Jurnal CoreIT, Vol.11, No.2, December 2025
ISSN 2460-738X (Print)
ISSN 2599-3321 (Online)

Benchmarking Various Machine Learning Models to Detect Lung Cancer

Iis Afrianty [1], Liza Afriyanti [1*]
[1]Dept. of Informatics Engineering, Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
iis.afrianty@uin-suska.ac.id, liza.afriyanti@uin-suska.ac.id*

Abstract. Lung cancer is one of the highest causes of death in the world, including in Indonesia, which is largely caused by delayed early detection. Conventional methods such as CT-scan, thoracic surgery, and histopathology have limitations in terms of cost, time, and accessibility. Therefore, machine learning (ML)-based approaches are a promising alternative to support early detection of lung cancer quickly and at a low cost. This study aims to benchmark the performance of various machine learning algorithms in detecting lung cancer using public datasets. A literature review was conducted on Scopus and Web of Science indexed articles over the past five years to identify trends and research gaps related to the selection of ML algorithms. The dataset used consisted of 310 data with 16 clinical symptom attributes and two classes, namely cancer and non-cancer, which had an unbalanced distribution. A total of nine ML algorithms were tested, including Random Forest, Support Vector Machine, Logistic Regression, Multi-Layer Perceptron, C4.5, Bayesian Network, RepTree, Naïve Bayes, and P.A.R.T, with a cross-validation scheme. Performance evaluation was conducted using the Accuracy, Precision, F-Measure, True Positive Rate, ROC, False Positive Rate, Precision-Recall Curve, and Matthews Correlation Coefficient metrics. The results of the experiment showed that the Support Vector Machine performed best in balanced data distribution, while Random Forest showed more stable performance in unbalanced data conditions. This analysis confirms that algorithm selection and data distribution greatly affect the quality of lung cancer detection, and emphasizes the importance of fair and standardized benchmarking in the development of machine learning-based detection systems.

Keywords: Benchmarking, Lung Cancer, Machine Learning

Received November 2025 / **Revised** December 2025 / **Accepted** December 2025

INTRODUCTION

Cancer is the deadliest disease in the world with a prevalence rate of 11.4 percent. Indonesia noted that since 2020, the cause of death due to lung cancer has increased by 18%. This is because the patient is late to check with the doctor, so when examined by the doctor, it turns out that the patient has been justified in having advanced stage cancer.

In general, cancer detection techniques in Indonesia such as *CT scan* [1], thoracic surgery [2], and histopathology [3]. Cancer detection techniques require expensive and expensive time, so a fast and low-cost alternative detection technique is needed. Several studies have used machine learning techniques as an alternative detection technique for lung cancer. This machine learning technique is very helpful in detecting the characteristics of lung cancer [4].

Many researchers are trying to find the best techniques for detecting lung cancer, such as [5], [6], [7], [8] and [9]. [5] using Support Vector Machines (SVM), Decision trees (DT), and Artificial Neural Networks techniques in the TCGA dataset. [6] using the SEER dataset to use Random Forests (RF), General Linear Regression (GL), RF, Gradient Boosted Machines (GBM), and Ensemble learning modifications. [7] compared machine learning techniques such as SVM, K-Nearest Neighbour, Decision Tree, Random Forest, Gradient Boosting Decision tree on private datasets as well as those done by [8] who used EHR datasets to be classified by random forest. It is different from [9] that uses public datasets, such as UCI Machine learning with the Lung Cancer dataset against the Naïve Bayes (NB) technique, C4.5 Decision Tree, and Support Vector Machine (SVM).

Benchmarking is one of the techniques used to get the best of the best techniques [10], especially for lung cancer cases. However, the various *machine learning* techniques proposed by various researchers to detect lung cancer are not all comparable [11], this is one of them due to differences in datasets to access to

DOI: 10.24014/coreit.v11i2.38590

datasets [12] (*public* or *private*). Benchmarking is not a technique by taking other research results and combining them in a tabulation. Benchmark is more about the research process of others can be compared to how to follow the same process done by them, such as using datasets, parameters, and configurations. According to [13], benchmarks should at least have aspects of relevance, reproducibility, fairness, verifiability, and usability.

In practice, the selection of ML algorithms for a given scientific problem is more complex than simply choosing one of the specific machine learning technologies and algorithms [14]. The selection of the most effective ML algorithm is based on many factors, including the type, amount, and quality of the training data, the availability of labeled data, the type of problem being addressed (prediction, classification, and so on), the accuracy and overall performance required, and the hardware systems available for training and inference [15]. Therefore, it is necessary to select the best technique by comparing various machine learning techniques in a standard dataset and performance evaluation.

We conducted a performance evaluation of various machine learning techniques using a public lung cancer dataset. This dataset contains a variety of symptoms that can be used to predict lung cancer. We compared nine machine learning techniques namely Random Forest, Support Vector Machine, logistic regression, Multi-Layer Perceptron, C4.5, Bayesian Network, Reptree, Naive Bayes and P.A.R.T using lung cancer datasets. Each machine learning technique was evaluated using *Accuracy, F-Measure, Precision, True Positive Rate, ROC, False Positive Rate, Precision Recall Curve, and Matthew's correlation coefficient*. Based on the performance tests carried out, the best machine learning techniques to detect lung cancer have been successfully identified.

The purpose of this study is to determine the performance of the best machine learning techniques on public datasets to detect lung cancer.

We evaluated the performance of various machine learning techniques of public datasets (lung cancer). Very few researchers have done the same comparison of machine learning techniques as we did. With multi-dimensional issues consisting of ML algorithm choices, hardware architectures, and various scientific issues, choosing the optimal ML algorithm for a given task is not trivial. This has become a significant barrier for many scientists who want to use modern ML methods in their scientific research [14].

