e-ISSN: 2614-8404
p-ISSN: 2776-3234

# Comparative Analysis of Machine Learning Methods in Predicting Diabetes Risk Based on Genetic Data

**Sekar Ayu Wijaya Kusumaningrum[1], Oleh Soleh[2], Muhamad Yusup[3]**
[1]Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Raharja
[2]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Raharja
[3]Program Studi Digital Bisnis, Fakultas Ekonomi dan Bisnis, Universitas Raharja
Email: [1]sekar.ayu@raharja.info, [2]oleh.soleh@raharja.info, [3]yusuf@raharja.info

***Abstract*** – Type 2 Diabetes Mellitus (T2DM) is a global chronic disease caused by the interaction of genetic and environmental factors. The use of genetic data offers great potential for early detection and personalized intervention. However, the complex analysis of genetic data requires sophisticated approaches like machine learning. This study aims to compare the performance of three machine learning algorithms Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN) in predicting T2DM risk based on genetic data. By using a Systematic Literature Review of studies published between 2019 and 2024, the accuracy data from each algorithm was compared. The analysis results show that Random Forest has the best performance with an accuracy of 99.3%. This algorithm excels due to its ability to handle high-dimensional datasets and reduce overfitting. In comparison, KNN achieved an accuracy of 87% and Logistic Regression 82%. These findings support the integration of machine learning into early detection systems and more precise and efficient clinical decision-making for T2DM management.

***Keywords – Type 2 Diabetes; Machine Learning; Random Forest; Logistic Regression; K-Nearest Neighbors; Genetic Data***

## I. INTRODUCTION

Type 2 Diabetes is a chronic disease that is currently one of the major global health problems. This metabolic disease, known as type 2 diabetes mellitus, has various etiologies and is characterized by high blood sugar levels. This condition also leads to abnormalities in carbohydrate, fat, and protein metabolism due to inadequate insulin function. Insulin plays a vital role in maintaining stable blood sugar levels and helping body cells absorb glucose [1]. According to data from the International Diabetes Federation (IDF) in 2021, 537 million adults aged 20–79 worldwide suffer from diabetes. This number is projected to increase to 643 million by 2030 and 783 million by 2045. Risk factors contributing to this disease include age, gender, obesity, hypertension, genetics, diet, and lifestyle [2]. Indonesia ranks fifth as the country with the most diabetes sufferers in the world, with 19.5 million cases in 2021 and an estimated 28.6 million cases by 2045.

The pathophysiology of type 2 diabetes primarily involves two mechanisms: insulin resistance and progressive pancreatic beta-cell dysfunction [3]. Insulin resistance is a condition in which the body's cells, especially in vital organs such as the liver, muscles, and fat tissue, do not respond effectively to insulin. This disorder leads to an increase in blood glucose levels or hyperglycemia, which prompts the pancreas to produce more insulin (hyperinsulinemia) as a compensatory response. Insulin resistance is often triggered by the accumulation of visceral fat, a sedentary lifestyle, and unhealthy eating habits [4].

In the field of artificial intelligence, Machine Learning (ML) is a branch that allows a system to learn from data and then make predictions or decisions without being explicitly programmed [5]. In the health sector, ML plays a crucial role in diagnosing diseases, predicting risks, and personalizing treatment therapies [6]. ML is ideal for analyzing genetic data due to its ability to identify complex patterns and non-linear relationships among diverse genetic features [7]. Specifically in predicting diseases like diabetes, ML algorithms can be used to identify high-risk individuals by analyzing genetic data and other factors, enabling earlier prevention and management efforts.

Type 2 diabetes is a complex condition resulting from the interaction between genetic and environmental factors. In recent years, several genetic variants have been identified that can increase the risk of developing type 2 diabetes. Mutations in certain genes, which are often found in individuals with a family history of type 2 diabetes mellitus, play an important role in glucose metabolism and insulin regulation. These changes can cause dysfunction in insulin secretion and cell resistance to it [8]-[3]. Some genes that have been identified to be linked to type 2 diabetes include:

a. Gen TCF7L2: plays a role in the process of insulin secretion.
b. Gen ABCC8: helps regulate insulin.
c. Gen GLUT2: supports glucose uptake in the pancreas.
d. Gen GCGR: involved in glucose regulation along with the hormone glucagon.

