http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 10, No.2, Juni 2022 ISSN : 2302-9706

DATA MINING USING A SUPPORT VECTOR MACHINE, DECISION TREE, LOGISTIC REGRESSION AND RANDOM FOREST FOR PNEUMONIA PREDICTION AND CLASSIFICATION

Bahtiar Imran[1], Zaeniah[2,*], Sriasih[3], Surni Erniwati[4], Salman[5]

[1,2,3,4,5] Universitas Teknologi Mataram, Mataram and 83115, Indonesia

[1]bahtiarimranlombok@gmail.com; [2] lp2mutm@gmail.com*; [3]sriasihlbk@gmail.com;
[4]mentari1990@gmail.com; [5]asal.lombok@gmail.com
* corresponding author

Article Info

Received : 10 May 2022
Revised : 30 May 2022
Accepted : 07 June 2022

This study uses Data Mining with four classification models. The object of this research is pneumonia data. The proposed models are Support Vector Machine (SVM), Decision Tree, Logistic Regression and Random Forest. Tests have been carried out using Cross-Validation Sampling and Stratified Sampling using several Folds of 3, 10 and 20. The results obtained are Logistic Regression models get the highest and most consistent accuracy results compared to SVM, Decision Tree and Random Forest. The tests evidence this carried out with the results of Number of Folds 3 getting the AUC value of 0.990, Accuracy 0.962, F1 0.962, Precision 0.962 and Recall 0.962. Number of Folds 10 gets the AUC value of 0.991, Accuracy 0.961, F1 0.961, Precision 0.961 and Recall 0.961. Number of Folds 20 gets AUC 0.991, Accuracy 0.965, F1 0.965, Precision 0.965 and Recall 0.965. From this study, Logistic Regression got good results for predicting and classifying pneumonia.

Keywords: prediction, data mining, classification pneumonia.

1. Introduction

Many lung disorders affect people worldwide, including chronic obstructive pulmonary disease (COPD), asthma, tuberculosis, fibrosis, and pneumonia [1]. Lung cancer is one of the leading causes of death in both women and men [2], and pneumonia accounts for most lung infection-related deaths after kidney transplantation [3]. WHO estimates that more than 4 million premature deaths occur yearly from diseases related to air pollution generated from household waste, such as pneumonia [4]. Ultrasound images, X-rays etc., are often used to diagnose conditions such as pneumonia. Image is the first procedure or primary process used to detect disorders in a person's body. Images provide a better view with the addition of a doctor's assistance to diagnose internal diseases [5].

Many studies have been carried out using x-ray results from a patient with pneumonia and with various methods, including using Machine Learning [1], [3], [5]–[8], Data Mining [2], [9], Deep Learning [4], [10], Image Analitics[11], Big Data [7], Support Vector Machine [12], Convolutions and Dynamic Capsule Routing [13], Convolutional Neural Network [14] and others. From previous research, researchers got different results, as in the study of Bahtiar Imran et al. [1] using a machine learning model for pneumonia classification by utilizing the results from X-Ray images. This study, Epoch 700, got the best results with 92% accuracy. Research V.Krishnaiah et a [2] used the Data Mining technique for lung cancer prediction. This study showed that using Naïve Bayes gets better results than other methods. You Luo et al. [3] used machine learning to predict severe pneumonia in hospitalized patients after kidney transplantation. The results of this study, the Random Forest method, got better results with a sensitivity of 0.67, specificity of 0.97, a positive likelihood ratio of 22.33, a negative likelihood ratio of 0.34, AUROC 0.91 AUPRC 0.72. Okeke Stephen et al. [4] used Deep Learning to classify pneumonia in health care. The results of this study that using the CNN model got better results. Aiyesha Sadiya et al. [5] used Machine Learning to diagnose tuberculosis and pneumonia. Of the proposed methods, such as Naïve Bayes, Decision Tree and Random Forest, the Random Forest method gave the best results. Hafiz Muneeb Ahmad et al. [11] this study uses Image Analytics from Orange tools for pneumonia prediction. The results obtained are that Logistic

Regression gives better results. K.R. Swetha et al. [7] proposes Big Data, Deep Learning and Machine Learning for pneumonia prediction. The results obtained in this study are that the Convolutional Neural Network (CNN) method provides very accurate results. Elina Naydenova et al. [9] this study uses Data Mining to diagnose pneumonia in children. This study shows that the proposed method can be used to support the diagnosis of pneumonia in children. Kuang Ming Kuo et al. [8] used machine learning to predict pneumonia in schizophrenic patients. The results obtained are that the proposed prediction model can serve as a helpful support tool for doctors in treating schizophrenic patients. Ansh Mittal et al. [13] used Convolutions and Dynamic Capsule Routing for pneumonia detection based on x-ray images. The results of this study were the E4CC model worked optimally and provided a test accuracy of 96.36%.