Many researchers are trying to find the best techniques for detecting lung cancer, such as [5], [6], [7], [8] and [9]. [5] using Artificial Neural Networks, Support Vector Machines (SVM), Decision trees (DT) techniques in the TCGA dataset. [6] using the SEER dataset to use Random Forests (RF), General Linear Regression (GL), RF, Gradient Boosted Machines (GBM), and Ensemble learning modifications. [7] compared machine learning techniques such as SVM, K-Nearest Neighbour, Decision Tree, Random Forest, and Gradient Boosting Decision tree on private datasets. The same is true for those who use EHR datasets to be classified by Random Forest, Logistic Regression, and XGBoost. It is different from [8] that uses public datasets, such as UCI Machine learning with the Lung Cancer dataset against Naïve Bayes (NB), C4.5, Decision Tree, and Support Vector Machine (SVM) techniques.

Table 1. Related research

Researchers	Method	Dataset	Info Paramater*	Performance Evaluation
(Yang et al., 2022)	Artificial Neural Networks, Support Vector Machines (SVM), Decision trees (DT)	Audience	Partial	AUC dan ROC
(Bartholomai & Frieboes, 2018)	learning Random Forests (RF), General Linear Regression (GL), RF, Gradient Boosted Machines (GBM), dan modifikasi Ensemble learning	Publik	Full	RMSE dan Accuracy
(Tang et al., 2018)	SVM, K-Nearest Neighbour, Decision	Private	Partial	Accuracy, Recall, dan F1 Measure

Researchers	Method	Dataset	Info Paramater*	Performance Evaluation
	Tree, Random Forest, dan Gradien Boosting Decision tree			
(Faisal et al., 2018)	Naïve Bayes (NB), C4.5, Decision Tree, and Support Vector Machine (SVM)	Publik	Partial	Accuracy, Precision, Recall, dan F1 Measure
(Alsinglawi et al., 2022)	Random Forest, Logistic Regression, dan XGBoost	Private	partial	Accuracy
This research	Random Forest, Support Vector Machine, logistic regression, Multi-Layer Perceptron, C4.5, Bayesian Network, Reptree, Naive bayes dan P.A.R.T	Publik	Full	Accuracy, F-Measure, Precision, True Positive Rate, ROC, False Positive Rate, Precision Recall Curve, dan Matthew's correlation coefficient

METHODS

A. Literature review

We reviewed articles indexed by Scopus and WoS using the criteria, namely Machine learning and Lung Cancer. Quality research based on quality references, which is why we chose articles that are indexed by Scopus and WoS. We take the last 5 years of articles, to ensure that the research we conduct is always up-to-date and provides a high contribution value, especially in detecting lung cancer.

B. Dataset

We used the most widely used data sources by researchers to find out which techniques are best used to detect lung cancer. We used datasets that were used to detect lung cancer. The dataset consisted of 16 fields consisting of *smoking, yellow_fingers, anxiety, peer_pressure, chronic disease, fatigue, allergy, wheezing, alcohol consuming, coughing, shortness of breath, swallowing difficulty, and chest pain*. This dataset consists of two classes, namely lung cancer and non-lung cancer. This dataset consists of 310 rows of data.

C. Experiment Setting

The techniques used for lung cancer detection are Nine machine learning techniques, namely Random Forest, Support Vector Machine, logistic regression, Multi-Layer Perceptron, C4.5, Bayesian Network, Reptree, Naive Bayes and P.A.R.T using lung cancer datasets. Also, we made sure the model formed by the machine learning technique was better, so we added a cross-validation technique.

D. Performance Evaluation

Each machine learning technique was evaluated using Accuracy, F-Measure, Precision, True Positive Rate, ROC, False Positive Rate, Precision Recall Curve, and Matthews correlation coefficient. The evaluation is very widely used by researchers in the field of lung cancer detection.

E. Result and Discussion

In this stage, we discuss the performance evaluation of machine learning techniques in lung cancer datasets. Evaluation is discussed based on the parameters, datasets and machine learning techniques used. So that it can be known that the performance of the proposed technique will be maximum in what parameters and evaluations it will be. As well as the identification of weaknesses of machine learning techniques can be known so that further research can be carried out to cover the weaknesses of the machine learning techniques found.

RESULT AND DISCUSSION

In this chapter, the results of the research are described and analyzed based on the performance evaluation of various classification algorithms on the dataset used. This dataset consists of two main classes, namely "Cancer" and "No Cancer" with an unbalanced distribution. This data was then tested using several machine

learning models, including Random Forest, Support Vector Machine (SVM), Logistic Regression, Multi-Layer Perceptron (MLP), and others.

The following is the distribution of data sets from this study. The dataset consists of 2 types of data, namely cancer and non-cancer, as shown in table 2. The dataset is broken down using no test set simply using K-fold, and with the default parameters.

Table 2. Dataset partition table

a. Cancer and non-cancer divisions

		90%	80%	70%	60%	50%
Cancer	270	243	216	189	162	135
No cancer	39	4	8	12	16	20
	309	10%	20%	30%	40%	50%
		247	224	201	178	155

b. Non-Cancer and Cancer Divisions

		90%	80%	70%	60%	50%
No cancer	39	35	31	27	23	20
Cancer	270	27	54	81	108	135
	309	10%	20%	30%	40%	50%
		62	85	108	131	155

From table 2, it can be seen that the first column shows the total number of cases (270 for Cancer and 39 for No Cancer). The next columns show the number of cases by various percentages. The values in the row show the number of actual case divisions by percentage. The bottom table is the same as the top but with the reverse category order (No Cancer and Cancer). Analysis Distribution The first dataset had a higher number of cases for Cancer than for No Cancer. For example, 243 cases (90%) were cancers compared to only 4 cases (10%) that were non-cancerous. The second dataset (bottom) had a higher number of cases for No Cancer than for Cancer. For example, 35 cases (90%) were No Cancer compared to only 27 cases (10%) were Cancer. With a dataset model like this, it may be more representative of a situation where Cancer cases are more common than No Cancer. The second dataset may be more representative of a situation where No Cancer cases are more common than Cancer. The condition of sharing this dataset is intended to detect or study Cancer cases with a higher focus, then the dataset at the top is more suitable, but if the goal is to detect or study No Cancer cases with a higher focus, the dataset at the bottom is more suitable.

