Institute of Research and Publication Indonesia (IRPI)
Public Research Journal of Engineering, Data Technology and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/predatecs
Vol. 2 Iss. 2 January 2025, pp: 60-67
ISSN(P): 3024-921X | ISSN(E): 3024-8043

Applying A Supervised Model for Diabetes Type 2 Risk Level Classification

Ahmad Dhani[1*], Danur Lestari[2], Meriana Prihati Ningrum[3], M. Andhika Fakhrizal[4], Ganis Lintang Gandini[5]

[1,2,3,4]Departemen of Information System, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[5]Department of Medicine, Faculty of Medicine, Al-Azhar University, Egypt

E-Mail: [1]12150311729@students.uin-suska.ac.id,
[2]danurlestari12@gmail.com, [3]merianaprihati@gmail.com,
[4]andhikafakhrizal123@gmail.com, [5]ganisgandini.stu153@azhar.edu.eg

Received Dec 27th 2023; Revised Aug 9th 2024; Accepted Oct 24th 2024; Available Online Jan 12th 2025, Published Jan 18th 2025
Corresponding Author: Ahmad Dhani

Abstract

Diabetes can lead to heart attacks, kidney failure, blindness, and increased risk of death. This research was conducted with the aim of classifying a diabetes risk dataset. In this context, performance comparison was carried out on three supervised learning algorithms: K-Nearest Neighbor (K-NN), Naive Bayes Classifier (NBC), and Random Forest, against a dataset containing information on specific indicators related to diabetes risk. Additionally, this study also aimed to evaluate the accuracy comparison of the results produced by these three algorithms. The results of this research show that Random Forest performs very well in detecting diabetes, prediabetes, and non-diabetes, with high precision, recall, and F1-score levels. Meanwhile, although the results are still below Random Forest, both Naive Bayes and K-NN still demonstrate significant performance, especially regarding prediabetes cases. In conclusion, from the comparison results, the Random Forest algorithm shows the highest accuracy level at 99%, followed by K-NN with an accuracy of 85%, while NBC has the lowest accuracy rate of 74%. This research indicates that the Random Forest algorithm excels in classifying data compared to the other two algorithms.

Keyword: Classification, Diabetes, K-Nearest Neighbor, Naive Bayes, Random Forest

1. INTRODUCTION

One of the most serious global health problems is diabetes, which affects people in all countries. The prevalence continues to increase alarmingly and is predicted to reach 578 million people in 2030 and 700 million people in 2045 [1]. Diabetes is one of the serious and long-term diseases. This disease occurs as a result of increased glucose levels in the bloodstream. This increase is caused by the inability of the pancreatic beta cells to produce insulin properly [2]. This situation can be fatal as it increases the risk of various complications such as heart disease, stroke, kidney failure, cancer, and blindness [3]. Diabetes has become one of the most common causes of death in recent years.

Approximately 4 million people die from diabetes each year. Therefore, the International Diabetes Federation (IDF) predicts that the global healthcare costs to treat this disease will reach $850 billion in 2017 [4]. Furthermore, in 2020, the IDF reported that the prevalence of diabetes in Indonesia exceeded 6% of the total adult population, around 172 million people [5]. This reflects the magnitude of the challenge in addressing diabetes. Additionally, the overall costs also indicate that diabetes has a significant economic impact, including prevention efforts, long-term care, and treatment expenses.

The high mortality rate associated with the severe outcomes of diabetes is a global health problem that requires serious attention. Cholesterol, body mass index (BMI), and negative habits such as smoking, as well as a lack of health monitoring, are some important variables that can increase the risk of diabetes. Moreover, the importance of the aforementioned figures reinforces the need to use data mining techniques to classify and predict the factors that may cause a person to develop diabetes. mTherefore, this research highlights the urgency of implementing machine learning techniques in developing an active approach to diabetes prevention. The application of advanced techniques such as data mining and supervised learning in diabetes risk classification contributes significantly by evaluating and comparing more effective machine learning models for disease classification, supporting both its prevention and treatment.

