EARLY PREDICTION OF DIABETES USING MACHINE LEARNING TECHNIQUES

1st Bhavesh Rathi
Dept of Computer Science
University of Debrecen
Debrecen, Hungary
bhavesh.rathi@mailbox.unideb.hu

2nd Filipe Madeira
Dept of Computer Science
Politechnic Institute of Santarem
Santarem, Portugal
filipe.madeira@esg.ipsantarem.pt

2023 Global Conference on Wireless and Optical Technologies (GCWOT) | 979-8-3503-3371-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/GCWOT57803.2023.10064682

***Abstract*—Diabetes is a chronic illness or group of metabolic disorders in which a person has a sustained rise in blood glucose levels (BG) because of a lack of, or an inability of, cells to respond to insulin. These days, this illness is causing severe health issues and long-term obstacles. Massive quantities of highly confidential material are present in the healthcare sector, and they must be handled properly. Diabetes Mellitus (DM) is regarded as one of the worst diseases in the world. For such examination of diabetes, clinical specialists require a trustworthy framework of objectives [20]. The collection includes data on 768 patients and the nine distinctive traits that correlate to them. To estimate blood sugar levels, we applied one ML systems to the sample [11]. We use KNN machine learning approach because it gives the perfect estimation of the dataset. Different Machine Learning (ML) techniques may be used to evaluate data from various perspectives and condense it into valuable evidence. The KNN method is used in this study to extrapolate diabetes [11]. In this paper, I have proposed future prediction of the diabetes in the human body.**

Keywords—Machine learning, KNN, Diabetes Predictions

I. INTRODUCTION

According to the WHO (World Health Organization), 1.9 million individuals worldwide lose their lives to diabetes each year. One condition that develops when blood glucose/blood sugar levels are extremely high in the body is diabetes. Diabetes, according to medical professionals, is a condition in which the pancreas of the human body either cannot create enough insulin (Type 1 diabetes) or the insulin that is produced cannot be used by the body's cells (Type 2 diabetes). Following the digestion of food, glucose is released when we eat. A blood hormone called insulin encourages cells to ingest blood glucose and use it as fuel by traveling from the blood to the cells [11]. Because cells cannot take up sucrose when the pancreas is producing insufficient insulin, the fructose stays in the circulation. As a result, the blood sugar levels rise to an extremely undesirable level.

The human body experiences several symptoms of high blood sugar, including acute hunger, strong thirst, and frequent urine. The typical range of glucose concentrations in an adult is 70 to 99 mg/dL. Diabetes is indicated if the glucose level is greater than 126 mg/dl. If a person's body sugar levels are between 100 and 125 mg/dl, they are deemed to have prediabetes.

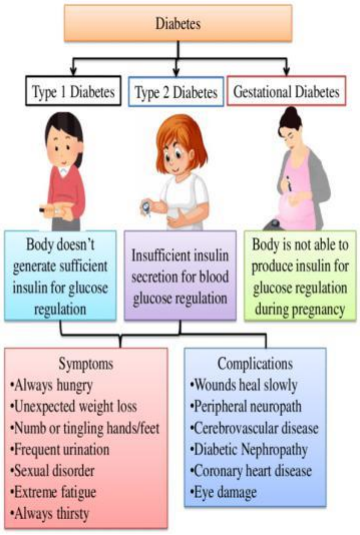

Fig:1 The Structure of Diabetic Mellitus and its symptoms [40]

When blood sugar levels in the body reach dangerously high levels, heart disease, renal failure, stroke, and nerve damage may result [5,6]. Diabetes cannot be cured permanently [11]. Long-term diabetes is a leading cause of macrovascular and microvascular complications, which are health issues. Damage to the heart, brain, and legs' big blood arteries constitutes the macrovascular problem. Small blood arteries are damaged by microvascular complications, which affects the kidneys, eyes, feet, and nerves. If diabetes is identified early, it can be effectively controlled. Diabetes can be avoided by following a healthy exercise routine and food plan [9]. Patients with prediabetes can reduce their risk of acquiring Type 2 diabetes by decreasing weight and engaging in physical activity. Several AI techniques are used in a range of sectors to conduct predictive analysis of massive data. Although prognostic health evaluation is a difficult task, it may eventually help specialists make prompt and knowledgeable decisions regarding the treatment and care of patients. The study's main goal is to use ML techniques to aid medical professionals in making the initial diagnosis of diabetes. The comparison of the several ML techniques employed here reveals which approach is most appropriate for diabetes forecast. The major goal is to identify novel trends and then evaluate these patterns to give consumers information that is pertinent to their needs. With the use of AI,