This analysis shows the existence of class imbalances in both datasets. In practice, it is important to account for these imbalances in machine learning models because they can affect the accuracy and performance of the model. Several techniques such as resampling, oversampling, or under sampling can be applied to deal with this problem.

The results of the analysis for each classification of the data that have been shared above can be seen as follows:

A. Random Forest

The following are the results of the performance evaluation of the Random Forest algorithm based on the class distribution and some test metrics.

Table 3. Random Forest evaluation results

a. Cancer Class: Non-Cancer Class

Cancer Class: Non-Cancer Class								
Distribution Class	**Performance Evaluation for Random Forest**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	91.2621	0.909	0.907	0.913	0.941	0.386	0.947	0.574
50:50	91.6129	0.908	0.909	0.916	0.918	0.438	0.934	0.578
60:40	93.2584	0.925	0.924	0.933	0.881	0.514	0.942	0.520
70:30	94.0299	0.926	0.922	0.940	0.944	0.784	0.964	0.265
80:20	94.1964	0.935	0.929	0.942	0.793	0.965	0.957	-0.029
90:10	98.3193	0	0	0.983	0.725	0.983	0.976	0

b. Non-Cancer Class: Cancer Class

Non-Cancer Class: Cancer Class								
Distribution Class	**Performance Evaluation for Random Forest**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	91.2621	0.909	0.907	0.913	0.941	0.386	0.947	0.574
50:50	86.4516	0.847	0.839	0.865	0.881	0.659	0.912	0.268
60:40	88.4615	0.879	0.877	0.885	0.914	0.366	0.931	0.573
70:30	91.6667	0.914	0.916	0.917	0.938	0.201	0.946	0.770
80:20	89.4118	0.892	0.897	0.894	0.917	0.157	0.911	0.770
90:10	87.0968	0.871	0.871	0.871	0.956	0.133	0.957	0.738

The results of the evaluation of the performance of the random forest algorithm based on the class distribution can be seen from table 3 above. The table above presents an evaluation of the performance of the Random Forest algorithm based on the various class distributions between Cancer and Non-Cancer. In this evaluation, several important metrics are used, including *Accuracy, F-Measure, Precision, True Positive Rate (*TPR*), Receiver Operating Characteristic (*ROC*), False Positive Rate (*FPR*), Precision-Recall Curve (*PRC*), and Matthews Correlation Coefficient* (MCC).

In table 3a, the performance of the Random Forest algorithm for the "Cancer" class shows that the accuracy increases with the imbalance of the class distribution. The 90:10 distribution resulted in the highest accuracy of 98.3193%, but with significant decreases in F-Measure, Precision, and MCC values. This suggests a trade-off between high accuracy and the ability of the model to correctly detect both classes, especially on highly unbalanced datasets.

In contrast, on more balanced distributions such as 70:30 and 80:20, the Precision, F-Measure, and MCC values remain more consistent, although the accuracy is slightly lower compared to the highly unbalanced distribution. This suggests that the Random Forest model works better at detecting the overall class on a more balanced distribution, while maintaining the trade-offs between various performance metrics.

In table 3b, which shows the results for the "No Cancer" class, it can be seen that a more balanced distribution of classes such as 70:30 and 80:20 gives more stable results, especially at MCC values that reach 0.770 in these two distributions. The 90:10 distribution shows a decrease in accuracy of up to 87.0968%, which indicates that in cases of extreme data imbalances, the model tends to have difficulty in classifying minority classes well.

Overall, these tables show that while accuracy can improve on unbalanced datasets, other metrics such as F-Measure and MCC can experience significant decreases, so it is important to consider the balance between classes when training classification models.

B. Support Vector Machine

The following are the results of the performance evaluation of the Support Vector Machine algorithm based on the class distribution and some test metrics.

Table 4. Support Vector Machine evaluation results table

a. Cancer Class: Non-Cancer Class

Cancer Class: Non-Cancer Class								
Distribution Class	**Performance Evaluation for Support Vector Machine**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	92.5566	0.921	0.920	0.926	0.782	0.362	0.886	0.632
50:50	95.4839	0.954	0.954	0.955	0.889	0.177	0.934	0.795
60:40	94.9438	0.945	0.946	0.949	0.775	0.399	0.916	0.654
70:30	93.5323	0.927	0.922	0.935	0.614	0.706	0.905	0.296
80:20	93.75	0.933	0.929	0.938	0.486	0.965	0.930	-0.032
90:10	97.8992	0.973	0.967	0.979	0.498	0.983	0.967	-0.008

b. Non-Cancer Class: Cancer Class

Non-Cancer Class: Cancer Class								
Distribution Class	**Performance Evaluation for Support Vector Machine**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	92.5566	0.921	0.920	0.926	0.782	0.362	0.886	0.632
50:50	89.6774	0.897	0.897	0.897	0.770	0.356	0.869	0.541
60:40	90	0.899	0.898	0.900	0.820	0.260	0.864	0.651
70:30	94.4444	0.944	0.944	0.944	0.926	0.093	0.921	0.852
80:20	90.5882	0.906	0.908	0.906	0.905	0.095	0.873	0.801
90:10	87.0968	0.871	0.874	0.871	0.873	0.125	0.827	0.741

Table 4 above shows the performance evaluation of the Support Vector Machine (SVM) algorithm based on the various class distributions between Cancer and Non-Cancer. In table 4a focusing on the "Cancer" class, it can be seen that the highest Accuracy is achieved at a 90:10 distribution with a value of 97.8992%. However, despite the high accuracy, the MCC value shows a drastic drop to -0.008, which signals an imbalance in the classification between the dominant and minority classes. On more balanced distributions, such as 50:50 and 60:40, SVM provides more stable results with more consistent F-Measure and Precision, as well as higher MCC values, such as 0.795 in a 50:50 distribution.