DOI: https://doi.org/10.57152/predatecs.v2i2.1105

Data mining is a powerful Artificial Intelligence (AI) tool, and has the ability to find useful information by analyzing data from various perspectives. Then, classify the information and summarize the relationships identified in the database [6]. Data mining can be used as a powerful data analysis tool to explore complex patterns in data sets (in this study, medical data sets related to diabetes). By applying supervised learning techniques, a diabetes risk classification model can be developed. This model can identify those who are at high risk of developing diabetes or those who are not, in order to take early preventive measures [7].

Classification of diabetes risk levels, taking into account common risk factors such as age, family history, ethnicity, waist circumference, and tension [8], is a suitable approach to identify individuals who are susceptible to diabetes. Through an in-depth analysis of these factors, classification techniques can produce more accurate predictions regarding an individual's diabetes risk. In addition, information on lifestyle [9], weight, smoking habits, etc [10] can be incorporated into the classification model to enrich the risk assessment. As such, this classification approach is not only based on genetic factors, but also includes lifestyle elements that could potentially affect a person's overall health. By considering a combination of these factors, classification techniques can find unique patterns that indicate whether a person is at risk of developing diabetes or otherwise.

This study uses classification techniques to predict whether people will develop diabetes. Classification techniques can use multiple algorithms. K-NN, Naive Bayes, and Random Forest algorithms are used in this study. The aim of this study is to compare the most accurate results of data mining classification related to diabetes risk level using these three algorithms. Previous studies have compared various supervised learning algorithms such as SVM, decision trees, and naive Bayes to predict diabetes risk. They used the Pima Indians Diabetes dataset available on the UCI repository. This dataset includes 768 examples and eight attributes. The results obtained from this study show that Naive Bayes is the best algorithm for predicting diabetes with an accuracy of 76.30% [11].

The authors in [12] used 6 machine learning classification algorithms such as Logistic Regression, Naive Bayes, SVM, Decision Tree, Random Forest, and Artificial Neural Network (ANN) for early prediction of type-II diabetes mellitus (T2DM). Random Forest achieved the highest accuracy of 93.75% compared to other models. This research uses the K-NN algorithm to classify diabetics' diet with the Kaggle dataset, achieving the highest accuracy of 80.34% compared to Naïve Bayes (77.34%) and Decision Tree (77.43%). This algorithm utilizes Euclidean distance for consistent classification results [13]. This study uses the K-NN algorithm to classify skin lesions into normal or malignant with 98% accuracy. This MATLAB-based system is efficient, accurate, and equipped with a GUI for user convenience[14]. This study used Naïve Bayes to predict obesity risk from the Kaggle dataset (2,111 data, 17 attributes), resulting in an accuracy of 77.48%. This algorithm is efficient for simple data, but less optimal on complex data [15]. This study evaluates the performance of Naïve Bayes Classifier in medical diagnosis on heart disease and diabetes datasets, showing competitive accuracy over methods such as Decision Tree and SVM, especially on datasets with uneven distribution [16]. This study compared Random Forest and SVM in heart failure classification, with RF achieving the highest accuracy of 83.33% compared to SVM 81.51%, showing the superiority of RF on a 90:10% hold out distribution [17]. This study evaluated reforestation in Austria for avalanche protection using orthoimages and Random Forest algorithm, achieving 87-98% accuracy with Kappa 0.81-0.93 [18].

Another study was conducted with the aim of analyzing the strengths and weaknesses of various algorithms in predicting diabetes. This study compared the performance of five algorithms, including K-Nearest Neighbors, Naive Bayes, Decision Trees, Regression, and SVM. The research findings indicate that SVM stands out with the highest accuracy rate, reaching 77.3%[19]. Based on the aforementioned background, this research classifies the diabetes dataset by comparing three algorithms: K-NN, Naive Bayes, and Random Forest. The objective sought through this study is to establish a solid foundation for selecting the optimal diabetes risk classification algorithm, serving as a reference for subsequent strategic steps. Therefore, this research will assist in adapting preventive and healthcare measures by adjusting recommendations based on accurate predictions.