From several studies that have been carried out, there have never been a study using Data Mining with the Support Vector Machine (SVM), Decision Tree, Logistic Regression and Random Forest models to predict and classify pneumonia. For this reason, this study proposes a Support Vector Machine (SVM), Decision Tree, Logistic Regression and Random Forest model to predict and classify pneumonia. The determination of this model has based on the reason that the SVM model is a classification model that gives good results with high accuracy [15], [16], and the Decision Tree is a good and accurate classification model [17], [18]. Logistic regression gives classification results with high accuracy [11], while Random Forest is a model that offers good results with high accuracy [19]–[22]. Before the prediction and classification stages are carried out, the data used as training and testing data is pre-processed using the Normalize to Interval method. The tools used for this prediction and classification are Orange Data Mining, for the prediction and classification phase using the prediction widget and Test and Score while evaluating the success of the classification results using the Confusion Matrix.

2. Method

The stages of the research process in this study are described in the flow chart of the research methodology, which is depicted in Figure 1.

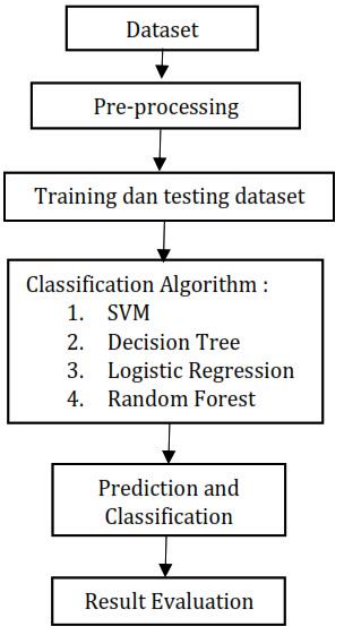


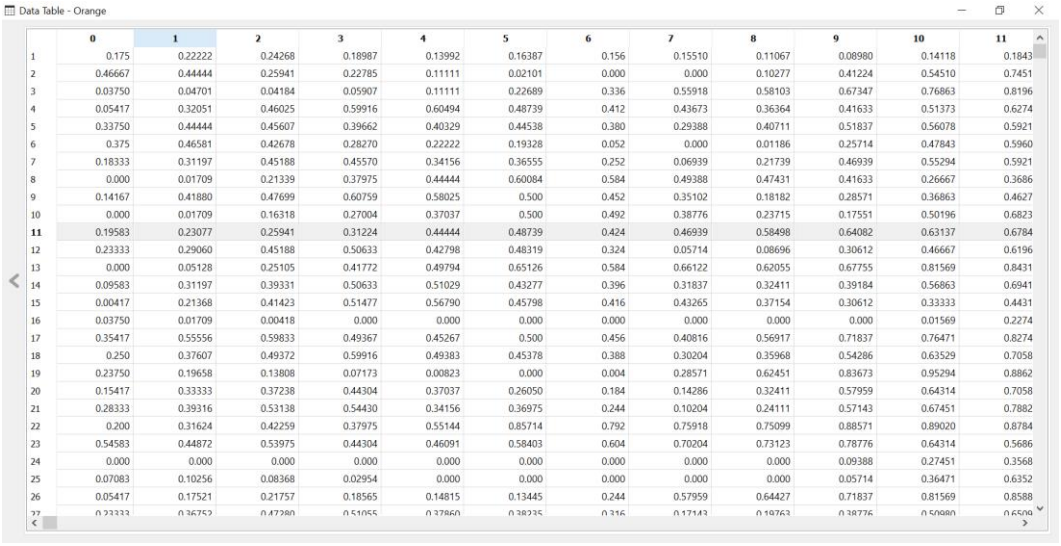
Figure 1. Research Stages

2.1. Dataset

The data used in this study was taken on a dataset sharing website, namely Kaggle. The dataset used in this study was 5216 data in an excel file. The attributes used are 783 attributes, while for the tools used to process datasets using Orange Data Mining, orange is one of the most effective tools used for data mining [23]–[26].

2.2. Pre-processing

After the data collection stage is carried out, there are still a lot of missing, null and redundant data [27]. It is necessary to do pre-processing to eliminate Null and Redundant data. The method used in this pre-processing is Normalize Features by utilizing the Normalize to Interval [1,1] feature. **F**igure 2 is an example of data that has been pre-processed.