In contrast, in highly unbalanced distributions such as 80:20 and 90:10, although the accuracy remains high, the FPR and MCC values drop dramatically, suggesting that the SVM model begins to have difficulty in precisely distinguishing classes when the class distribution becomes highly unbalanced.

In table 4b focusing on the "No Cancer" class, similar results can be observed. In the 50:50 distribution, SVM showed better performance with an Accuracy of 89.6774%, and an MCC value of 0.541. However, as the distribution becomes more unbalanced as in 90:10, the MCC value drops to 0.741, while the Accuracy also decreases to 87.0968%.

In addition, table 4 shows that a more balanced class distribution such as 70:30 results in better performance overall, especially in the case of MCC, which reaches a value of 0.852, suggesting that the SVM model can better differentiate classes in this distribution.

Overall, table 4 indicates that although the SVM algorithm has a high accuracy capability, especially on unbalanced datasets, a more balanced class distribution in general provides more stable and fair performance results in classifying both classes.

In both cases, a more balanced class distribution tends to provide better performance in terms of MCC, which indicates better classification quality. Better class balance seems to be essential to get optimal results with the SVM algorithm.

C. Logistic Regression

The following are the results of the performance evaluation of the Logistic Regression algorithm based on class distribution and several test metrics.

Table 5. Logistic Regression evaluation results

a. Cancer Class: Non-Cancer Class

Cancer Class: Non-Cancer Class								
Distribution Class	**Performance Evaluation for Logistic Regression**							
	Accuracy	**F-Measure**	**Precion**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	93.2039	0.930	0.929	0.932	0.934	0.295	0.946	0.676
50:50	90.9677	0.908	0.906	0.910	0.865	0.354	0.918	0.581

Cancer Class: Non-Cancer Class								
Distribution Class	**Performance Evaluation for Logistic Regression**							
	Accuracy	**F-Measure**	**Precion**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
60:40	92.1348	0.921	0.921	0.921	0.717	0.402	0.885	0.519
70:30	91.0448	0.916	0.923	0.910	0.731	0.552	0.929	0.314
80:20	92.4107	0.935	0.948	0.924	0.713	0.605	0.952	0.237
90:10	95.7983	0.962	0.966	0.958	0.449	0.984	0.965	-0.021

b. Non-Cancer Class: Cancer Class

Non-Cancer Class: Cancer Class								
Distribution Class	**Performance Evaluation for Logistic Regression**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	93.2039	0.930	0.929	0.932	0.934	0.295	0.946	0.676
50:50	89.6774	0.899	0.901	0.897	0.864	0.313	0.913	0.560
60:40	89.2308	0.892	0.892	0.892	0.832	0.262	0.887	0.630
70:30	89.8148	0.896	0.896	0.898	0.896	0.207	0.913	0.720
80:20	82.3529	0.825	0.829	0.824	0.857	0.184	0.827	0.629
90:10	90.3226	0.903	0.903	0.903	0.935	0.100	0.924	0.803

The table above shows the performance evaluation of the Logistic Regression algorithm based on the class distribution between Cancer and Non-Cancer. In table 5a, which focuses on the "Cancer" class, the results show that the 90:10 distribution results in the highest Accuracy, which is 95.7983%, followed by the F-Measure and Precision values which are also high. However, despite the very high accuracy, the MCC value became negative (-0.021), indicating an imbalance in the model's performance when detecting a minority class, although the overall accuracy seemed satisfactory.

In more balanced distributions such as 50:50 and 60:40, MCC values tend to be more stable, with the highest value on the default distribution (0.676). However, a more balanced distribution results in slightly lower accuracy compared to a 90:10 distribution, which indicates that the model may experience trade-offs in detection between classes when the dataset becomes more unbalanced.

For table 5b, which shows the evaluation of the "Non-Cancer" class, a more balanced distribution such as 70:30 yields more consistent results, with an Accuracy of 89.8148% and the highest MCC value of 0.803. The 90:10 distribution results in an increase in Accuracy to 90.3226%, but at the 80:20 distribution, the Accuracy value decreases to 82.3529%, although the MCC remains at a stable level.

From these results, it can be seen that the Logistic Regression algorithm shows a fairly good performance in both classes, with more stable performance metrics in a more balanced distribution. However, as with other algorithms, Logistic Regression tends to face challenges in maintaining the consistency of metrics such as MCCs on highly unbalanced distributions.

D. Multi-Layer Perceptron

The following are the results of the performance evaluation of the Perceptron Multi-layer algorithm based on the class distribution and some test metrics.

