Turnitin

by Jonson Manurung

Submission date: 08-Apr-2025 09:43AM (UTC+0700)
Submission ID: 2613223324
File name: 389-Perbaikan_1.docx (250.62K)
Word count: 4325
Character count: 24271

Jurnal Mandiri IT
Vol. xx No. xx, Month (202x), pp. xx-xx
ISSN 2301-8984 (Print)

Comparison of decision tree and naive bayes methods in glioma classification based on clinical and molecular factors

Ni Wayan Emmy Rosiana Dewi[1], I Made Suwija Putra[2], Erwinsyah Simanungkalit[3], Franky Gerald Cliford Manoppo[4]
[1,2] Teknologi Informasi, Universitas Udayana, Bali, Indonesia
[3] Manajemen Bisnis, Politeknik Negeri Medan, Medan, Indonesia
[4] Teknik Informatika, Politeknik Negeri Manado, Manado, Indonesia

ARTICLE INFO

Article History:

Received August xx, 202x
Revised August xx, 202x
Accepted August xx, 202x

Keywords:

Clinical Factors
Decision Tree Classifier
Glioma Classification
Molecular Factors
Naive Bayes Classifier

ABSTRACT

This study compares Decision Tree and Naive Bayes machine learning methods in classifying gliomas based on clinical and molecular factors. The dataset used consisted of 839 patient records with features such as Grade, Gender, Age, Race, and gene mutation status. This study evaluated the accuracy of each method, where Decision Tree Classifier achieved 98% accuracy on training data and 76% on test data, while Naive Bayes Classifier obtained 74% and 71% accuracy. Both models showed good predictive ability based on clinical and molecular features. Feature importance analysis revealed that IDH1 gene mutation was a significant factor in glioma classification in both models. This comprehensive approach combining clinical and molecular factors aims to identify the most optimal method to support clinical decision-making in glioma diagnosis. This research contributes to the development of medical decision support systems and provides insight into the effectiveness of Decision Tree and Naive Bayes in processing complex medical data, especially in identifying the impact of molecular markers such as IDH1.

Corresponding Author:

Ni Wayan Emmy Rosiana Dewi,
Teknologi Informasi,
Universitas Udayana,
Jl.Kampus Bukit UNUD Jimbaran ,Badung-Bali, 80361, Indonesia..
Email: emmyrosiana@unud.ac.id

1. INTRODUCTION

Glioma is a type of brain tumor that has a high degree of heterogeneity, both in clinical and molecular aspects (Barthel et al., 2022; Becker et al., 2021). Accurate diagnosis and proper classification are essential for determining effective therapeutic strategies and improving patient prognosis (Mirbabaie et al., 2021; Zubair et al., 2021). In recent years, machine learning has become a widely used tool in the medical field, especially in the analysis and classification of complex data (Jayatilake & Ganegoda, 2021; Saturi, 2023). Decision Tree and Naive Bayes methods are two approaches that are frequently used in medical data classification due to their respective advantages in interpretability and ability to handle probabilistic data (Kopitar et al., 2020; Vellido, 2020). However, the effectiveness of these two methods in glioma classification that considers clinical and molecular factors simultaneously still needs to be further investigated.

Although many studies have explored the application of machine learning in glioma classification, there is no consensus on the most optimal method to use in a clinical context. Decision Tree is known for its ability to generate models that are easy to interpret, but often lacks accuracy in

handling high-dimensional data (Syahputri & Hasibuan, 2024; Wang et al., 2020). Meanwhile, Naive Bayes can handle probabilistic data well, but has limitations in managing correlations between variables (Wickramasinghe & Kalutarage, 2021). Therefore, this study aims to compare the performance of both methods in glioma classification based on clinical and molecular factors.

Research conducted by Widya et al. (2023) shows that Decision Tree C4.5 algorithm for lung cancer classification to aid early detection. The dataset consists of 309 samples with 16 symptom attributes. The test results showed 89% accuracy, 70% precision, and 74.5% recall, indicating that the algorithm is effective for lung cancer prediction. On the other hand, a study by Yang et al. (2020) sdeveloped a neoadjuvant chemotherapy (NAC) response prediction model in breast cancer using the Naive Bayes algorithm. Data from 287 patients were analyzed based on the expression of 17 candidate genes. The results showed that the model had a sensitivity of 84.5% and specificity of 62.0%, with a higher pathologic complete response (pCR) rate in patients classified as sensitive (42.3% vs 7.6%). This model could potentially aid the prediction of NAC effectiveness for more precise treatment. In addition, research by Reddy et al. (2022) which compared Decision Tree and Naïve Bayes algorithms in predicting cardiovascular disease using Heart Failure Dataset. The results show that Naïve Bayes has higher accuracy (86%) than Decision Tree (82%), with better precision and recall. This proves that Naïve Bayes is more effective for cardiovascular disease prediction, aiding data-driven early diagnosis. Previous research has shown that Decision Tree and Naïve Bayes are effective in the classification of various diseases. Therefore, this study will apply both algorithms to glioma classification, to assess their accuracy and effectiveness in supporting early diagnosis.

