

Implementing brain tumor detection using various machine learning techniques

Rani Puspita, Cindy Rahayu

Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Jul 23, 2024

Revised Jan 4, 2025

Accepted Jan 16, 2025

Keywords:

Brain tumor

Detection

Evaluation

Machine learning

Random forest

ABSTRACT

The brain is a very complex organ of the human body. One of the brain diseases is a tumor. Brain tumors are caused by uncontrolled cell growth. Early recognition, classification and analysis of brain tumors is very important to find out whether there is a tumor in a person's brain so it is important for us to do this in order to treat the tumor thoroughly. Machine learning (ML) techniques that have the highest accuracy in detecting the health sector are extreme gradient boosting (XGBoost), logistic regression, random forest, k-nearest neighbor (KNN), naive Bayes, and support vector machine (SVM). In this research, data collection and exploration were carried out, data training using six methods, and evaluation using a confusion matrix. After conducting the experiment, it was obtained that random forest had the highest accuracy, namely 98.41%. Where XGBoost obtained an accuracy of 98.14%, logistic regression obtained an accuracy of 97.34%, KNN and naive Bayes of 97.34%, and SVM of 97.88%.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Rani Puspita

Computer Science Department, School of Computer Science, Bina Nusantara University

Jakarta, Indonesia 11480

Email: rani.puspita@binus.ac.id

1. INTRODUCTION

The brain is a very complex human body organ. The brain has billions of cells. One of the diseases of the brain is a tumor. Brain tumors are caused by uncontrolled cell growth [1]. This is the same as research that states that there are all kinds of abnormalities in the brain that endanger human health. One of them is a brain tumor [2]. This can be supported by research conducted by Siar and Teshnehlal [3] that when most cells are old and damaged, they will be destroyed and replaced by newer cells. With that, problems arise and lead to tumor growth in the brain.

According to the World Health Organization (WHO), around 700,000 people are affected by brain tumors, and since 2019, around 86,000 patients have been diagnosed with brain tumors [4]. Of the 700,000, 69.1% of people were diagnosed with benign tumors and 30.1% were diagnosed with malignant tumors [5]. Tumors are malignant cells produced by the uncontrolled development of cancer cells. There are two types of tumors, namely malignant and benign. Malignant brain tumors originate in the brain, grow quickly, and aggressively attack the surrounding areas and affect the central nervous system. Then benign brain tumors is a mass of cells that grow relatively slowly in the brain [6].

This is in line with other research which states that tumors are a serious cancer and can attack adults and children. Analysis and classification of brain tumors are very important to find out whether there really is a tumor in a person's brain so it is important for us to do this in order to treat the tumor adequately [7], and machine learning (ML) techniques are considered a good basis for carrying out classification and mining tasks. In ML, the features selected as input for the model have a good impact on the results of the model used [8]. The

diagnosis of a brain tumor needs to be made clear by classifying it so that a positive brain tumor diagnosis can be treated immediately. In this case, the radiologist must diagnose the brain tumor as early as possible, and then check again to see if the results are correct [9].

In this way, the problem with this research is due to the lack of early detection of brain tumors, so many patients only find out that they actually suffer from a brain tumor. So, it is necessary to carry out research using ML techniques. ML is a subset of artificial intelligence (AI). AI is a term that refers to the application of algorithms that can analyze large datasets to classify, recommend, predict, and others. Under the AI is ML. ML is the process of learning models using real data to categorize, predict, and detect based on training observations made by humans. The results will apply to future data [10]. There are many methods in ML, including support vector machine (SVM), k-nearest neighbor (KNN), random forest, decision trees (DTs), and multi-layer perceptron (MLP) [11]. Apart from that, in ML there are several other methods that can be used, namely logistic regression and naive Bayes [12]. Then according to other research, XGBoost is also included in the ML method [13]. According to Ankit and Kole [14] SVM, logistic regression, KNN, naive Bayes, random forest classifier, and XGBoost classifier are methods that can be used to detect brain tumors.

Naive Bayes is an efficient and effective algorithm in ML and data mining. Naive Bayes was proposed by Thomas Bayes from England. Naive Bayes is proven to be fast and accurate. But sometimes naive Bayes requires training information in the form of mean and variance to carry out classification. Then KNN. KNN is an algorithm for classification [15]. Then there is SVM which is a technique in ML. SVM is a supervised learning model with associated learning algorithms. SVM aims to analyze data used for classification and regression. Additionally, SVM can also perform non-linear classification using a kernel trick that implicitly inserts it into a high-dimensional feature space [16]. Then another ML technique is random forest. Random forest is the most commonly used ML algorithm due to its simplicity and variety. They are usually used for classification and regression. Random forest is an ML approach that collects results from a single decision tree so that it can increase predictive efficiency through voting [17]. Apart from naive Bayes, KNN, SVM, and random forest, we also use logistic regression and XGBoost methods. Logistic regression is used to model binary outcomes. Logistic regression involves estimation tasks. The response only has two possible outcomes, namely that it can be represented by a binary variable with the values 0 and 1. For XGBoost, boosting involves collecting all weak classifications to produce a strong classification. Usually what is done in the first step is to organize the data. All forms of data will be converted to numeric because XGBoost only works with vector numeric. Then clean the data and run through the features [18]. Based on the background that has been explained, this research will focus on applying ML techniques using XGBoost, logistic regression, random forest, KNN, naive Bayes, and SVM to detect brain tumors using features as criteria. Each piece of data in this research will be divided into training and testing data. After that, each ML method will be used, trained and evaluated.

