Indonesian Journal of Data and Science

Volume 5 Issue 3
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v5i3.190

Research Article

Grid Search Hyperparameter Analysis in Optimizing The Decision Tree Method for Diabetes Prediction

Desi Anggreani [1,*] Hamdani [2,] Nurmisba [3,] Lukman [4,]

[1] Universitas Muhammadiyah Makassar, Makassar, Indonesia, desianggreani@unismuh.ac.id
[1] Universitas Muhammadiyah Makassar, Makassar, Indonesia, hamdaniunismuh@gmail.com
[1] Universitas Muhammadiyah Makassar, Makassar, Indonesia, nurmisba2307@gmail.com
[1] Universitas Muhammadiyah Makassar, Makassar, Indonesia, lukman@unismuh.ac.id

Correspondence should be addressed to Desi Anggreani; desianggreani@unismuh.ac.id

Received 11 September 2024; Accepted 20 November 2024; Published 31 December 2024

Abstract:

Diabetes is a global health issue that continues to rise, especially in Indonesia, caused by unhealthy lifestyles, poor diets, and genetic factors. Early detection of diabetes risk is crucial to prevent serious complications, and machine learning offers innovative predictive solutions. This research focuses on the development of a diabetes risk prediction model using the Decision Tree algorithm with hyperparameter optimization through the Grid Search technique. The research methodology includes the collection of patient medical data with key attributes such as glucose levels, blood pressure, skin health, insulin, body mass index (BMI), diabetes pedigree, age, and health history. The hyperparameter tuning process is carried out by varying key parameters such as the maximum tree depth (max_depth), the minimum number of samples required to split a node (min_samples_split), and the minimum number of samples required at a leaf node (min_samples_leaf). Grid Search is used to systematically explore hyperparameter combinations in order to find the optimal configuration that can improve the model's performance. The research process includes data preprocessing, splitting the dataset into training and testing sets, model training, and evaluation using accuracy metrics, confusion matrix, and ROC AUC curve. The initial results show a model accuracy of 76%, which was then improved to 81% after hyperparameter optimization using Grid Search. The visualization of the decision tree reveals that glucose levels and BMI have the most significant contributions in predicting diabetes risk. This research demonstrates the potential of machine learning in supporting the early detection of diabetes, with the Decision Tree algorithm showing promising predictive capabilities. Nevertheless, further research with larger datasets and the integration of other algorithms is highly recommended to improve the accuracy and generalization of the model. The main contribution of this research is the development of a machine learning-based approach that can assist medical personnel in screening for diabetes risk more efficiently and accurately.

Keywords: Diabetes, Decision Tree, Hyperparameter Optimization, Grid Search.
Dataset link: http://bit.ly/3VfMr1J

1. Introduction

Health is a key aspect of a successful life, with a healthy body as a sign of avoiding disease. Health problems have experienced a pattern transition where diseases were initially dominated by infectious diseases and have now moved to Non-Communicable Diseases (NCDs). Non-communicable diseases include cancer, kidney disease, hypertension, diabetes and so on [1]. In addition, the number of physical and mental health patients increases and decreases over time. Events like this will have a negative impact on the preparation of the hospital or health workers. Lack of preparation in terms of early treatment for people with diseases [2].

Diabetes mellitus has become one of the increasingly complex and concerning global health issues. According to data from the International Diabetes Federation (IDF), the number of diabetes sufferers worldwide continues to increase significantly [3]. In 2021, it was recorded that 537 million people suffered from diabetes, with projections predicting an increase to 643 million by 2030 and reaching 783 million by 2045 [4]. The situation in Indonesia is not

much different. Our country ranks fifth in the world with the highest number of diabetes sufferers. In 2021, Indonesia recorded 19.5 million sufferers, with projections increasing to 28.6 million by 2045 [5]. This figure indicates an urgent need for a comprehensive approach to the prevention, early detection, and management of diabetes [6]. The complexity of diabetes lies not only in the increasing number of sufferers but also in the health consequences it causes. This disease is not just a metabolic disorder, but can also trigger various serious complications such as cardiovascular disease, kidney failure, nerve damage, and eye disorders [7]. Contributing risk factors include unhealthy lifestyles, poor dietary habits, lack of physical activity, obesity, and genetic predisposition [8].