In addition, the CAPN10 gene has been linked to the incidence of type 2 diabetes mellitus, especially in populations in America, Mexico, and the Javanese ethnic group in Indonesia [9]. Changes in these genes can affect various biological processes, including glucose metabolism, sensitivity, and insulin secretion by pancreatic beta cells [10]. Therefore, using genetic data to predict diabetes risk opens up great opportunities for the

e-ISSN: 2614-8404
p-ISSN: 2776-3234

development of early detection strategies and personalized interventions.

Genetic data has high complexity due to its characteristics, such as large size and non-linear interactions between genes or significant variation among individuals. Therefore, the analysis of genetic data requires a careful and sophisticated approach, such as the use of Machine Learning (ML) techniques.

ML, as a branch of Artificial Intelligence, is a field of science that focuses on the application and development of algorithms that enable computers to adapt and learn from empirical data [11]. Its ability to process complex data makes ML an ideal method for analyzing genetic data.

This research has high relevance in the era of global health technology, where data-based risk prediction is a key focus in the development of health systems that adopt a precision medicine approach. Early detection in high-risk individuals can facilitate the implementation of more efficient and effective preventive measures [12]. This approach not only has the potential to alleviate the economic burden associated with treatment but also to improve the overall quality of life of the community [13].

Based on this background, this study aims to conduct a comparative analysis of the performance of three machine learning algorithms, namely Logistic Regression, Random Forest, and K-Nearest Neighbors, in predicting the risk of type 2 diabetes using genetic data.

Logistic Regression is one of the machine learning algorithms used for classification tasks. This algorithm is a special form of regression analysis that uses a binary response variable and predictors that can be continuous, categorical, or a mixture of both [14]. The advantage of this analysis is that it does not require assumptions of normal multivariate distribution or equality of the covariance matrix, and can be applied to various data scales [15].

Random Forest is a machine learning technique that combines several decision trees to make predictions [16]. Each tree in this model is built separately, using a randomly selected subset of the training data. The predictions generated by each tree are then combined to produce a final prediction [17].

K-Nearest Neighbors (KNN) is a non-parametric classification algorithm that classifies a sample based on its proximity to other samples in the feature space. This algorithm is generally used to classify objects based on training data that has the closest distance to its "neighbors". This proximity can be calculated using Euclidean distance [18].

This research project was carried out systematically by reviewing literature published between 2019–2024. The purpose of this review is to gain a deep understanding of the effectiveness of each algorithm under various conditions and datasets. The results of this study are expected to contribute to the development of accurate and flexible genetic data-based prediction systems that are adaptable to population trends. In addition, this analysis also supports the integration of machine learning technology into a precise clinical decision-making system, thus helping in the early detection and prevention of diseases more optimally.

## II. RESEARCH METHODOLOGY

The research method used is a Systematic Literature Review (SLR), which is a structured approach to collecting, identifying, and evaluating articles or research relevant to a specific topic [19]. This method was chosen because it has several important advantages over traditional literature reviews.

The SLR (Systematic Literature Review) method allows research to reduce bias because the process follows strict and transparent protocols at every stage, from searching to data analysis. This differs from narrative literature reviews, which are often subjective and do not have standard selection criteria [20]. By applying clear inclusion and exclusion criteria, SLR can increase the validity and reliability of the findings. The results also become more reliable and can be replicated by other researchers.

The choice of the SLR (Systematic Literature Review) method is very effective for providing a comprehensive synthesis of existing research, answering research questions in depth, and helping to identify research gaps that can be the basis for future studies [21].

The overall research flow can be illustrated through the flowchart in Figure 1. Figure 1 presents a comprehensive overview of all stages of this research methodology, which uses the SLR approach. Each step is designed to ensure the research process runs systematically and objectively, in order to obtain accurate and accountable results.



e-ISSN: 2614-8404
p-ISSN: 2776-3234

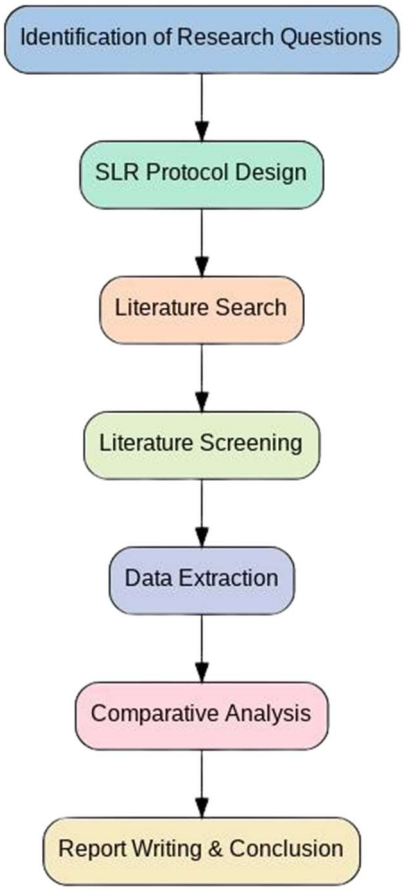


**Figure. 1** Research flow diagram

The detailed explanation for each stage in the research flowchart is as follows:

1. **Identification of Research Questions:** This first stage forms the foundation of the entire research. Here, the research questions are formulated specifically and clearly. These questions serve as the main guide for the literature search and analysis process that will be conducted.

