Analyze Important Features of PIMA Indian Database For Diabetes Prediction Using KNN

Aziz Perdana[1]*, Arief Hermawan[2], Donny Avianto[3]
Magister Teknologi Informasi[1], [2], [3]
Universitas Teknologi Yogyakarta
Sleman, Indonesia
azizperdana@gmail.com[1], ariefdb@uty.ac.id[2], donny@uty.ac.id [3]

***Abstract*— Diabetes is a chronic, non-communicable disease, and a long-term health condition that affects how the body uses glucose, the type of sugar that gives energy. In Indonesia, diabetes ranks as the sixth highest cause of death, following conditions related to childbirth. In 2021, Indonesia has a total of 19.5 million diabetes patients, making it the fifth-highest in the world. Some machine learning research has used data from the PIDD (PIMA Indian Diabetes Dataset) to predict diabetes. In this research, in addition to prediction accuracy, data complexity is also important. This research analyzes important features in the PIMA Indian database using the KNN (k-nearest neighbor) method for classification. The results show that using KNN with k=22 value results in the highest accuracy of 83.12%. The analysis also found that the important features required by the KNN method to achieve high accuracy from the PIMA Indian database, in order of importance, are glucose, age, insulin, blood pressure, Body Mass Index, pregnancy, skin thickness, and diabetes pedigree function. However, when used in the KNN classification method, the diabetes pedigree function feature was found to be unnecessary, not relevant, and can be reduced.**

***Keywords*— diabetes prediction, knn, pidd, importance features, machine learning**

***Abstrak*—Diabetes adalah penyakit kronis, tidak menular, dan kondisi kesehatan jangka panjang yang memengaruhi cara tubuh menggunakan glukosa, jenis gula yang memberi energi. Di Indonesia, diabetes menempati urutan keenam penyebab kematian tertinggi, menyusul kondisi terkait persalinan. Pada tahun 2021, Indonesia memiliki total 19,5 juta penderita diabetes, menjadikannya urutan kelima terbanyak di dunia. Beberapa penelitian pembelajaran mesin telah menggunakan data dari PIDD (PIMA Indian Diabetes Dataset) untuk memprediksi diabetes. Dalam penelitian ini, selain akurasi prediksi, kompleksitas data juga penting. Penelitian ini menganalisis fitur-fitur penting dalam database PIMA Indian menggunakan metode KNN (k-nearest neighbor) untuk klasifikasi. Hasil penelitian menunjukkan bahwa penggunaan KNN dengan nilai k=22 menghasilkan akurasi tertinggi sebesar 83,12%. Analisis juga menemukan bahwa fitur penting yang diperlukan oleh metode KNN untuk mencapai akurasi tinggi dari database PIMA India, berdasarkan urutan kepentingannya, adalah *glucose, age, insulin, blood pressure, Body Mass Index, pregnancy, skin thickness,* dan *diabetes pedigree function*. Namun, ketika digunakan dalam metode klasifikasi KNN, fitur *diabetes pedigree function* ditemukan tidak diperlukan, tidak relevan, dan dapat dihilangkan pada saat proses pembelajaran maupun prediksi.**

***Kata Kunci*—prediksi diabetes, knn, pidd, fitur penting, mesin pembelajaran**

I. INTRODUCTION

As demonstrated by Figure 1, the application of AI (Artifficial Intelligence) and ML (Machine Learning) in medical devices is increasing rapidly. In 2021, FDA or US Food and Drug Administration already approved 100 AI and ML-based commercial health devices as reported in [1]. This is a significant increase compared to the number of commercial health devices that equipped with AI and ML that were approved by the FDA in 2017, which was only 26. Out of the 100 commercial health devices equipped with AI and ML that are approved by the FDA, the majority are utilized in the fields of radiology and heart health. Additionally, there are 5 devices specifically for diabetes, as reported in [2]. In Indonesia, diabetes is the 6th leading cause of death, with 19.5 million people affected in 2021. This ranks Indonesia as having the 5th highest diabetes patient population globally [3]. The use of AI and ML in commercial health devices for diabetes can help to improve the diagnosis and management of this disease, which can have a substantial effect on the overall health of the population. The increasing number of AI and ML-based commercial health devices approved by the FDA in recent years suggests that this trend will persist in the future.

