Jurnal TAM (*Technology Acceptance Model*)
Jurnal TAM, Volume 14, Number 2, December 2023
E ISSN: 2579-4221; P ISSN: 2339-1103, pp. 152-157
Accredited SINTA 4 Number 225/E/KPT/2022
https://jurnal.ftikomibn.ac.id/index.php/JurnalTam/index

COMPARISON OF DECISION TREE AND NAÏVE BAYES ALGORITHMS IN CLASSIFICATION MODELS TO DETERMINE LECTURER PERFORMANCE USING K FOLD CROSS VALIDATION

Erna Kumalasari Nurnawati[1*], Muhammad Sholeh[2], Renna Yanwastika Ariyana[3], Eska Almuntaha[4]
[1,2,3]Prodi Informatika, Institut Sains & Teknologi Akprind, Yogyakarta
[4]Prodi Bisnis Digital, Institut Sains & Teknologi Akprind, Yogyakarta
[1,2,3,4]Jl Kalisahak 28 Balapan, Yogyakarta, Indonesia
E-mail: ernakumala@akprind.ac.id[1*], muhash@akprind.ac.id[2],renna@akprind.ac.id[3], eska@akprind.ac.id[4]

Article history:
Received: August 19, 2023
Revised: November 09, 2023
Accepted: December 02, 2023

Corresponding authors
*ernakumala@akprind.ac.id

Keywords:
Classification;
Lecture Performance;
Decision Tree;
Naïve Bayes;
K-Fold Validation;

Abstract
Lecturer performance is very important to support the progress of higher education. Determination of lecturer performance is based on Tri Dharma activities, including: teaching, research and community service. This study aims to build a model that can predict the predicate of lecturers from the activities carried out. The best model is obtained by comparing the use of two algorithms, namely Decision Tree and Naive Bayes. Data mining methods use the CRISP-DM method, namely business understanding, data understanding, data preparation, modeling, evaluation, and development. Performance testing of training data using K Fold Cross Validation. The modeling results with this performance show that the Decision Tree algorithm has better performance with 94.70%, accuracy, 93.24% precision and 96.33% recall, while Naïve Bayes algorithm has performance with 92.95%, accuracy 90.08% and 96.33%. This shows that modeling using the Decision Tree algorithm can be used as a model in determining lecturer performance.

1. INTRODUCTION

Data mining is an evolution of information technology and is an interdisciplinary subject and is often referred to as knowledge discovery from data [1]. Data mining is a process of finding information from data stored in a database or datasheet with a certain algorithm. The data mining process uses various techniques such as techniques in statistical, mathematical, and machine learning. These techniques will identify and process data into a model that can be used as a reference in decision making. Many policies and decisions are made based on data. Data is an asset and an important element in an institution, both government and private, education-based institutions, banking, military, disaster management, tourism, and so on. Data becomes an asset that can be used to find patterns that can be used in decision making. The information obtained from the model can be used in projecting strategies or policies carried out for the business development process. The search for big data must be done carefully. The bigger the data, the bigger the process needed to sort the data according to the needs. Managing large amounts of data with many attributes and classifying is an important step so that the required information can be presented as needed. Currently, data mining is growing rapidly, due to the growing use of non-structured data that is increasingly being used [2]. Prediction is one of the methods in data mining that is used to make model predictions using historical data. Classification is the work of data analysis, namely how to find a model that describes and distinguishes classes, identifies a set of categories on the basis of a training data set whose categories are known [1]. The algorithms used in the classification include Decision Tree, Neural Network, Support Vector Machine, kNN and Naïve Bayes.

Lecturers are professional teachers and scientists whose job is to transform and develop education through education, research, and community service based on the "Tri Dharma" of Higher Education [3]. Lecturer performance is measured by the performance of lecturers in the tri dharma, namely education, research and community service. Each of these tasks is further derived in several parameters. This study aims to compare the performance of the Decision Tree and Naïve Bayes algorithms to get a better classification model in determining the classification model that can predict the classification of lecturer performance.