The main challenge in managing diabetes is the limitation of effective early detection systems. The diagnostic process often takes a long time, requires complex manual analysis, and depends on the availability of limited medical resources [9]. This condition requires a more innovative approach that is faster, more accurate, and can be widely implemented. The development of machine learning technology offers innovative solutions to address these constraints. Artificial intelligence algorithms are capable of quickly analyzing large amounts of medical data, identifying risk patterns, and providing standardized predictions [10]. This approach not only accelerates the detection process but also improves the accuracy of identifying individuals at high risk of diabetes.

In the context of diabetes risk detection, the Decision Tree algorithm is one of the most promising machine learning approaches. Decision Tree is a non-parametric classification method capable of producing predictive models with a decision tree structure that is easy to understand and interpret [11]. The main advantage of this algorithm lies in its ability to capture non-linear relationships between predictor variables and target variables, as well as to identify the most significant risk factors [12]. The optimization of Decision Tree performance heavily relies on the hyperparameter tuning process, which is a critical aspect in the development of machine learning models. Hyperparameters are configuration parameters that cannot be directly learned from the training data but have a significant impact on the algorithm's performance [13]. In the context of Decision Trees, some key hyperparameters include the maximum tree depth (max_depth), the minimum number of samples required to split an internal node (min_sample_split), and the minimum number of samples required to be at a leaf node (min_samples_leaf).

The Grid Search technique has proven effective in systematically exploring hyperparameter combinations. This method performs an exhaustive search through a predetermined subset of hyperparameters, allowing for the identification of optimal configurations that maximize model performance [14]. Previous research has shown that hyperparameter optimization can significantly improve the accuracy of diabetes risk prediction. Several studies report an improvement in model performance by 10-15% through proper tuning processes [15]. However, the main challenge remains the model's ability not only to achieve high accuracy but also to provide meaningful interpretations for healthcare practitioners.

Previous research has demonstrated the effectiveness of machine learning in predicting diabetes risk, with several studies showing prediction accuracy above 80% [16]. The Decision Tree algorithm has specifically shown significant potential in health risk classification, including diabetes [17]. his research focuses on the development of a diabetes risk prediction model using the Decision Tree algorithm. Through a machine learning approach, the study aims to explore the algorithm's ability to classify risk based on key medical variables such as blood glucose levels, body mass index (BMI), blood pressure, age, and family history. This research aims to deeply explore the potential of Decision Tree in predicting diabetes risk, with a particular focus on hyperparameter optimization using Grid Search. Through this systematic approach, we hope to develop a predictive model that is not only accurate but also capable of providing meaningful insights in the efforts to prevent and manage diabetes.

2. Method:

This research is designed with a quantitative approach using the Decision Tree algorithm to predict diabetes risk [18]-[19]. The research process utilizes various tools and technologies, including the Python programming language and the Scikit-Learn library for the implementation of the Decision Tree algorithm, as well as visualization tools such as Matplotlib and Seaborn to understand data patterns. The dataset was processed through a cleaning process, such as handling missing values and ensuring data format consistency. Next, the data is divided into two parts 90% as training data to build the model and 10% as test data to evaluate the model's performance.

The research design encompasses several key stages conducted systematically, starting from data collection to the visualization of the final model [20]. The data collection process aims to gather patient medical information relevant to diabetes risk factors. The collected data then undergoes preprocessing, which includes data cleaning, transformation, and normalization to ensure optimal data quality before use. Subsequently, the dataset is split into training and testing sets to ensure the model's ability to generalize effectively. The next stage involves model training,

where machine learning algorithms are employed to learn patterns within the data. The performance of the trained model is then evaluated using specific metrics to verify its reliability in identifying diabetes risk factors. The results of the analysis and evaluation are visualized to provide a clear and in-depth understanding of the model's performance and its relevance to medical data. This structured approach is designed to ensure that the research includes a comprehensive analysis, as illustrated in **Figure 1**.