Data Table - Orange

	0	1	2	3	4	5	6	7	8	9	10	11
1	0.175	0.22222	0.24268	0.18987	0.13992	0.16387	0.156	0.15510	0.11067	0.08980	0.14118	0.1843
2	0.46667	0.44444	0.25941	0.22785	0.11111	0.02101	0.000	0.000	0.10277	0.41224	0.54510	0.7451
3	0.03750	0.04701	0.04184	0.05907	0.11111	0.22689	0.336	0.55918	0.58103	0.67347	0.76863	0.8196
4	0.05417	0.32051	0.46025	0.59916	0.60494	0.48739	0.412	0.43673	0.36364	0.41633	0.51373	0.6274
5	0.33750	0.44444	0.45607	0.39662	0.40329	0.44538	0.380	0.29388	0.40711	0.51837	0.56078	0.5921
6	0.375	0.46581	0.42678	0.28270	0.22222	0.19328	0.052	0.000	0.01186	0.25714	0.47843	0.5960
7	0.18333	0.31197	0.45188	0.45570	0.34156	0.36555	0.252	0.06939	0.21739	0.46939	0.55294	0.5921
8	0.000	0.01709	0.21339	0.37975	0.44444	0.60084	0.584	0.49388	0.47431	0.41633	0.26667	0.3686
9	0.14167	0.41880	0.47699	0.60759	0.58025	0.500	0.452	0.35102	0.18182	0.28571	0.36863	0.4627
10	0.000	0.01709	0.16318	0.27004	0.37037	0.500	0.492	0.38776	0.23715	0.17551	0.50196	0.6823
11	0.19583	0.23077	0.25941	0.31224	0.44444	0.48739	0.424	0.46939	0.58498	0.64082	0.63137	0.6784
12	0.23333	0.29060	0.45188	0.50633	0.42798	0.48319	0.324	0.05714	0.08696	0.30612	0.46667	0.6196
13	0.000	0.05128	0.25105	0.41772	0.49794	0.65126	0.584	0.66122	0.62055	0.67755	0.81569	0.8431
14	0.09583	0.31197	0.39331	0.50633	0.51029	0.43277	0.396	0.31837	0.32411	0.39184	0.56863	0.6941
15	0.00417	0.21368	0.41423	0.51477	0.56790	0.45798	0.416	0.43265	0.37154	0.30612	0.33333	0.4431
16	0.03750	0.01709	0.00418	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.01569	0.2274
17	0.35417	0.55556	0.59833	0.49367	0.45267	0.500	0.456	0.40816	0.56917	0.71837	0.76471	0.8274
18	0.250	0.37607	0.49372	0.59916	0.49383	0.45378	0.388	0.30204	0.35968	0.54286	0.63529	0.7058
19	0.23750	0.19658	0.13808	0.07173	0.00823	0.000	0.004	0.28571	0.62451	0.83673	0.95294	0.8862
20	0.15417	0.33333	0.37238	0.44304	0.37037	0.26050	0.184	0.14286	0.32411	0.57959	0.64314	0.7058
21	0.28333	0.39316	0.53138	0.54430	0.34156	0.36975	0.244	0.10204	0.24111	0.57143	0.67451	0.7882
22	0.200	0.31624	0.42259	0.37975	0.55144	0.85714	0.792	0.75918	0.75099	0.88571	0.89020	0.8784
23	0.54583	0.44872	0.53975	0.44304	0.46091	0.58403	0.604	0.70204	0.73123	0.78776	0.64314	0.5686
24	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.09388	0.27451	0.3568
25	0.07083	0.10256	0.08368	0.02954	0.000	0.000	0.000	0.000	0.000	0.05714	0.36471	0.6352
26	0.05417	0.17521	0.21757	0.18565	0.14815	0.13445	0.244	0.57959	0.64427	0.71837	0.81569	0.8588
27	0.23333	0.36752	0.47280	0.51055	0.37860	0.38235	0.316	0.17143	0.19763	0.38776	0.50980	0.6509

Figure 2. Data After Pre-Processing

2.3. Training and Testing Dataset

The widget used for training and testing is the Data Sampler at this stage. This data sampler divides the data used for training and testing. The training data used in this study were 80% = 4174 data and 20% = 1043 data as testing data. In the data set, one attribute is used as a target, in this case, the Prediction attribute. Prediction attribute has two classes, including Pneumonia class and Normal class.

2.4. Prediction and Classification

The classification model used in this study uses four models: Support Vector Machine (SVM), Decision Tree, Logistic Regression, and Random Forest.

The steps are as follows.

1. Import dataset
2. Pre-processing using Normalize Features
3. Select Columns = Enter the attributes used and determine the features that will be used as targets.
4. Data Sampler = distribution of training and testing data
5. Test and Score = receive testing and training data from the data sampler. Test and Score also receive learning from classification models, including SVM, Decision Tree, Logistics Regression and Random Forest.

2.5. Evaluation Result

Test and Score has an output in the form of predictions and evaluation results. This widget is used to measure the level of results from the classification evaluation using the Confusion Matrix. The confusion matrix in this study was used to evaluate and calculate the performance of the classification model [28]—table 1 example Confusion Matrix.

Table 1. Confusion matrix

	PREDICTED

Actual		Normal	Pneumonia
	Normal	True Positives (TP)	False Negatives (FN)
	Pneumonia	False Positives (FP)	True Negatives (TN)

Accuracy is measured using [29]:

$$\textbf{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

Precision is measured using [22]:

$$\textbf{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall is measured using [22]:

$$\textbf{Recall} = \frac{TP}{TP+FN} \quad (3)$$

3. Results and Disccusion

The test results used a data set of 5216, and the number of attributes is 783. The distribution of training and testing data uses a data sampler widget, where 80% training data = 4174 data and 20% = 1043 data as testing data. In this test, the target attribute is a prediction with two classes: the Pneumonia class and the Normal class. Sampling used in this study uses Cross-Validation and Stratified Sampling using a Number of Folds of 3, 10 and 20.

