reduce the effects of diabetes, this work is expected to be helpful for healthcare.

2. Material and Method

2.1. Dataset Information

Early stage diabetes risk prediction dataset (UCI Machine Learning Repository, 2020) is prepared by Islam et al. (Islam et al, 2020). Data have been collected from the records of the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. There are 520 instances and 16 features that are related to diabetes. 15 attributes are categorical and 1 attribute is continuous. There are some attributes that are terms of medicine so, these attributes are explained below.

- Polydipsia is excessive thirst and is one of the first symptoms of diabetes (Hickman, 2020).
- Unexplained weight loss is when someone drops a significant amount of weight without a change in diet or exercise. It can occur in people who have type 2 diabetes, but it is more common at type 1 (Berkley, 2021).

ISSN: 2057-5688

- Polyphagia is excessive hunger that increases appetite significantly and persistently. It is one of the main symptoms of diabetes (Jones, 2021).
- Thrush is a yeast infection (candida albicans). High sugar levels supply better conditions for candida albicans to grow (Thrush, 2019).
- Blurred vision means the loss of sharpness of vision and it makes it impossible to see fine details. Instability of blood sugar is known as a reason for blurred vision (Coelho, 2021).
- Paresis is the condition of weakness of voluntary movement (Petrie, 2021). It can be a symptom of diabetes.
- Muscle stiffness is the inability of the muscles to relax normally. It can affect any part of the body and it causes difficulty of moving (Cirino, 2019).
- Diabetic patients are more likely to have alopecia areata. Alopecia leads to hair loss on any part of the body (Watson, 2018).

Since there are no missing values at the dataset, encoding categorical values was enough for dataset preparation. Categorical values which are yes/positive are denoted by 1 and no/negative are assigned as 0 in the analysis.

2.2. Methodology

In the presented study, classical and advanced machine learning methods are applied in order to predict diabetes in the early stages. To determine the most convenient technique which gives the highest classification performance result especially for the diabetic prediction purpose, eight well known machine learning techniques are applied to the same dataset and the results are compared with the same metrics. These selected machine learning methods are decision tree, random forest, support vector machine, XGBoost, K-nearest neighbor, Naive Bayes, artificial neural network and convolutional neural network. Results are evaluated in terms of accuracy, precision, recall and f score by applying cross validation schema. Methodology of this research is illustrated in Figure 1 and implemented methods are explained in the subsections.

2.2.1. Decision Tree

Decision tree is a machine learning method that visualizes how the created model predicts data. It builds a tree in which nodes of the tree represent features, branches represent which direction must be taken after each node and leaves represent predictions. Classes of given data can be predicted by traversing from root to leaves by choosing regarding branches. Decision tree also shows the importance of features, the most important and elective feature takes place at the root node. In the presented study, models have been created by using Gini Information Gain with two-level pruning settings.

ISSN: 2057-5688

2.2.2. Random Forest

When a part of the decision tree is built incorrectly, it causes the model to make false predictions. Random forest is a machine learning technique that aims to solve this overfitting issue. In this approach, predictions of several randomly created decision trees are combined and the most voted label is returned as a label of given data. The voting process over the decision of multiple tree models is illustrated in Figure 2 (Ampadu, 2021).

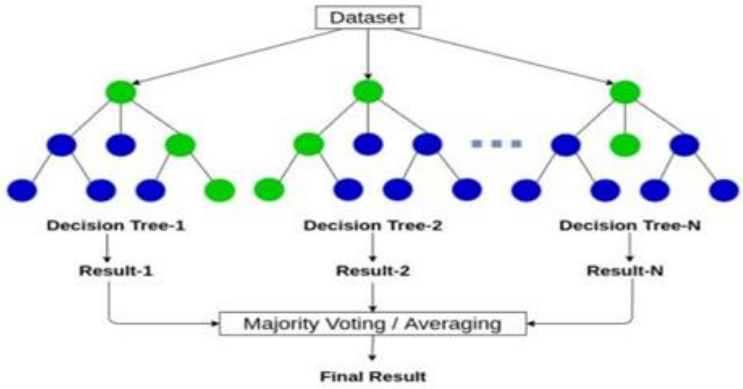

Figure 1: Voting Process at Random Forest

2.2.3. Support Vector Machine

Support vector machine maps each instance in a space and divides that space into hyperplanes. Each hyperplane represents a class and classification is done by mapping each data. Training time and cost may be too much for large size datasets so it is better to use SVM for small datasets. In this study, a polynomial kernel with a degree of 3 and a regularization parameter as 0.1 is implemented.