Research related to the classification model has been carried out by Bilal et al. to classify Roman-Urdu Sentiment mining using the Naïve Bayes, Decision Tree and k-NN methods. In this case, the Naïve Bayes algorithm has the best performance [4]. Meanwhile Fitri et al conducted a sentiment analysis classification on Twitter social media using the Naïve Bayes, Decision Tree and Random Forest methods. In this study, the Naïve Bayes algorithm has the best performance [5]. Comparison of Decision Tree and Naïve Bayes algorithms was also used to classify the registration of Diabetes patients with HbA1c measurements of prospective patients by Pujianto et al. [6]. The use of Naïve Bayes and Random Forest methods was also used to predict individual survival until the second lactation in dairy cows carried out by Van der Heide et al. [7]. Meanwhile, Ashari et al compared the performance of the Naïve Bayes algorithm, Decision Tree and k-NN to determine alternative building designs. Best performance is to use Decision Tree [8]. Rahmadani et al also compared the performance of the Naïve Bayes algorithm and Decision Tree to select features in the genetic algorithm [9]. Meanwhile, Suryadi et al conducted a comparative analysis of the Decision Tree and Naïve Bayes algorithms to determine the classification of university-level new student profiles [10]. The comparison of the performance of Naïve Bayes and Decision Tree used to measure the performance of 480 students in India was carried out by Yadav et al [11]. Meanwhile, Farhana classified academic performance in evaluating research on academic staff in Kuala Lumpur using Naïve Bayes algorithm [12].

The development of data mining classification models can be built with various applications, both based on programming languages and visual programming. As done by Ashari et al [8], Yadav et al [11] made a model using programming, while Farhana [12], Puspita et al [13], Pujianto [6]and Rosandy [14] used visual programming.

In this research, Naïve Bayes algorithm and Decision Tree are used to classify lecturer performance. Teaching performance parameters include: lecturer questionnaire, quantitative learning achievement, qualitative learning achievement, accuracy in submitting student grades and number of lecture attendance. While the research performance parameters include: the number of research activities, publications, patents/intellectual property rights. While the community service performance parameters are calculated from the number of community service activities, journal publications and the application of appropriate technology produced. The next attribute is attendance at meetings of study programs, faculties and universities.

II. LITERATURE

2.1. Classification

Classification is the process of finding a pattern or function that describes and distinguishes classes of data or a concept [1]. This is a data analysis task, i.e., the process of finding a pattern that describes and distinguishes data classes and concepts. Classification is the problem of determining to which set of categories (subpopulations) a new observation belongs, based on a training dataset containing observations and members of the known type. Formula of the classification problem is presented in the following formulation [15]. If given data labelled which contains n-samples and each sample is represented by k-features,

$$x^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_k^{(i)} \end{pmatrix} \in X \text{ where } X \in R^k \text{ and } y^{(i)} \in Y = \{0,1,\ldots k\}\, k : \text{constant}$$

The classification model is a supervised learning model that aims to approximate the mapping f: X → Y with the classification model f. In addition, the classification model f is used to predict the label (y) of the experimental data.

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix} \in$$

R^k *so that the prediction of the test data label y=f(x)*

To measure success in classification problems are accuracy, precision, recall and F score. Comparison of the performance of various algorithms seen from the percentage of the four parameters above.

2.2. Decision Tree Algorithms

A decision tree is a classification methodology in which the classification process is modelled using a set of hierarchical decisions about feature variables arranged in a tree-like structure. Decisions at specific nodes of the tree, called split criteria, are usually conditions on one or more feature variables in the training data. A splitting criterion divides the training data into two or more parts [15]. Decision tree guidance algorithms have two types of nodes, called interior nodes and leaf nodes. Each leaf node is labelled with the dominant class of that node. A special interior node is the root node, which

corresponds to the entire feature space. A typical decision tree induction algorithm starts with the full training data set at the root node and recursively splits the data into lower-level nodes based on splitting criteria.