2. MATERIAL AND METHOD

This research is a Diabetes risk classification using a dataset taken from kaggle in the scope of the entire country. The dataset consists of 22 variables that have 3 classes. O for non-diabetics, 1 for prediabetics, and 2 for diabetics. There is a class imbalance in this dataset. This research uses Google Collabs as the chosen tool to compare the three algorithms namely K-Nearest Neighbors (K-NN), NBC, and Random Forest. Combining, Naive Bayes Classifier (NBC), and Random Forest in data analysis yields many advantages. This is largely because each algorithm has the ability to work synergistically to complement each other [20]. K-NN classification classifies samples directly rather than making obvious generalizations on training data. To use K-NN on a large scale, it requires calculating the distance between the input sample and every sample in the training data; this leads to lower computation time and memory usage. This limitation is the main problem in using K-NN on a large scale [21]. Naïve Bayes Classifier is a statistical classification method based on Bayes' theorem that can be used to estimate the probability of class membership. Naïve Bayes Classifier is one of the best algorithms in data mining classification techniques that can handle irrelevant data and large datasets [22]. One of the main advantages of the Random Forest model is its ability to limit overfitting without significantly reducing its prediction accuracy. However, a lack of research has been conducted regarding the assessment of the prediction accuracy of this model, and the results are not fully understood [23]. With that, the researcher conducted the research stages as shown figure 1.

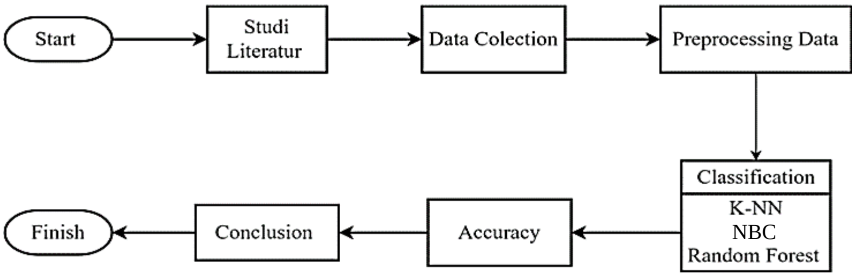

Figure 1. Research Methodology

The research stages begin with (1) Literature Study, modules, scientific papers, and other relevant sources related to the research topic, (2) Data Collection obtained from kaggle, (3) Classification can be decided with the K-NN, Naïve Bayes, Random forest algorithms, (4) Accuracy refers to measuring the performance of the classification model after the training data process, (5) From the accuracy that has been obtained using three algorithms, researchers draw conclusions from the research results.

2.1. Data Collection

The data used in this research is classification data obtained through the kaggle website. The attributes contained in this data are related to the classification of diabetes risk levels such as HighChol, CholCheck, BMI, HeartDiseasorAttack, Healthcare, etc. The data collection can view table 1.

Table 1. Data Collection

HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	…	Income
High BP	HighChol	CholCheckin5years1	40.0	Yes	No	…	3.0
NoHigh	No HighChol	NoCholCheckin5Years1	25.0	Yes	No	…	1.0
High BP	HighChol	CholCheckin5Years1	28.0	No	No	…	8.0
High	No HighChol	CholCheckin5Years1	27.0	No	No	…	6.0
…	…	…	…	…	…	…	…
High BP	HighChol	CholCheckin5Years1	25.0	No	No	…	2.0

2.2. Data Preprocessing

Before conducting an analysis using a machine learning model, a step that must be taken is to prepare the data thoroughly. This step is done by cleaning and changing the data to make it more suitable for the training model to be used. And this step is known as Data Preprocessing.

2.3. Diabetes

Diabetes is a disease characterized by impaired glucose homeostasis that is designated as an increasing epidemic [24]. The signs of diabetes are often ignored by people because the process is not immediately visible, even the damage caused by diabetes can occur before the appearance of symptoms. The common symptoms that occur in people with diabetes are significant weight loss, easy fatigue, anxiety and frequent body pain [25].