3.1. Number of Folds 3

By using Number of Folds 3, the test results obtained are, using Decision Tree AUC 0.812, Accuracy 0.892, F1 0.891, Precision 0.891 and Recall 0.892, SVM model with AUC value 0.986, Accuracy 0.953, F1 0.952, Precision 0.952 and Recall 0.953. Random Forest with AUC 0.973, Accuracy 0.935, F1 0.934, Precision 0.934 and Recall 0.935. As for Logistic Regression with AUC value 0.990, Accuracy 0.962, F1 0.962, Precision 0.962 and Recall 0.962. Figure 3 test results using Number of Folds 3.

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Tree	0.812	0.892	0.891	0.891	0.892
SVM	0.986	0.953	0.952	0.952	0.953
Random Forest	0.973	0.935	0.934	0.934	0.935
Logistic Regression	0.990	0.962	0.962	0.962	0.962

Figure 3. Number of Folds 3

Evaluation of the results using the Confusion Matrix using Number of Folds 3 on the Logistic Regression model got the classification results for the Normal class = 1003 and the Pneumonia class = 3010, while for the failed classification results for the Normal type = 84 and the Pneumonia class = 76. Figure 4 results Confusion Matrix on Logistic Regression model with Number of Folds 3.

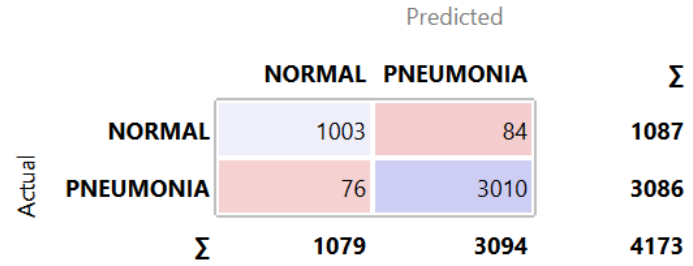


Figure 4. Confusion Matrix with Logistic Regression Model

Evaluation of the results using the Confusion Matrix using Number of Folds 3 in the Decision Tree model got the classification results for the Normal class = 825 and Pneumonia class = 2899, while for the failed classification results for the Normal type = 262 and the Pneumonia class = 187. Figure 5 results Confusion Matrix on the Decision Tree model with Number of Folds 3.

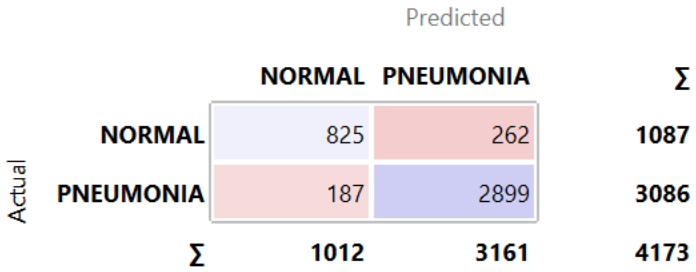


Figure 5. Confusion Matrix with Decision Tree Model

Evaluation of the results using the Confusion Matrix using Number of Folds 3 on the SVM model got the classification results for the Normal class = 965 and the Pneumonia class = 3011, while for the failed classification results for the Normal class = 122 and the Pneumonia class = 75. Figure 6 Confusion results Matrix on SVM model with Number of Folds 3.

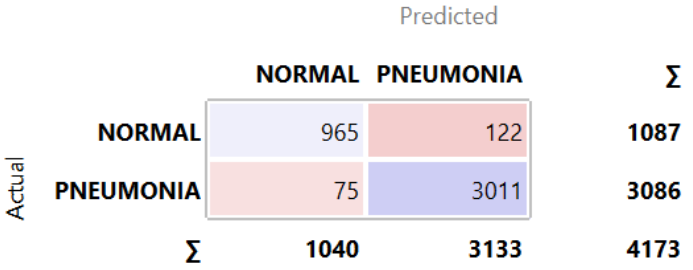


Figure 6. Confusion Matrix With SVM Model

Evaluation of the results using the Confusion Matrix using Number of Folds 3 in the Random Forest model got the classification results for the Normal class = 913 and Pneumonia class = 2987, while for the failed classification results for the Normal type = 174 and the Pneumonia class = 99. Figure 7 results Confusion Matrix on Random Forest model with Number of Folds 3.

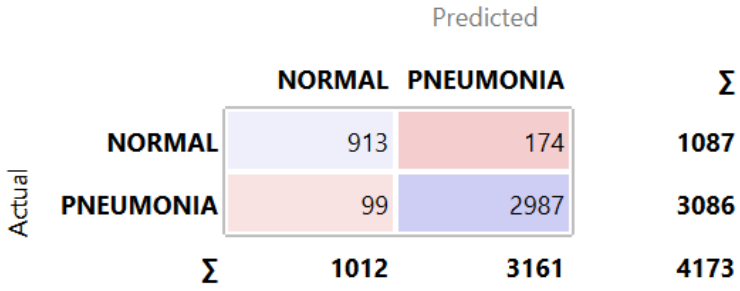
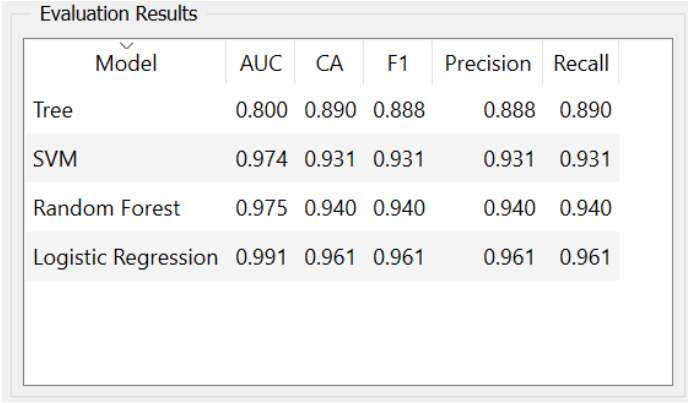