2. LITERATURE REVIEW

In this section, researchers use literature studies in international journals or conferences related to the researched topic. In this chapter, the researcher aims to study and also look for various references regarding brain tumor detection, ML techniques, and others to see the problems that exist today and also find solutions to these problems. The results of the literature study regarding brain tumor detection can be seen in Table 1 [19]–[25] (see in Appendix) and the results of the literature study regarding ML techniques can be seen in Table 2 [26]–[31] (see in Appendix).

3. METHOD

In this chapter, researchers will explain what ML techniques are used in research. There are 6, namely XGBoost, logistic regression, random forest, KNN, naive Bayes, and SVM. The explanation will start in sections 3.1 to 3.6.

3.1. XGBoost

XGBoost was first proposed by Chen and Guestrin in 2016 [32]. XGBoost is a tree-based method that is popular in ML. Mostly XGBoost is used as a base model to solve ML tasks. XGBoost is a mixture of bagging and boosting algorithms that can improve accuracy sequentially [33]. XGboost can identify effective and efficient ways to combine local and global contextual patterns. XGBoost is strong against overfitting which makes it very easy for the model to make choices. By using XGBoost, researchers can identify strong patterns and be able to differentiate between patients and healthy people [34].

This is in line with other research that the XGBoost model performs classification very well and has the same proportions for both classes. According to other research, XGBoost is a better model than single

ML models such as logistic regression, SVM, decision trees, and others. Cross-validation between models is powerful and integrated in XGBoost. According to this research, XGBoost is a model that is built sequentially to reduce accuracy in iterations. Gradient boosting is the original model of XGBoost. Where the way it works is to combine the weak so that they become stronger.

3.2. Logistic regression

Logistic regression models the relationship between categorical variables and covariates. This method combines linear independent variables with log-odds of a probability in a logistic model. Logistic regression can be considered ML [35]. Logistic regression is used to analyze data in statistics that determine the relationship between one or more independent variables and the dependent variable. This method is used for prediction. The dataset contains the target class. In this case, let's say X is a binary response variable. $X_i = 1$ if present, and $X_i = 0$ if absent, then the data $[X_1, X_2, \dots, X_n]$ are independent. Let π_i be the probability of success. Then the logistic function for π_i is:

$$\text{Logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} \quad (1)$$

$$\text{where, } \pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})} = \frac{\exp(x' \beta)}{1 + \exp(x' \beta)} = \Lambda(x' \beta) \quad (2)$$

3.3. Random forest

Random forest is a popular technique in ML. Random forest has successfully demonstrated the success of its method in various fields, including the health sector. Random forests have the ability to handle high-dimensional data, capture non-linear relationships, and so on, so random forests are the right choice of method for predictions in the health sector [36]. The random forest method was suggested by Breiman in 2001. Random forest produces two types of information significance, namely measuring the significance including the response variable, and measuring the internal data model [37]. There are several phases of the random forest algorithm, namely:

- Determine a trained model: The point is to determine the training model. This technique is used in collecting classification models from original data.
- Built random forest structure: The point is to collect data from all bootstraps and produce n trees.
- Voting phase: What this means is the voting or pooling phase. Where this phase helps us determine the right and wrong features for each tree.

This is in line with the opinion of other researchers that random forest is a tree-based ML technique. Random forest estimates correctly and the relationships are stable between variables [38]. Random forest is able to handle numeric or categorical variables in prediction situations.

3.4. KNN

The basic ML model for classification and regression is KNN. The goal is to determine the distance between new unlabeled data points and the training data set stored in feature space. The k value in KNN is a hyperparameter used to sort the k -nearest data points. In the optimized KNN, neighbors only have positive relationships with requestors. Nearby neighbors can influence better than distant neighbors. The KNN algorithm uses a distance function to determine the weight of the influence of close neighbors. Euclidean distance, Pearson correlation, Manhattan distance, and spearman correlation is a distance function used for continuous variables. There are pros and cons to using KNN. The advantage is that it uses a nonparametric approach, is easy to understand and easy to apply. But the con is that KNN is not explicitly trained quickly. Figure 1. is an illustration of how KNN can do its job using Euclidean distance. An example of a k value is six data points. The new sample data points fall into the blue class in this illustration [39].