The stages of the decision tree algorithm are as follows:

a. Prepare data training
b. Select attribute as root

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

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

Which S is the set of cases, n is the number of partitions, and is the fraction of in S.

c. Create a branch for each value
d. Repeat the process for each branch until all cases on the branch have the same class

2.3 Naïve Bayes Algorithms

The Naïve Bayes algorithm is a probabilistic model that aims to predict the category of the sample data expressed by probability [1]. Bayes classifiers are based on Bayes' conditional probability theorem. This theorem quantifies the conditional probability of a random variable (class variable) given known observations about the values of another set of random variables (characteristic variables). Bayes' theorem is widely used in probability and statistics [15].

$$P(Y=Y|X=(x_1,x_2,.....,x_k))$$

Prediction results of data sample category x= using Naïve Bayes model is y* obtained by maximizing the value of or expressed by the equation:

$$y^*=argmax\ P(Y=Y|X=(x_1,x_2,.....,x_k)$$

Bayes' Theorem:

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)}$$

Where the probability values P(X|Y) and P(Y) are calculated from the training data. The stages of the Naive Bayes algorithm are [16]:

a. Preparing for training data
b. Fit naive Bayes classifier to training set
c. Predicting the test set results
d. Conclusion

III. RESEARCH METHODS

This study included eight main phases: data collection and understanding, preprocessing, data cleansing, developing models, validation test, analysis and training, testing, and outcome analysis. The research was conducted using a datasheet that was processed from the data used to assess the performance indicators of lecturers. Data mining methods use the CRISP-DM (Cross-Industry Standard Process for Data Mining) method. CRISP-DM is a data mining methodology which consists of six stages, namely business understanding, data understanding, data preparation, modeling, evaluation, and development.

3.1. Research Stages

The initial stage of the research is to understand the data. The data stored in the datasheet is analysed. The analysis process includes checking the attributes needed in classification modeling, checking incomplete data, empty data and others. Next, the model is made using Visual Programming with both algorithms. Furthermore, a performance analysis is carried out by looking at the parameter values for the success of the model. The stages of the research are presented in Figure 1.

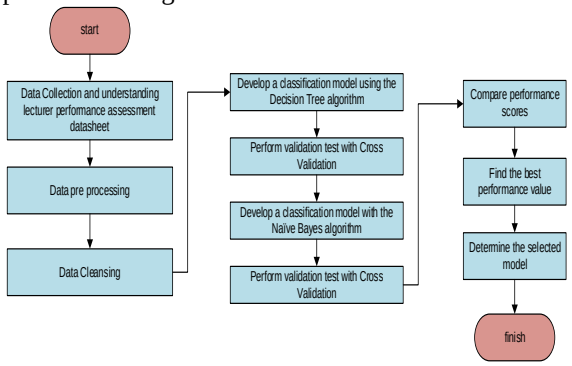

Figure 1. Research Stages

3.2. Dataset

The datasheet used is processed from various units which are a collection of those processed from research data, scientific publications, implementation of community service, teaching and learning processes and lecturer activities in participating in activities in study programs. The results of data processing are a process of various data stored in a database and for the purposes of data mining processes, a collection of data from various databases is stored in the form of a CSV file. The datasheet consists of lecturer performance data of Institut Sains & Teknologi Akprind in 2019-2021 consisting of 113 rows and 8 attributes. The k-fold is a way to evaluate model performance or algorithms [17]. The modeling is done using Rapid Miner Studio Educational 9.10.008 Visual Programming.

3.3. Analysis

The analysis in this study compared training, validation, and test performance, based on three models using confusion matrices. Table 1 shows the confusion matrix of lecturer performance predicates [18].

Table 1. Confusion Matrix of Lecture Predicate

	Actual Good	**Actual Poor**
Predicted Good	True Positive (TP)	False Negative (FN)

Predicted Poor	False Positive (FP)	True Negative (TN)

Table 2 presents the performance of the classification matrix on the model [18] and [19]:

Table 2. Classification Task Performance Metrics

Metrics	**Formula**
Accuracy	(TP+TN)/(TP+FP+TN+FN)
Precision	TP/(TP+FP)
Recall	TP/(TP+TN)
F-Score	2*Precision*Recall/ (Precision Recall)

3.4. K-Fold Cross Validation

Cross validation is a statistical method for evaluating and comparing learning algorithms by dividing the data into two, namely training data and testing data. K-Fold Cross Validation could be an approval strategy by partitioning the information into k-subsets, at that point rehashing it k times for learning and testing [17]. In each reiteration, one subset is utilized as test information and the other subset as learning information. The testing division prepare is displayed in table 3.