Figure 7. Confusion Matrix with Random Forest Model

3.2. Number of Folds 10

Tests using Number of Folds 10 get the following results, with the Decision Tree model with AUC value 0.800, Accuracy 0.890, F1 0.888, Precision 0.888 and Recall 0.890, SVM model with AUC value 0.974, Accuracy 0.931, F1 0.931 , Precision 0.931 and Recall 0.931. Random Forest Model with AUC value of 0.975, Accuracy 0.940, F1 0.940, Precision 0.940 and Recall 0.940 and Logistic Regression Model with AUC value 0.991, Accuracy 0.961, F1 0.961, Precision 0.961 and Recall 0.961. Figure 8 test results using Number of Folds 10.



Model	AUC	CA	F1	Precision	Recall
Tree	0.800	0.890	0.888	0.888	0.890
SVM	0.974	0.931	0.931	0.931	0.931
Random Forest	0.975	0.940	0.940	0.940	0.940
Logistic Regression	0.991	0.961	0.961	0.961	0.961

Figure 8. Number of Folds 10

Evaluation of the results using the Confusion Matrix using Number of Folds 10 on the Logistic Regression model got the classification results for the Normal class = 1000 and the Pneumonia class = 3010, while for the failed classification results for the Normal class = 87 and the Pneumonia class = 76. Figure 9 results Confusion Matrix on Logistic Regression model with Number of Folds 10.

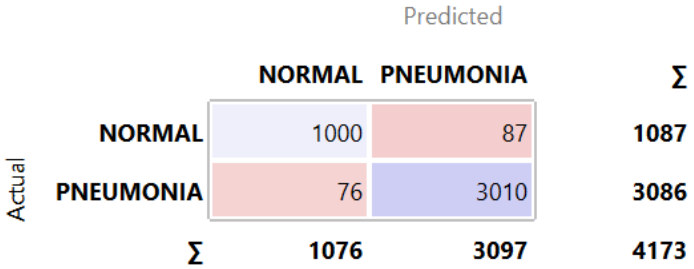


Figure 9. Confusion Matrix with Logistic Regression Model

Evaluation of the results using the Confusion Matrix using Number of Folds 10 in the Decision Tree model got the classification results for the Normal class = 795 and Pneumonia class = 2915, while for the failed classification results for the Normal class = 268 and the Pneumonia class = 195. Figure 10 results Confusion Matrix on Decision Tree model with Number of Folds 10.

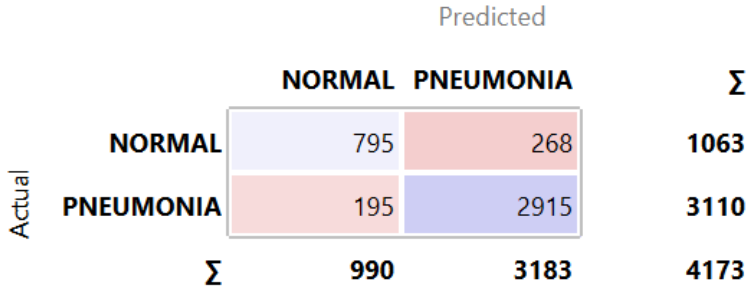


Figure 10. Confusion Matrix with Decision Tree Model

Evaluation of the results using the Confusion Matrix using Number of Folds 10 on the SVM model got the classification results for the Normal class = 924 and Pneumonia class = 2931, while for the failed classification results for the Normal class = 139 and the Pneumonia class = 179. Figure 11 Confusion results Matrix on SVM model with Number of Folds 10.

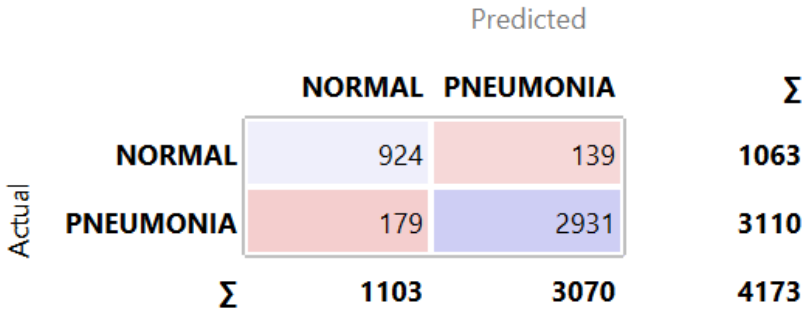


Figure 11. Confusion Matrix With SVM Model

Evaluation of the results using the Confusion Matrix using Number of Folds 10 in the Random Forest model got the classification results for the Normal class = 901 and Pneumonia class = 3003, while for the failed classification results for the Normal class = 162 and the Pneumonia class = 107. Figure 12 results Confusion Matrix on Random Forest model with Number of Folds 10.