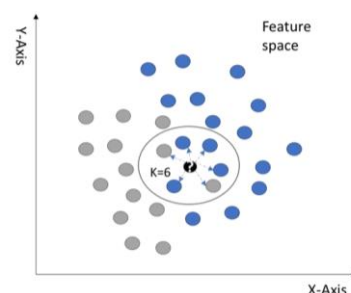


Figure 1. Illustration using KNN

3.5. Naive Bayes

Naive Bayes is a method used for probabilistic classification which was developed based on Bayes theorem. Naive Bayes aims to improve the learning efficiency of classification tasks. Naive Bayes can be trained to understand parameters by extracting basic statistics per class from features. This model may not depend on Bayesian algorithms. In the real world, naive Bayes can be used for real-time prediction and text classification. The weakness of naive Bayes is that this method is simple and this method only requires a limited amount of data to estimate the required categorization. Then this method does not need to evaluate the entire covariance matrix. The following is the formula in naive Bayes:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (3)$$

3.6. SVM

SVM is a supervised ML method that develops a decision boundary or hyperplane. SVM allows predicting labels based on a vector's single or multiple features. SVM comes from the name of the closest points which are known as support vectors. Based on this research, SVM is most often used for biomedica practice to automatically categorize profiles. For example, deoxyribonucleic acid (DNA) sequences, gene expression profiles, and so on. All of this can be used with the SVM method. Figure 2 is an example of scenario classification using SVM techniques.

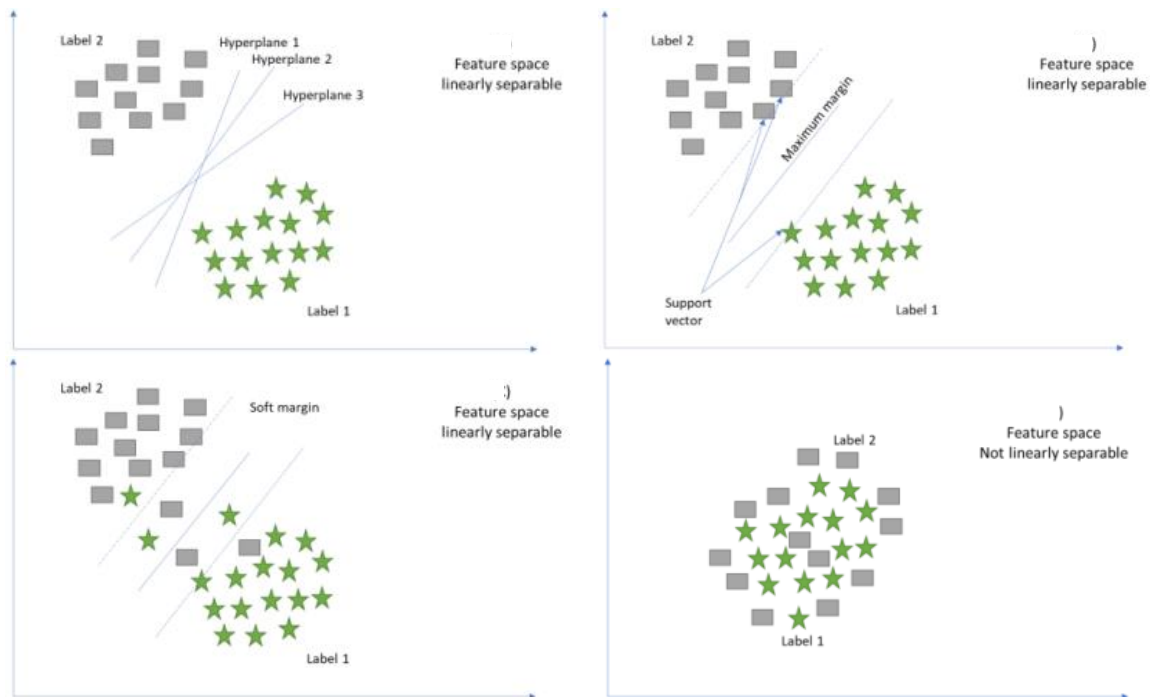


Figure 2. Classification using SVM techniques

4. PROPOSED SOLUTION METHOD

Proposed method aims to find solutions to research problems that have been explained in the background. This research method is written in framework form to make it easier to understand so as to achieve the goals in a structured manner. The research methodology for solutions based on brain error detection will be explained in Figure 3.

Based on the framework in Figure 3, the first stage in research is data collection and data exploration. In data exploration, researchers look for any classes in the data, then check for nulls, use MinMaxScaler and divide the data into 80 training data and 20 testing data. After that, researchers will conduct experiments using six methods, namely XGBoost, logistic regression, random forest, KNN, naive Bayes, and SVM and obtain their accuracy. After that, the researcher can see the results report in the form of precision, recall, and F1-score for each research method carried out.