2.4. Google Colab and Python Programming

Google offers access to Google Collaboratory, commonly known as Google Colab, to anyone with a Gmail account. Google colab provides a user-friendly and intuitive programming environment. Jupyter notebook-based cloud pursues as platform Google Colab specially designed for Python programming languages without the need for a local installation. Python is a highly interpreted programming language known for its simple syntax and ease of learning [26]. Python's pre-installed library in Colab also makes the development process much easier, eliminating the need for users to prepare the development environment manually before the start of the project.

2.5. K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN is a simple algorithm that is effectively used to classify data. The concept of the K-NN algorithm is to find the closest distance between the evaluated data and the nearest neighbor in the training data [27]. The advantage of comparing the K-NN algorithm with other classification algorithms such as Decision Tree, and Neural Network is that it has the ease of overcoming problems with large class sizes and does not require many parameters to produce high accuracy [28], the following is equation 1.

$$d_{\mathrm{Euclidian}} = \sqrt{\sum_{i=1}^{n} (x_{i2} - x_{i1})^2} \tag{1}$$

In this equation, $d_{\mathrm{Euclidian}}$ represent the Euclidean distance between two data, while n represents the number of attributes of data. The variable x_{i1} represents the attribute value of the sample data, and x_{i2} represents the attribute value for the test data. By applying this formula, the K-NN algorithm can determine the nearest neighbor of the evaluated data so that the classification process can be carried out based on data that has similarities.

2.6. Naïve Bayes Classifier

Naive Bayes is a simple classification algorithm used to calculate a set of probabilities by summing the frequencies and combinations of values from a given data set. This algorithm has class conditional independence, which is a very strong assumption about the independence of each condition or event [29]. The advantage contained in this algorithm is that it facilitates the performance of classifiers in determining classes [30]. The equation of the Naïve Bayes theorem are 2.

$$P(C_i|X) = \frac{P(X|C_i) * P(C_i)}{P(X)} \tag{2}$$

In this formula, C_i represents the solution class of the i-th pattern, where i indicates the number of class labels, and X refers to the criteria defined based on the given input data. The term $P(C_i|X)$ indicates the probability of the class label C_i occuring given the input criteria X. Furthermore, $P(C_i)$ represents the prior probability of the class label C_i, while $P(X)$ denotes the probability of the input criteria X.

2.7. Random Forest

An ensemble classifier called a random forest (RF) classifier creates several decision trees using a subset of training samples and variables that is chosen at random [31]. The equation of the Random Forest are 3 [34].

$$A_{k,\mathrm{pred}} = \frac{1}{T} \sum_{t=1}^{T} A_{k,t,\mathrm{pred}} \tag{3}$$

In this formula $A_{k,\mathrm{pred}}$ represents the predicted activity value for the k-th compound, as determined by the Random Forest model. The parameter T refers to the total number of decision trees within the Random Forest. Furthermore, $A_{k,t,\mathrm{pred}}$ represents the predicted activity value for the k-th compound as determined by the t-th decision tree.

2.8. Confusion Matrix

An assessment technique called a confusion matrix is used to classify performance according to right and wrong. There is recall, precision, and accuracy in the confusion matrix. The four outputs of this formula's computations are recall, precision, accuracy, and error rate. Assessing the test items correctness and falsity forms the basis for evaluating the categorization model [32].

3. RESULTS AND DISCUSSION

3.1. K-Nearest Neighbor

In this study, the researchers applied the K-NN algorithm to perform data classification. The test results showed that the classification accuracy rate reached 85%. This finding indicates that K-NN is an effective choice for classification with significant accuracy. The following are the classification evaluation results obtained from applying the K-NN algorithm.

Table 2. Classification Results of K-NN Algorithm

	True No Diabetes	True Prediabetes	True Diabetes	Class Precision
No Diabetes Prediction	5578	5	1584	0.87

	True No Diabetes	True Prediabetes	True Diabetes	Class Precision
Prediabetes Prediction	1232	35	110	0.52
Diabetes Prediction	6846	27	3518	0.67

From the data analysis on table 2, it can be noted that the classification model with the K-NN algorithm is able to provide very accurate predictions for No Diabetes cases, with a precision rate of around 87%. This indicates that the model rarely gives false positive predictions for individuals who do not actually have diabetes. However, when facing Prediabetes cases, the precision drops to around 52%, indicating a tendency for the model to give false positive predictions. In terms of detecting Diabetes cases, the model shows an adequate level of precision, which is around 68.98%.