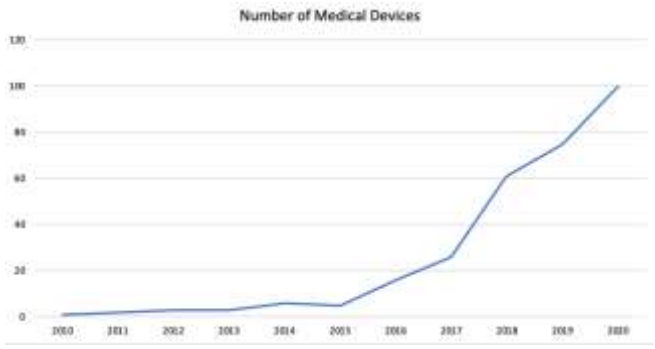

Fig. 1. Research stages

Research on AI and ML regarding diabetes, mostly use PIDD. PIDD is data information from 768 women patients aged 21 years or older from the Phoenix, Arizona, United States area

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i1.1598, Copyright ©2023
Submitted : January 25, 2023, Revised : March 6, 2023, Accepted : March 7, 2023, Published : March 13, 2023

[4]. There have been many studies related to PIDD that aim to compare methods that can produce high accuracy. In terms of prediction, in addition to accuracy, the performance of calculations in training also needs to be improved.

There are several datasets commonly used in machine learning research for diabetes, such as (1) the NHANES or National Health and Nutrition Examination Survey, a dataset that contains information on diabetes and other health conditions, obtained by CDC or the Centers for Disease Control and Prevention [5], (2) Global Health Observatory (GHO) data, a dataset belong to World Health Organization (WHO) that contains information on diabetes prevalence and other health indicators for countries around the world [6] (3) the ELSA (English Longitudinal Study of Ageing) database [7], (4) the Diabetes Data Set of 130-US hospitals from years 1999 to 2008, which contains over 100,000 hospital visits for diabetes and includes information on patient demographics, diagnosis, medications, and hospital outcomes [8], (5) The Framingham Heart Study, a long-term study that has collected data on cardiovascular disease risk factors, including diabetes, in a large population sample [9, 10], (6) the German Diabetes Risk Score (DRS) dataset from the German National Cohort (NAKO Health Study) that contains information on diabetes risk factors such as age, sex, and weight, as well as lab test results and other health information [11], (7) The Pima Indian Diabetes dataset, which contains data from over 800 patients of Pima Indian heritage with diabetes [12], (8) Early Classification of Diabetes, a dataset that comprises of 520 observations, including 17 characteristics that are obtained from the Bangladesh patients at the Sylhet Diabetes Hospital through direct questionnaires and diagnosis results [13, 14], (9) The National Diabetes Data Group (NDDG) dataset, which contains data from over 1,200 patients with diabetes and (10) the Hospital Frankfurt Germany Diabetes Data Set [15]. These datasets can be found on different sources, such as UCI Machine Learning Repository, Kaggle, and from the institutions that collected the data.

II. Related Work

A. Research to Increase Accuracy and Compare Accuracy Results with Different Classification Methods

Research in this category was carried out using the Standard Deviation K Nearest Neighbour Classifier with a training data to test data composition of 9:1, resulting in an average accuracy of 83.76% [16]. Nada Ali compared the level of prediction accuracy of PIMA data generated by Logistic Regression (accuracy=81.2%), K-Nearest Neighbour (accuracy=79.1%), Support Vector Machine (accuracy=83%), Random Forest (accuracy=81.4%) and Naive Bayes (81.1%) [17]. Nada Ali used a training data to test data composition of 7:3. The highest level of accuracy from the research conducted by Nada Ali is 83% which was generated by the Support Vector Machine (SVM) method.