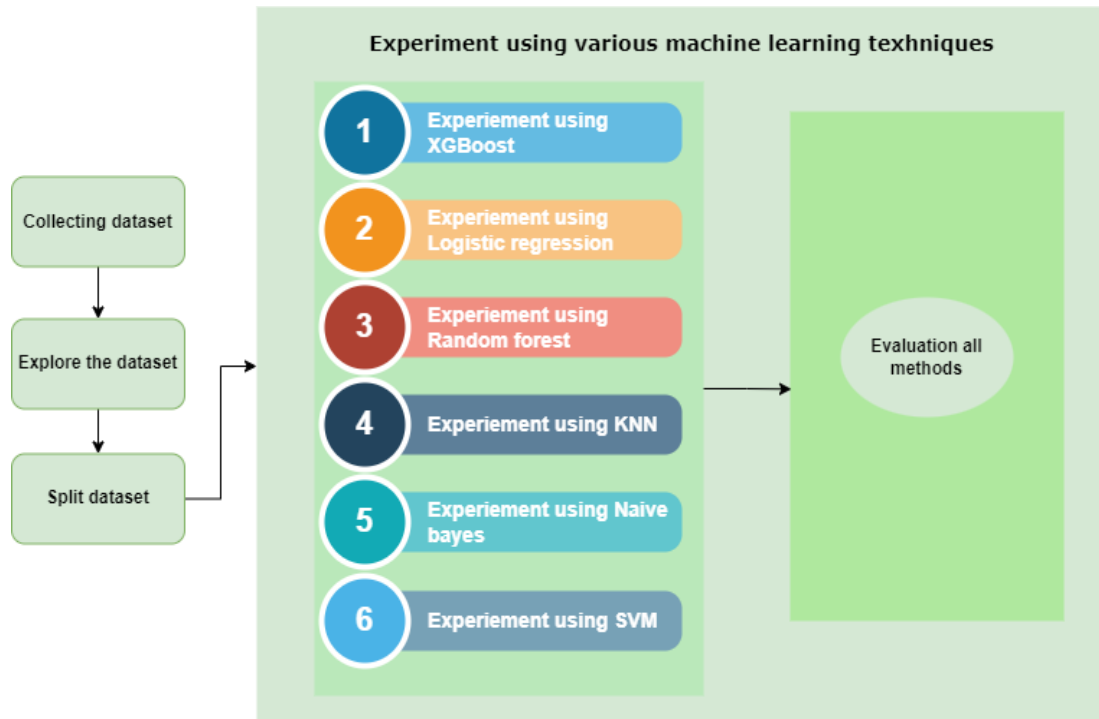


Figure 3. Proposed method

5. RESULTS AND DISCUSSION

In this section, the researcher divides it into two sections, namely sections 5.1 and 5.2. In section 5.1, the researcher explains and discusses about dataset. Researchers collect data and also explore the data that has been collected. And in section 5.2, the researcher explains the results of experiments that have been carried out using several methods chosen based on a literature review. The methods are XGBoost, logistic regression, random forest, KNN, naive Bayes, and SVM.

5.1. Collecting and exploring dataset

Researchers used data from the *Kaggle.com* site. This dataset is about brain tumor feature including five first order features and eight texture features with the target level. There were 3,762 data in this research. There is no duplicate data so the data remains 3,762 and is not reduced. After collecting the dataset, we continue with dataset exploration. Researchers carry out exploration by looking at class, variance, standard deviation, entropy, skewness, kurtosis, contrast, energy, angular second moment (ASM), homogeneity, and dissimilarity. The first order features are mean, variance, standard deviation, skewness, and kurtosis. Then the second order features are Contrast, energy, ASM, entropy, homogeneity, dissimilarity, correlation, and coarseness. Then the researchers also carried out the MinMaxScaler before carrying out experiments and evaluations. MinMaxScaler is usually used to create data in the range 0-1. The class column defines whether the image has a tumor or not. If there is a tumor, it is labeled 1; if not, it is labeled 0.

5.2. The results

Based on the steps that have been taken before getting results such as collecting and exploring datasets, dividing datasets, and others, in this chapter, the researcher will present the results of their experiments. Table 3 is an explanation of the method used and its accuracy. Then there is also a confusion matrix obtained from the six methods in Table 4. Confusion matrix is an important performance evaluation in artificial intelligence that provides an overall picture of the prediction results of the model used. Then, Figure 4 is a comparison of several methods used in this research experiment.

Based on the six methods used, including XGBoost, logistic regression, random forest, KNN, naive Bayes, and SVM, random forest is the method that produces the highest accuracy, namely 98.41%. XGBoost got an accuracy of 98.14%, logistic regression got an accuracy of 97.34%, KNN and naive Bayes of 97.34%, and SVM of 97.88%. All models achieved good accuracy, but when compared, random forest is the best method for this research case.

Table 3. Evaluation result for the research using various ML techniques

Methods	Accurations
XGBoost	98.14%
Logistic regression	97.34%
Random forest	98.41%
KNN	97.34%
Naïve Bayes	97.34%
SVM	97.88%