3.2. Naïve Bayes Classifier

In addition to using the K-NN algorithm, in this study researchers carried out the classification process by applying the NBC algorithm model. The test results of the NBC algorithm revealed a classification accuracy rate of 74%. This finding indicates that the application of the NBC algorithm can be an effective alternative in carrying out classification with a satisfactory level of accuracy. The following are the results of the NBC algorithm evaluation.

Table 3. Classification Results of Naïve Bayes Classification Algorithm

	True No Diabetes	True Prediabetes	True Diabetes	Class Precision
No Diabetes Prediction	45322	252	11593	0.90
Prediabetes Prediction	794	16	567	0.04
Diabetes Prediction	4342	89	5960	0.33

Based on the table 3, the Naive Bayes algorithm shows superiority in identifying individuals without diabetes (No Diabetes) with a high level of precision, reaching 90%. This means that about 90% of those predicted to have no diabetes are actually free from the condition. However, it should be noted that the performance of the model dropped significantly in the "Prediabetes" class with a precision of only 4%, indicating the difficulty of the model in classifying prediabetes. Similar results were seen in the "Diabetes" class with a precision of around 33%, indicating the challenge in identifying individuals who actually have diabetes. Most likely, there are many false positives in these two classes. This analysis highlights that the model tends to be better at identifying healthy individuals than those who may have pre-diabetes or diabetes.

3.3. Random Forest

Within the scope of this research, the final algorithm applied is Random Forest to run the classification process. The test results show that the classification accuracy rate reaches 99%. However, it should be noted that the accuracy rate is still at a relatively low level. The following is an overview of the test results from the application of the Random Forest algorithm.

Table 4. Classification Results of Random Forest Algorithm

	True No Diabetes	True Prediabetes	True Diabetes	Class Precision
No Diabetes Prediction	57092	11	64	0.99
Prediabetes Prediction	83	1284	10	0.99
Diabetes Prediction	356	5	le+04	0.99

The Random Forest algorithm performed very well in classifying each class with a precision of 0.99 for all three classes. This means that most of the predictions for these classes are accurate and reliable. Overall, the analysis results show that the model has good potential to identify "No Diabetes", "Prediabetes", and "Diabetes" cases.

3.4. Comparison of K-Nearest Neighbor (K-NN), Naïve Bayes Classifier (NBC), and Random Forest Algorithms

In dealing with the complexity of data mining, the selection of supervised learning algorithms is a key aspect. A comparison between the three main algorithms, namely KNN, NBC, and Random Forest, is essential to identify the most effective approach for data processing in this study. At this point, focusing on assessing the accuracy, precision, and recall values of each algorithm is the first step in understanding their strengths and weaknesses. By doing this comparison, the performance of supervised learning algorithms in the context of data mining can be more clearly illustrated. Comparison of KNN, NBC and Random Forest can view figure 2.