B. Research Related to Dimensionality Reduction

Research related to dimensionality reduction, one of which was carried out by Shamriz Nahzat using KNN by adding 2 new features (blood pressure greater than 80 and glucose level greater than 105) outside of the 8 features in the PIDD, using only the PIDD features, and by removing the skin thickness and diabetes Pedigree Function features resulting in accuracy in sequence of 81%, 82%, 83% [18]. The comparison of training data with test data carried out by Shamriz is 70:30. Thammi Reddy compared the Support Vector Machine classification method (accuracy with PCA=68.8% and accuracy without PCA=68%), Naive Bayes (accuracy with PCA=77.36% and accuracy without PCA=76%), and Decision Tree (accuracy with PCA=76.22% and accuracy without PCA=75%), with the combination of Naive Bayes and PCA obtaining the highest accuracy, which is 77.36% [19].

C. Research on Important Features in the PIMA Indian Database

Research on important features in the PIMA Indian database has been carried out by Choudhary with a training data to test data ratio of 85:15 using the Logistic Regression, Support Vector Machine, Random Forest (300 forest) methods resulting in accuracy in sequence of 73%, 75%, and 88% [20]. Choudhary also determined the important features of the PIMA Indian Database starting from the most important as follows: Glucose, BMI, Age, Diabetes Pedigree Function, Pregnancies, Blood Pressure, Skin Thickness, and Insulin. In addition, Choudary has analyzed that having a low number of false negatives is comparatively more crucial for the model (as it can be risky to categorize high-risk patients as low-risk). As a result, Choudary focuses on Precision and Recall.. Hafsa Binte Kibria used a ratio of training data : test data is 7:3. In her research, it was found that when using the Random Forest method, the irrelevant feature was glucose, blood pressure, and pregnancy [21]. Vaishali, tried to find important features by using feature selection based on genetic algorithms. The results of this feature selection were then tested with the Naive Bayes, J48 graft, and MLP algorithms with a ratio of training data : test data is 7:3, which resulted in accuracy of 79.13%, 76.95%, 79.56%, all of which showed increased accuracy after feature selection. In her research, genetic algorithm-based feature selection resulted in 4 important features: Glucose, BMI, Diabetes Pedigree Function, and age [22]. Sanghyuck conducted research by performing correlation analysis on the PIMA Indian Database and found that only with the glucose, BMI, age features processed with the Support Vector Machine (SVM) algorithm resulted in accuracy=70.4%, precision=66.7%, recall=43.7%, and F1 score=52.8% [23].

The above research have not yet mapped the important features of the PIMA Indian Database using the KNN classification method. This research attempts to map important features in the PIMA Indian Database when using the KNN classification method.

III. Theoritical Basis

Dimensionality reduction is an initial step that is used to enhance the precision of learning features and decrease the duration of training [24]. Dimensionality reduction helps to eliminate irrelevant, noisy, and excessive features from the data [25]. Computational performance can be improved by

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i1.1598, Copyright ©2023
Submitted : January 25, 2023, Revised : March 6, 2023, Accepted : March 7, 2023, Published : March 13, 2023

simplifying the existing features. Redundant and irrelevant features to accuracy can be eliminated.

Dimensionality reduction divided to feature extraction and feature selection. There are 16 feature selection methods that can be used, namely (1) chi-square score, (2) t-test score, (3) wilcoxon, (4) Least absolute shrinkage and selection operator, (5) relief, (6) mutual information, (7) minimum redundancy maximum relevance ensemble, (8) random forest, (9) extra tree ensemble, (10) gradient boosting decision tree, (11) xgboost, (12) Elastic net, (13) L-based linear support vector machine, (14) variance, (15) L-based logistic regression, and (16) Principal Component Analysis (PCA) [26].

The KNN algorithm is a type of non-parametric classification technique. This algorithm does not assume anything about the underlying data distribution and is known for its ease and effectiveness. KNN is a method of supervised learning, which means that it requires a labeled training dataset where data points have been assigned to different classes. The KNN algorithm then inputs new unlabeled data into a class based on the nearest neighbor class in the training dataset [27]. KNN itself has many variants, including Adaptive, Fuzzy, Locally adaptive, Ensemble and Generalized mean distance, Mutual, k-means clustering, Classic one, Hassanat [28].