2.2.4. XGBoost

Extreme gradient boosting is a tree based machine learning framework, which starts by building weak models and ends up with a strong model. This process is done parallelly by adding more nodes to decision trees by considering the gradient of loss function. When classifying an instance, the result of each tree is considered and the most voted result is returned as the output of the model.

2.2.5. K- Nearest Neighbour

K nearest neighbor is a simple yet effective machine learning algorithm. Training data is represented in a graph and an assumption that examples of the same classes will be closely positioned. When predicting the label of an instance, position of that instance at graph is determined by using its features and the k neighbors that are closest to that point is found. Labels of these neighbors are considered and prediction of the model is returned as the most seen label among neighbours.

ISSN: 2057-5688

2.2.6. Gaussian Naïve Bayes

Naive Bayes is a machine learning algorithm that is based on Bayes theorem. It makes an assumption that all attributes are independent so it does not produce good results when the dataset size is large and it has a lot of features. Gaussian Naive Bayes is a variant of Naive Bayes that implements Gauss normal distribution.

2.2.7. Artificial Neural Network

Artificial neural networks aim to imitate the functions of the human brain to solve complex problems. It is the process of creating a network that includes nodes and connections to make predictions. For initialization, random weights will be given to each connection and weights will be updated by calculating the loss of train data. To make predictions for a problem that has n classes, n nodes are set to the output of the network and each output of those n nodes represent the likelihood of given data to be related to that class.

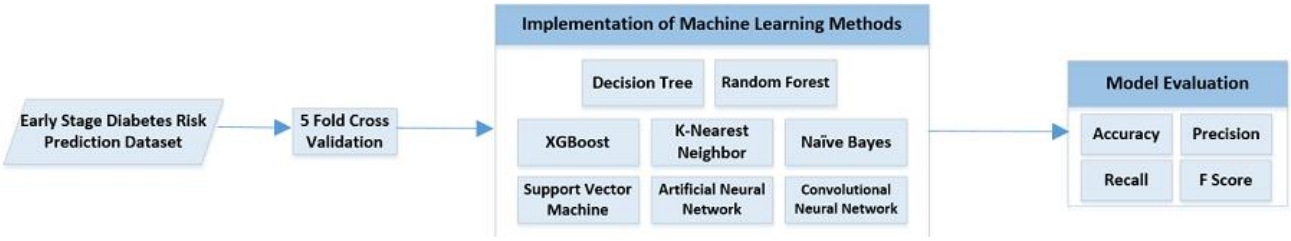

Figure 2: The Flowchart of the Study

2.2.8. Convolutional Neural Network

Convolutional neural network is a deep learning method that filters input data before feeding into the neural network. It is beneficial for reducing input shape by eliminating unrelated attributes. Convolution, pooling and flatten layers are applied then output of these layers are fed into dense layers. It is often used for images, but it can also be applied to numerical data. While 2 dimensional convolutional neural network architecture can be used for images, 1 dimensional convolutional neural network architecture can be used for numerical data. In this study, 1 dimensional convolutional neural network having 2 convolutional layers was applied over the classification of diabetic dataset. Convolution layers have 8 and 4 filters. A dropout layer was assigned to the output of convolution layers for avoiding overfitting, then pooling and flattening layers were located at the end of the 1D CNN architecture. These results are fed into a neural network which has 500 neurons and a sigmoid output function.

2.3. Performance Metrics

Several metrics to calculate performance of machine learning techniques have been used in the proposed study. These metrics were calculated over the confusion matrix.

ISSN: 2057-5688

In the confusion matrix, True positive (TP) is the count of positive labeled data that are correctly classified, true negative is the count of negative labeled data that are correctly classified, false positive is the count of negative labeled data that are classified as positive by mistake and false negative is the count of positive labeled data that are classified as positive by mistake. Mostly used performance metrics are accuracy, recall, precision and f-score.