This study aims to compare the performance of Decision Tree and Naive Bayes methods in classifying gliomas based on clinical and molecular factors. The comparison is based on evaluation metrics such as accuracy, sensitivity, specificity and processing time. Thus, this study is expected to identify the most optimal method in supporting the diagnosis and clinical decision-making process. Most previous studies have only focused on analyzing clinical or molecular factors separately in the classification of a disease. This has led to a lack of a thorough understanding of the interaction between clinical and molecular factors in the context of prediction and diagnosis Meizoso et al. (2021). In addition, there is no comprehensive study that directly compares Decision Tree and Naive Bayes in the same study for glioma classification by considering both factors. Therefore, this study contributes to filling the gap by exploring both methods simultaneously.

The novelty of this study lies in the comprehensive approach of comparing two popular machine learning methods for glioma classification by considering clinical and molecular factors simultaneously. The results of this study are expected to significantly contribute to the development of medical decision support systems in glioma diagnosis, as well as provide new insights into the effectiveness of Decision Tree and Naive Bayes methods in processing complex medical data. The justification for this research is based on the urgent need for classification methods that are not only accurate but also efficient and easy to interpret in clinical practice.

2. RESEARCH METHOD

In this research, the method used is machine learning with Decision Tree Classifier and Naive Bayes Classifier algorithms. The research flow is presented in Figure 1.

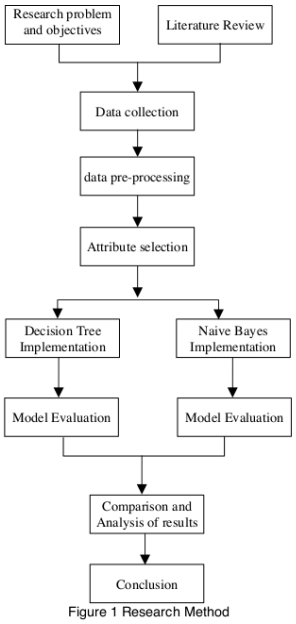

Figure 1 Research Method

Data

The dataset used in this study was obtained from a validated public medical database related to diabetes risk factors from UC Irivine Machine Learning Repository (Tasci et al., 2022). The dataset consists of 839 patient data with relevant features for glioma classification.

Pemilihan Data

The data obtained was then processed using the Python programming language version 3.10.0. and Google Colab. Snippets of data to be further processed are shown in Table 2 and Table 3.

Table 1 Pieces of Research Dataset

Grade	Gender	Age_at_diagnosis	Race	IDH1	TP53	ATRX	PTEN	EGFR	CIC	MUC16	PIK3CA
0	0	51.3	0	1	0	0	0	0	0	0	1
0	0	38.72	0	1	0	0	0	0	1	0	0
0	0	35.17	0	1	1	1	0	0	0	0	0
0	1	32.78	0	1	1	1	0	0	0	1	0
0	0	31.51	0	1	1	1	0	0	0	0	0

Table 2 Continued Pieces of Research Dataset

NF1	PIK3R1	FUBP1	RB1	NOTCH1	BCOR	CSMD3	SMARCA4	GRIN2A	IDH2	FAT4	PDGFRA
0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	1	0

Data Pre-Processing

In the pre-processing stage, there is no need to make many changes to the data that will be used in the machine learning model because the data is clean, neat, and ready to be processed in the next stage. There are few changes made, namely to the Age_at_diagnosis variable which previously had a float data type to int with age rounding. Then the Age_at_diagnosis variable is also normalized using MinMaxScaler which produces a value between 0 and 1. After that, the next stage the Grade variable is selected as the target, and the other variables become Features.