979-8-3503-3371-8/23/$31.00 ©2023 IEEE

this article aims to assist experts and professionals in the early diagnosis of diabetes [11]. One of the simplest Supervised Machine Learning techniques, K-Nearest Neighbors is frequently utilized for classification issues. KNN searches for related objects. Comparing the feature vectors yields a measurement of the separation or similarity. A distance metric, such as Euclidian distance or Cosine similarity, is applied when comparing the feature vectors. With the non-parametric K-Nearest Neighbors approach, we may determine the parameters without assuming a distribution for the raw data [18].

II. Related Work

The presented clarification that can be used for a timely diabetes forecast for a patient with high precision and recommended a KNN strategy for diabetes estimate. The outcome has demonstrated the effectiveness and primary significance of the probability system. The anticipated model yields outstanding diabetes calculation outcome. Information on diabetic infection is pre-handled after numerous archives have been prepared. The dataset contains a total of 769 patient records, 6 of which cause them to miss estimations. The remaining 763 patient records are used for pre-handling after the 6 such records were removed from the dataset. This is a pictorial display of a variety of ideas that are crucial to the architecture, including its attributes, elements, and sections. Here, we show the proposed model diagram for results computation [20]. Several strategies could be helpful in the fields of data science and computing to create the appropriate data-driven predictive models based on machine learning methods [10]. One of the most popular approaches in machine learning is classification. Finding a group of models that enables the recognition and differentiation of data types is what this procedure entails. The goal of classification is to establish the category of upcoming data items using knowledge from the past. In classification, the model is typically taught on a training set, and the model is then evaluated on the test set. Since there isn't a perfect method for all data sets, several categorization techniques have been created in the literature to date [23].

III. Methodology

A. Dataset

The dataset's goal is to predict if a patient has diabetes based on a few symptomatic estimates that were kept in mind. The dataset includes an outcome objective variable and few clinical pointer concerns. The degree of pregnancy, BMI, insulin level, patient age, and other indicators are examples of indicator factors [20].

Pregnancy: Antenatal births in total

- BP: maximum blood pressure (mm Hg)
- Insulin: (mu U/ml) Two-hour plasma glucagon.
- Glucose: Plasma glucose level during a 2-hour glucose tolerance test.
- Mass index: weight in kilograms (height in m) 2

TABLE I.

Attributes Mean	Standard Deviation	Min/Max Value
Number of pregnancies 3.8	3.5	1/18
2 A plasma glucose level of 120.9	33	57/198
69.1 is the diastolic blood pressure.	19.5	25/120
Skinfold thickness of the fourth triceps (mm): 20.5	17	8/55
Insulin 2-hour serum level 79.8	116.2	16/848
Body mass index (kg/m2):	8.0	19.2/58.3
Pedigree function for diabetes of 0.5	0.6	0.0852/2.33
Age 33.2	11.9	22/82
Class	Tested Positive:	Diabetic
Tested Negative:	Non-Diabetic	

B. *KNN Machine Learning Approach*

- Choose the neighbor's number for the prediction.
- Find the Euclidean separation between each of the m neighbors.
- Using the estimated Euclidean distance, select the m neighbors that are closest to you.
- Then, among these m neighbors, count the number of data points in each category.

❖ Assign the extra data points to the group with the highest neighbor count.

Our structure is finished.

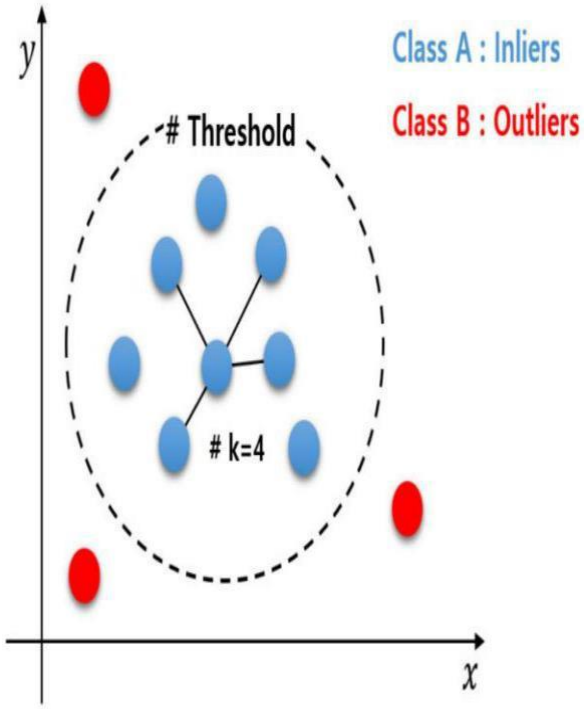

Fig.2 This Data points show the Category for the neighbors [41]