Accuracy is a performance metric that is the ratio of correctly classified data over all data. It is a commonly used metric but it does not give detailed information about model performance. Precision is the ratio of true positive over all of the positive classified data. For diabetes classification, precision shows the ability of the model to predict patients and not labeling healthy people as patients. Recall is the ratio of true positive over all of the positive data. For diabetes classification, it shows how many patients can be detected by the model. F score is the harmonic mean of precision and recall so it is a strong metric to calculate model performance. In the calculation of F-Score, precision and recall is denoted as P and R, respectively. All the formulas are given below:

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

$$\text{Precision} = TP/(TP+FP) \quad (2)$$

$$\text{Recall} = TP/(TP+FN) \quad (3)$$

Model performance may also be shown in a confusion matrix. Confusion matrix is an n x n matrix that n denotes the number of labels of a given dataset. Each row represents actual labels and each column represents predicted labels. Confusion matrix shows the performance of the model as illustrated in figure 3.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 3: Standard Form of Confusion Matrix

For evaluating the techniques over the dataset, these metrics have been calculated by applying 5 fold cross validation. When k fold cross validation is applied, the dataset will be divided into k parts. Training process will be done k times and at each time k-1 folds are used for training and 1 fold is used for testing. This approach prevents unbalanced data problem so model metrics can be measured in a more trustworthy way (Hawkins, Subhash & Mills, 2003).

ISSN: 2057-5688

Method	Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)
K-Nearest Neighbor	89.42	88.92	90.34	89.09
Decision Tree	95.58	95.17	95.58	95.34
Random Forest	97.69	97.69	97.53	97.57
Support Vector Machine	94.62	94.18	94.51	94.31
Naïve Bayes	88.85	88.30	88.11	88.16
XGBoost	97.89	97.80	97.77	97.76
Artificial Neural Network	92.31	91.82	92.40	92.07
Convolutional Neural Network	**99.04**	**100.00**	**98.63**	**99.31**

Table 1. The Classification Performances of the Utilized Machine Learning Techniques for the Diabetes Risk Prediction Dataset

3. Results and Discussion

All of the machine learning models that are described above are implemented for diabetes risk prediction dataset and the results in terms of accuracy, precision, recall and f-score are presented in Table 1. Results are validated by using a 5-fold cross validation schema in which training and testing sets were arranged as 416 and 104 samples for each fold. General performance scores listed in Table 1 were calculated by averaging the fold results.

According to the results, the highest classification accuracy is achieved by 1 dimensional convolutional neural network. An accuracy of 99.04% has been measured for the classification of the dataset.

In terms of precision, convolutional neural network achieved 100% performance. This shows that when the model created by using convolutional neural network classifies a sample as patient, the possibility of it being healthy is very low. Since a health issue is the topic of this research, recall is also a very important metric. Determining a healthy person as a patient can be tolerated in some cases but diagnosing a diabetic patient as healthy may cause serious issues. In terms of recall, all of the implemented machine learning models resulted in a satisfactory level. The highest recall score is measured as 98.63% when using the designed 1 dimensional convolutional neural network over the dataset. Also, the F-score metric should be considered to balance precision and recall scores. The highest F-score is 99.31% and it is also obtained by using the same designed 1 dimensional convolutional neural network. By considering these results, the convolutional neural network model is the most successful model to predict diabetes.

While the proposed convolutional neural network achieved the maximum classification accuracy as shown in Table 1, the lowest performance is obtained by using Naïve Bayes method. All of the performance metrics are measured below 90%. Since the features of the dataset are not independent, assumption of

independent features (Rish, 2001) may lead to the obtained low performance. This model may misinterpret given samples and label them wrong.

In order to define the most informative features of the dataset, XGBoost and Decision Tree models were investigated. After the training phase of XGBoost, the importance of each attribute to the performance of classification is illustrated in Figure 4.