Table 4. Confusion matrix of the models

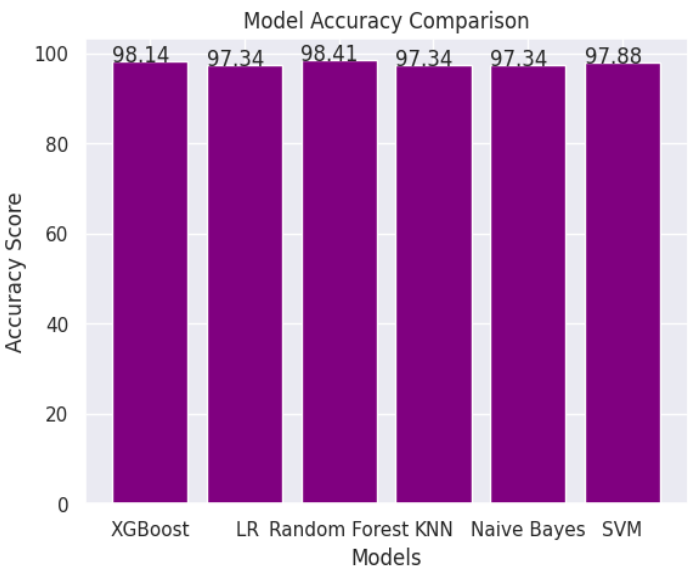
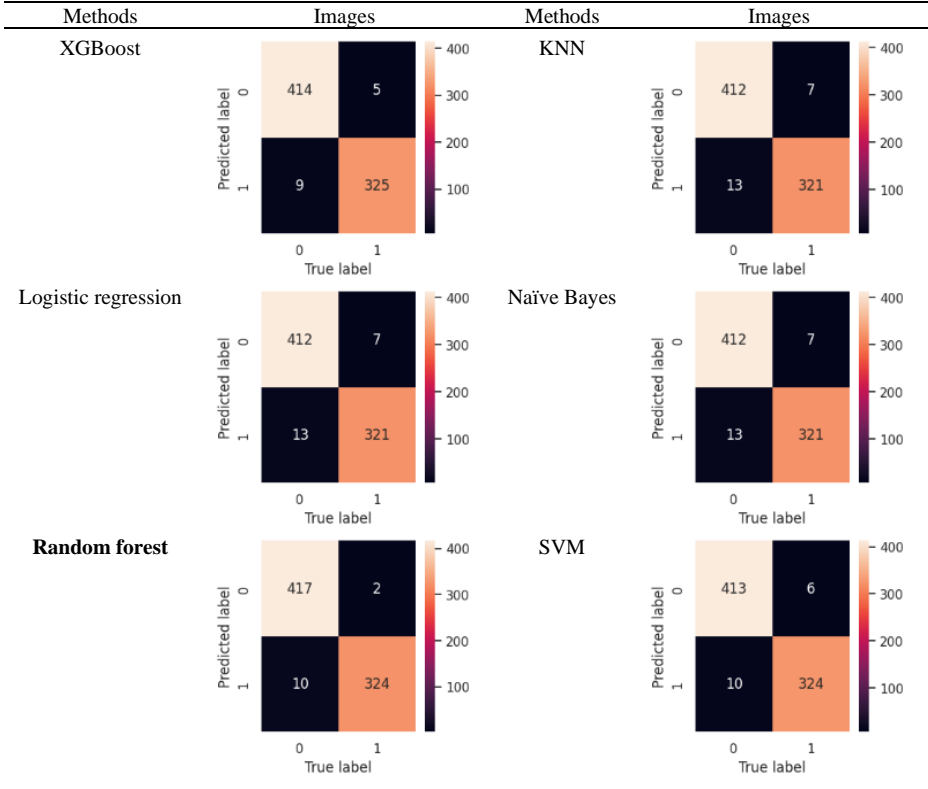


Figure 4. Image comparison of all accuracy

6. CONCLUSION

Brain tumor detection was carried out using six methods, namely XGBoost, Logistic regression, random forest, KNN, naive Bayes, and SVM. The data used comes from Kaggle.com where there are first-order and second order features. The first order features are mean, variance, standard deviation, skewness, and kurtosis. Then the second order features are contrast, energy, ASM, entropy, homogeneity, dissimilarity, correlation, and coarseness. For this research, data collection and exploration were carried out, then experiments and evaluation were carried out using a confusion matrix. In this research, random forest obtained the highest accuracy, namely 98.41%.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude for the support of Bina Nusantara University in financing of this scientific article's publication. They would also like to thank the Computer Science Department, School of Computer Science, for its support in writing this article.

FUNDING INFORMATION

This research was supported by Bina Nusantara University, which provided funding to publish this research in the International Journal of Electrical and Computer Engineering (IJECE).

AUTHOR CONTRIBUTIONS STATEMENT

The authors' contributions to this study are outlined below using the Contributor Roles Taxonomy (CRediT). Each author's role in the research process is listed to ensure transparency and clarity in authorship.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Rani Puspita	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Cindy Rahayu	✓				✓	✓	✓	✓			✓		✓	✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

The research related to human use has been complied with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

- [1] J. Amin, M. Sharif, M. Raza, T. Saba, and M. A. Anjum, "Brain tumor detection using statistical and machine learning method," *Computer Methods and Programs in Biomedicine*, vol. 177, pp. 69–79, Aug. 2019, doi: 10.1016/j.cmpb.2019.05.015.
- [2] J. Kang, Z. Ullah, and J. Gwak, "MRI-based brain tumor classification using ensemble of deep features and machine learning classifiers," *Sensors*, vol. 21, no. 6, pp. 1–21, Mar. 2021, doi: 10.3390/s21062222.
- [3] M. Siar and M. Teshnehlal, "Brain tumor detection using deep neural network and machine learning algorithm," in *2019 9th*