Table 3 Variable Description

Variable Name	Role	Description	Interpretation
Grade	Target	Glioma grade class information	0 = "LGG" 1 = "GBM"
Gender	Feature	Gender	0 = "male" 1 = "female"
Age_at_diagnosis	Feature	Age at diagnosis	Years
Race	Feature	Excessive thirst	0 = "white" 1 = "black or african American" 2 = "asian" 3 = "american indian or alaska native"
IDH1	Feature	isocitrate dehydrogenase	0 = NOT_MUTATED 1 = MUTATED
TP53	Feature	tumor protein p53	0 = NOT_MUTATED 1 = MUTATED
ATRX	Feature	ATRX chromatin remodeler	0 = NOT_MUTATED 1 = MUTATED
PTEN	Feature	phosphatase and tensin homolog	0 = NOT_MUTATED 1 = MUTATED
EGFR	Feature	epidermal growth factor receptor	0 = NOT_MUTATED 1 = MUTATED
CIC	Feature	capicua transcriptional repressor	0 = NOT_MUTATED 1 = MUTATED
MUC16	Feature	mucin 16, cell surface associated	0 = NOT_MUTATED 1 = MUTATED
PIK3CA	Feature	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha	0 = NOT_MUTATED 1 = MUTATED
NF1	Feature	neurofibromin 1	0 = NOT_MUTATED 1 = MUTATED
PIK3R1	Feature	phosphoinositide-3-kinase regulatory subunit 1	0 = NOT_MUTATED 1 = MUTATED
FUBP1	Feature	far upstream element binding protein 1	0 = NOT_MUTATED 1 = MUTATED
RB1	Feature	RB transcriptional corepressor 1	0 = NOT_MUTATED 1 = MUTATED
NOTCH1	Feature	notch receptor 1	0 = NOT_MUTATED 1 = MUTATED
BCOR	Feature	BCL6 corepressor	0 = NOT_MUTATED 1 = MUTATED
CSMD3	Feature	CUB and Sushi multiple domains 3	0 = NOT_MUTATED 1 = MUTATED
SMARCA4	Feature	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4	0 = NOT_MUTATED 1 = MUTATED
GRIN2A	Feature	glutamate ionotropic receptor NMDA type subunit 2A	0 = NOT_MUTATED 1 = MUTATED
IDH2	Feature	isocitrate dehydrogenase	0 = NOT_MUTATED 1 = MUTATED
FAT4	Feature	FAT atypical cadherin 4	0 = NOT_MUTATED 1 = MUTATED
PDGFRA	Feature	platelet-derived growth factor receptor alpha	0 = NOT_MUTATED 1 = MUTATED

Table 3 menampilkan informasi mengenai variabel yang digunakan pada penlitian ini, terdiri dari variabel target dan fitur-fitur pendukung. Variabel target adalah Grade, yang mengklasifikasikan tingkat glioma menjadi LGG (Low-Grade Glioma) dan GBM (Glioblastoma Multiforme). Fitur-fitur lainnya mencakup data klinis seperti Gender, Age_at_diagnosis, dan Race, serta berbagai fitur molekuler yang menunjukkan status mutasi gen, seperti IDH1, TP53, EGFR, dan lainnya, dengan nilai 0 untuk NOT_MUTATED dan 1 untuk MUTATED.

Split Data and Training Data

Before conducting data training, it is necessary to divide the training data and test data with a ratio of 80% training data, namely 671 records and 20% test data, namely 168 records. Next, perform the training process on the training data with the Naïve Bayes Classifier algorithm to get a classification pattern that will be applied to the test data.

Decision Tree Classifier

Decision Tree is a tree-shaped model used for classification and decision making (Nikita & Nikitas, 2020; Suresh et al., 2020). Nodes represent the test data, branches show the results, and leaf nodes are the output classes (Buntine, 2020). The root node is the main node in the tree. This algorithm calculates the entropy value of each attribute to measure uncertainty, and then compares the Gain value to select the best attribute to split the data (Priyanka & Kumar, 2020; Thomas et al., 2020).
Entropy is calculated by :

$$Entropy(S) = \sum_{i=1}^{n} -pi * \log_2 pi \quad (1)$$

The higher the entropy, the more disorganized the data, meaning that the data is evenly spread across the various classes. Conversely, if the entropy is low, the data is more concentrated in one class (Abrori & Fatah, 2025).
While Gain is formulated as:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (2)$$

Gain indicates the effectiveness of attribute A in dividing the data. The larger the Gain value, the better the attribute is at reducing uncertainty. The attribute with the highest Gain is selected as the node in the Decision Tree (Abrori & Fatah, 2025).