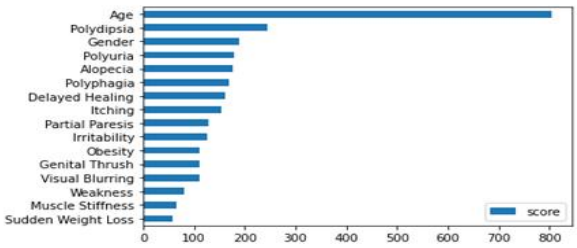

Figure 4: The importance of features for XGBoost Model

According to the XGBoost classification performance which resulted in 97.89%, 97.80% and 97.77% in accuracy, precision and recall scores, respectively, age of the patients have crucial importance in the classification. Most patients have been correctly classified as diabetic or non-diabetic by using "Age" information in the dataset. Then, the polydipsia which indicates the status of excessive thirst condition of patients is the second crucial parameter in the definition of diabetes. The least important features were determined as sudden weight loss and muscle stiffness situations of the patients as those can happen because of many other reasons such as stress or excessive sportive activity.

Decision tree is another easy understandable technique with its outcomes. After the pruning, the tree that is created for diagnosis of diabetes is illustrated in Figure 5. It shows that the most important feature is polyuria for the created decision tree model.

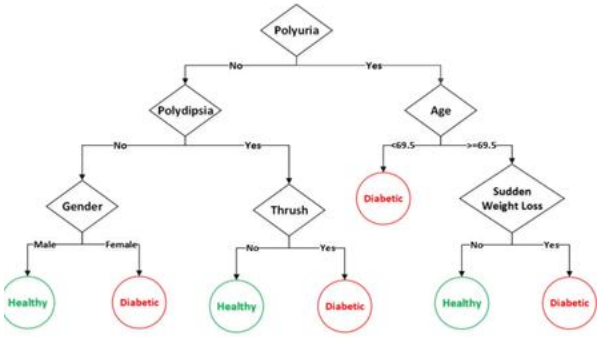

Figure 5: Importance of features for Decision Tree Model

ISSN: 2057-5688

4. Conclusions and Recommendations

In this research, eight machine learning methods are applied to the early stage diabetes risk prediction dataset. The results are compared in terms of performance metrics such as accuracy, recall, precision and f-score. The most successful model is the designed 1 dimensional convolutional neural network model, and it has an accuracy of 99.04% over data set over the 5 fold cross validation schema. In literature, there was no research that XGBoost or Convolutional Neural Networks applied to the early stage diabetes risk prediction dataset. The results explained in this paper show that both of these two methods perform satisfactorily in detecting early stage diabetes risk. Since evaluated metrics are high, further work of creating an early stage diabetes risk prediction application may be done by using the created Convolutional Neural Network model.

References

[1]. Ampadu, H. (2021, May 01). Random Forests Understanding. AI Pool. https://ai-pool.com/a/s/random-forests-understanding

[2]. Berkley, C. (2021, May 18). How Is Rapid Weight Loss Related to Diabetes. Verywell Health.

[3]. https://www.verywellhealth.com/rapid-weight-loss-5101064

[4]. Bilgin, G. (2021). Makine Öğrenmesi Algoritmaları Kullanarak Erken Dönemde Diyabet Hastalığı Riskinin Araştırılması. Zeki Sistemler Teori ve Uygulamaları Dergisi, 4(1), 55-64. https://doi.org/10.46387/bjesr.790225

[5]. Cirino, E. (2019, July 6). What Causes Muscle Rigidity. Healthline. https://www.healthline.com/ health/ musclerigidity

[6]. Coelho, S. (2021, April 28). What Is Blurred Vision. Verywell Health. https:// www.very wellhealth.com/blurred-vision-5114184

[7]. Draelos, R. (2019). Measuring Performance: The Confusion Matrix. Glass Box Medicine. https://glassboxmedicine.com/2019/02/17/measuringperformance-the-confusion-matrix/

[8]. Harris, M. I., Klein, R., Welborn, T. A. & Knuiman, M. W. (1992). Onset of NIDDM occurs at least 4–7 yr before clinical diagnosis. Diabetes Care, 15(7), 815-819. DOI: 10.2337/diacare.15.7.815

[9]. Hawkins, D. M., Subhash, C. B. & Mills, D. (2003). Assessing Model Fit by Cross-Validation. Journal of Chemical Information and Computer Sciences, 43(2), 579–586. https://doi.org/10.1021/ci025626i