Implementing brain tumor detection using various machine learning techniques (Rani Puspita)

- International Conference on Computer and Knowledge Engineering, ICCKE 2019*, Oct. 2019, pp. 363–368. doi: 10.1109/ICCKE48569.2019.8964846.
- [4] U. Zahid *et al.*, “BrainNet: Optimal deep learning feature fusion for brain tumor classification,” *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–13, Aug. 2022, doi: 10.1155/2022/1465173.
 - [5] A. Rehman, M. A. Khan, T. Saba, Z. Mehmood, U. Tariq, and N. Ayesha, “Microscopic brain tumor detection and classification using 3D CNN and feature selection architecture,” *Microscopy Research and Technique*, vol. 84, no. 1, pp. 133–149, Sep. 2021, doi: 10.1002/jemt.23597.
 - [6] A. B. Ifra and M. Sadaf, “Automatic brain tumor detection using convolutional neural networks,” in *Lecture Notes in Networks and Systems*, vol. 494, Springer Nature Singapore, 2023, pp. 419–427. doi: 10.1007/978-981-19-4863-3_41.
 - [7] N. Noreen, S. Palaniappan, A. Qayyum, I. Ahmad, M. Imran, and M. Shoaib, “A deep learning model based on concatenation approach for the diagnosis of brain tumor,” *IEEE Access*, vol. 8, pp. 55135–55144, 2020, doi: 10.1109/ACCESS.2020.2978629.
 - [8] M. Shahajad, D. Gambhir, and R. Gandhi, “Features extraction for classification of brain tumor MRI images using support vector machine,” in *Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering*, Jan. 2021, pp. 767–772. doi: 10.1109/Confluence51648.2021.9377111.
 - [9] A. Younis, L. Qiang, C. O. Nyatega, M. J. Adamu, and H. B. Kawuwa, “Brain tumor analysis using deep learning and VGG-16 ensembling learning approaches,” *Applied Sciences*, vol. 12, no. 14, p. 7282, Jul. 2022, doi: 10.3390/app12147282.
 - [10] L. Rubinger, A. Gazendam, S. Ekhtiari, and M. Bhandari, “Machine learning and artificial intelligence in research and healthcare,” *Injury*, vol. 54, pp. S69–S73, May 2023, doi: 10.1016/j.injury.2022.01.046.
 - [11] C. Zhang, Y. Liu, and N. Tie, “Forest land resource information acquisition with Sentinel-2 image utilizing support vector machine, k-nearest neighbor, random forest, decision trees and multi-layer perceptron,” *Forests*, vol. 14, no. 2, p. 254, Jan. 2023, doi: 10.3390/f14020254.
 - [12] A. Tariq *et al.*, “Modelling, mapping and monitoring of forest cover changes, using support vector machine, kernel logistic regression and naive bayes tree models with optical remote sensing data,” *Heliyon*, vol. 9, no. 2, p. e13212, Feb. 2023, doi: 10.1016/j.heliyon.2023.e13212.
 - [13] Z. Nabavi, M. Mirzei, H. Dehghani, and P. Ashtari, “A hybrid model for back-break prediction using XGBoost machine learning and metaheuristic algorithms in Chadormalu Iron Mine,” *Journal of Mining and Environment*, vol. 14, no. 2, pp. 689–712, 2023, doi: 10.22044/jme.2023.12796.2323.
 - [14] G. Ankit and A. Kole, “A comparative study of enhanced machine learning algorithms for brain tumor detection and classification,” *Authorea Preprints*, Oct. 2023, doi: 10.36227/techrxiv.16863136.
 - [15] A. Nurdina and A. B. I. Puspita, “Naïve Bayes and KNN for airline passenger satisfaction classification: Comparative analysis,” *Journal of Information System Exploration and Research*, vol. 1, no. 2, Jul. 2023, doi: 10.52465/joiser.v1i2.167.
 - [16] G. Nguyen *et al.*, “Machine learning and deep learning frameworks and libraries for large-scale data mining: A survey,” *Artificial Intelligence Review*, vol. 52, no. 1, pp. 77–124, 2019, doi: 10.1007/s10462-018-09679-z.
 - [17] H. B. Ly, T. A. Nguyen, and B. T. Pham, “Estimation of soil cohesion using machine learning method: A random forest approach,” *Advances in Civil Engineering*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/8873993.
 - [18] N. H. N. B. M. Shahri, S. B. S. Lai, M. B. Mohamad, H. A. B. A. Rahman, and A. Bin Rambli, “Comparing the performance of adaboost, xgboost, and logistic regression for imbalanced data,” *Mathematics and Statistics*, vol. 9, no. 3, pp. 379–385, May 2021, doi: 10.13189/ms.2021.090320.