Euclidean Distance is a method of measuring the distance between two points in a straight line using the Pythagorean theorem. It is one of the most commonly used distance calculation methods in machine learning processes. The formula for Euclidean distance is obtained by taking the square root of the difference between two vectors [29] as in

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad (1)$$

IV. RESEARCH METHOD

The research stages are depicted as in Figure 2. The dataset is prepared by PIDD that is downloaded from the Machine Learning and Data Science community website kaggle with the address www.kaggle.com.

The information is then preprocessed by eliminating one feature and subsequently, a ratio of 9:1 are being used to split data training and data testing. Then, with KNN Mixed Euclidean Distance, accuracy will be sought for each value of k. The results of each change in k value will be recorded to obtain irrelevant and redundant features that can be removed.

Fig. 2. Research stages

V. RESULT AND DISCUSSION

PIDD is a compilation of medical diagnostic records for patients from the Pima Indian community and gathered by the National Institute of Diabetes and Digestive and Kidney Diseases. PIDD is a binary classification problem, with 268 data indicated as diabetes (labeled as 1) and 500 data indicated as non-diabetes (labeled as 0). The features in the dataset include Pregnancies (refers to the number of pregnancies a person has had), Glucose (the concentration of glucose in the blood 2 hours after taking an oral glucose tolerance test), BloodPressure (refers to Diastolic blood pressure measurement in millimeters of mercury (mm Hg)), SkinThickness (refers to the thickness of the skin on the triceps in millimeters (mm)), Insulin (refers to the amount of insulin in the blood 2 hours after a test, measured in microunits per milliliter (mu U/ml)), BMI (refers to Body mass index, calculated as weight in kilograms divided by height in meters squared), Diabetes Pedigree Function (refers to a measure of the likelihood of developing diabetes based on family history), Age (refers to the age of the person in years), and Outcome (is a binary classification indicating whether the person has diabetes or not (0 or 1)) [30]. PIDD is available for download from different online sources, such as kaggle.com.

Rapidminer are being used in this research, we can see the statistics of PIDD from rapidminer as shown in table 1.

TABLE I. STATISTIC OF PIDD

Name	*Type*	*Min*	*Max*	*rata-rata*
Outcome (label)	Binomial	0	1	0 (500), 1 (268)
Pregnancies	Integer	0	17	3,845
Glucose	Integer	0	199	120,895
BloodPressure	Integer	0	122	69,105
skinThickness	Integer	0	99	20,536
Insulin	Integer	0	846	79,799
BMI	Real	0	67,1	31,993
Diabetes Pedigree Function	Real	0,078	2,42 0	0,472
Age	Integer	21	81	33,241

to process KNN, we build a process sequence in rapidminer as seen like in figure 3. First, select attribute operator were being used for filter feature that will be used (or not used) for the training and testing process. Then, using split data, we determine the ratio of data training and data data testing as seen in figure 5. After splitting data, we train data using k-nn operator with k, weighted vote, measure types, and mixed measure parameter as seen in figure 6. We changed k to find the best accuracy.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i1.1598, Copyright ©2023
Submitted : January 25, 2023, Revised : March 6, 2023, Accepted : March 7, 2023, Published : March 13, 2023

