Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 5, October 2025, Page. 3456-3467
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4724

Prediction of Life Expectancy of Lung Cancer Patients After Thoracic Surgery Using Decision Tree Algorithm and Adaptive Synthetic Sampling

Muhammad Erdi[1], Muhammad Itqan Mazdadi[*2], Radityo Adi Nugroho[3], Andi Farmadi[4], Triando Hamonangan Saragih[5], Hasri Akbar Awal Rozaq[6]

[1,2,3,4,5]Faculty of Mathematics and Natural Science, Department of Computer Science, Lambung Mangkurat University, Kalimantan, Indonesia
[6]Graduate School of Informatics, Department of Computer Science, Gazi University, Ankara, Türkiye

Email: [1]mazdadi@ulm.ac.id

Received : May 13, 2025; Revised : Jun 24, 2025; Accepted : Jun 30, 2025; Published : Oct 16, 2025

Abstract

This research focuses on predicting the life expectancy of lung cancer patients after undergoing thoracic surgery, using a decision tree classification algorithm (C4.5) combined with adaptive synthetic sampling to handle data imbalance. Data imbalance in the lung cancer patient dataset is a major obstacle in obtaining accurate prediction results, especially in identifying minority classes. Data imbalance in the lung cancer patient dataset is a major obstacle in obtaining accurate prediction results, especially in identifying minority classes. By applying ADASYN, the data distribution becomes more even, thus improving the performance of the C4.5 model. The results showed that combining these methods increased the prediction accuracy from 67% to 87%. In addition, the precision, recall, and f1-score for minority classes have significantly improved, which were previously difficult to identify by the model. Thus, combining the C4.5 algorithm and the ADASYN technique proved effective in dealing with the challenge of data imbalance and resulted in better prediction in the case of lung cancer. This study is expected to contribute to the field of medical classification and serve as a reference for further research on similar cases.

Keywords : *Algorithm Optimization, Healthcare AI, Survival Prediction, Synthetic Sampling, Thoracic Prognostics.*

1. INTRODUCTION

Lung cancer is currently considered one of the deadliest diseases and has been in the spotlight as a global cause of death. Recent data shows that the disease has affected the human population more significantly than previously estimated. Currently, lung cancer occupies the seventh position in the mortality index, causing approximately 1.5% of total deaths worldwide [1]. According to data from the World Health Organization (WHO), lung cancer is responsible for about 2.09 million new diagnoses each year and causes about 1.76 million deaths. Projections for 2020 show an increase in the number of incident cases to approximately 2.21 million [2]. Understanding and predicting the survival rate of lung cancer patients is crucial for developing more effective treatment strategies and improving patients' quality of life. Thoracic surgery, which involves diagnosing and treating conditions of the lungs, esophagus, and chest organs, is central to this effort. Accurate survival predictions following thoracic surgery not only aid in determining patient prognosis but also guide treatment planning and post-operative care [3][4].

In medical science, predictive modeling has become a cornerstone for prognosis and treatment planning. Machine learning algorithms, categorized into supervised, unsupervised, and reinforcement learning, are increasingly applied in this domain. Supervised learning leverages labeled data for training models, while unsupervised learning identifies hidden patterns in unlabeled datasets. Reinforcement learning, based on trial-and-error methods, enables adaptive decision-making in complex scenarios [5]. Among the various machine learning algorithms, the C4.5 algorithm is recognized for its high accuracy in classification tasks. It builds decision trees using entropy and information gain to optimize splits, offering interpretable and efficient rules. Studies have shown its reliability, particularly in medical applications. For example, C4.5 achieved an accuracy of 90.835% in predicting stroke outcomes, outperforming other methods like Naïve Bayes [6][7][8]. Additionally, C4.5 was chosen over algorithms like Random Forest or XGBoost due to its interpretability, which is crucial in medical contexts [referensi]. While ensemble methods such as Random Forest and XGBoost often yield higher accuracy, they are more complex and difficult to interpret, potentially limiting their use in clinical settings where transparency is vital. Moreover, C4.5 requires fewer computational resources, making it suitable for hospitals or medical facilities with limited infrastructure. These considerations highlight why C4.5 is a practical and effective choice for the task of predicting lung cancer survival.