Table 6. Multi-layer Perceptron evaluation results

a. Cancer Class: Non-Cancer Class

Cancer Class: Non-Cancer Class								
Distribution Class	**Performance Evaluation for Multi-Layer Perceptron**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	92.233	0.921	0.919	0.922	0.938	0.318	0.942	0.633
50:50	94.8387	0.948	0.948	0.948	0.952	0.178	0.960	0.770
60:40	92.1348	0.921	0.921	0.921	0.925	0.402	0.958	0.519
70:30	94.0299	0.944	0.950	0.940	0.962	0.316	0.970	0.547
80:20	91.9643	0.930	0.942	0.920	0.818	0.725	0.962	0.149
90:10	96.6387	0.966	0.966	0.966	0.490	0.983	0.967	-0.017

b. Non-Cancer Class: Cancer Class

Non-Cancer Class: Cancer Class								
Distribution Class	**Performance Evaluation for Multi-Layer Perceptron**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	92.233	0.921	0.919	0.922	0.938	0.318	0.942	0.633
50:50	90.9677	0.915	0.923	0.910	0.920	0.184	0.929	0.651
60:40	89.2308	0.896	0.901	0.892	0.940	0.194	0.956	0.657
70:30	91.6667	0.918	0.921	0.917	0.957	0.102	0.962	0.788
80:20	87.0588	0.872	0.875	0.871	0.934	0.129	0.931	0.728
90:10	93.5484	0.935	0.937	0.935	0.979	0.075	0.981	0.870

The following are the results of the performance evaluation of the Multi-Layer Perceptron (MLP) algorithm for classification in two categories: Cancer Class and Non-Cancer Class. The evaluation was conducted using several different data distributions, namely 50:50, 60:40, 70:30, 80:20, and 90:10, as well as one default setting.

Accuracy: The 90:10 distribution in both tables shows high accuracy, especially in the Cancer Class: Non-Cancer Class with a value of 96.64%, although a high False Positive Rate (FPR) value in this distribution may indicate a classification imbalance.

F-Measure and Precision: These metrics are generally high at the 50:50 and 90:10 distributions, which indicates that the model is quite good at handling positive predictions. However, at certain distributions such as 80:20, even though the Precision is high, a low MCC value indicates a problem in overall performance, especially for Cancer Class: Non-Cancer Class.

True Positive Rate (TPR) and ROC: In more balanced data distributions such as 50:50 and 70:30, the model shows a good ability to detect correct predictions. However, at the 90:10 distribution for Cancer Class: Non-Cancer Class, the ROC value dropped significantly, indicating a decrease in performance in distinguishing between the two classes.

False Positive Rate (FPR): This value varies quite significantly across multiple distributions. The 90:10 distribution for the Cancer Class: The Non-Cancer Class shows very high FPR values, which means the model often misclassifies negative samples as positive. In contrast, at the same distribution for Non-Cancerous Class: Cancer Class, the FPR value was very low, showing good performance in reducing negative misclassification.

MCC (Matthews Correlation Coefficient): This metric shows the overall balance between positive and negative class predictions. In the Non-Cancerous Class: Cancer Class, the highest MCC is at the 90:10 distribution, while in the Cancer Class: Non-Cancer Class, the MCC is very low at the same distribution, indicating that the model works much better in the situation of the Non-Cancer class as a minority class than in the case of Cancer being a minority class.

Overall, these algorithms show the MLP model is heavily influenced by the distribution of data. The models tend to perform best on more balanced distributions, and some metrics such as MCC and FPR are strongly influenced by unbalanced distributions, especially in Cancer Class: Non-Cancer Class.

The MLP model works better with a more balanced distribution (50:50 or 70:30). Higher MCC and stable accuracy indicate that the model is capable of handling the minority and majority classes well. Highly unbalanced distributions tend to significantly degrade model performance, as seen from the negative MCC. Overall, a more balanced class distribution tends to provide better performance in terms of MCC, suggesting that better classification quality can be achieved by maintaining a balance between classes in the dataset.

E. C4.5

The following are the results of the evaluation of the performance of the C4.5 algorithm based on the class distribution and some test metrics.

Table 7. C4.5 Evaluation

a. Cancer Class: Non-Cancer Class

Cancer Class: Non-Cancer Class								
Distribution Class	**Performance Evaluation for C4.5**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	90.2913	0.899	0.897	0.903	0.787	0.409	0.882	0.531
50:50	89.0323	0.886	0.884	0.890	0.744	0.442	0.870	0.481
60:40	86.5169	0.869	0.873	0.865	0.633	0.633	0.867	0.220
70:30	93.5323	0.923	0.916	0.935	0.721	0.785	0.929	0.229
80:20	96.4286	0	0	0.964	0.394	0.964	0.924	0
90:10	98.3193	0	0	0.983	0.197	0.983	0.955	0

b. Non-Cancer Class: Cancer Class

Non-Cancer Class: Cancer Class								
Distribution Class	**Performance Evaluation for C4.5**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	90.2913	0.899	0.897	0.903	0.787	0.409	0.882	0.531
50:50	87.0968	0.871	0.871	0.871	0.740	0.445	0.863	0.426
60:40	86.9231	0.863	0.860	0.869	0.712	0.404	0.803	0.514
70:30	91.6667	0.917	0.918	0.917	0.892	0.127	0.895	0.781
80:20	72.9412	0.728	0.728	0.729	0.754	0.320	0.760	0.412
90:10	75.8065	0.757	0.757	0.758	0.755	0.254	0.718	0.506

Cancer Distribution: No Cancer:
The C4.5 algorithm shows a decrease in performance in MCC values when the distribution becomes more unbalanced (80:20 and 90:10), where MCC values reach 0. The best performance is seen in the 70:30 distribution with an MCC of 0.229.

Distribution of No Cancer: Cancer:
The C4.5 algorithm showed an increase in performance at MCC values when the distribution was more balanced (50:50 and 70:30), with the highest MCC value at the 70:30 distribution (0.781). Performance decreases on more unbalanced distributions (80:20 and 90:10), but MCC values remain above 0.

For a more balanced distribution (No Cancer : Cancer 70:30): The C4.5 algorithm shows excellent performance with the highest MCC. This suggests that this distribution is more optimal for the C4.5 algorithm.

For more unbalanced distributions (Cancer: No Cancer 80:20 and 90:10): The C4.5 algorithm showed a significant decrease in performance. Imbalance handling techniques such as oversampling, undersampling, or the use of customized evaluation metrics may be necessary to improve performance.