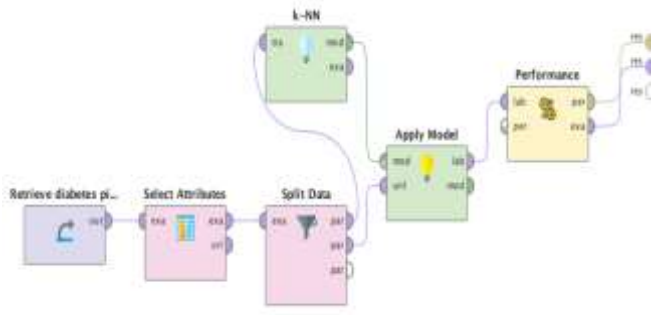

Fig. 3. RapidMiner Process

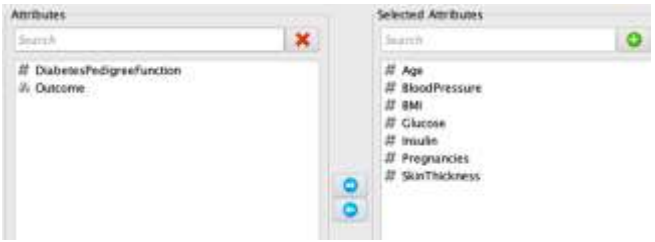

Fig. 4. KNN Process Without Diabetes Pedigree Function

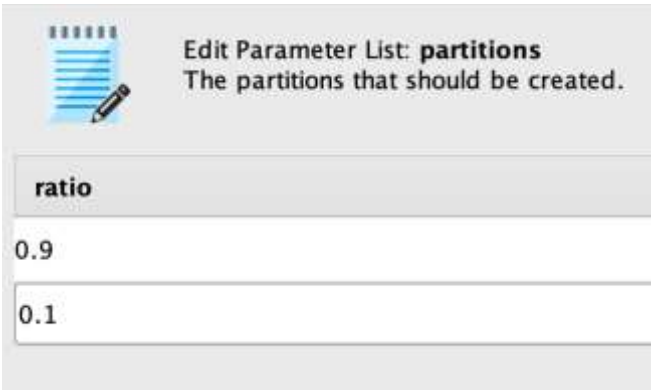

Fig. 5. Ratio of data training and data testing

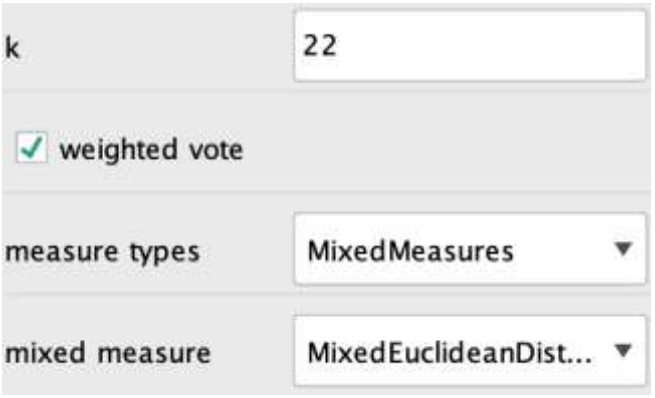

Fig. 6. k-NN Operator in Rapidminer

In the last process in figure 3, performance operator are being used to show the accuracy of the KNN. An example of accuracy result from rapidminer is shown in figure 7. Figure 7 show accuracy result of KNN with k=22 without Diabetes Pedigree Function feature.

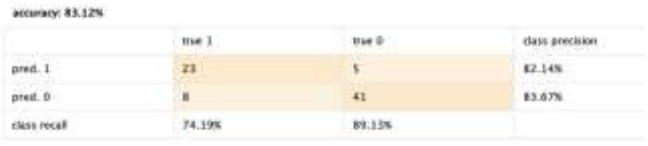

Fig. 7. Table View of KNN Accuracy Result With k=22 Without Diabetes Pedigree Function

To know recall value, we compute Figure 7 as in Figure 8 with formula 2

$$Recall = \frac{Tp}{(Tp + Fn)} \quad (2)$$

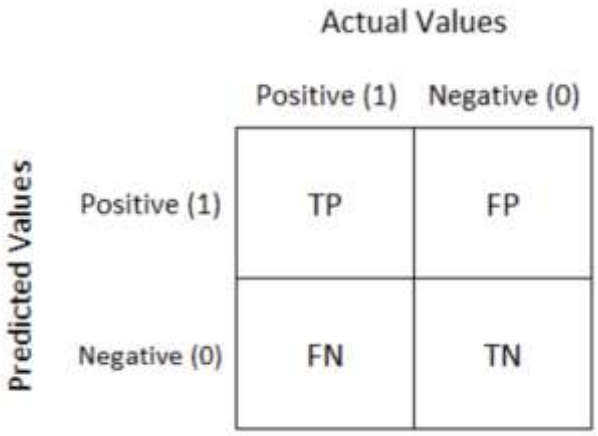

Fig. 8. Confusion Matrix as described in rapidminer official site

Table 2 shows that the highest accuracy level using KNN is at a value of K=22 for all features and without the diabetes Pedigree Function feature. This study also shows that the accuracy for all features is the same as the accuracy without the diabetes Pedigree Function feature for all values of K except for K=5.