Tabel 3. Fold Cross Validation Division [17]

Testing 1	Test	Train	Train	Train	Train	Train
Testing 2	train	Test	Train	Train	Train	Train
Testing 3	Train	Train	Test	Train	Train	Train
Testing 4	Train	Train	Train	Test	Train	Train
Testing 4	Train	Train	Train	Train	Test	Train

IV. RESULTS

4.1. Business Understanding

At this stage, an analysis is carried out to develop a classification model that can be used to predict lecturers' performance predicates. Predictions can be made by attribute data from research results, publications, community service activities, teaching and learning processes and the activeness of lecturers in activities in the study program. Based on these attributes, a classification model can be made that can predict the performance of lecturers based on available data.

4.2. Data Understanding

Data understanding is done by understanding the existing datasheet. The existing attributes were analysed before clustering. At the data understanding stage, identification of the attributes in the datasheet is also carried out. There are existing attributes that do not match the criteria in the classification, so the attribute will be deleted. Attributes that match the criteria will be used in the classification process. Figure 2 presents the imported datasheet into Rapid Miner Studio.

4.3. Data Preparation

Data preparation is carried out so that the processed data has been avoided from data containing errors or unnecessary data. The steps taken include:

a. Checking data type. Attributes that can be used in the classification process must be of real type. Real type attributes are PBM, PD, PM and KR. Data type checking must be done before the classification process.
b. Checking missing value. The datasheet used must be free from missing values. The inspection is done by looking at the missing column. If in the missing column all attributes contain 0, then there is no missing value in the dataset.
c. Define data labels. One of the requirements of the classification model is the existence of an attribute that becomes a label. In Rapid Miner, set role and predicate set operators can be used. The labels that become the predicate are divided into two categories: Good or Poor.
d. Selecting the attributes used in model generation. Not all of the attributes in the datasheet are used in modeling. Only the most influential attributes will be used. Attributes that are not used include the name and department code attribute. Figure 2 presents the attribute selection process in Rapid Miner with the attribute select operator.

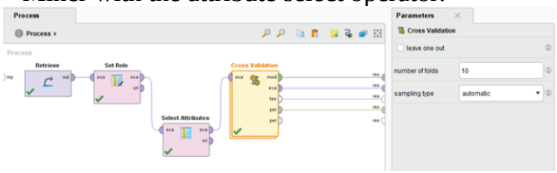

Figure 2. Selection of attributes used in modeling

4.4. Model Development

Next step is to develop a classification model with Decision Tree and Naïve Bayes algorithms. Evaluation results are used as a comparison model. The results of the comparison are done by looking at the performance results. The model selected is the model with the best performance. The validation process is performed using K-fold cross-validation.

a. Classification model using Decision Tree algorithm

Development of a classification model with a Decision Tree algorithm on Rapid Miner using the Decision Tree operator and validation using the Cross Validation operator using 10 folds Cross Validation. Figure 3 presents the Decision Tree algorithm selection process on Rapid Miner.

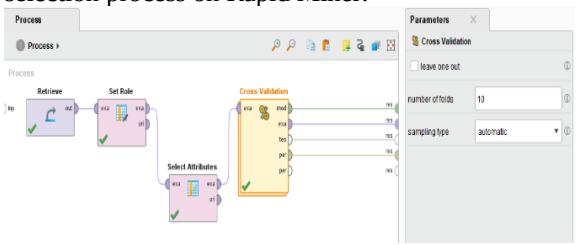

Figure 3. Classification Modelling using Decision Tree Algorithm

The model generated with training data must be tested to determine its performance. Figure 4 presents the process of testing model performance using Rapid Miner.