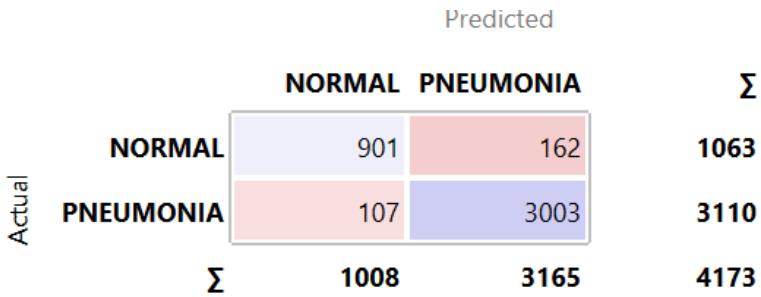


Figure 12. Confusion Matrix With SVM Model

3.3. Number of Folds 20

Tests using Number of Folds 20 get the following results, with the Decision Tree model with AUC values of 0.800, Accuracy 0.889, F1 0.888, Precision 0.887 and Recall 0.889, SVM models with AUC values 0.976, Accuracy 0.935, F1 0.935, Precision 0.935 and Recall 0.935. Random Forest Model with AUC value of 0.978, Accuracy 0.942, F1 0.941, Precision 0.941 and Recall 0.942 and Logistic Regression Model with AUC value 0.991, Accuracy 0.965, F1 0.965, Precision 0.965 and Recall 0.965. Figure 13 test results using Number of Folds 20.

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Tree	0.800	0.889	0.888	0.887	0.889
SVM	0.976	0.935	0.935	0.935	0.935
Random Forest	0.978	0.942	0.941	0.941	0.942
Logistic Regression	0.991	0.965	0.965	0.965	0.965

Figure 13. Number of Folds 20

Evaluation of the results using the Confusion Matrix using Number of Folds 20 on the Logistic Regression model got the classification results for the Normal class = 979 and the Pneumonia class = 3047. For the failed classification results, the Normal class = 84 and for the Pneumonia class = 63. Figure 14 results Confusion Matrix on Logistic Regression model with Number of Folds 20.

Predicted

Actual	NORMAL	PNEUMONIA	Σ
NORMAL	979	84	1063
PNEUMONIA	63	3047	3110
Σ	1042	3131	4173

Figure 14. Confusion Matrix with Logistic Regression Model

Evaluation of the results using the Confusion Matrix using Number of Folds 20 in the Decision Tree model got the classification results for the Normal class = 788 and Pneumonia class = 2923, while for the failed classification results for the Normal class = 275 and the Pneumonia class = 187. Figure 15 results Confusion Matrix on Decision Tree model with Number of Folds 20.

Predicted

Actual	NORMAL	PNEUMONIA	Σ
NORMAL	788	275	1063
PNEUMONIA	187	2923	3110
Σ	975	3198	4173

Figure 15. Confusion Matrix with Decision Tree Model

Evaluation of the results using the Confusion Matrix using Number of Folds 20 on the SVM model got the classification results for the Normal class = 937 and Pneumonia class = 2964, while for the failed classification results for the Normal class = 126 and for the Pneumonia class = 146. Figure 16 Confusion results Matrix on SVM model with Number of Folds 20.

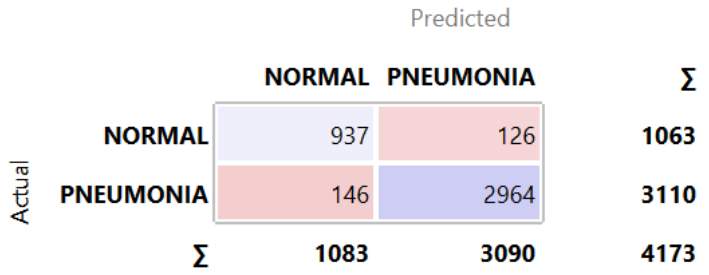


Figure 16. Confusion Matrix With SVM Model

Evaluation of the results using the Confusion Matrix using Number of Folds 20 in the Random Forest model got the classification results for the Normal class = 916 and Pneumonia class = 3014, while for the failed classification results for the Normal class = 147 and the Pneumonia class = 96. Figure 17 results Confusion Matrix on Random Forest model with Number of Folds 20.

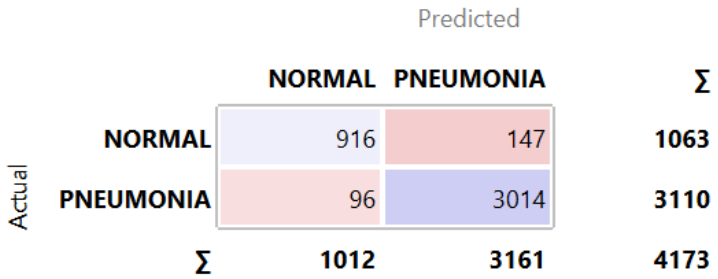


Figure 17. Confusion Matrix with Random Forest Model

From the results of tests that have been carried out using four classification models with 783 attributes, predictions and classifications are good, but for further development, model predictions and classifications can be developed using different numbers of folds to get maximum results.