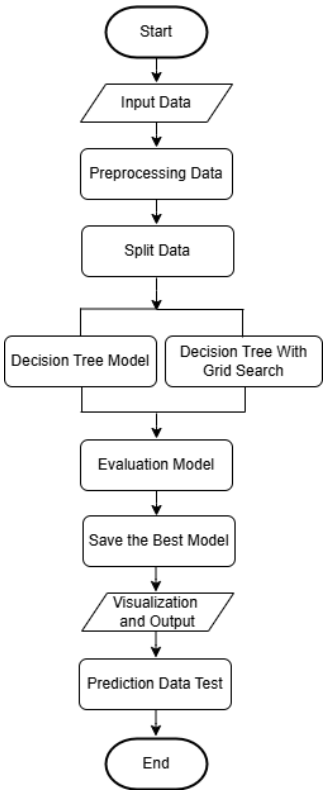

Figure 1. Flowchart of Decision Tree Optimization with Grid Search Hyperparameter method

The dataset used in this study was obtained from credible medical reference sources and consists of 768 records. It includes eight medical attributes, such as glucose levels, blood pressure, BMI, insulin levels, skin health, diabetes pedigree function, age, and a diagnostic label (positive or negative). This dataset was selected because it encompasses relevant medical attributes for analysing diabetes risk. The data is presented in **Table 1** below.

Table 1. Dataset

Glucose	Blood Pressure	Skin Health	Insulin	BMI	Diabetes Genealogy Function	Age	Prediction
148	72	35	0	33.6	0.627	50	1
85	66	29	0	26.6	0.351	31	0
183	64	0	0	23.3	0.672	32	1
89	66	23	94	28.1	0.167	21	0
137	40	35	168	43.1	2.288	33	1
116	74	0	0	25.6	0.201	30	0
78	50	32	88	31	0.248	26	1
115	0	0	0	35.3	0.134	29	0
197	70	45	543	30.5	0.158	53	1
…	…	…	…	…	…	…	…

The data analysis method involves constructing a decision tree using the Decision Tree algorithm based on attributes with the highest Information Gain or Gini Index [21]. The model is evaluated using metrics such as accuracy, confusion matrix, and ROC AUC to assess its predictive performance. To enhance performance, hyperparameter tuning is conducted using the Grid Search method. Grid Search is a hyperparameter optimization technique that explores various combinations of hyperparameter values to identify the best configuration based on evaluation metrics such as accuracy or F1-score [22]. By performing hyperparameter tuning, the resulting model achieves a more optimal accuracy level. This approach aims to produce a predictive model that can accurately and efficiently assist in the early detection of diabetes risk. **Figure 2** illustrate Decision Tree model.