2. **SLR Protocol Design:** At this stage, the researcher establishes a strict and transparent protocol to ensure the article screening process runs systematically. This protocol includes inclusion criteria (articles to be included) and exclusion criteria (articles to be excluded) that will be used to filter articles, thereby increasing the validity and reliability of the findings. Below is a summary of the criteria used:

**Table 1**. Table Inclusion and Exclusion

| Criteria | Inclusion | Exclusion |
|---|---|---|
| Timeframe | Scientific journals published between 2019-2024. | Journals published outside the 2019-2024 range. |
| Topic | Studies applying Logistic Regression, Random Forest, or K-Nearest Neighbors algorithms for type 2 diabetes risk prediction. | Studies that do not discuss these three algorithms or do not focus on diabetes risk prediction. |
| Data Source | Journals from leading academic databases such as Google Scholar and ResearchGate. | Publications not from these specified academic databases. |

3. **Literature Search:** Based on the designed protocol, a literature search is conducted in several leading academic databases, such as Google Scholar and ResearchGate. These databases were chosen because they provide access to a wide range of scientific journals relevant to computer science and data science topics, ensuring a comprehensive and systematic literature review. The search is performed using a combination of specific keywords to effectively filter the results [22].

4. **Literature Screening:** In this stage, the articles found are filtered according to the inclusion and exclusion criteria set in the previous stage.

5. **Data Extraction:** The data analyzed in this study did not come from direct experiments, but rather from reviews of scientific journals published over the past five years (2019-2024). Each reviewed journal contains the results of applying the three algorithms to genetic datasets for diabetes risk prediction.

   The datasets commonly used in the analyzed journals include clinical and genetic attributes, such as:

   - Number of pregnancies
   - Glucose levels
   - Blood pressure
   - Skin thickness
   - Insulin levels
   - Body Mass Index (BMI)
   - Diabetes pedigree function
   - Age
   - Diabetes status (outcome)

   These attributes are used as input variables in the machine learning model training process [23]. The accuracy data from each method is then compared to determine the algorithm with the best performance in predicting diabetes risk based on genetic data.

6. **Comparative Analysis:** In this stage, the data extracted from the selected articles will be

e-ISSN: 2614-8404
p-ISSN: 2776-3234

analyzed and compared to evaluate the performance of the three machine learning algorithms: Random Forest, K-Nearest Neighbors, and Logistic Regression. This analysis will focus on key performance metrics, including accuracy, precision, recall, and F1-score. The results of this comparison will form the basis for determining which algorithm demonstrates the most optimal performance in predicting type 2 diabetes risk.

7. **Report Writing & Conclusion:** The final stage of this research flow is the preparation of the report and the formulation of a conclusion. All findings obtained from the Comparative Analysis stage will be synthesized and presented systematically to answer the research questions established at the beginning. This report will include an in-depth discussion of the algorithm performance comparison results, interpretation of each finding, and the formulation of concise and solid conclusions. Additionally, this section will also contain suggestions for future research as a contribution to the advancement of knowledge.

## III. RESULTS AND DISCUSSION

Based on a comprehensive literature study and analysis of the performance of each machine learning method in predicting diabetes risk based on genetic data, the comparative results presented in Table 1 were obtained.