4. Conclusion

This study uses Data Mining with four classification models, namely SVM, Decision Tree, Logistic Regression and Random Forest, to predict and classify pneumonia. From the results of the tests that have been carried out using Test and Score, the Logistic Regression model results get the highest and most consistent accuracy results compared to SVM, Decision Tree, and Random Forest. The test results can be proven by using the Number of Folds that have been carried out, Number of Folds 3 get results for Decision Tree AUC 0.812, Accuracy 0.892, F1 0.891, Precision 0.891 and Recall 0.892, SVM model with AUC value 0.986, Accuracy 0.953, F1 0.952, Precision 0.952 and Recall 0.953. Random Forest AUC value is 0.973, Accuracy is 0.935, F1 is 0.934, Precision is 0.934 and Recall is 0.935. Logistic Regression AUC value is 0.990, Accuracy is 0.962, F1 is 0.962, Precision is 0.962 and Recall is 0.962. Number of Folds 10 got the following results, with the Decision Tree model with AUC value of 0.800, Accuracy 0.890, F1 0.888, Precision 0.888 and Recall 0.890, SVM model AUC value 0.974, Accuracy 0.931, F1 0.931, Precision 0.931 and Recall 0.931. Random Forest AUC value 0.975, Accuracy 0.940, F1 0.940, Precision 0.940 and Recall 0.940 and Logistic Regression AUC value 0.991, Accuracy 0.961, F1 0.961, Precision 0.961 and Recall 0.961. Number of Folds 20 gets the following results, with the Decision Tree model with AUC values of 0.800, Accuracy 0.889, F1 0.888, Precision 0.887 and Recall 0.889, SVM models with AUC values 0.976, Accuracy 0.935, F1 0.935, Precision 0.935 and Recall 0.935. Random Forest Model with AUC value of 0.978, Accuracy 0.942, F1 0.941, Precision 0.941 and Recall 0.942 and Logistic Regression Model with AUC value 0.991, Accuracy 0.965, F1 0.965, Precision 0.965 and Recall 0.965.

References

[1] B. Imran and L. D. Bakti, “Implementation of Machine Learning Model for Pneumonia Classification Based on X-Ray Images,” *J. Mantik*, vol. 5, no. 3, pp. 2101–2107, 2021.

[2] V. Krishnaiah, D. Narsimha, and D. Chandra, “Diagnosis of lung cancer prediction system

using data mining classification techniques," *Int. J. Comput. Sci. Inf. Technol.*, vol. 4, no. 1, pp. 39–45, 2013.

[3] Y. Luo *et al.*, "Machine learning for the prediction of severe pneumonia during posttransplant hospitalization in recipients of a deceased-donor kidney transplant," *Ann. Transl. Med.*, vol. 8, no. 4, pp. 82–82, 2020, doi: 10.21037/atm.2020.01.09.

[4] O. Stephen, M. Sain, U. J. Maduh, and D. U. Jeong, "An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare," *J. Healthc. Eng.*, vol. 2019, 2019, doi: 10.1155/2019/4180949.

[5] A. Sadiya, A. V. Illur, A. Nanda, E. Rao, K. P. Vidyashree, and M. Ahmed, "Differential diagnosis of tuberculosis and pneumonia using machine learning," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 6 Special Issue 4, pp. 245–250, 2019, doi: 10.35940/ijitee.F1049.0486S419.

[6] Y. Erdaw and E. Tachbele, "Machine learning model, applied on chest X-ray images enables automatic detection of COVID-19 cases with high accuracy," *Int. J. Gen. Med.*, vol. 14, pp. 4923–4931, 2021, doi: 10.2147/IJGM.S325609.

[7] K. R. Swetha, M. Niranjanamurthy, M. P. Amulya, and M. Y. Manu, "Prediction of Pneumonia Using Big Data, Deep Learning and Machine Learning Techniques," *Proc. 6th Int. Conf. Commun. Electron. Syst. ICCES 2021*, no. August, pp. 1697–1700, 2021, doi: 10.1109/ICCES51350.2021.9489188.

[8] K. M. Kuo, P. C. Talley, C. H. Huang, and L. C. Cheng, "Predicting hospital-acquired pneumonia among schizophrenic patients: A machine learning approach," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–8, 2019, doi: 10.1186/s12911-019-0792-1.

[9] E. Naydenova, A. Tsanas, S. Howie, C. Casals-Pascual, and M. De Vos, "The power of data mining in diagnosis of childhood pneumonia," *J. R. Soc. Interface*, vol. 13, no. 120, 2016, doi: 10.1098/rsif.2016.0266.

[10] M. Zandehshahvar, M. van Assen, H. Maleki, Y. Kiarashi, C. N. De Cecco, and A. Adibi, "Toward understanding COVID-19 pneumonia: a deep-learning-based approach for severity analysis and monitoring the disease," *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, 2021, doi: 10.1038/s41598-021-90411-3.