Naive Bayes Classifier

Naive Bayes classification uses the principle of maximum likelihood estimation to classify samples into the most likely category (Chen et al., 2021; Ren et al., 2022). The following are the steps in training the model and making predictions using naive bayes. For each class $P(C_i)$, the prior probability is calculated based on the proportion of classes in the training data (Jogo et al., 2023):

$$P(C_i) = \frac{Number\ of\ data\ in\ the\ C_i\ class}{Total\ number\ of\ data} \quad (3)$$

The likelihood for data X=(x1,x2,...,xn)X = (x_1, x_2, ..., x_n)X=(x1,x2,...,xn) given class C_i is calculated by assuming that each feature x_k is independent. Then:

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) \quad (4)$$

For Gaussian Naïve Bayes, $P(x_k \mid C_i)$ is calculated using a normal (Gaussian) distribution for each feature x_k in class C_i

$$P(x_k \mid C_i) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \exp\left(-\frac{(x_k - \mu_{ik})^2}{2\sigma_{ik}^2}\right) \quad (5)$$

Description:
μ_{ik} is the mean of feature x_k in class C_i,
σ_{ik}^2 is the variance of feature x_k in class C_i.

The posterior probability for class C_i , given the data X, is calculated using Bayes' Theorem:

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \quad (6)$$

Description:
$P(X \mid C_i)$ is the likelihood (as calculated above),
$P(C_i)$ is the prior probability for class C_i,
$P(X)$ is the evidence (probability of X), which is the same for all classes and is often ignored when comparing between classes.

The model will predict the class that has the highest posterior probability. In other words, the model chooses the class C_i that maximizes the value of $P(C_i \mid X)$:

$$\hat{C}arg\frac{max}{C_i}[P(X \mid C_i)P(C_i)] \quad (7)$$

Classification

After training the data with the Decision Tree Classifier and Naive Bayes Classifier algorithms and obtaining a classification model, the next step is to perform classification using both algorithms on the test data.

Evaluation

After the classification process is complete, the next step is to evaluate the classification results. At this stage, the accuracy, precision, recall, and f1-score values are calculated by utilizing the confusion matrix. Confusion matrix is a matrix that shows the actual and predicted classification results with a size of LxL, where L is the number of labels or classifications, and L is the number of classes (Rayhan & Setyohadi, 2021). In this research, the confusion matrix used is 2x2 because it has 2 (two) labels/classes (Tharwat, 2018). The terms of the 2x2 confusion matrix are shown in Figure 2 below.

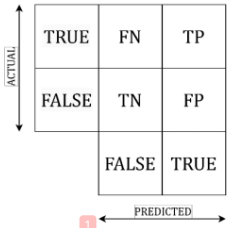

Figure 2 Confusion Matrix 2x2

The accuracy, precision, recall, and f1-score values can be calculated with the following equations (Maulana et al., 2024) (8) - (11).

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (8)$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (9)$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (10)$$

$$\text{F1-Score} = \frac{2 * Precision * Recall}{\text{Precision} + \text{Recall}} \quad (11)$$

To identify the most influential features in classifying glioma disease, we use Feature Importance and Feature Impact graphs. The Feature Importance graph, generated by the Decision Tree algorithm, measures how often a feature is used in decision-making (Saarela & Jauhiainen, 2021). Features frequently used for data splitting play a greater role in prediction. The contribution of each feature is evaluated based on how much the tree's uncertainty decreases with each split. Gini impurity is calculated by the formula (Arya Darmawan et al., 2023):

$$Gini(t) = 1 - \sum p_i^2 \quad (12)$$

Where p_i is the proportion of samples for each class i. After the split is performed, the impurity reduction is calculated by comparing the impurity before and after the split, and then this reduction is summed up for all nodes where the feature is used. The importance of a feature is then obtained from the total impurity reduction multiplied by the proportion of samples affected at each node (Scornet, 2023).

The Feature Impact graph, generated by the Naive Bayes algorithm, measures each feature's impact on classification (Wickramasinghe & Kalutarage, 2021). Features with higher impact values play a more significant role in influencing the model's predictions. This impact is calculated using the average parameter (θ), which represents the distribution of feature values across different classes.

Assuming feature independence, the impact of a feature is determined by the absolute difference in the mean values of that feature between classes:

$$Impact(f) = \left|\mu_{f|class=1|} - \mu_{f|class=0|}\right| \quad (13)$$

The larger this difference, the greater the feature's impact on distinguishing between classes in the Naïve Bayes model. This approach highlights how much each feature's distribution varies between the two classes, making it easier to identify the most influential features for classification

3. RESULT AND DISCUSSIONS

The data training process carried out on the train data of 416 records using the Decision Tree Classifier and Naive Bayes Classifier algorithms and the classification applied to the test data of 104 records resulted in accuracy as presented in Figure 3 below.