From the average accuracy, it can be observed that the highest accuracy ranking is by using all features, without the diabetes Pedigree Function feature, without the skin thickness feature, without the diabetes Pedigree Function and pregnancy features, without the pregnancy feature, without the BMI feature, without the blood pressure feature, without the insulin feature, and without the glucose feature.

TABLE II. TABLE OF ACCURACY BY REMOVING ONE OR TWO FEATURES (IN PERCENT)

Feature	*k=5*	*k=21*	*k=22*	*k=23*
without Glucose	64,94	61,04	61,04	57,14
without Age	68,83	71,43	71,43	72,23
without insulin	74,03	71,43	72,73	74,03

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i1.1598, Copyright ©2023
Submitted : January 25, 2023, Revised : March 6, 2023, Accepted : March 7, 2023, Published : March 13, 2023

without bloodpressure	68,83	75,32	74,03	75,32
without BMI	74,03	76,62	77,92	77,92
without pregnancy	70,13	79,22	81,82	77,92
without PF - Pregnancy	70,13	79,22	81,82	77,92
without Skin	76,62	76,62	77,92	77,92
without PF	75,32	81,82	83,12	79,22
all feature	76,62	81,82	83,12	79,22

From the recall result as shown in Table 3, it can be observed that the highest accuracy ranking is by using all features, without the diabetes Pedigree Function feature, without pregnancy, without the diabetes Pedigree Function and pregnancy features, without BMI, without the skin, without the insulin feature, without the blood pressure feature, without the age feature,and without the glucose feature.

TABLE III. TABLE OF RECALL VALUE BY REMOVING ONE OR TWO FEATURES (IN PERCENT)

Feature	***k=5***	***k=21***	***k=22***	***k=23***
without Glucose	51,61	35,48	35,48	32,26
without Age	45,16	48,39	48,39	51,61
without insulin	64,52	58,06	58,06	58,06
without bloodpressure	58,06	61,29	58,06	61,29
without BMI	58,06	61,29	61,29	61,29
without pregnancy	58,06	67,74	67,74	64,52
without PF - Pregnancy	58,06	67,74	67,74	64,52
without Skin	61,29	61,29	61,29	64,52
without PF	64,52	70,97	74,19	67,74
all feature	67,74	70,97	74,19	67,74

VI. CONCLUSION

According to Table 2 (accuracy result) in this research, the important features for the K-Nearest Neighbors (KNN) method to produce high accuracy from PIDD were ranked in the following order: (1) Glucose, (2) Age, (3) Insulin, (4) Blood Pressure, (5) BMI, (6) Pregnancy, (7) Skin Thickness, and (8) Diabetes Pedigree Function. The study also concluded that the Diabetes Pedigree Function feature is a redundant and irrelevant feature for the KNN method. In addition, important features according Table 3 (recall result), ranked in order of importance is (1) Glucose, (2) Age, (3) Blood pressure, (4) Insulin, (5) Skin Thickness, (6) BMI, (7) Pregnancy, (8) Diabetes Pedigree function. However, it is important to note that this conclusion is based on the specific classification method used in this study, which is KNN. In order to gain a more comprehensive understanding of the importance of these features, future research needs to be conducted using other classification methods. Additionally, research can be done by removing two features at once to find redundant feature combinations, in order to determine whether there are any combinations of features that are not necessary for accurate classification. This could help to simplify the process of identifying important features and improve the overall efficiency of the classification method.

REFERENCES

[1] M. G. Simeng Zhu, Indrin Chetty, Farzan Siddiqui, "The 2021 landscape of FDA-approved artificial intelligence/machine learning-enabled medical devices: An analysis of the characteristics and intended use," *International Journal of Medical Informatics*, 2022, doi: https://doi.org/10.1016/j.ijmedinf.2022.104828.

[2] R. D. S. Khariri, "Transisi Epidemiologi Stroke sebagai Penyebab Kematian pada Semua Kelompok Usia di Indonesia," in *Seminar Nasional Riset Kedokteran*, 2021, vol. 2, no. 1: Fakultas Kedokteran, UPN Veteran Jakarta.