**Table 2.** Comparison of Machine Learning Algorithm Performance

| Model | Accuracy | Recall | Presisi | F1-Score |
|---|---|---|---|---|
| Random Forest [24]. | 99,3% | 99,5% | 99,1% | 99% |
| KNN [25]. | 87% | 77% | 95% | 85% |
| Logistic Regression [26] | 82% | 79% | 81% | 80% |

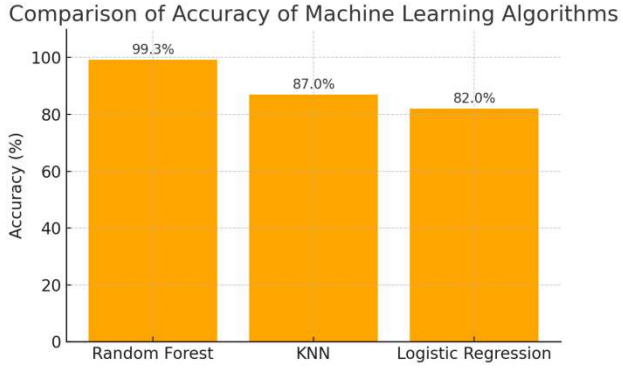


**Figure. 2** Algorithm Performance Comparison Bar Chart.

As depicted in Table 2 and Figure 2, from the results displayed, it is clear that the Random Forest (RF) shows the most superior performance in predicting diabetes risk. With an accuracy of 99.3%, RF significantly outperforms K-Nearest Neighbors (KNN) at 87% and Logistic Regression (LR) at 82%. This superior performance is also reflected in its other metrics, such as recall (99.5%), precision (99.1%), and F1-score (99%).

The main advantage of Random Forest (RF) is its ability to handle complex, high-dimensional datasets, such as genetic data, while minimizing the problem of overfitting. This occurs because the algorithm combines the results from many decision trees to produce a more stable and accurate prediction. The high recall and precision values indicate that the RF model is highly effective. It is not only proven accurate in identifying individuals at risk of diabetes (positive cases) but also capable of accurately distinguishing them from healthy individuals.

In contrast, K-Nearest Neighbors (KNN), although achieving high precision at 95%, shows a lower recall value of 77%. This indicates that while KNN is quite effective at identifying positive cases, it has limitations in its sensitivity to minority data or underrepresented cases in the dataset. This limitation can be a drawback in clinical scenarios, where high sensitivity is crucial to avoid potentially harmful false negatives.

Meanwhile, Logistic Regression (LR) shows the lowest performance among the three algorithms, with an accuracy of 82%. Nevertheless, LR has an important advantage: its ease of interpretation. In a medical context, understanding the predictor factors is crucial. A well-interpretable model, like LR, allows doctors to understand each variable's contribution to predicting diabetes risk.

It is important to note, however, that the variation in performance of Logistic Regression and K-Nearest Neighbors in this study could be due to differences in dataset characteristics, such as sample size, data attributes used, or patient population. Overall, the results of this discussion reaffirm the main findings of literature studies that demonstrate the effectiveness of machine learning algorithms in genetic data analysis for clinical purposes. Random Forest proves to be the most optimal choice for type 2 diabetes risk prediction, and this finding directly addresses the research objectives set.

.

## IV. CONCLUSION

Type 2 Diabetes Mellitus is a complex global health challenge where early detection through genetic data analysis is crucial. This study, using a Systematic Literature Review (SLR) approach, aimed to evaluate and compare the performance of machine learning algorithms as a predictive solution. The analysis confirms the main finding that the Random Forest algorithm achieves the highest accuracy at 99.3%. Additionally, Random Forest also shows very high recall, precision, and F1-scores, all above 99%. This superiority is due to its excellent ability to classify data accurately and consistently, especially when handling high-dimensional datasets and minimizing overfitting.

For comparison, K-Nearest Neighbors (KNN) shows good performance in terms of precision (95%) but has limitations with a lower recall (77%). This indicates that

e-ISSN: 2614-8404
p-ISSN: 2776-3234

while effective at identifying positive cases, its sensitivity to minority data is limited. Meanwhile, although Logistic Regression (LR) has lower accuracy (82%), this algorithm remains relevant due to its advantage in model interpretation. This capability is very important in prediction cases that require a high understanding of the factors contributing to the outcome.

Overall, this research demonstrates that the integration of Machine Learning, particularly Random Forest, into genetic data analysis has great potential for developing more precise and efficient early detection systems and clinical decision-making in managing type 2 diabetes. For future research, there are several suggestions that can be explored further to strengthen the validity and contribution of these findings. It is recommended to validate the existing models with more diverse genetic datasets, from different populations or larger and more detailed clinical data centers to ensure the models generalizability and robustness.

Second, subsequent studies shouls focus on methodological refinement. This includes exploring more advanced Machine Learning algorithms, such as Deep Learning (e.g., Convolutional Neural Networks or Recurrent Neural Networks), which may better capture complex interactions within genetic features, as well as incorporating sophisticated feature selection methods to further enhance predictive performance and model interpretability.