Overall the C4.5 algorithm is more effective on a more balanced distribution and shows excellent performance in identifying cancer cases when the data distribution is more evenly distributed.

F. Bayesian Network

The following are the results of the evaluation of the performance of the Bayesian Network algorithm based on the class distribution and some test metrics.

Table 8. Bayesian Network evaluation

a. Kelas Kanker: Kelas Tidak Kanker

Cancer Class: Non-Cancer Class								
Distribution Class	**Performance Evaluation for Bayesian Network**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	87.055	0.868	0.865	0.871	0.855	0.501	0.898	0.387
50:50	85.1613	0.826	0.813	0.852	0.746	0.746	0.864	0.151
60:40	91.0112	0	0	0.910	0.437	0.910	0.814	0
70:30	94.0299	0	0	0.940	0.729	0.940	0.928	0
80:20	96.4286	0	0	0.964	0.428	0.964	0.921	0
90:10	98.3193	0	0	0.983	0.445	0.983	0.963	0

b. Kelas Tidak Kanker: Kelas Kanker

Non-Cancer Class: Cancer Class								
Distribution Class	**Performance Evaluation for Bayesian Network**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	87.055	0.868	0.865	0.871	0.855	0.501	0.898	0.387
50:50	83.871	0.811	0.792	0.839	0.691	0.791	0.834	0.069
60:40	81.5385	0.812	0.809	0.815	0.834	0.483	0.855	0.344
70:30	79.6296	0.801	0.807	0.796	0.869	0.290	0.871	0.485
80:20	75.2941	0.755	0.759	0.753	0.827	0.265	0.831	0.479
90:10	74.1935	0.741	0.741	0.742	0.740	0.275	0.736	0.472

A decrease in accuracy and MCC, indicates a more balanced class distribution but a decrease in performance in Bayesian Network algorithms. At the 60:40 to 90:10 dataset split, the F-Measure and Precision are 0, indicating the model cannot make valid predictions on these distributions. This may be because the model has become highly biased towards the majority class. The MCC is 0, indicating no correlation between prediction and reality. Further decline in accuracy but the MCC still showed a moderate correlation between prediction and reality.

In the 50:50 Class distribution, the Bayesian Network showed a significant decrease in MCC. The model may not be robust enough to handle a well-balanced distribution of classes. In the second table, the MCC shows a significant increase in the 70:30 distribution, suggesting that the model is better able to handle moderate class imbalances. In table 8a, the 90:10 distribution yields an MCC of 0, indicating that the model is highly biased towards the majority class and cannot make valid predictions.

In table 8b, a moderate MCC shows that the model still has some predictive capabilities despite the significant class imbalance. The Bayesian Network model performs best in a 70:30 class distribution with a higher MCC, suggesting that this distribution is better suited to handling class imbalances. Extreme imbalances such as 90:10 severely affect the performance of Bayesian Network models, making them ineffective in making valid predictions.

Overall, the Bayesian Network model appears to be more sensitive to class imbalances than other models analyzed previously. A more moderate class distribution such as 70:30 gives better results in terms of MCC, indicating better classification quality.

G. Reptree

The following are the results of the evaluation of the performance of the Bayesian Network algorithm based on the class distribution and some test metrics.

Table 9. Reptree Evaluation

a. Cancer Class: Non-Cancer Class

Cancer Class: Non-Cancer Class								
Distribution Class	**Performance Evaluation for Reptree**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	85.7605	0.839	0.828	0.858	0.739	0.701	0.859	0.207
50:50	87.0968	0.865	0.860	0.871	0.691	0.530	0.836	0.375
60:40	90.4494	0.892	0.885	0.904	0.639	0.685	0.863	0.286
70:30	93.5323	0.909	0.884	0.935	0.614	0.941	0.912	-0.018
80:20	96.4286	0	0	0.964	0.394	0.964	0.924	0
90:10	98.3193	0	0	0.983	0.197	0.983	0.955	0

b. Non-Cancer Class: Cancer Class

Non-Cancer Class: Cancer Class								
Distribution Class	**Performance Evaluation for Reptree**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	85.7605	0.839	0.828	0.858	0.739	0.701	0.859	0.207
50:50	85.8065	0.823	0.809	0.858	0.629	0.788	0.814	0.122
60:40	86.9231	0.866	0.863	0.869	0.801	0.369	0.858	0.529
70:30	84.2593	0.839	0.838	0.843	0.807	0.299	0.827	0.566
80:20	74.1176	0.739	0.738	0.741	0.715	0.313	0.703	0.434
90:10	70.9677	0.710	0.710	0.710	0.738	0.300	0.710	0.410

From table 9 showing an Accuracy value of 85.7605 and an MCC value of 0.207, this is a baseline with MCC showing a weak correlation between prediction and reality. At the 50:50 class distribution, the Accuracy is 87.0968 and the MCC: 0.375, a slight improvement in accuracy and MCC, indicating a more balanced class distribution improves performance. The values of Accuracy 90.4494 and MCC: 0.286 were obtained in the 60:40 class distribution, the Accuracy increased, but the MCC showed an insignificant increase, indicating the imbalance still affected the performance.

An accuracy of 93.5323 and an MCC of -0.018 was obtained at a class distribution of 70:30, The MCC decrease indicates that despite the high accuracy, the class imbalance greatly affects performance. The 80:20 and 90:10 class distributions, F-Measure and Precision are 0, indicating the model cannot make valid predictions on these distributions. This may be because the model has become highly biased towards the majority class. The MCC is 0, indicating no correlation between prediction and reality.

Analysis Based on the RepTree Model:

1. 50:50 Class Distribution:

The model shows a moderate improvement in performance with a balanced distribution of classes, but the MCC improvement shows that the model still has some difficulties in handling class imbalances efficiently.