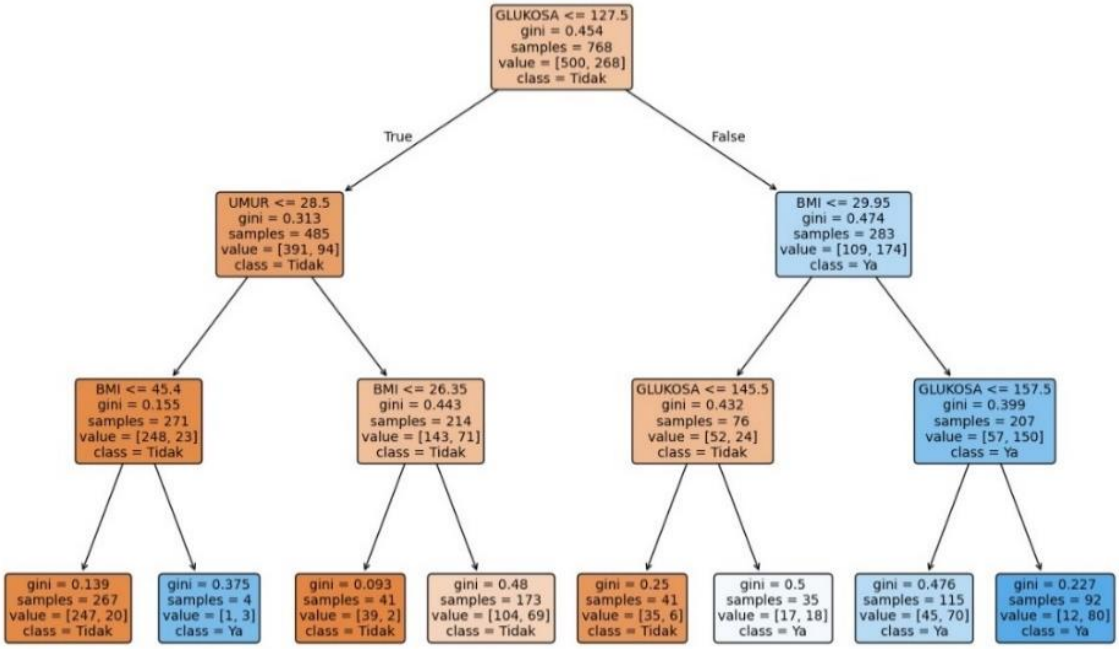

Figure 2. Visualization of The Decision Tree Method

The decision to predict diabetes risk is based on patients' medical attributes. This algorithm utilizes threshold values for each attribute to make decisions. At the root node, the algorithm examines blood glucose levels with a threshold of 127.5 mg/dL. Patients with glucose levels ≤ 127.5 mg/dL are further assessed based on age, categorized as either ≤ 28.5 years or > 28.5 years. Depending on age and BMI, patients are classified as either at risk of diabetes or not. For patients with glucose levels > 127.5 mg/dL, the algorithm checks BMI. If BMI ≤ 29.95, glucose levels are re-examined. Patients with glucose levels ≤ 145.5 mg/dL are classified as not at risk, whereas those with glucose levels > 145.5 mg/dL are classified as at risk. If BMI > 29.95, glucose levels are re-evaluated with a threshold of 157.5 mg/dL. Overall, this decision tree employs threshold values of medical attributes to classify patients into risk or non-risk categories for diabetes. Mathematically, the Decision Tree algorithm uses Entropy and Information Gain formulas to determine the optimal split at each tree node [22]. Entropy (S) is calculated using the formula:

$$Entropy(X) = -\sum_{i}^{n} p(Xi)\ log2\ p(Xi)$$

Description:

Entropy(X) : The information set of an event x.

P(X) : Probability of occurrence of event x.

n : Number of classes in the dataset.

$$G(D,t) = -\sum P(Ci)\log(Ci) + P(t)\sum P(Ci|t) + P(T)\sum P(Ci|T)\log P\,(Ci|T)$$

Description:

C : The information set of an event C.

$P(Ci)$: Probability of occurrence of a category.

$P(t)$: Probability of occurrence of word t in the text.

$P(T)$: Probability of non-occurrence of word t in the article.

$P(Ci|t)$: Probability of word t appearing in class i.

$P(Ci|T)$: Probability of non-occurrence of word t in class i.

3. Results and Discussion

Results

Based on the comparison graph of the initial model's accuracy (Initial Model) and the model after hyperparameter optimization using Grid Search (Post Grid Search), a significant improvement in model accuracy is evident. The initial model achieved an accuracy of approximately 76%, indicating that the model was not yet optimal in predicting diabetes risk. However, after undergoing hyperparameter optimization with Grid Search, the model's accuracy increased substantially, reaching around 81%. This improvement demonstrates the effectiveness of hyperparameter tuning in enhancing the model's predictive performance.