Finally, a long-term goal would be the development of a user-friendly clinical decision support system prototype based on the validated model. Such a system could enable clinicians to directly utilize this genetic risk prediction tool in real-time healthcare settings, facilitating personalized and preventative management of type 2 diabetes.

Furthermore, the exploration of more advanced feature selection methods can identify the genes or attributes most contributing to diabetes risk, making the models not only more efficient but also more interpretable. Finally, a comparison of Random Forest's performance with increasingly popular Deep Learning algorithms, such as Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN), can be conducted to explore the potential for higher performance, especially in handling complex gene interactions that traditional models might not capture.

## REFERENCES

[1] Hovi, H. S. W., Id Hadiana, A., & Rakhmat Umbara, F. (2023). Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM). Informatics and Digital Expert (INDEX), 4(1), 40–45, doi: 10.36423/index.v4i1.895

[2] Widiasari, K. R., Made, I., Wijaya, K., & Suputra, P. A. (2021). Diabetes Melitus Tipe 2: Faktor Risiko, Diagnosis, dan Tatalaksana. In Ganesha Medicina Journal (Vol. 1), doi: 10.23887/gm.v1i2.40006

[3] Fatmona, F. A., Permana, D. R., & Sakurawati, A. (2023). Gambaran Tingkat Pengetahuan Masyarakat tentang Pencegahan Diabetes Melitus Tipe 2 di Puskesmas Perawatan Siko. MAHESA: Malahayati Health Student Journal, 3(12), doi: 10.33024/mahesa.v3i12.12581

[4] Mulya Harahap, Raja & Rostini, Tiene & Suraya, Nida. (2024). Pemeriksaan Laboratorium pada Resistansi Insulin. Action Research Literate. 8. 3625-3632. 4 doi: 10.46799/arl.v8i12.2569

[5] Hidayat, R., Sy, YS, Sujana, T., Husnah, M., Saputra, HT, & Okmayura, F. (2024). Implementasi Machine Learning Untuk Prediksi Penyakit Jantung Menggunakan Algoritma Support Vector Machine. BIOS : Jurnal Teknologi Informasi Dan Rekayasa Komputer , 5 (2), 161-168, doi:10.37148/bios.v5i2.152

[6] Wardhana, R. G., Wang, G., & Sibuea, F. (2023). Penerapan Machine Learning Dalam Prediksi Tingkat Kasus Penyakit di Indonesia. Journal of Information System Management (JOISM), 5(1). doi: 10.24076/joism.2023v5i1.1136

[7] Siswaja, H. D., & Ramdhani, Y. (2024). Pendekatan Algoritma Neural Network dan Genetic Algorithm Untuk Prediksi Penyakit Ginjal Kronis.. Jurnal Responsif: Riset Sains dan Informatika, 6(2), 232-239, doi: 10.51977/jti.v6i2.1778

[8] Ardika, O. B., Larasati, T. A., Suharmanto, & Kurniati, I. (2024). Impaired Insulin Secretion and Sensitivity in Adolescents with Family History of Type 2 Diabetes Mellitus. Medical Profession Journal of Lampung, 14(1). PDF: https://journalofmedula.com/index.php/medula/article/download/943/744/5640

[9] Tursinawati, Y., Hakim, R. F., Rohmani, A., Kartikadewi, A., & Sandra, F. (2020). CAPN10 SNP-19 is Associated with Susceptibility of Type

e-ISSN: 2614-8404
p-ISSN: 2776-3234

2 Diabetes Mellitus: A Javanese Case-control Study. The Indonesian Biomedical Journal, 12(2), doi: 10.18585/inabj.v12i2.984

[10] Angria, N. A. (2024). Polimorfisme Gen VDR Fok1 Pada Penderita Diabetes Melitus Menggunakan PCR-RFLP. Journal of Nursing and Health, 9(2), 259–267, doi: 10.52488/jnh.v9i2.362

[11] Silalahi, A. P., Simanullang, H. G., & Hutapea, M. I. (2023). Supervised Learning Metode K-Nearest Neighbor Untuk Prediksi Diabetes Pada Wanita. METHOMIKA: Jurnal Manajemen Informatika & Komputerisasi Akuntansi, 7(1), 144–149, doi: 10.46880/jmika.Vol7No1.pp144-149

[12] Olina, Y. B., Ernawati, E., Aisah, S., Al Jihad, M. N., Setyawati, D., Baidhowy, A. S., & Arifianto, N. (2024). Meningkatkan Kesadaran Hidup Sehat Melalui Skrining Deteksi Dini Penyakit Tidak Menular di Lingkungan Universitas Muhammadiyah Semarang. SALUTA: Jurnal Pengabdian Kepada Masyarakat, 4(1), doi: 10.26714/sjpkm.v4i1.16404