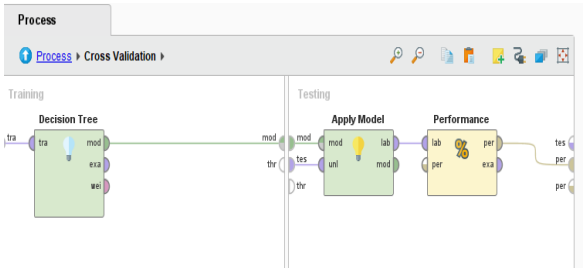

Figure 4. Model Performance Testing

b. Classification model using Naïve Bayes algorithm

Developing a classification model with the Naive Bayes algorithm in Rapid Miner using the Naive Bayes operator and validation using the cross-validation operator with 10-fold cross-validation. Figure 5 shows the Naive Bayes algorithm selection process in Rapid Miner. As in the Decision Tree algorithm, the modeling results are also tested for performance with performance operators.

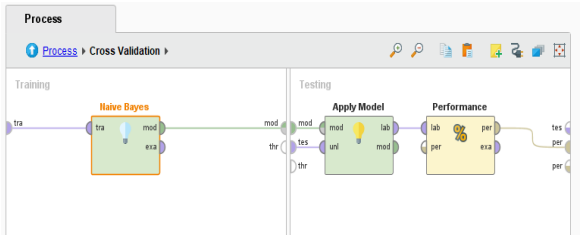

Figure 5. Classification Modelling using Naïve Bayes Algorithm

4.5. Model Evaluation

A classification model is selected based on the best performance of the resulting model. Evaluation is done by cross-validation. Figure 6 and Figure 7 shows the performance results of the Decision Trees and Naive Bayes algorithms. Based on Figure 6 and Figure 7, the accuracy of the Decision Tree algorithm is 94.70%, while the accuracy of the Nave Bayes algorithm is 92.95%.

Table View Plot View

accuracy: 94.70% +/- 6.27% (micro average: 94.78%)

	true Good	true Poor	class precision
pred. Good	56	2	96.55%
pred. Poor	4	53	92.98%
class recall	93.33%	96.36%	

Figure 6. Precision Value for Decision Tree

Table View Plot View

accuracy: 92.95% +/- 5.68% (micro average: 93.04%)

	true Good	true Poor	class precision
pred. Good	54	2	96.43%
pred. Poor	6	53	89.83%
class recall	90.00%	96.36%	

Figure 6. Precision Value for Naïve Bayes

Table 3 presents a comparison of the performance results of the two algorithms:

Tabel 3. Performance Comparison between two algorithms

	Decision Tree Algorithm	Naïve Bayes Algorithm
Accuracy	94.70% +/- 6.27% (micro average: 94.78%)	92.95% +/- 5.68% (micro average: 93.04%)
Precision	93.24% +/- 8.83% (micro average: 92.98%) (positive class: Poor)	90.14% +/- 8.63% (micro average: 89.83%) (positive class: Poor)
Recall	96.33% +/- 7.77% (micro average: 96.36%) (positive class: Poor)	96.33% +/- 7.77% (micro average: 96.36%) (positive class: Poor)

According to table 3, the accuracy and precision of the Decision Tree algorithm is higher than the Naïve Bayes algorithm. While the recall value of the two algorithms is the same. So, it can be concluded that, in making a model to determine the predicate of lecturer performance results, it can be recommended to use a Decision Tree algorithm.

V. CONCLUSION

The development of a classification model to determine the predicate of lecturers using the Decision Tree and Naive Bayes methods produces a model whose performance can be measured with performance matrices parameters, namely accuracy, precision, recall and F-Score. The modeling results with this performance show that the Decision Tree algorithm has better performance with 94.70%, accuracy, 93.24% precision and 96.33% recall, while Naïve Bayes algorithm has performance with 92.95%, accuracy 90.08% and 96.33%.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, "Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)," 2022.

[2] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," IJCSI International Journal of Computer Science Issues, vol. 4, no. 1, 2009.