 - [19] H. V. Thanh *et al.*, “Hydrogen storage on porous carbon adsorbents: rediscovery by nature-derived algorithms in random forest machine learning model,” *Energies*, vol. 16, no. 5, p. 2348, Feb. 2023, doi: 10.3390/en16052348.
 - [20] U. K. Lilhore *et al.*, “Hybrid model for precise hepatitis-C classification using improved random forest and SVM method,” *Scientific Reports*, vol. 13, no. 1, Aug. 2023, doi: 10.1038/s41598-023-36605-3.
 - [21] A. Samad, S. Taze, and M. Kürsad Uçar, “Enhancing milk quality detection with machine learning: a comparative analysis of KNN and distance-weighted KNN algorithms,” *International Journal of Innovative Science and Research Technology (IJISRT)*, pp. 2021–2029, Apr. 2024, doi: 10.38124/ijisrt/ijisrt24mar2123.
 - [22] H. Gonaygunta, “Machine learning algorithms for detection of cyber threats using logistic regression,” *International Journal of Smart Sensor and Adhoc Network*, pp. 36–42, Jan. 2023, doi: 10.47893/ijssan.2023.1229.
 - [23] W. Chang, X. Chen, Z. He, and S. Zhou, “A prediction hybrid framework for air quality integrated with W-BiLSTM(PSO)-GRU and XGBoost methods,” *Sustainability (Switzerland)*, vol. 15, no. 22, p. 16064, Nov. 2023, doi: 10.3390/su152216064.
 - [24] J. Gu and S. Lu, “An effective intrusion detection approach using SVM with naïve Bayes feature embedding,” *Computers and Security*, vol. 103, p. 102158, Apr. 2021, doi: 10.1016/j.cose.2020.102158.
 - [25] R. D. Joshi and C. K. Dhakal, “Predicting type 2 diabetes using logistic regression and machine learning approaches,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 14, p. 7346, Jul. 2021, doi: 10.3390/ijerph18147346.
 - [26] D. N. George, H. B. Jehloli, A. Subhi, and A. Oleiwi, “Brain tumor detection using shape features and machine learning algorithms,” *International Journal of Scientific & Engineering Research*, vol. 6, no. 12, p. 454, 2015.
 - [27] J. Amin, M. Sharif, M. Raza, and M. Yasmin, “Detection of brain tumor based on features fusion and machine learning,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 15, no. 1, pp. 983–999, Nov. 2024, doi: 10.1007/s12652-018-1092-9.
 - [28] C. L. Choudhury, C. Mahanty, R. Kumar, and B. K. Mishra, “Brain tumor detection and classification using convolutional neural network and deep neural network,” *Computers, Materials and Continua*, vol. 73, no. 3, pp. 4501–4518, 2022, doi: 10.32604/cmc.2022.030392.
 - [29] A. H. Khan *et al.*, “Intelligent model for brain tumor identification using deep learning,” *Applied Computational Intelligence and Soft Computing*, vol. 2022, pp. 1–10, Jan. 2022, doi: 10.1155/2022/8104054.
 - [30] Z. Jia and D. Chen, “Brain tumor identification and classification of MRI images using deep learning techniques,” *IEEE Access*, pp. 1–1, 2024, doi: 10.1109/ACCESS.2020.3016319.
 - [31] E. M. Senan, M. E. Jadhav, T. H. Rassem, A. S. Aljaloud, B. A. Mohammed, and Z. G. Al-Mekhlafi, “Early diagnosis of brain tumour MRI images using hybrid techniques between deep and machine learning,” *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1–17, May 2022, doi: 10.1155/2022/8330833.
 - [32] T. Chen and C. Guestrin, “XGBoost,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
 - [33] A. Sharma and W. J. M. I. Verbeke, “Improving diagnosis of depression with XGBoost machine learning model and a large biomarkers Dutch dataset (n = 11,081),” *Frontiers in Big Data*, vol. 3, Apr. 2020, doi: 10.3389/fdata.2020.00015.
 - [34] L. Torlay, M. Perrone-Bertolotti, E. Thomas, and M. Baci, “Machine learning–XGBoost analysis of language networks to classify patients with epilepsy,” *Brain Informatics*, vol. 4, no. 3, pp. 159–169, Apr. 2017, doi: 10.1007/s40708-017-0065-7.
 - [35] S. Nusinovic *et al.*, “Logistic regression was as good as machine learning for predicting major chronic diseases,” *Journal of*