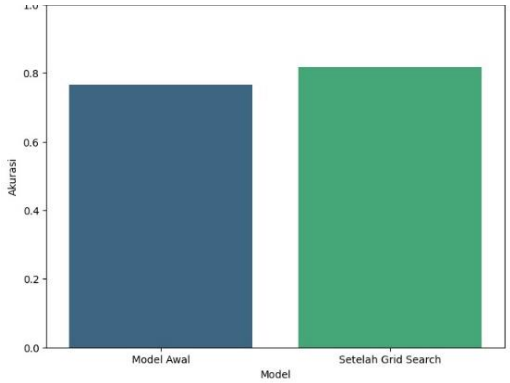

Figure 3. Comparison of Accuracy of the initial model and after grid search

Figure 3 demonstrates that the Grid Search technique is effective in identifying the optimal hyperparameter configuration, thereby enhancing the predictive capability of the Decision Tree model. Such optimization is crucial for producing an accurate and reliable model for the early detection of diabetes risk. The significant improvement in accuracy from the initial model to the post-Grid Search model highlights the potential of machine learning, particularly the Decision Tree algorithm with hyperparameter optimization, in developing an effective tool for diabetes risk detection.

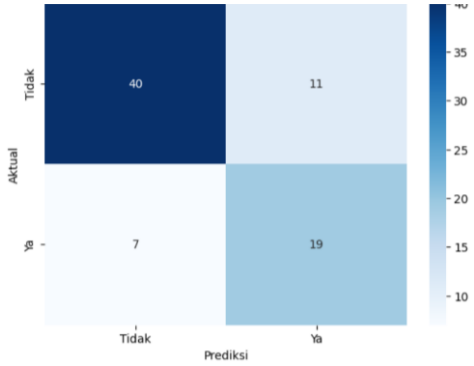

Figure 4. Visualisation of Confusion Matrix

Figure 4 provides an evaluation of the model's performance in classifying diabetes risk based on the dataset used. The analysis shows that 40 samples were correctly classified as "Not at Risk" (True Negative/TN), while 11 samples were incorrectly classified as "At Risk" (False Positive/FP) despite actually being in the "Not at Risk" category. Additionally, 7 samples were misclassified as "Not at Risk" (False Negative/FN) when they actually belonged to the "At Risk" category. On the other hand, 19 samples were correctly classified as "At Risk" (True Positive/TP). These results indicate that the model demonstrates reasonably good capability in classifying diabetes risk, correctly classifying a total of 59 samples out of 77 tested. However, there are still 18 instances of misclassification, highlighting areas for further improvement in the model's predictive accuracy.

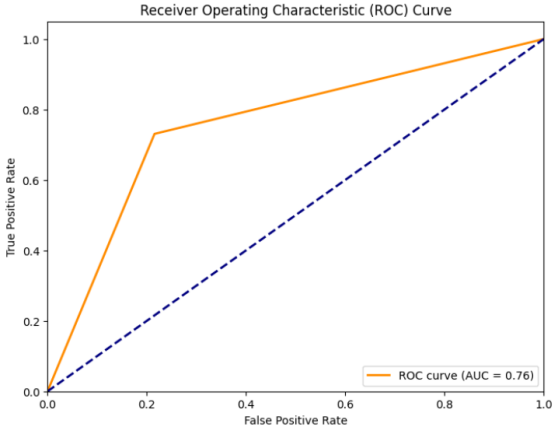

Figure 5. Receiver Operating Characteristic (ROC) Curve

Figure 5 illustrates the model's ability to differentiate between the "At Risk" and "Not at Risk" classes across various thresholds. The orange curve depicts the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR), with an Area Under the Curve (AUC) value of 0.76. This AUC score indicates that the model performs reasonably well, with adequate positive discrimination capability. Based on the evaluation results, the Confusion Matrix reveals that the model can predict diabetes risk with moderate accuracy, although some misclassifications occur, particularly in the False Positive (FP) category. Combining the solid AUC value with the Confusion Matrix results, the Decision Tree algorithm demonstrates significant potential as an early detection tool for diabetes risk. This potential can be further realized through appropriate hyperparameter tuning to enhance model performance.