2. 70:30 Class Distribution:

In table 9b, the MCC shows a significant increase at the 70:30 distribution, suggesting that the model is better able to handle moderate class imbalances.

3. 90:10 Class Distribution:

In table 9a, the 90:10 distribution results in an MCC of 0, indicating that the model is highly biased towards the majority class and cannot make valid predictions.

In table 9b, a moderate MCC shows that the model still has some predictive capabilities despite the significant class imbalance.

The RepTree model shows best performance at a 70:30 class distribution with a higher MCC, suggesting that this distribution is better suited to handle class imbalances. Extreme imbalances such as 90:10 severely affect the performance of the RepTree model, making it ineffective in making valid predictions.

Overall, the RepTree model appears to be more sensitive to class imbalances than some of the other models analyzed previously. A more moderate class distribution such as 70:30 gives better results in terms of MCC, indicating better classification quality.

H. Naive Bayes

The following are the results of the evaluation of the performance of the Naïve Bayesian algorithm based on the class distribution and some test metrics.

Table 10. Naïve Bayes Evaluation

a. Cancer Class: Non-Cancer Class

Cancer Class: Non-Cancer Class								
Distribution Class	**Performance Evaluation for Naïve Bayes**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	89.644	0.893	0.890	0.896	0.902	0.432	0.926	0.500
50:50	90.9677	0.903	0.901	0.910	0.910	0.439	0.936	0.550
60:40	92.6966	0.921	0.918	0.927	0.830	0.514	0.925	0.490
70:30	92.5373	0.924	0.922	0.925	0.925	0.629	0.956	0.309
80:20	95.5357	0.958	0.961	0.955	0.752	0.483	0.951	0.424
90:10	98.3193	0	0	0.983	0.679	0.983	0.976	0

b. Non-Cancer Class: Cancer Class

Non-Cancer Class: Cancer Class								
Distribution Class	**Performance Evaluation for Naïve Bayes**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	89.644	0.893	0.890	0.896	0.902	0.432	0.926	0.500
50:50	86.4516	0.866	0.867	0.865	0.876	0.446	0.912	0.410
60:40	86.9231	0.870	0.872	0.869	0.891	0.301	0.907	0.559
70:30	89.8148	0.894	0.896	0.898	0.949	0.231	0.956	0.717
80:20	84.7059	0.847	0.846	0.847	0.927	0.184	0.933	0.668
90:10	82.2581	0.822	0.822	0.823	0.916	0.188	0.919	0.638

This is a good baseline with the MCC showing a moderate correlation between prediction and reality. At 50:50 data distribution: Accuracy: 86.4516, MCC: 0.410. A decrease in accuracy and MCC, indicating a more balanced distribution of classes decreases performance. At 60:40 data distribution: Accuracy: 86.9231, MCC: 0.559. Accuracy is slightly improved with MCC also improved, suggesting a more balanced class distribution improves classification quality. At 70:30 data distribution: Accuracy: 89.8148, MCC: 0.717. A significant increase in MCC showed a better correlation between prediction and reality. Distribution 80:20: Accuracy: 84.7059, MCC: 0.668. The decrease in accuracy but the MCC remains quite high, demonstrating the model's ability to handle class imbalances. And at a 90:10 distribution: Accuracy: 82.2581, MCC: 0.638. Further decline in accuracy but the MCC still showed a moderate correlation between prediction and reality.

Overall, the Naïve Bayes model appears to be more sensitive to class imbalances than some of the other models analyzed previously. A more balanced class distribution tends to give better results in terms of MCC, indicating better classification quality.

I. P.A.R.T.

The following are the results of the performance evaluation of the P.A.R.T. algorithm based on class distribution and several test metrics.

Table 11. P.A.R.T. Evaluation

a. Cancer Class: Non-Cancer Class

Cancer Class: Non-Cancer Class								
Distribution Class	**Performance Evaluation for P.A.R.T.**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	89.9676	0.898	0.897	0.900	0.834	0.387	0.902	0.530
50:50	91.6129	0.913	0.912	0.916	0.819	0.353	0.904	0.604
60:40	91.573	0.909	0.905	0.916	0.682	0.572	0.891	0.409
70:30	93.0348	0.930	0.930	0.930	0.777	0.551	0.938	0.380
80:20	92.8571	0.932	0.936	0.929	0.709	0.845	0.949	0.075
90:10	95.7983	0.962	0.966	0.958	0.686	0.984	0.974	-0.021

b. Non-Cancer Class: Cancer Class

Non-Cancer Class: Cancer Class								
Distribution Class	**Performance Evaluation for P.A.R.T.**							
	Accuracy	**F-Measure**	**Precision**	**TPR**	**ROC**	**FPR**	**PRC**	**MCC**
Default	89.9676	0.898	0.897	0.900	0.834	0.387	0.902	0.530
50:50	83.2258	0.832	0.832	0.832	0.691	0.579	0.835	0.254
60:40	82.3077	0.814	0.809	0.823	0.661	0.516	0.775	0.339
70:30	90.7407	0.908	0.910	0.907	0.933	0.130	0.918	0.760
80:20	78.8235	0.788	0.788	0.788	0.735	0.245	0.709	0.543
90:10	79.0323	0.790	0.790	0.790	0.797	0.221	0.762	0.572

Accuracy is very high but the MCC is negative, suggesting the model may only predict the majority class.

Analysis Based on the P.A.R.T. Model:

1. 50:50 Class Distribution:

The model shows an improvement in performance with a balanced class distribution in table 11a, particularly with a significantly increased MCC. Accuracy also improved at a 50:50 distribution, suggesting that the P.A.R.T model works better with balanced data.

2. 70:30 Class Distribution:

In table 11B, the MCC shows a significant improvement at the 70:30 distribution, suggesting that the model is better able to handle moderate class imbalances. Accuracy remains high, indicating that the model does not only predict the majority class.