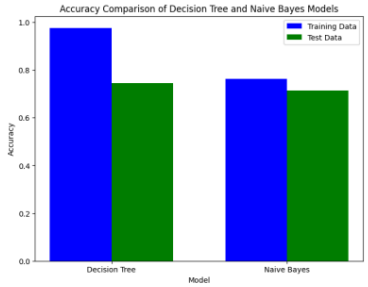

Figure 3 Accuracy Comparison of Decision Tree and Naive Bayes Models

Figure 3 shows that the classification of training data and test data with the Decision Tree Classifier algorithm shows quite good results with the achievement of accuracy on training data of 98% and test data of 76% and with the Naive Bayes Classifier algorithm shows quite good results with the

achievement of accuracy on training data of 74% and test data of 71%. The following is a confusion matrix to evaluate the classification results that have been carried out with the two algorithms, as shown in Figure 4 and Figure 5.

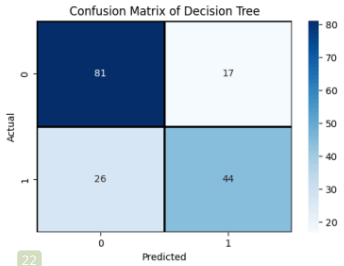

Figure 4 Confusion Matrix of Decision Tree Classifier

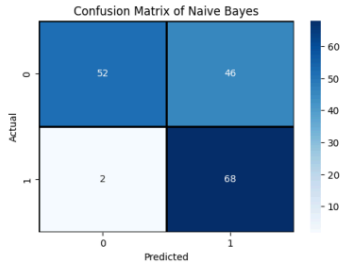

Figure 5 Confusion Matrix of Naive Bayes Classifier

Based on the 2x2 confusion matrix in Figures 4 and 5, the prediction to determine the classification is done on 168 records. For the Decision Tree classifier (Figure 4), the prediction results for class 0 are 81 true negatives and 17 false negative, while class 1 has 44 true positives and 26 false positives. For the Naïve Bayes classifier (Figure 5), the prediction results for class 0 are 52 true negative and 46 false negative, while class 1 has 68 true positives and 2 false positives.

Performance of Decision Tree:
Accuracy: 0.7440476190476191
Classification Report:

	precision	recall	f1-score	support
0	0.76	0.83	0.79	98
1	0.72	0.63	0.67	70
accuracy			0.74	168
macro avg	0.74	0.73	0.73	168
weighted avg	0.74	0.74	0.74	168

Figure 6 Performance of Decision Tree Classifier

Performance of Naive Bayes:
Accuracy: 0.7142857142857143
Classification Report:

	precision	recall	f1-score	support
0	0.96	0.53	0.68	98
1	0.60	0.97	0.74	70
accuracy			0.71	168
macro avg	0.78	0.75	0.71	168
weighted avg	0.81	0.71	0.71	168

Figure 7 Performance of Naive Bayes Classifier

To determine the performance of the algorithm in classifying glioma disease, an evaluation was conducted by calculating accuracy, precision, recall, and f1-score. For the Decision Tree (Figure 6), the resulting accuracy was 74%, with average values for precision, recall, and f1-score of 74%, 74%, and 74%, respectively. For the Naive Bayes classifier (Figure 7), the resulting accuracy is 71%, with average values for precision, recall, and f1-score of 81%, 71%, and 71%, respectively. These results show that the Decision Tree classifier performs slightly better in classifying glioma diseases with higher accuracy and consistency compared to the Naive Bayes classifier.

Furthermore, to find out which features have the most impact on the classification of glioma disease, we can look at two graphs, namely Feature Importance and Feature Impact. Both graphs in

Figure 8 and Figure 9 illustrate how much each feature or variable contributes in classifying glioma disease.

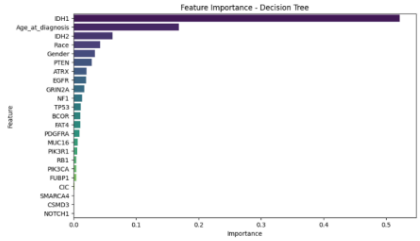

Figure 8 Feature Importance Decision Tree Model

In the graph in Figure 8, the IDH1 feature has the highest importance, followed by features such as Age_at_diagnosis, IDH2, Race, and Gender. This shows that the IDH1 gene mutation is the most influential factor in the classification of glioma diseases according to the Decision Tree model.