[3] I. D. Federation, *IDF Diabetes Atlas*, 10 ed., 2021. [Online]. Available: https://diabetesatlas.org/idfawp/resource-files/2021/07/IDF_Atlas_10th_Edition_2021.pdf.

[4] M. M. R. Sajratul Yakin Rubaiat, Md.Kamrul Hasan, "Important Feature Selection & Accuracy Comparisons of Different Machine Learning Models for Early Diabetes Detection," presented at the 2018 International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh, 2018.

[5] X. Y. Nana Zhang, Xiaolin Zhu, Bin Zhao, Tianyi Huang and Qiuhe Ji, "Type 2 diabetes mellitus unawareness, prevalence, trends and risk factors: National Health and Nutrition Examination Survey (NHANES) 1999–2010," *Journal of International Medical Research*, vol. 45, no. 2, 2017, doi: https://doi.org/10.1177/0300060517693178.

[6] E. B. B. Ama G. Ampofo, "Beyond 2020: Modelling obesity and diabetes prevalence," *Diabetes Research and Clinical Practice*, vol. 167, 2020, doi: https://doi.org/10.1016/j.diabres.2020.108362.

[7] O. K. NIKOS FAZAKIS, ELIAS DRITSAS, SOTIRIS ALEXIOU, NIKOS and A. K. M. FAKOTAKIS, "Machine Learning Tools for Long-term Type 2 Diabetes Risk Prediction," *IEEE Access*, vol. 9, 2021, doi: http://dx.doi.org/10.1109/ACCESS.2021.3098691.

[8] S. S. Sarthak, and Surya Prakash Tripathi, "EmbPred30: Assessing 30-days Readmission for Diabetic Patients using Categorical Embeddings," *arXiv*, 2020, doi: https://doi.org/10.48550/arXiv.2002.11215.

[9] X. G. Lanxin Miao, Hasan T Abbas, Khalid A Qaraqe, and Qammer H Abbasi, "Using Machine Learning to Predict the Future Development of Disease," presented at the 2020 International Conference on UK-China Emerging Technologies (UCET), Glasgow, UK, 2020.

[10] M. S. Sajida Perveen, Karim Keshavjee, & AzizGuergachi "Prognostic Modeling and Prevention of Diabetes Using Machine Learning Technique," *Scientific Reports*, 2019, doi: https://doi.org/10.1038/s41598-019-49563-6.

[11] E. S. Yochai Edlitz, "Prediction of type 2 diabetes mellitus onset using logistic regression-based scorecards," *eLife*, 2022, doi: https://doi.org/10.7554/eLife.71862.

[12] V. S. Salliah Shafi Bhat, Gufran Ahmad Ansari, Mohd Dilshad Ansari, and Md Habibur Rahman, "Prevalence and Early Prediction of Diabetes Using Machine Learning in North Kashmir: A Case Study of District Bandipora," *Computational Intelligence and Neuroscience*, 2022, doi: https://doi.org/10.1155/2022/2789760.

[13] R. F. M. M. Faniqul Islam, Sadikur Rahman & Humayra Yasmin Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," *Computer Vision and Machine Intelligence in Medical Image Analysis*, pp. 113-125, 2019, doi: http://dx.doi.org/10.1007/978-981-13-8798-2_12.

[14] E. D. a. M. Trigka, "Data-Driven Machine-Learning Methods for Diabetes Risk Prediction," *Sensors*, vol. 22, 2022, doi: https://doi.org/10.3390/s22145304.