3. 90:10 Class Distribution:

In the first table, the 90:10 distribution produces a negative MCC, indicating that the model is highly biased towards the majority class and cannot make valid predictions. In table 11b, a moderate MCC shows that the model still has some predictive capabilities despite the significant class imbalance.

The P.A.R.T. model shows best performance at 50:50 and 70:30 class distributions with higher MCCs, suggesting that these distributions are better suited to handling class imbalances. Extreme imbalances such as 90:10 severely affect the performance of P.A.R.T. models, making them ineffective in making valid predictions.

Overall, the P.A.R.T. model shows that a more balanced class distribution tends to give better results in terms of MCC, indicating better classification quality. This model appears to be more sensitive to class imbalances than some of the other models analyzed previously.

CONCLUSION

Based on the results of the evaluation, the Support Vector Machine algorithm showed excellent performance in detecting both classes, with the highest accuracy recorded in the 50:50 and 70:30 dataset divisions, which were 95.48% and 94.44%. These results suggest that SVM is more suitable for use in datasets with a more balanced distribution. In contrast, on highly unbalanced dataset divisions such as 90:10, SVM accuracy tends to decrease.

The Random Forest algorithm shows significant performance stability, especially at a 90:10 dataset split with an accuracy of 98.32%. This shows that Random Forest is able to handle data imbalances well.

On the other hand, algorithms such as Naive Bayes and C4.5 also perform well on more unbalanced datasets. Naive Bayes had the highest accuracy at 90:10 with a value of 98.32%, while C4.5 achieved the best accuracy at 80:20 with a value of 96.43%.

However, it is important to note that certain algorithms, such as Bayesian Network and Reptree, show significant performance degradation when faced with data imbalances. The Bayesian Network, for example, experienced a decrease in accuracy at a 90:10 distribution, with a low MCC, indicating that this algorithm is more sensitive to class imbalances.

From this analysis, it can be concluded that a more balanced distribution of the dataset results in more optimal model performance, especially for algorithms such as Support Vector Machine and Logistic Regression. To address dataset imbalances, techniques such as oversampling or undersampling can be applied to improve the overall performance of the model.

REFERENCES

[1] Nurhidayah, B. Abdul Samad, and B. Abdullah, "Perbandingan Metode Contrast Enhancement pada Citra CT-Scan Kanker Paru-paru," *Gravitasi*, vol. 19, no. 2, pp. 24–28, Dec. 2020.

[2] R. T. Prasetio and S. Susanti, "Prediksi Harapan Hidup Pasien Kanker Paru Pasca Operasi Bedah Toraks Menggunakan Boosted k-Nearest Neighbor".

[3] J. Joseph and L. W. A. Rotty, "Kanker Paru: Laporan Kasus," *Med. Scope J.*, vol. 2, no. 1, Jul. 2020.

[4] M. S. Kumar and K. V. Rao, "Prediction of Lung Cancer Using Machine Learning Technique: A Survey," in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India: IEEE, Jan. 2021, pp. 1–5.

[5] Y. Yang, L. Xu, L. Sun, P. Zhang, and S. S. Farid, "Machine learning application in personalised lung cancer recurrence and survivability prediction," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 1811–1820, 2022.

[6] J. A. Bartholomai and H. B. Frieboes, "Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques," in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Louisville, KY, USA: IEEE, Dec. 2018, pp. 632–637.

[7] H. Tang, J. Zhao, and X. Yang, "Explore Machine Learning for Analysis and Prediction of Lung Cancer Related Risk Factors," in *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence*, Shenzhen China: ACM, Dec. 2018, pp. 41–45.

[8] B. Alsinglawi *et al.*, "An explainable machine learning framework for lung cancer hospital length of stay prediction," *Sci. Rep.*, vol. 12, no. 1, p. 607, Jan. 2022.

[9] M. I. Faisal, S. Bashir, Z. S. Khan, and F. Hassan Khan, "An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer," in *2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST)*, Karachi, Pakistan: IEEE, Dec. 2018, pp. 1–4.

[10] S. M. Kulkarni and G. Sundari, "COMPARATIVE ANALYSIS OF PERFORMANCE OF DEEP CNN BASED FRAMEWORK FOR BRAIN MRI CLASSIFICATION USING TRANSFER LEARNING," vol. 16, 2021.

[11] P. J. Mahadevia, L. A. Fleisher, K. D. Frick, J. Eng, S. N. Goodman, and N. R. Powe, "Lung Cancer Screening With Helical Computed Tomography in Older Adult Smokers: A Decision and Cost-effectiveness Analysis," *JAMA*, vol. 289, no. 3, p. 313, Jan. 2003.

[12] M. R. Raharjo and A. P. Windarto, "Penerapan Machine Learning dengan Konsep Data Mining Rough Set (Prediksi Tingkat Pemahaman Mahasiswa terhadap Matakuliah)," *J. MEDIA Inform. BUDIDARMA*, vol. 5, no. 1, p. 317, Jan. 2021.

[13] J. V. Kistowski, J. A. Arnold, K. Huppler, K.-D. Lange, J. L. Henning, and P. Cao, "How to Build a Benchmark," in *Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering*, Austin Texas USA: ACM, Jan. 2015, pp. 333–336.

[14] J. Thiyagalingam, M. Shankar, G. Fox, and T. Hey, "Scientific machine learning benchmarks," *Nat. Rev. Phys.*, vol. 4, no. 6, pp. 413–420, Apr. 2022.

[15] D. N. Haddad, K. L. Sandler, L. M. Henderson, M. P. Rivera, and M. C. Aldrich, "Disparities in Lung Cancer Screening: A Review," *Ann. Am. Thorac. Soc.*, vol. 17, no. 4, pp. 399–405, Apr. 2020.