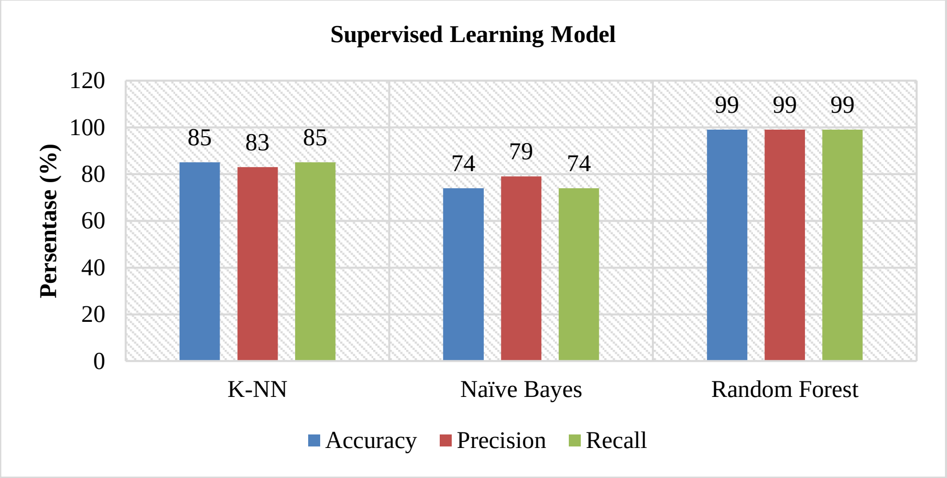

Figure 2. Bar Chart Comparison of KNN, NBC, and Random Forest

From the classification results of the three algorithms, Random Forest showed excellent performance with high precision, recall, and F1-score for each class, signifying its ability to provide accurate and balanced predictions. In particular, the model was effective in detecting diabetes and prediabetes cases, and had a low false positive rate. On the other hand, Naive Bayes showed lower performance, especially in the Prediabetes class, with a very low precision rate. K-NN, although performing well in the No diabetes class, faced difficulties in detecting prediabetes cases, characterized by low precision and recall in the Prediabetes class. Overall, Random Forest emerges as a better choice for classification purposes in this context, while Naive Bayes and K-NN show some weaknesses, especially with regard to prediabetes cases. Model selection depends on the specific objectives and data characteristics, but the results of this evaluation provide an initial view of the relative performance of each algorithm in diabetes classification tasks.

Previous research used three classifiers, namely Naive Bayes, Random Forest, and K-NN to predict CKD. This study shows that the Random Forest classifier can be used to predict CKD with high accuracy. This classifier can be used to help doctors diagnose CKD early, so that treatment can be done early and prevent the disease from getting worse [33].

This study examines the performance of supervised learning algorithms for diabetes risk classification. From the results, the Random Forest algorithm showed the most superior performance with an accuracy rate of 99%, outperforming K-NN (85%) and NBC (74%). This confirms the reliability of Random Forest in handling datasets with high dimensions and complexity. As a comparison, research by authors in [12] also tested supervised learning algorithms for early prediction of type II diabetes. The study found that Random Forest had the highest accuracy of 93.75%, outperforming other algorithms such as Logistic Regression and Support Vector Machine. This study supports our research findings that Random Forest is the optimal choice of algorithm for chronic disease risk classification, including diabetes [12].

These two studies illustrate that the utilization of Random Forest is not only effective for medical datasets but also relevant for other scenarios that require accurate classification considering various variables. Thus, the results of this study can be an important reference in the development of artificial intelligence-based systems to support clinical decisions and disease prevention.

4. CONCLUSION

In the context of this diabetes classification research, the performance comparison results among the K-Nearest Neighbor (K-NN), Naive Bayes Classification (NBC), and Random Forest algorithms indicate that Random Forest leads with high accuracy and precision, reaching 99%. This algorithm successfully provides accurate and balanced predictions for each class, including Without Diabetes, Prediabetes, and Diabetes. In contrast, although Naive Bayes excels in identifying individuals without diabetes with high precision, its precision significantly decreases in prediabetes and diabetes cases, reflecting challenges in classifying these conditions. Meanwhile, K-NN, effective in classifying cases without diabetes, faces difficulty in detecting prediabetes, as indicated by its low precision. Therefore, while model selection depends on specific goals, the comparison results emphasize the dominance of Random Forest in providing accurate predictions for diabetes, while Naive Bayes and K-NN exhibit some limitations. Positive support from previous studies regarding the effectiveness of Random Forest in predicting chronic diseases also provides additional positive considerations regarding the potential for early diagnosis and timely intervention.

REFERENCES

[1] P. Saeedi et al., “Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition,” Diabetes Res Clin Pract, vol. 157, Nov. 2019, doi: 10.1016/j.diabres.2019.107843.