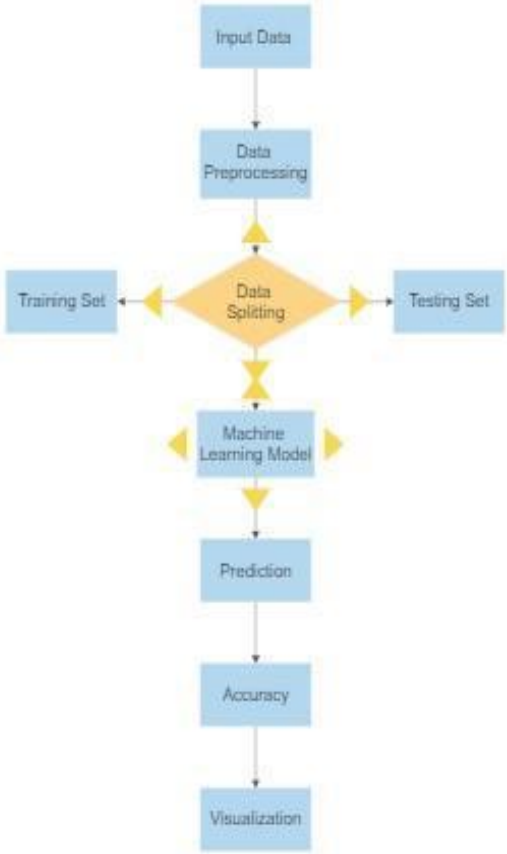

Fig. 3. Flow chart of Methodology.

It is easy to put into practice. It can withstand noisy training data. If there is a lot of training data, it could work better.

 K's value must always be determined, and sometimes that can be difficult. The high computing cost is caused by the need to determine the separation between each data point for each training sample.

Data cleansing: data for all four nations were extracted and saved as a.xlsx files. Although the collection had a few entities, however not every field was required for the application. When predicting survival days, it is necessary to consider characteristics including ages, sexuality, changes, and modified genetics. Dropping the remaining fields; ii) Splitting the data: To successfully develop an ML model, the data must be split into a training and test set. However, difficulties like; might arise if the data is wrongly broken.

1) Overfitting—the model performs poorly on fresh, or test data while being trained extraordinarily well on training data

2) underfitting, the model is unable to make a connection between the input due to a lack of data. As a result, it struggles with training data. To determine the split ratio, tests were run for three categories: 60:40, 50:50, and 70:30. The relative MAE scores were 7.35, 0.015, and 0.090. If the MAE score is low, a model is seen as being good. As a result, the experimental study's standard split ratio was 80:20. iii) Normalization The dataset's several fields each have a variety of values. For example, age is expressed as an integer and

tumor stage as an alphanumeric value. Several columns also contained null or no weight values. To remedy this difference, a uniform scale that represents all the fields was needed. Before adding the training and test data to the model, normalization was carried out; iv) platform for implementation: The recommended approaches' code was created using the Python programming language [27].

IV. Results

By computing statistical parameters like selectivity, tolerance, and efficiency, which are covered in, the outcomes are contrasted. The KNN Algorithm is employedto compute the results.