[11] H. Muneeb Ahmad, M. Sohail, M. Muneeb Ahmad, S. Iqbal, A. Sarfaraz, and K. Noor, "Predictions of Pneumonia Disease using Image Analytics in Orange Tool," *GS Int. Conf. Comput. Sci. Eng. 2020 (GSICCSE 2020*, no. August 2020, 2020.

[12] S. Guhathakurata, S. Kundu, A. Chakraborty, and J. S. Banerjee, "A novel approach to predict COVID-19 using support vector machine," *Data Sci. COVID-19*, pp. 351–364, 2021, doi: 10.1016/b978-0-12-824536-1.00014-9.

[13] A. Mittal *et al.*, "Detecting pneumonia using convolutions and dynamic capsule routing for chest X-ray images," *Sensors (Switzerland)*, vol. 20, no. 4, pp. 1–30, 2020, doi: 10.3390/s20041068.

[14] Y. S. Taspinar, I. Cinar, and M. Koklu, "Classification by a stacking model using CNN features for COVID-19 infection diagnosis," *J. Xray. Sci. Technol.*, vol. 30, no. 1, pp. 73–88, 2021, doi: 10.3233/xst-211031.

[15] F. R. Lumbanraja, E. Fitri, Ardiansyah, A. Junaidi, and R. Prabowo, "Abstract Classification Using Support Vector Machine Algorithm (Case Study: Abstract in a Computer Science Journal)," *J. Phys. Conf. Ser.*, vol. 1751, no. 1, pp. 1–12, 2021, doi: 10.1088/1742-6596/1751/1/012042.

[16] A. Abubakar *et al.*, "A support vector machine classification of computational capabilities of 3D map on mobile device for navigation aid," *Int. J. Interact. Mob. Technol.*, vol. 10, no. 3, pp. 4–10, 2016, doi: 10.3991/ijim.v10i3.5056.

[17] H. H. Patel and P. Prajapati, "Study and Analysis of Decision Tree-Based Classification Algorithms," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 56–61, 2018.

[18] B. T. Jijo and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[19] M. A. Abd-Elrazek, A. A. Othman, M. H. Abd Elaziz, and M. N. Abd-Elwhab, "Intelligent

Prediction of Breast Cancer: A Comparative Study," *Egypt. Comput. Sci. J.*, vol. 42, no. 3, pp. 29–43, 2018, [Online]. Available: http://ecsjournal.org/Archive/Volume42/Issue3/3.pdf.
[20] A. More, S. Mhatre, V. Kamble, V. Patil, and S. Bhairnallykar, "Breast Cancer Prediction Using Classification Techniques of Machine Learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 1, 2022, doi: 10.26483/ijarcs.v10i5.6464.
[21] R. Rawal, "BREAST CANCER PREDICTION USING MACHINE LEARNING," *J. Emerg. Technol. Innov. Res.*, vol. 7, no. 5, pp. 13–24, 2020, doi: 10.2478/acss-2020-0018.
[22] S. M. Ayyoubzadeh, A. Almasizand, and ..., "Early Breast Cancer Prediction Using Dermatoglyphics: Data Mining Pilot Study in a General Hospital in Iran," *Heal. Educ. Heal. Promot.*, vol. 9, no. 3, pp. 279–285, 2021, [Online]. Available: https://biot.modares.ac.ir/article-5-53673-en.html.
[23] E. R. Kaur and V. Chopra, "Implementing Adaboost and Enhanced Adaboost Algorithm in Web Mining," *Int. J. Advanced Res. Comput. Commun. Eng.*, vol. 4, no. 7, pp. 306–311, 2015, doi: 10.17148/IJARCCE.2015.4771.
[24] M. K. KELEŞ, "Breast cancer prediction and detection using data mining classification algorithms: A comparative study," *The. Vjesn.*, vol. 26, no. 1, pp. 149–155, 2019, doi: 10.17559/TV-20180417102943.
[25] V. Y. Kulkarni and P. K. Sinha, "Effective Learning and Classification using Random Forest Algorithm," *Int. J. Eng. Innov. Technology*, vol. 3, no. 11, pp. 267–273, 2014.
[26] A. Sinha, B. Sahoo, S. S. Rautaray, and M. Pandey, "Analysis of Breast Cancer Dataset Using Big Data Algorithms for Accuracy of Diseases Prediction," *Lect. Notes Data Eng. Commun. Technol.*, vol. 44, pp. 271–277, 2020, doi: 10.1007/978-3-030-37051-0_31.
[27] M. H. Krishna and D. K. N. Rao, "PREDICTION OF BREAST CANCER USING MACHINE LEARNING TECHNIQUES," *Int. J. Manag. Technol. Eng.*, vol. 8, no. 12, pp. 150–153, 2018, doi: 10.2174/2213275912666190617160834.
[28] K. Swetha and R. Ranjana, "Breast Cancer Prediction Using Machine Learning and Data Mining," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 6, no. 3, pp. 610–615, 2020, [Online]. Available: https://www.academia.edu/download/65272215/CSEIT206219.pdf.
[29] V. Chaurasia, S. Pal, and B. B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *J. Algorithms Comput. Technol.*, vol. 12, no. 2, pp. 119–126, 2018, doi: 10.1177/1748301818756225.