[2] E. Kalyva, E. Malakonaki, C. Eiser, and D. Mamoulakis, “Health-related quality of life (HRQoL) of children with type 1 diabetes mellitus (T1DM): Self and parental perceptions,” Pediatr Diabetes, vol. 12, no. 1, pp. 34–40, Feb. 2011, doi: 10.1111/j.1399-5448.2010.00653.x.

[3] V. Vijayan and A. Ravikumar, “Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus,” 2014.

[4] F 2019 Classification of diabetes mellitus. 2019. [Online]. Available: http://apps.who.int/bookorders.

[5] J. Tanoey and H. Becher, “Diabetes prevalence and risk factors of early-onset adult diabetes: results from the Indonesian family life survey,” Glob Health Action, vol. 14, no. 1, 2021, doi: 10.1080/16549716.2021.2001144.

[6] A. Algarni, “Data Mining in Education.” [Online]. Available: www.ijacsa.thesai.org

[7] L. Chaves and G. Marques, “Data mining techniques for early diagnosis of diabetes: A comparative study,” Applied Sciences (Switzerland), vol. 11, no. 5, pp. 1–12, Mar. 2021, doi: 10.3390/app11052218.

[8] B. Shivananda Nayak et al., “The association of age, gender, ethnicity, family history, obesity and hypertension with type 2 diabetes mellitus in Trinidad,” Diabetes and Metabolic Syndrome: Clinical Research and Reviews, vol. 8, no. 2, pp. 91–95, 2014, doi: 10.1016/j.dsx.2014.04.018.

[9] W. K. Grylls, J. E. McKenzie, C. C. Horwath, and J. I. Mann, “Lifestyle factors associated with glycaemic control and body mass index in older adults with diabetes,” Eur J Clin Nutr, vol. 57, no. 11, pp. 1386–1393, Nov. 2003, doi: 10.1038/sj.ejcn.1601700.

[10] B. C. K. Choi and F. Shi, “Risk factors for diabetes mellitus by age and sex: results of the National Population Health Survey,” Diabetologia, vol. 44, no. 10, pp. 1221–1231, Oct. 2001, doi: 10.1007/s001250100648.

[11] D. Sisodia and D. S. Sisodia, “Prediction of Diabetes using Classification Algorithms,” in Procedia Computer Science, Elsevier B.V., 2018, pp. 1578–1585. doi: 10.1016/j.procs.2018.05.122.

[12] S. M. Ganie and M. B. Malik, “Comparative analysis of various supervised machine learning algorithms for the early prediction of type-II diabetes mellitus,” Int J Med Eng Inform, vol. 14, no. 6, pp. 473–483, 2022.

[13] A. F. Lubis et al., “Classification of Diabetes Mellitus Sufferers Eating Patterns Using K-Nearest Neighbors, Naïve Bayes and Decission Tree,” Public Research Journal of Engineering, Data Technology and Computer Science, vol. 2, no. 1, pp. 44–51, Apr. 2024, doi: 10.57152/predatecs.v2i1.1103.

[14] M. Q. Hatem, “Skin lesion classification system using a K-nearest neighbor algorithm,” Vis Comput Ind Biomed Art, vol. 5, no. 1, Dec. 2022, doi: 10.1186/s42492-022-00103-6.

[15] A. I. Putri et al., “Implementation of K-Nearest Neighbors, Naïve Bayes Classifier, Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction,” Public Research Journal of Engineering, Data Technology and Computer Science, vol. 2, no. 1, pp. 26–33, Apr. 2024, doi: 10.57152/predatecs.v2i1.1110.

[16] X. Liu and Z. Wang, “Deep Learning in Medical Image Classification from MRI-based Brain Tumor Images,” Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.00636

[17] A. Rahmah, N. Sepriyanti, M. H. Zikri, I. Ambarani, and M. Yusuf Bin Shahar, “Implementation of Support Vector Machine and Random Forest for Heart Failure Disease Classification,” Public Research Journal of Engineering, Data Technology and Computer Science, vol. 1, no. 1, pp. 34–40, 2023, [Online]. Available: https://journal.irpi.or.id/index.php/predatecs/article/view/816

[18] T. Grätz, S. Vospernik, and C. Scheidl, “Evaluation of afforestations for avalanche protection with orthoimages using the random forest algorithm,” Eur J For Res, vol. 143, no. 2, pp. 581–601, Apr. 2024, doi: 10.1007/s10342-023-01640-2.