- Clinical Epidemiology*, vol. 122, pp. 56–69, Jun. 2020, doi: 10.1016/j.jclinepi.2020.03.002.
- [36] M. S. A. Nasser and S. S. Abu-naser, "Predictive modeling of obesity and cardiovascular disease risk: a random forest approach," vol. 7, no. 12, pp. 26–38, 2023.
- [37] S. K. Sharma, U. K. Lilhore, S. Simaiya, and N. K. Trivedi, "An improved random forest algorithm for predicting the COVID-19 pandemic patient health," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 1, pp. 67–75, 2021.
- [38] G. Haripriya, K. Abinaya, N. Aarthi, and P. P. Kumar, "Random forest algorithms in health care sectors: a review of applications," *International Journal of Recent Development in Computer Technology & Software Applications*, vol. 5, no. July 2021, 2021.
- [39] A. Aldahiri, B. Alrashed, and W. Hussain, "Trends in using IoT with machine learning in health prediction system," *Forecasting*, vol. 3, no. 1, pp. 181–206, Mar. 2021, doi: 10.3390/forecast3010012.

APPENDIX

Table 1. Related works – ML techniques




Reference	Process	Method	Result
Thanh <i>et al.</i> [19]	Sample data collection, implementation ML model, nature-inspired model, evaluation method	Random forest	This research received a relevance score of 1 and 0.48. root mean square error (RMSE) and mean absolute error (MAE) reaches 0.6 to 1, and 0.38 to 0.52
Lilhore <i>et al.</i> [20]	Data visualization, data pre-processing and splitting (data cleaning, data normalization, and verify attributes), ranker method, check data imbalance, classification, and prediction outcome	Random forest and SVM	In experiment 1, the proposed method achieved an accuracy of 95.89% for k=5 and 96.29% for k=10. For the second attempt, we got 91.24% for 80:20 and 92.39% for 70:30
Samad <i>et al.</i> [21]	Data preprocessing, model creation with KNN dan distance-weighted KNN (DW-KNN), performance evaluation criteria	KNN and DW-KNN	Shows that DW-KNN and KNN have almost similar accuracy, namely 99.53% and 98.58%
Gonaygunta [22]	Data exploration, extraction of the features, training, classification, and evaluation	Logistic regression	Logistic regression is the first of three ML subcategories. Therefore, logistic regression is an important algorithm and can be applied to cases in ML
Chang <i>et al.</i> [23]	Data description and preprocessing, correlation analysis, experiment with the methods, evaluation	Bidirectional long short-term memory, particle swarm algorithm-gated recurrent unit (W-BiLSTM PSO-GRU) and XGBoost methods	The results show that the accuracy is 0.94 and the MAE and RMSE values are lower than 0.02 and 0.03, respectively
Gu and Lu [24]	Intrusion or normal, intrusion detection model with naïve Bayes-SVM, train SVM classifier, transformed data, naïve Bayes feature embedding, evaluation	SVM with naïve Bayes feature embedding	The experiment got accurate and good results, namely 93.75% accuracy on the UNSW-NB15 dataset, 98.92% accuracy on the CICIDS2017 dataset, 99.35% accuracy on the NSL-KDD dataset and 98.58% accuracy on the Kyoto 2006 + dataset
Joshi and Dhakal [25]	Data exploration, experiment with logistic regression, model selection, validation and cross validation method, classification tree, evaluation	Logistic regression	This research succeeded in getting an accuracy of 78.26% and an error rate of 21.74%. This model is good enough to make predictions about type 2 diabetes, so it can be used as an early preventive measure

Table 2. Related works-brain tumor detection




Reference	Process	Method	Result
George <i>et al.</i> [26]	Image acquisition, pre-processing, segmentation, feature extraction, classification, and evaluation	Decision tree algorithm and MLP algorithm	This research succeeded in obtaining precision with an accuracy of 95% by considering 174 brain image samples using MLP
Amin <i>et al.</i> [27]	Lesion enhancement, lesion segmentation, feature extraction, classification, evaluation	RF	This research succeeded in getting good accuracy, namely 0.91 complete, 0.89 non-enhancing, and 0.90 enhancing dice similarity coefficient
Choudhury <i>et al.</i> [28]	Convolutional neural networks (CNN), model description, experiment with CNN, evaluation	CNN	The research achieved the accuracy 96.08%, and the f-score 97.3
Khan <i>et al.</i> [29]	Data preprocessing, splitting data into training and validation, experiment with CNN, and evaluation	CNN	This research obtained an accuracy of 92.13% for precision and a miss rate of 7.87%
Jia and Chen [30]	Preprocessing, skull stripping, segmentation, morphological operation, feature extraction, SVM for classification, testing	Fully automatic heterogeneous segmentation using SVM (FAHS-SVM)	In this research, the numerical accuracy results were 98.51%
Senan <i>et al.</i> [31]	Collecting dataset, using Residual neural network 18 (ResNet18) and AlexNet, classification with SoftMax, SVM, and KNN, evaluation	SoftMax, SVM, KNN	This research got very good results. The AlexNext + SVM technique obtained the best results with an accuracy of 95.10%, sensitivity of 95.25%, and specificity 98.50%

Implementing brain tumor detection using various machine learning techniques (Rani Puspita)

BIOGRAPHIES OF AUTHORS

Rani Puspita    is a lecturer at the School of Computer Science at Bina Nusantara University. Focus on artificial intelligence, multimedia technology, and software engineering. Her undergraduate education background is Informatics Engineering at Syarif Hidayatullah State Islamic University Jakarta, and her graduate education background is computer science at Bina Nusantara University. She is the best graduate of Informatics Engineering Major at Syarif Hidayatullah State Islamic University Jakarta (115th Graduation). And she received Summa Cum Laude of Master of Computer Science at Binus University (65th Graduation). She can be contacted via email: rani.puspita@binus.ac.id.



Cindy Rahayu    is a lecturer at Bina Nusantara University with a focus on artificial intelligence. Her undergraduate education background is informatics engineering at Syarif Hidayatullah State Islamic University Jakarta and her graduate education background is Computer Science at Bina Nusantara University. She can be contacted at email: cindy.rahayu@binus.ac.id.