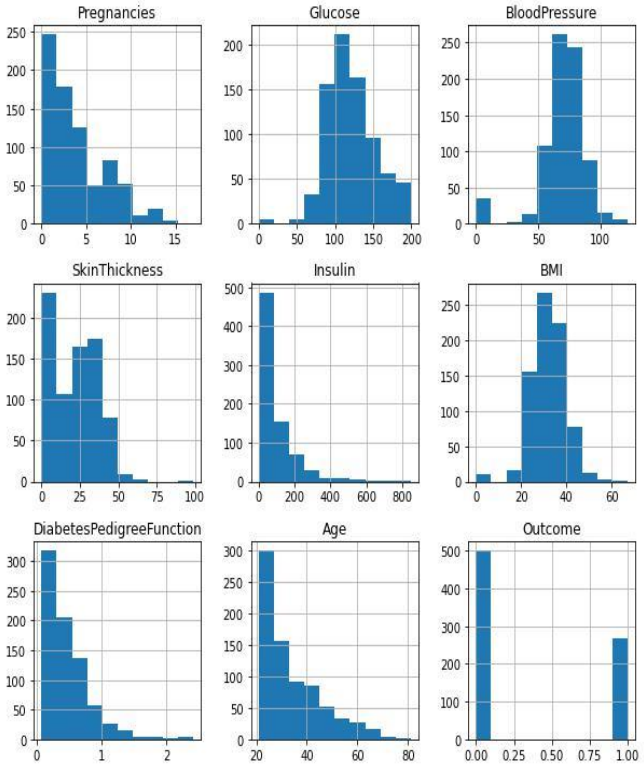

Fig.5 The visualization of the Diabetes Mellitus

A precise distribution of data is required. The class detected by scans and present in the information is the one that is represented in the confusion matrix, the expected class is the one that algorithms predict. The number of classes B1 samples that were successfully classified makes up the total percentage (True positive). The True Negative is the total number of correctly categorized class B2 samples. False Negative is the quantity of class B1 samples that were incorrectly classified as C2. False Positive is the quantity of class B2 samples that were misclassified as C1 samples.

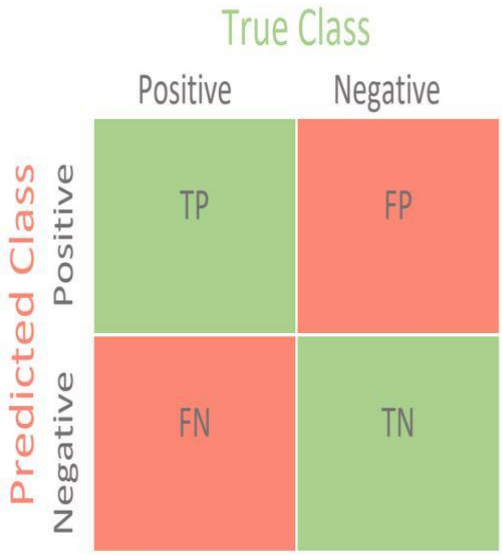

Fig.4 The Confusion Matrix

Using the Indian Diabetes dataset, we empirically assessed the accuracy of weighted estimates. It was split in a train data (70%) and a test data (30%). In this essay, we thus evaluate values of m (3, 5, 7 and 10) on train data to approximate the inaccuracy level to shorten execution time in python programming. The K-nearest Neighbor method's selected values for K are shown in the table. For instance, K=5 was selected in the table below since it produces the highest accuracy. The performance results for k-nearest neighbor techniques with subjective polling are clearly superior to those without subjective polling in the table below [33]. For instance, when using the KNN technique without weights, the accuracy of the dataset without missing data. When we apply weight, this strategy gets (80.12 percent).

TABLE II.

KNN Method				
K=10	**K=7**	**K=5**	**K=3**	
67.26 %	70.90 %	73.89%	69.50 %	K-Nearest Neighbor without weight
89.65 %	86.95 %	92.95 %	92.55 %	Proposed K-Nearest Neighbor with weigh