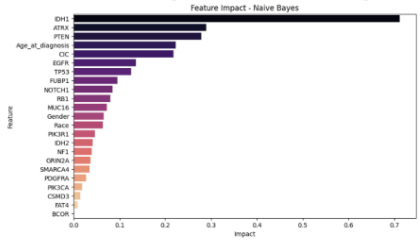

Figure 9 Feature Impact Naive Bayes Model

In Figure 9, IDH1 also takes the top spot, followed by features such as ATRX, PTEN, and Age_at_diagnosis. This confirms that the presence of a gene mutation in IDH1 is very influential in increasing the likelihood of someone being diagnosed with glioma.

By comparing the two models, the IDH1 feature is consistently the most influential factor in glioma classification. In both the Decision Tree and Naive Bayes models, IDH1 always took the top spot in terms of contribution to the prediction process. This confirms that the IDH1 gene mutation plays a significant role in distinguishing patients diagnosed with low-grade glioma (LGG) from those diagnosed with high-grade glioma (GBM). This feature suggests that IDH1 can be used as a key indicator in detecting and predicting the likelihood of glioma. Therefore, further analysis of IDH1 gene mutations is essential to deepen our understanding of the genetic factors that impact this disease.

4. CONCLUSIONS

This study demonstrate that the Decision Tree Classifier outperforms the Naive Bayes Classifier in classifying glioma disease, achieving higher accuracy (74%) and balanced precision, recall, and F1-score compared to the Naive Bayes model's 71% accuracy. This research contributes by highlighting the critical role of the IDH1 gene mutation as the most influential feature in distinguishing between low-grade glioma (LGG) and high-grade glioma (GBM), suggesting its potential as a key biomarker for glioma diagnosis and prediction. The findings imply that employing the Decision Tree Classifier can improve classification reliability in medical diagnostics, supporting more accurate and consistent patient assessments. However, the study is limited by the dataset size and the specific features considered, which may not capture the full complexity of glioma classification. Future research should

focus on expanding the dataset, incorporating additional clinical and molecular features, and exploring advanced machine learning models to enhance classification performance and deepen our understanding of glioma pathogenesis

REFERENCES

Abrori, S., & Fatah, Z. (2025). *Implementasi Metode Decission Tree Dalam Mengklasifikasi Depresi Menggunakan Rapidminer*. *5*(2), 123–132.

Arya Darmawan, M. B., Dewanta, F., & Astuti, S. (2023). Analisis Perbandingan Algoritma Decision Tree, Random Forest, dan Naïve Bayes untuk Prediksi Banjir di Desa Dayeuhkolot. *TELKA - Telekomunikasi Elektronika Komputasi Dan Kontrol*, *9*(1), 52–61. https://doi.org/10.15575/telka.v9n1.52-61

Barthel, L., Hadamitzky, M., Dammann, P., Schedlowski, M., Sure, U., Thakur, B. K., & Hetze, S. (2022). Glioma: molecular signature and crossroads with tumor microenvironment. *Cancer and Metastasis Reviews*, 1–23.

Becker, A. P., Sells, B. E., Haque, S. J., & Chakravarti, A. (2021). Tumor Heterogeneity in Glioblastomas: From Light Microscopy to Molecular Pathology. In *Cancers* (Vol. 13, Issue 4). https://doi.org/10.3390/cancers13040761

Buntine, W. (2020). Learning classification trees. In *Artificial Intelligence frontiers in statistics* (pp. 182–201). Chapman and Hall/CRC.

Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved naive Bayes classification algorithm for traffic risk management. *EURASIP Journal on Advances in Signal Processing*, *2021*(1), 30.

Jayatilake, S. M. D. A. C., & Ganegoda, G. U. (2021). Involvement of Machine Learning Tools in Healthcare Decision Making. *Journal of Healthcare Engineering*, *2021*(1), 6679512. https://doi.org/https://doi.org/10.1155/2021/6679512

Jogo, M. M. S., Biddinika, M. K., & Fadlil, A. (2023). Klasifikasi Penyakit Diabetes dengan Algoritma Decision Tree dan Naïve Bayes. *RESISTOR (Elektronika Kendali Telekomunikasi Tenaga Listrik Komputer) Vol.*, *6*(2), 113–118.