[19] M. Radja and A. W. R. Emanuel, “Performance Evaluation of Supervised Machine Learning Algorithms Using Different Data Set Sizes for Diabetes Prediction,” in 2019 5th International Conference on Science in Information Technology (ICSITech), IEEE, Oct. 2019, pp. 252–258. doi: 10.1109/ICSITech46713.2019.8987479.

[20] R. Kamali, Y. S. Sari, I. Aldmour, and R. Budiarto, “Verification of Covid-19 Social Assistance Recipients using Naïve Bayes Classifier,” International Journal of Emerging Multidisciplinaries: Computer Science & Artificial Intelligence, vol. 1, no. 2, pp. 1–12, Sep. 2022, doi: 10.54938/ijemdcsai.2022.01.2.100.

[21] A.-J. Gallego, J. Calvo-Zaragoza, and J. R. Rico-Juan, “Insights into efficient k-nearest neighbor classification with convolutional neural codes,” IEEE Access, vol. 8, pp. 99312–99326, 2020.

[22] M. H. Effendy, D. Anggraeni, Y. S. Dewi, and A. F. Hadi, “Classification of Bank Deposit Using Naïve Bayes Classifier (NBC) and K–Nearest Neighbor (K-NN),” in International Conference on

Mathematics, Geometry, Statistics, and Computation (IC-MaGeStiC 2021), Atlantis Press, 2022, pp. 163–166.

[23] X. Zhou, P. Lu, Z. Zheng, D. Tolliver, and A. Keramati, “Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree,” Reliab Eng Syst Saf, vol. 200, p. 106931, 2020.

[24] N. Rachdaoui, “Insulin: The friend and the foe in the development of type 2 diabetes mellitus,” Mar. 01, 2020, MDPI AG. doi: 10.3390/IJMS21051770.

[25] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “KNN Model-Based Approach in Classification.”

[26] S. Raschka, J. Patterson, and C. Nolet, “Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence,” Information, vol. 11, no. 4, p. 193, 2020.

[27] Ramachandran A, “Know the signs and symptoms of diabetes,” Indian J Med Res, 2014.

[28] I. , S. D. , & K. A. Saini, “QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases,” J Adv Res, 2013.

[29] T. R. Patil and M. S. S. Sherekar, “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification,” International Journal Of Computer Science And Applications, vol. 6, no. 2, 2013, [Online]. Available: http://www.cs.bme.hu/~kiskat/adatb/bank-data-

[30] J. D. M. Rennie, “massachusetts institute of technology-artificial intelligence laboratory Improving Multi-class Text Classification with Naive Bayes Improving Multi-class Text Classification with Naive Bayes,” 2001.

[31] Y. Xia, “Correlation and association analyses in microbiome study integrating multiomics in health and disease,” 2020, pp. 309–491. doi: 10.1016/bs.pmbts.2020.04.003.

[32] Z. C. Dwinnie et al., “Application of the Supervised Learning Algorithm for Classification of Pregnancy Risk Levels,” vol. 1, no. 1, pp. 26–33, 2023, [Online]. Available: https://journal.irpi.or.id/index.php/predatecs/article/view/806

[33] R. , A. S. V. , & S. V. Devika, “Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest. In 2019 3rd International conference on computing methodologies and communication (ICCMC),” 3rd International conference on computing methodologies and communication (ICCMC), 2019.

[34] N. T. Luchia, M. Mustakim, N. Noviarni, K. Sussolaikah and T. Arifianto, "Feature Selection In Support Vector Machine And Random Forest Algorithms For The Classification Of Recipients Of The Smart Indonesia Program," 2024 International Conference on Circuit, Systems and Communication (ICCSC), Fes, Morocco, 2024, pp. 1-6, doi: 10.1109/ICCSC62074.2024.10616886.