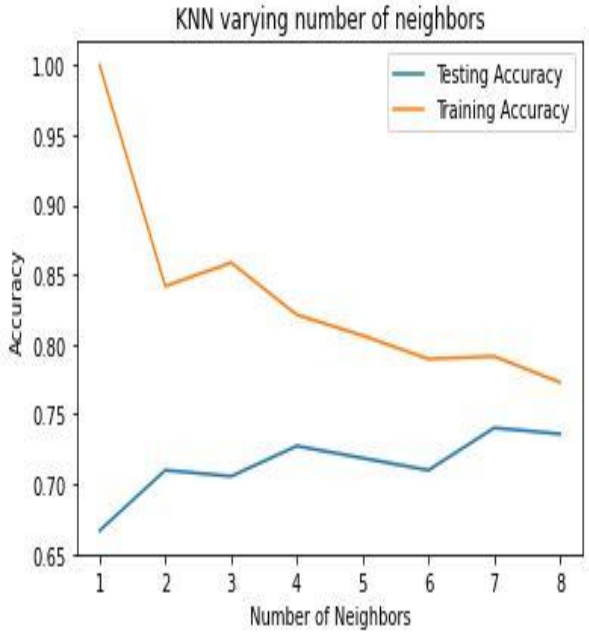

Fig.6 Knn varying number of neighbors

CONCLUSION

The accurate identification of diseases is one of the successful applications of machine learning. Due to poor eating, diabetes is more frequently encountered in adults. Early diagnosis of this serious condition and appropriate treatment lowers the prevalence of diabetics. In this study, we developed the Balanced Nearest neighbor technique for type 2 diabetes prediction, and the results were contrasted with those of the existing algorithms. This can also be used to determine whether someone has cancer or heart problems. Although machine learning methods are used to diagnose diseases, little study has been done to forecast how those diseases will be treated [33]. Any person who has an unhealthy lifestyle, an excessive amount of stress, and extra body weight can develop diabetes mellitus. In this paper, diabetes is predicted using the lazy k-nearest neighbor algorithm. We developed a method using pre-processing techniques and parameter optimization for the lazy classifier. We discovered that using the suggested methods on K - means resulted in an accuracy improvement of 8.48 percent, which is a significant improvement [29]. For the prediction and diagnosis of diabetes, several methods have been developed. Compared to the current standard systems, these algorithms offer greater accuracy. However, we have considered a novel classification approach called KNNB1R in this work to increase the KNN method's classification accuracy. One-attribute-Rule is the algorithm used in this method. This algorithm uses the one-attribute-rule algorithm to give each attribute a weight. Next, the selectivity, selectivity, and efficiency of classification performance are assessed. It is found that the accuracy of the suggested KNNB1R algorithm has significantly improved. Hybrid classification models that combine KNN and other data mining techniques may be the subject of future research [02].

REFERENCES

1. Aldi Nugroho, O. R. (2020). Identification of Student Academic Performance using the KNN Algorithm. *Engineering, MAthematics and Computer Science Vol.2 No.3*.
2. Amal H. Khaleel, G. A.-S. (2017). A Weighted Voting of K-Nearest Neighbor Algorithm for Diabetes Mellitus. *IJCSMC, Vol. 6, Issue. 1*.
3. AMEER ALI, M. A.-J. (2020). DIABETES CLASSIFICATION BASED ON KNN. *IIUM Engineering Journal, Vol. 21, No. 1*.
4. Andreas Lüschow, C. W. (2016). Classifying Medical Literature Using k-Nearest-Neighbours Algorithm. *CEUR Workshop Proceedings*.
5. CHIKALKAR, S. N. (2020). K -NEAREST NEIGHBORS MACHINE LEARNING ALGORITHM. *IJCRT | Volume 8, Issue 12*.
6. Chinmoy Ghosh, S. S. (2019). MACHINE LEARNING BASED SUPPLEMENTARY PREDICTION SYSTEM USING K NEAREST NEIGHBOUR ALGORITHM. *INTERNATIONAL CONFERENCE ON INDUSTRY INTERACTIVE*

INNOVATIONS IN SCIENCE, ENGINEERING AND TECHNOLOGY.

7. Georgios Chatzigeorgakidis, S. K. (2018). FML-kNN: scalable machine learning on Big Data using k-nearest neighbor joins. *Journal of Big Data*.

8. Gufron, B. S. (2019). Implementation of the K-Nearest Neighbor Method to determine the Classification of the Study Program Operational Budget in Higher Education. *1st International Conference of Health,Science & Technology(ICOHETECH)*.

9. Haq, M. R. (2019). Machine Learning for Load Profifile Data Analytics and Short-term Load Forecasting. *Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange*.

10. Iqbal H. Sarker, M. F. (2020). K-Nearest Neighbor Learning based Diabetes Mellitus Prediction and Analysis for eHealth Services. *EAI Endorsed Transactions on Scalable Information Systems*.

11. Jobeda Jamal Khanam, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *Korean Institute of Communications and Information Sciences (KICS)* .

12. Junyu Zhang, J. C. (2019). Study of Employment Salary Forecast using KNN Algorithm. *Advances in Computer Science Research, volume 91 International Conference on Modeling, Simulation and Big Data Analysis*.

13. K. Govindasamya, T. V. (2017). A Study on Classifi cation and Clustering Data Mining Algorithms based on Students Academic Performance Prediction. *International Journal of Control Theory and Applications Volume 10*.

14. Kadir Sabancı, M. K. (2015). The Classification of Eye State by Using kNN and MLP Classification Models According to the EEG Signals. *International Journal of Intelligent Systems and Applications in Engineering*.

15. Khalid Alkhatib, H. N. (2013). Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm. *International Journal of Business, Humanities and Technology Vol. 3 No. 3*.