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, *10*(1), 1–12. https://doi.org/10.1038/s41598-020-68771-z

Maulana, R., Narasati, R., Herdiana, R., Hamonangan, R., & Anwar, S. (2024). Komparasi Algoritma Decision Tree Dan Naive Bayes Dalam Klasifikasi Penyakit Diabetes. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(6), 3865–3870. https://doi.org/10.36040/jati.v7i6.8265

Meizoso, J. P., Moore, H. B., & Moore, E. E. (2021). Fibrinolysis Shutdown in COVID-19: Clinical Manifestations, Molecular Mechanisms, and Therapeutic Implications. *Journal of the American College of Surgeons*, *232*(6), 995–1003. https://doi.org/10.1016/j.jamcollsurg.2021.02.019

Mirbabaie, M., Stieglitz, S., & Frick, N. R. J. (2021). Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. In *Health and Technology* (Vol. 11, Issue 4). Springer Berlin Heidelberg. https://doi.org/10.1007/s12553-021-00555-5

Nikita, E., & Nikitas, P. (2020). Data mining and decision trees. In *Statistics and probability in forensic anthropology* (pp. 87–105). Elsevier.

Priyanka, & Kumar, D. (2020). Decision tree classifier: a detailed survey. *International Journal of Information and Decision Sciences*, *12*(3), 246–269.

Rayhan, Y., & Setyohadi, D. B. (2021). Classification of grape leaf disease using convolutional neural network (CNN) with pre-trained model VGG16. *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, 1–5.

Reddy, V. S. K., Meghana, P., Reddy, N. V. S., & Rao, B. A. (2022). Prediction on Cardiovascular disease using Decision tree and Naïve Bayes classifiers. *Journal of Physics: Conference Series*, *2161*(1). https://doi.org/10.1088/1742-6596/2161/1/012015

Ren, Q., Zhang, H., Zhang, D., Zhao, X., Yan, L., Rui, J., Zeng, F., & Zhu, X. (2022). A framework of active learning and semi-supervised learning for lithology identification based on improved naive Bayes. *Expert Systems with Applications*, *202*, 117278.

Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations

for classification models. *SN Applied Sciences*, *3*(2), 272.
Saturi, S. (2023). Review on Machine Learning Techniques for Medical Data Classification and Disease Diagnosis. *Regenerative Engineering and Translational Medicine*, *9*(2), 141–164. https://doi.org/10.1007/s40883-022-00273-y
Scornet, E. (2023). Trees, forests, and impurity-based variable importance in regression. *Annales de l'Institut Henri Poincare (B) Probabilites et Statistiques*, *59*(1), 21–52.
Suresh, A., Udendhran, R., & Balamurgan, M. (2020). Hybridized neural network and decision tree based classifier for prognostic decision making in breast cancers. *Soft Computing*, *24*(11), 7947–7953.
Syahputri, C. N., & Hasibuan, M. S. (2024). OPTIMASI KLASIFIKASI DECISION TREE DENGAN TEKNIK PRUNING UNTUK MENGURANGI OVERFITTING. *JSiI (Jurnal Sistem Informasi)*, *11*(2), 87–96. https://doi.org/10.30656/jsii.v11i2.9161
Tasci, E., Camphausen, K., Krauze, A., & Zhuge, Y. (2022). *Glioma Grading Clinical and Mutation Features [Dataset]*. https://doi.org/10.24432/C5R62J
Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003
Thomas, T., P. Vijayaraghavan, A., Emmanuel, S., Thomas, T., P. Vijayaraghavan, A., & Emmanuel, S. (2020). Applications of decision trees. *Machine Learning Approaches in Cyber Security Analytics*, 157–184.
Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, *32*(24), 18069–18083.
Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning——a case study of bank loan data. *Procedia Computer Science*, *174*, 141–149.
Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277–2293. https://doi.org/10.1007/s00500-020-05297-6
Widya, H., Surya, N., Atina, V., & Maulindar, J. (2022). *Penerapan Algoritme Decision Tree Pada Klasifikasi Penyakit Kanker Paru-Paru*.
Yang, L., Fu, B., Li, Y., Liu, Y., Huang, W., Feng, S., Xiao, L., Sun, L., Deng, L., Zheng, X., Ye, F., & Bu, H. (2020). Prediction model of the response to neoadjuvant chemotherapy in breast cancers by a Naive Bayes algorithm. *Computer Methods and Programs in Biomedicine*, *192*. https://doi.org/10.1016/j.cmpb.2020.105458
Zubair, M., Wang, S., & Ali, N. (2021). Advanced Approaches to Breast Cancer Classification and Diagnosis. *Frontiers in Pharmacology*, *11*(February), 1–24. https://doi.org/10.3389/fphar.2020.632079