[13] Maliangkay, K. S., Rahma, U., Putri, S., & Istanti, N. D. (2023). Analisis Peran Promosi Kesehatan Dalam Mendukung Keberhasilan Program Pencegahan Penyakit Tidak Menular Di Indonesia. Jurnal Medika Nusantara, 1(2), 108–122, doi: 10.59680/medika.v1i2.284

[14] Rafiq, M., Rahmadani, A. A., Putri, A. A., Happy, D. M., Julia, J., Dala, M. A. D., Angka, M. T., & Wasono, W. (2023). Analisis Regresi Logistik Biner Untuk Memprediksi Faktor-Faktor Internal Yang Memengaruhi Keharmonisan Rumah Tangga Menurut Provinsi Di Indonesia Pada Tahun 2021. Prosiding Seminar Nasional Matematika dan Statistika, 3(1).

[15] Suhendra, M. A., Ispriyanti, D., & Sudarno, S. (2020). Ketepatan Klasifikasi Pemberian Kartu Keluarga Sejahtera di Kota Semarang Menggunakan Metode Regresi Logistik Biner dan Metode Chaid. Jurnal Gaussian, 9(1), 64–74, doi: 10.14710/j.gauss.v9i1.27524

[16] Susanto, T., Rahmaniar, F., Lestari, D. W., & Abdullah, K. (2020). Thermal aging and chemical resistance evaluation of carbon black filled natural rubber blending: effect of the composition of acrylo nitrile and styrene butadiene rubber. IOP Conference Series: Materials Science and Engineering, 980, 012002, doi: 10.1088/1757-899X/980/1/012002

[17] Salsabil, M., Azizah, N. L., & Eviyanti, A. (2024). Implementasi Data Mining Dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest Dan Xgboost. Jurnal Ilmiah Komputasi, 23(1), 51–58, doi: 10.32409/jikstik.23.1.3507

[18] Deng, S., Wang, L., Guan, S., Li, M., & Wang, L. (2023). Non-parametric Nearest Neighbor Classification Based on Global Variance Difference. International Journal of Computational Intelligence Systems, 16(1), 26, doi: 10.1007/s44196-023-00200-1

[19] Triandini, E., Jayanatha, S., Indrawan, A., Putra, G. W., & Iswara, B. (2019). Metode Systematic Literature Review untuk Identifikasi Platform dan Metode Pengembangan Sistem Informasi di Indonesia. Indonesian Journal of Information Systems, 1(2), 63–77, doi: 10.24002/ijis.v1i2.1916

[20] Simamora, S. C., Gaffar, V., & Arief, M. (2024). Systematic Literatur Review Dengan Metode Prisma: Dampak Teknologi Blockchain Terhadap Periklanan Digital. JURNAL ILMIAH M-PROGRESS, 14(1), 1–11, doi: 10.35968/m-pu.v14i1.1182

[21] Kolaski, K., Logan, L. R., & Ioannidis, J. P. A. (2023). Guidance to best tools and practices for systematic reviews. Systematic Reviews, 12(1), 96, doi: 10.1186/s13643-023-02255-9

e-ISSN: 2614-8404
p-ISSN: 2776-3234

[22] Klopfenstein, D. V., & Dampier, W. (2021). Commentary to Gusenbauer and Haddaway 2020: Evaluating retrieval qualities of Google Scholar and PubMed. Research Synthesis Methods, 12(2), 126–135, 10.1002/jrsm.1456

[23] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. ICT Express, 7(4), 432–439, doi: 10.1016/j.icte.2021.02.004

[24] Sriyanto, & Supriyatna, A. R. (2023). Prediksi Penyakit Diabetes Menggunakan Algoritma Random Forest. TEKNIKA, 8051410, doi: 10.5281/zenodo.8051410

[25] Buani, D. C. P. (2024). Deteksi Dini Penyakit Diabetes dengan Menggunakan Algoritma Random Forest. EVOLUSI: Jurnal Sains dan Manajemen, 12(1), 1–8, doi: 10.31294/evolusi.v12i1.21005

[26] Erlin, E., Marlim, Y. N., Junadhi, J., Suryati, L., & Agustina, N. (2022). Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm. Jurnal Nasional Teknik Elektro dan Teknologi Informasi, 11(2), 88–96, doi: 10.22146/jnteti.v11i2.3586