16. Kilian Q.Weinberger, L. K. (2009). Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research 10*.

17. Krati Saxena, D. Z. (2014). Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm . *International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 4*.

18. Medicherla, S. (2019). Machine Learning: KNN Algorithm. *International Journal of Software & Hardware Research in Engineering Volume 7*.

19. Mohammad Rezwanul Huq, A. A. (2017). Sentiment Analysis on Twitter Data using KNN and SVM. *(IJACSA) International Journal of Advanced Computer Science and Applications Vol. 8, No. 6*.

20. N.Krishnamoorthy, D. M. (2021). DIABETES PREDICTION IN HEALTHCARE USING KNN ALGORITHM. *INTERNATIONALJOURNALOF MULTIDISCIPLINARY EDUCATIONAL RESEARCH VOLUME:10*.

21. NirmalaDevi.M, A. a. (2013). An amalgam KNN to predict Diabetes Mellitus . *IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology*.

22. Nivethitha. A, P. K. (2021). Smart Disease Prediction Using Machine Learning. *International Journal of Innovative Science and Research Technology Volume 6 , Issue 6*.

23. Noori, B. (2021). Classification of Customer Reviews Using Machine Learning Algorithms. *Applied Artificial Intelligence An International Journal VOL. 35, NO. 8*.

24. Nurul E'zzati Md Isa, A. A. (2017). The Performance Analysis of K-Nearest Neighbors (K-NN) Algorithm for Motor Imagery Classification Based on EEG Signal. *MATEC Web of Conferences 140, 01024 ICEESI*.

25. Prasannavenkatesan Theerthagiri, I. J. (2021). Prediction of COVID-19 Possibilities using KNN Classification Algorithm. *International Journal of Current Research and Review*.

26. R. Madhuri, S. S. (2021). Application of machine learning algorithms for flood susceptibility assessment and risk management. *Journal of Water and Climate Change Volume 12.6* .

27. Rashmi Siddalingappa, S. K. (2022). K-nearest-neighbor algorithm to predict the survival time and classification of various stages of oral cancer: a machine learning approach. *F1000Research*.

28. Robert A. Sowah, A. A.-A.-M. (2020). Design and Development of Diabetes Management System Using Machine Learning. *Hindawi International Journal of Telemedicine and Applications Article ID 8870141*.

29. Roshi Saxena, D. S. (2021). Role of K-nearest neighbour in detection of Diabetes Mellitus. *Turkish*

Journal of Computer and Mathematics Education Vol.12 No.10.

30. Sadegh Bafandeh Imandoust, M. B. (2013). Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background. *S B Imandoust et al. Int. Journal of Engineering Research and Applications Vol. 3, Issue 5*.

31. Shahadat Uddin, I. H. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports volume 12*.

32. Slamet Wiyono, T. A. (2019). COMPARATIVE STUDY OF MACHINE LEARNING KNN, SVM, AND DECISION TREE ALGORITHM TO PREDICTSTUDENT'SPERFORMANCE. *International Journal of Research - GRANTHAALAYAH(IJRG) Vol.7*.

33. Sreeja Vishaly M, U. k. (2022). Type 2 Diabetic Prediction Using Machine Learning Algorithm. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS) Vol 88 No 1*.

34. Thein Yu, K. T. (2020). Comparing SVM and KNN Algorithms for Myanmar News Sentiment Analysis System. *ICCDE '20 Sanya, China*.

35. Tlamelo Emmanuel, T. M. (2021). A survey on missing data in machine learning. *Journal of Big Data*.

36. Tran, H. (2016). A SURVEY OF MACHINE LEARNING AND DATA MINING TECHNIQUES USED IN MULTIMEDIA SYSTEM. *Researchgate publication journal* .

37. Tran, H. (2019). A SURVEY OF MACHINE LEARNING AND DATA MINING TECHNIQUES USED IN MULTIMEDIA SYSTEM.

38. Yun-lei Cai, D. J.-f. (2010). A KNN Research Paper Classification Method Based on Shared Nearest Neighbor. *NTCIR-8 Workshop Meeting*.

39. Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Ann Transl Med Vol 4, No 11*.

40. Jain, P. (2021). Everything You Wanted to Know About Noninvasive Glucose Measurement and Control.

41. Lee, J. ,. (2020). Automatic Bridge Design Parameter Extraction for Scan-to-BIM. *Applied Sciences. 10. 7346. 10.3390/app10207346.*