Turnitin

ORIGINALITY REPORT

31%	23%	15%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	jurnal.iaii.or.id Internet Source	7%
2	assets.researchsquare.com Internet Source	2%
3	ejournal.isha.or.id Internet Source	2%
4	Submitted to City University Student Paper	2%
5	Submitted to The Scientific & Technological Research Council of Turkey (TUBITAK) Student Paper	1%
6	www.mdpi.com Internet Source	1%
7	kupdf.net Internet Source	1%
8	mafiadoc.com Internet Source	1%
9	web.realinfo.tv Internet Source	1%
10	Submitted to École Polytechnique Fédérale de Lausanne Student Paper	1%
11	Teuku Rizky Noviandy, Ghalieb Mutig Idroes, Irsan Hardi. "Integrating explainable artificial intelligence and light gradient boosting	1%

machine for glioma grading", Informatics and Health, 2025
Publication

12	www.frontiersin.org Internet Source	1%
13	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Computer Science Engineering", CRC Press, 2024 Publication	1%
14	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P. Prasad. "Algorithms in Advanced Artificial Intelligence - Proceedings of International Conference on Algorithms in Advanced Artificial Intelligence (ICAAAI-2024)", CRC Press, 2025 Publication	1%
15	Hasbanur Hafidz, M. Fakhriza. "Comparison of Naive Bayes Algorithms and Decision Tree for Classifying Hero Fighter Items in the Mobile Legends", Journal of Applied Science, Engineering, Technology, and Education, 2024 Publication	1%
16	ejournal.nusamandiri.ac.id Internet Source	<1%
17	Iskander Ben Rjiba, Georgina Tóth-Nagy, Ágnes Rostási, Petra Gyurácz-Németh, Viktor Sebestyén. "How should climate actions be planned? Model lessons from published action plans", Journal of Environmental Management, 2024 Publication	<1%
18	Submitted to Taylor's Education Group Student Paper	<1%

19	cuebic.co.jp Internet Source	<1%
20	Submitted to University of Central Florida Student Paper	<1%
21	enrichment.iocspublisher.org Internet Source	<1%
22	www.coursehero.com Internet Source	<1%
23	Padmaja Jonnalagedda, Brent Weinberg, Taejin L. Min, Shiv Bhanu, Bir Bhanu. "Computational modeling of tumor invasion from limited and diverse data in Glioblastoma", Computerized Medical Imaging and Graphics, 2024 Publication	<1%
24	Submitted to Udayana University Student Paper	<1%
25	Submitted to Georgia State University Student Paper	<1%
26	H L Gururaj, Francesco Flammini, V Ravi Kumar, N S Prema. "Recent Trends in Healthcare Innovation", CRC Press, 2025 Publication	<1%
27	Submitted to University of Southampton Student Paper	<1%
28	ebin.pub Internet Source	<1%
29	digilib.unila.ac.id Internet Source	<1%
30	arno.uvt.nl Internet Source	<1%
	downloads.hindawi.com	

31	Internet Source	<1%
32	repository.ubhara.id Internet Source	<1%
33	www.ijraset.com Internet Source	<1%
34	Mfundo Mandla Masuku, Nomakhosi Nomathemba Sibisi. "Combating School-Based Violence Using African Indigenous Knowledge Systems - Implications for Educational Safety", Routledge, 2025 Publication	<1%
35	ieomsociety.org Internet Source	<1%
36	math.mit.edu Internet Source	<1%
37	www.seruvenyayinevi.com Internet Source	<1%
38	Charu C. Aggarwal. "Data Classification - Algorithms and Applications", Chapman and Hall/CRC, 2019 Publication	<1%
39	Ruzhun Zhao, Yuchang Zhu, Yuanhong Li. "CLA: A self-supervised contrastive learning method for leaf disease identification with domain adaptation", Computers and Electronics in Agriculture, 2023 Publication	<1%
40	academic.oup.com Internet Source	<1%
41	download.bibis.ir Internet Source	<1%
	ojs.unitama.ac.id	

42	Internet Source	<1%
43	Keerthi Varadhi, Chinta Someswara Rao, GNVG Sirisha, Butchi Raju katari. "Recognizing human activities using light-weight and effective machine learning methodologies", F1000Research, 2024 Publication	<1%
44	Nur Adha Pasaribu, Sriani. "The Shopee Application User Reviews Sentiment Analysis Employing Naïve Bayes Algorithm", International Journal Software Engineering and Computer Science (IJSECS), 2023 Publication	<1%
45	core.ac.uk Internet Source	<1%

Exclude quotes	On	Exclude matches	< 1 words
Exclude bibliography	On		

Turnitin

GRADEMARK REPORT

FINAL GRADE

/100

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11