[15] J. P. L. Amin Ul Haq , Jalaluddin Khan , Muhammad Hammad Memon, Shah Nazir , Sultan Ahmad , Ghufran Ahmad Khan and Amjad Ali, "Intelligent Machine Learning Approach for Effective Recognition of

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i1.1598, Copyright ©2023
Submitted : January 25, 2023, Revised : March 6, 2023, Accepted : March 7, 2023, Published : March 13, 2023

Diabetes in E-Healthcare Using Clinical Data," *Sensors*, 2020, doi: https://doi.org/10.3390/s20092649.
[16] R. P. a. B. khuntia, "Analysis and Prediction Of Pima Indian Diabetes Dataset Using SDKNN Classifier Technique," presented at the IOP Conference Series: Materials Science and Engineering, Tamil Nadu, India, 2020.
[17] A. A. Y. Nada Ali Noori, "A Comparative Analysis for Diabetic Prediction Based on Machine Learning Techniques," *Journal of Basrah Researches ((Sciences))* 2021.
[18] M. Y. a. l. Shamriz Nahzat, "Diabetes Prediction Using Machine Learning Classification Algorithms," presented at the 2nd International Conference on Access to Recent Advances in Engineering and Digitalization (ARACONF), 10–12 March 2021, 2021.
[19] M. N. A.Thammi Reddy, "Minimal Rule-Based Classifiers using PCA on Pima-Indians-Diabetes-Dataset," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 12, 2022, doi: https://doi.org/10.35940/ijitee.l2476.1081219.
[20] I. Choudhary. "PIMA Indian Diabetes Prediction - Predicting the onset of diabetes." https://towardsdatascience.com/pima-indian-diabetes-prediction-7573698bd5fe (accessed December 16, 2022).
[21] M. N. Hafsa Binte Kibria, Md. Omaer Faruq Goni, Mominul Ahsan and Julfikar Haider, "An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI," *Sensors*, 2022, doi: https://doi.org/10.3390/s22197268.
[22] D. R. S. Vaishali R, S Ramasubbareddy, S Remya, Sravani Nalluri, "Genetic algorithm based feature selection and MOE fuzzy classification algorithm on Pima Indians Diabetes dataset," presented at the International Conference on Computing Networking and Informatics (ICCNI), 2017.
[23] M. K. Sanghyuck YOU, "A Study on Methods to Prevent Pima Indians Diabetes using SVM," *Korea Journal of Artificial Intelligence*, vol. 8, no. 2, pp. 7-10, 2020, doi: http://dx.doi.org/10.24225/kjai.2020.vol8.no2.7.
[24] S. A. S.Velliangiria, S Iwin Thankumar joseph, "A Review of Dimensionality Reduction Techniques for Efficient Computation," presented at the INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING (ICRTAC), 2019.
[25] A. M. A. Rizgar R. Zebari, Diyar Qader Zeebaree, Dilovan Asaad Zebari, Jwan Najeeb Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," *Journal of Applied Science and Technology Trends*, vol. 01, no. 02, pp. 56-70, 2020, doi: http://dx.doi.org/10.38094/jastt1224.
[26] D. W. PAN SUN, VINCENT CT MOK, LIN SHI, "Comparison of Feature Selection Methods and Machine Learning Classifiers for Radiomics Analysis in Glioma Grading," *IEEE Access*, 2019, doi: http://dx.doi.org/10.1051/matecconf/20164206002.
[27] S. D. Kashvi Taunk, Srishti Verma, Aleena Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," presented at the International Conference on Intelligent Computing and Control Systems (ICCS), 2019.
[28] I. H. Shahadat Uddin, Haohui Lu, Mohammad Ali Moni & Ergun Gide "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Scientific Reports*, 2022, doi: https://doi.org/10.1038/s41598-022-10358-x.
[29] N. K. Ishan Arora, Mayank Bansal, "Effect of Distance Metric and Feature Scaling on KNN Algorithm while Classifying X-rays," presented at the The 10th Seminary of Computer Science Research at Feminine, 2021.
[30] S. K. P. Sourav Kumar Bhoi, Kalyan Kumar Jena, P. Anshuman Abhisekh, Kshira Sagar Sahoo, Najm Us Sama, Shweta Supriya Pradhan, Rashmi Ranjan Sahoo, "Prediction of Diabetes in Females of Pima Indian Heritage: A Complete Supervised Learning Approach," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 10, 2021, doi: https://doi.org/10.17762/turcomat.v12i10.4958.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v12i1.1598, Copyright ©2023
Submitted : January 25, 2023, Revised : March 6, 2023, Accepted : March 7, 2023, Published : March 13, 2023