Discussion

This study underscores the potential of machine learning, particularly the Decision Tree algorithm, in facilitating early detection of diabetes risk. The model demonstrated moderate initial accuracy, which was significantly improved to 81% through hyperparameter optimization using Grid Search, highlighting the critical role of tuning in developing robust predictive models. The identification of glucose levels and BMI as dominant risk factors aligns with established medical evidence, demonstrating the Decision Tree's capability to pinpoint clinically relevant variables.Nevertheless, this research has certain limitations, notably the relatively small dataset size (768 samples), which may restrict the model's generalizability across broader populations. Employing more sophisticated algorithms, such as Random Forest or Neural Networks, alongside comprehensive cross-validation across diverse demographic cohorts, could enhance the model's predictive performance and reliability.

4. Conclusion

This research successfully implemented the Decision Tree algorithm to detect diabetes risk with commendable performance. Based on the Confusion Matrix, the model demonstrated effective classification capabilities, achieving a total of 59 correct predictions out of 77 samples. Evaluation results using the ROC curve revealed an AUC value of 0.76, indicating the model's adequate ability to distinguish between "At Risk" and "Not at Risk" patients. Hyperparameter optimization using Grid Search improved the model's accuracy to 81%, emphasizing the critical role of tuning in enhancing model performance. Blood glucose levels and BMI were identified as the most dominant variables in predicting diabetes risk, consistent with existing medical literature. However, this research has certain limitations, particularly the relatively small dataset size, which may affect the model's generalizability to broader populations. Future studies are recommended to utilize larger datasets, conduct cross-validation on diverse populations, and explore the use of more complex algorithms such as Random Forest or Neural Networks. Such

approaches hold significant potential to support healthcare professionals in detecting diabetes risk more efficiently and accurately.

References:

[1] D. Anggreani, I. A. E. Zaeni, A. N. Handayani, H. Azis and A. R. Manga', "Multivariate Data Model Prediction Analysis Using Backpropagation Neural Network Method," 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT), Surabaya, Indonesia, 2021, pp. 239-243, doi: 10.1109/EIConCIT50028.2021.9431879.

[2] D. Anggreani, Nurmisba, D. Setiawan, and Lukman, "Optimization of K-Means Clustering Method by Using Elbow Method in Predicting Blood Requirement of Pelamonia Hospital Makassar," Internet of Things and Artificial Intelligence Journal, vol. 4, no. 3, pp. 541–550, Aug. 2024, doi: 10.31763/iota.v4i3.755.

[3] International Diabetes Federation (IDF), "Diabetes facts and figures," 2021. Available: https://idf.org/about-diabetes/diabetes-facts-figures/. [Accessed: Nov. 30, 2024].

[4] World Health Organization, "Global Report on Diabetes," 2022. Available: https://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257_eng.pdf. [Accessed: Nov. 30, 2024].

[5] Directorate of Non-Communicable Disease Prevention and Control, Webinar on World Diabetes Day 2024. [Online]. Available: https://lms.kemkes.go.id/courses/799f17f7-c509-4577-92fb-315a4c7b9983. [Accessed: Nov. 30, 2024].

[6] Lancet Diabetes & Endocrinology, Diabetes Prevalence and Management: A Global Update. [Online]. Available: https://www.thelancet.com/issue/S2213-8587(24)X0012-1. [Accessed: Nov.30,2024].

[7] S. S. Reddy, N. Sethi, and R. Rajender, "A comprehensive analysis of machine learning techniques for incessant prediction of diabetes mellitus," International Journal of Grid and Distributed Computing, vol. 13, no. 1, pp. 1-22, 2020. doi: 10.33832/ijgdc.2020.13.1.01.

[8] R. Qasim, F. Moin, M. Ashraf, A. Khan, B. Sarwar, and A. Liaqat, "Risk factors, prevention, and treatment of type 2 diabetes," International Journal of Health Sciences, vol. 6, no. S6, pp. 8822–8832, 2022. doi: 10.53730/ijhs.v6nS6.12362.

[9] A. Hashmi, M. T. Nafis, S. Naaz, and I. Hussain, "A Machine Learning Approach for Diabetes Prediction in Women," International Journal of Food and Nutritional Science, vol. 11, no. 12, pp. 295–408, Feb. 2024.