[10]. Hickman, R. J. (2020, July 28). What Is Polydipsia. Verywell Health. https://www.very wellhealth.com/polydipsia-4783881

[11]. IBM Cloud Education. (2020, July 15). What is machine learning. IBM. https://www.ibm.com/ cloud/learn/machine-learning Islam,

[12]. M. M., Ferdousi, R., Rahman, S. & Bushra, H. Y. (2020). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. Computer Vision and Machine Intelligence in Medical Image Analysis, 113-125. DOI:10.1007/ 978-981-13-8798 2_12

[13]. Jones, H. (2021, April 5). Causes of Polyphagia. Verywell Health. verywellhealth. com/ polyphagia-5114624

[14]. Le, T. M., Vo, T. M., Pham, T. N. & Dao, S. V. T. (2020). A Novel Wrapper–Based Feature Selection for Early Diabetes Prediction Enhanced With a Metaheuristic. IEEE Access, 9, 7869-7884.DOI:10.1109/ACCESS.2020.3047942

[15]. Nahzat, S ,Yağanoğlu, M . (2021). Diabetes Prediction Using Machine Learning Classification Algorithms . Avrupa Bilim ve Teknoloji Dergisi , Ejosat Özel Sayı 2021 (ARACONF) , 53-59 . DOI: 10.31590/ ejosat.899716

[16]. Oladimeji, O. O., Oladimeji, A. & Oladimeji, O. (2021). Classification Models for Likelihood Prediction of Diabetes at Early Stage Using Feature Selection. Applied Computing and Informatics. https://doi.org/10.1108/ACI-01-2021-0022

[17]. Oleiwi, A. K., Shi, L., Tao, Y. & Wei, L. (2020). A Comparative Analysis and Risk Prediction of Diabetes at Early Stage using Machine Learning Approach. International Journal of Future Generation Communication and Networking, 13(3), 41514163.

[18]. Özer, İ. (2020). Uzun Kısa Dönem Bellek Ağlarını Kullanarak Erken Aşama Diyabet Tahmini. Mühendislik Bilimleri ve Araştırmaları Dergisi, 2(2), 50-57. https://doi.org/ 10. 38016/jista.877292

[19]. Petrie, T. (2021, June 07). What Is Paresis. Verywell Health. https://www.verywellhealth .com/paresis-5184820

[20]. Ramachandran, A. & Chamukuttan, S. (2008). Early Diagnosis and Prevention of Diabetes in Developing Countries. Reviews in Endocrine and Metabolic Disorders, 9(3), 193-201. DOI: 10.1007/s11154-008-9079-z

[21]. Rish, I. (2001). An Empirical Study of the Naïve Bayes Classifier. IJCAI Workshop on Empirical Methods in AI, 3(22). 41-46.

[22]. Sadhu, A. & Jadli, A. (2021). Early-Stage Diabetes Risk Prediction: A Comparative Analysis of Classification Algorithms. International Advanced Research Journal in Science, Engineering and Technology (IARJSET), 8(2), 193- 201. DOI: 10.17148/IARJSET.2021.8228

[23]. Thrush. (2019, January 15). Diabetes UK. https:// www.diabetes.co.uk/diabetes compli cations/diabetesand-yeast-infections.html.

ISSN: 2057-5688

[24]. UCI Machine Learning Repository. (2020, July 12). Early stage diabetes risk prediction dataset. https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes +risk+prediction+dataset.

[25]. U.S. Department of Health & Human Services. (2004, January 12). Diabetes: A National Plan For Action. The Importance Of Early Diabetes Detection. https://aspe.hhs. gov/ report/diabetes-national-planaction/importance-early-diabetes-detection

[26]. Watson, S. (2018, September 29). Does Diabetes Cause Hair Loss. Healthline. https://www .healthline. com/health/doesdiabetes-cause-hair-loss

[27]. WHO. (n.d.). Diabetes. Retrieved July 15, 2021, from https://www.who.int/health-topics/diabetes Wood, T. (n.d.). What is a Random Forest. DeepAI. Retrieved August 01, 2021, from https://deepai.org/machine-learningglossary-and-terms/random-forest