[3] Government of Republic Indonesia, "PERATURAN PEMERINTAH REPUBLIK INDONESIA," 2009. Accessed: Aug. 14, 2022. [Online]. Available: https://peraturan.bpk.go.id/Home/Details/4956/pp-no-37-tahun-2009

[4] M. Bilal, H. Israr, M. Shahid, and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques," Journal of King Saud University - Computer and Information Sciences, vol. 28, no. 3, pp. 330–344, Jul. 2016, doi: 10.1016/j.jksuci.2015.11.003.

[5] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm," in Procedia Computer Science, 2019, vol. 161, pp. 765–772. doi: 10.1016/j.procs.2019.11.181.

[6] U. Pujianto, A. L. Setiawan, H. A. Rosyid, and A. M. M. Salah, "Comparison of Naïve Bayes Algorithm and Decision Tree C4.5 for Hospital Readmission Diabetes Patients using HbA1c Measurement," Knowledge Engineering and Data Science, vol. 2, no. 2, p. 58, Dec. 2019, doi: 10.17977/um018v2i22019p58-71.

[7] E. M. M. van der Heide, R. F. Veerkamp, M. L. van Pelt, C. Kamphuis, I. Athanasiadis, and B. J. Ducro, "Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle," Journal of Dairy Science, vol. 102, no. 10, pp. 9409–9421, Oct. 2019, doi: 10.3168/jds.2019-16295.

[8] A. Ashari and A. M. Tjoa, "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool," 2013. [Online]. Available: www.ijacsa.thesai.org

[9] S. Rahmadani, A. Dongoran, M. Zarlis, and Zakarias, "Comparison of Naive Bayes and Decision Tree on Feature Selection Using Genetic Algorithm for Classification Problem," in Journal of Physics: Conference Series, Mar. 2018, vol. 978, no. 1. doi: 10.1088/1742-6596/978/1/012087.

[10] A. Setyanto dan Hanif Al Fattah, "Analisis Perbandingan Algoritma Decision Tree (C4.5) Dan K-Naive Bayes Untuk Mengklasifikasi Penerimaan Mahasiswa Baru Tingkat Universitas," 2017.

[11] K. Yadav and R. Thareja, "Comparing the Performance of Naive Bayes And Decision Tree Classification Using R," International Journal of Intelligent Systems and Applications, vol. 11, no. 12, pp. 11–19, Dec. 2019, doi: 10.5815/ijisa.2019.12.02.

[12] S. Farhana, "Classification of Academic Performance for University Research Evaluation by Implementing Modified Naive Bayes Algorithm," in Procedia Computer Science, 2021, vol. 194, pp. 224–228. doi: 10.1016/j.procs.2021.10.077.

[13] R. Puspita and A. Widodo, "Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS," Jurnal Informatika Universitas Pamulang, vol. 5, no. 4, p. 646, Dec. 2021, doi: 10.32493/informatika.v5i4.7622.

[14] T. Rosandy, "Perbandingan Metode Naive Bayes Classifier Dengan Metode Decision Tree (C4.5) Untuk Menganalisa Kelancaran Pembiayaan (Study Kasus : Kspps / Bmt Al-Fadhila)," vol. 02, 2016.

[15] C. C. Aggarwal, Data Mining. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-14142-8.

[16] R. Adrian, M. A. J. S. Perdana, A. Asroni, and S. Riyadi, "Applying the Naive Bayes Algorithm to Predict the Student Final Grade," Emerging Information Science and Technology, vol. 1, no. 2, 2020, doi: 10.18196/eist.127.

[17] Y. Jung, "Multiple predicting K-fold cross-validation for model selection," Journal of Nonparametric Statistics, vol. 30, no. 1, pp. 197–215, Jan. 2018, doi: 10.1080/10485252.2017.1404598.

[18] A. S. Nejad and R. Tavoli, "A Method for Estimating the Cost of Software Using Principle Components Analysis and Data Mining," Journal of Electrical and Computer Engineering Innovations, vol. 6, no. 1, pp. 33–42, 2018, doi: 10.22061/JECEI.2018.811.

[19] S. Riyadi, Y. Lestari, C. Damarjati, and K. H. Ghazali, "Performance Comparison of Deep Learning Models to Detect Covid-19 Based on X-Ray Images," 2022.