[10] T. Gautier, L. B. Ziegler, M. S. Gerber, E. Campos-Náñez, and S. D. Patek, "Artificial intelligence and diabetes technology: A review," Metabolism, vol. 124, p. 154872, Nov. 2021. doi: 10.1016/j.metabol.2021.154872.

[11] H. Tanveer, M. A. Adam, M. A. Khan, M. A. Ali, and A. Shakoor, "Analyzing the performance and efficiency of machine learning algorithms, such as deep learning, decision trees, or support vector machines, on various datasets and applications," The Asian Bulletin of Big Data Management / Data Science, vol. 3, no. 2, 2023. doi: 10.62019/abbdm.v3i2.83.

[12] M. Arifuzuzaman, M. R. Hasan, T. J. Toma, S. B. Hassan, and A. K. Paul, "An advanced decision tree-based deep neural network in nonlinear data classification," Technologies, vol. 11, no. 1, p. 24, Feb. 2023. doi: 10.3390/technologies11010024.

[13] N. Saum, S. Sugiura, and M. Piantanakulchai, "Hyperparameter optimization using iterative decision tree (IDT)," *IEEE Access*, vol. 10, pp. 3212387, Oct. 2022, doi: 10.1109/ACCESS.2022.3212387.

[14] M. Ahmad, M. A. Ali, M. R. Hasan, F. D. Mobo, and S. I. Rai, "Geospatial Machine Learning and the Power of Python Programming: Libraries, Tools, Applications, and Plugins," in Ethics, Machine Learning, and Python in Geospatial Analysis, IGI Global, 2024, p. 31. doi: 10.4018/979-8-3693-6381-2.

[15] Z. S. Dunias, B. Van Calster, D. Timmerman, A.-L. Boulesteix, and M. van Smeden, "A comparison of hyperparameter tuning procedures for clinical prediction models: A simulation study," Statistics in Medicine, vol. 43, no. 6, pp. 1119–1134, Jan. 2024, doi: 10.1002/sim.9932.

[16] E. O. Paul, "Hybrid decision tree-based machine learning models for diabetes prediction," *SCIREA* Journal of Information Science and Systems Science, vol. 8, no. 1, Feb. 2024, doi: 10.54647/isss120327.

[17] V.R. Modhugu and S.Ponnusamy,"Comparative analysis of machine learning algorithms for liver disease prediction: SVM, logistic regression, and decision tree," Asian Journal of Research in Computer Science, vol. 17, no. 6, pp. 188–201, 2024, doi: 10.9734/ajrcos/2024/v17i6467.

[18] A. K. Rahimi, O. J. Canfell, W. Chan, B. Sly, J. D. Pole, C. Sullivan, and S. Shrapnel, "Machine learning models for diabetes management in acute care using electronic medical records: A systematic review," International Journal of Medical Informatics, vol. 162, Jun. 2022, Art. no. 104758, doi: 10.1016/j.ijmedinf.2022.104758.

[19] S. Tangirala, "Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 2, pp. 295-408, 2020, doi: 10.14569/IJACSA.2020.0110277.

[20] G. S and S. Brindha, "Hyperparameters Optimization using Gridsearch Cross Validation Method for machine learning models in Predicting Diabetes Mellitus Risk," 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), Chennai, India, 2022, pp. 1-4, doi: 10.1109/IC3IOT53935.2022.9768005.

[21] O. Rahmati, M. Avand, P. Yariyan, J. P. Tiefenbacher, A. Azareh, and D. T. Bui, "Assessment of Gini-, Entropy- and Ratio-Based Classification Trees for Groundwater Potential Modelling and Prediction," Geocarto International, vol. 37, no. 12, pp. 3397–3415, 2021, doi: 10.1080/10106049.2020.1861664.

[22] D. M. Belete and M. D. Huchaiah, "Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results," International Journal of Computers and Applications, 2021, doi: 10.1080/1206212X.2021.1974663.