Despite advancements in machine learning, significant challenges remain, particularly in handling imbalanced data. Medical datasets are often skewed, with outcomes like survival being disproportionately represented compared to less frequent outcomes, such as mortality. This imbalance biases predictive models towards the majority class, reducing their generalization capability [9][10]. The impact of this imbalance is particularly pronounced in medical decision-making contexts. For example, during a one-year follow-up period, the number of lung cancer patients who survive is often much higher than those who do not. This discrepancy creates challenges in accurately classifying minority outcomes, such as post-operative mortality [11][12]. Effective strategies for addressing this imbalance are critical for improving prediction accuracy.

To address data imbalance, researchers have explored techniques like SMOTE (Synthetic Minority Oversampling Technique) and ADASYN (Adaptive Synthetic Sampling). ADASYN, in particular, adaptively synthesizes data points in regions that are difficult to classify, improving model performance. Studies have demonstrated the effectiveness of ADASYN in combination with various machine learning algorithms, such as Random Forest and XGBoost, achieving superior classification accuracy [13][14]. In medical contexts, ADASYN has shown promise when combined with decision tree algorithms like C4.5. Research comparing oversampling techniques found that ADASYN paired with C4.5 significantly improved accuracy in handling imbalanced datasets, outperforming traditional methods [4][15]. These findings highlight the potential of this approach for lung cancer prognosis and similar applications.

Given the challenges posed by imbalanced datasets and the need for accurate survival predictions in lung cancer patients, this research proposes a novel integration of the interpretable C4.5 decision tree algorithm with the Adaptive Synthetic Sampling technique (ADASYN). While prior studies have explored various machine learning methods for survival prediction [2], most have overlooked the issue of class imbalance [10][12] or employed simpler balancing techniques like SMOTE [13]. Moreover, the use of interpretable models such as C4.5 remains underexplored in conjunction with adaptive sampling methods [15]. This study addresses that gap by demonstrating how ADASYN can enhance C4.5's performance in identifying minority class outcomes in thoracic surgery cases. The goal is to produce not only accurate but clinically interpretable results that support better treatment decisions and patient care.

2. METHOD

The main objective of this study is to predict lung cancer patients' survival after thoracic surgery. The data used is not balanced, so this study wants to know the effect of ADASYN on classification results using C4.5. The proposed system for predicting the survival of postoperative thoracic lung cancer patients can be seen in Figure 1.

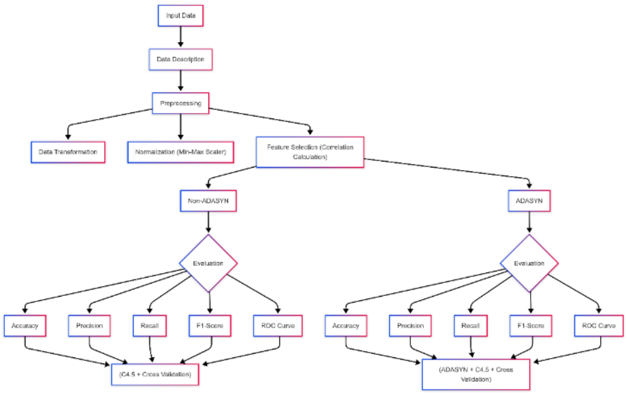

Figure 1. The Proposed Research Flow

2.1. Data Collection

The thoracic surgery dataset used in this study was obtained from Kaggle on June 9, 2025, based on the UCI Machine Learning Repository [https://www.kaggle.com/datasets/sid321axn/thoraric-surgery]. The dataset consists of 470 patient records and 16 features, including both nominal and numerical variables such as diagnostic codes, preoperative conditions (e.g., pain, dyspnoea), and physiological metrics (e.g., FEV1, FVC, age). The target variable, Risk1Y, represents patient survival status one year after surgery, labeled as “T” for deceased and “F” for survived. Since the dataset is publicly available, anonymized, and does not contain personally identifiable information, no ethical approval was required for its use in this research.

Table 1. Thoracic Surgery Dataset's Descriptions

ID	Description	Category	Range Values
DGN	ICD-10 code diagnostic-specific combinations for primary and secondary tumors as well as multiple tumors, if present	Nominal	{DGN1, DGN2, DGN4, DGN5, DGN6, DGN8, DGN3}
PRE4	FVC - Forced vital capacity	Numeric	{1.44, 6.3}
PRE5	FEV1 - Volume is exhaled at the end of the first second of forced expiration.	Numeric	{0.96, 86.3}
PRE6	Performance status - Zubrod scale	Nominal	{PRZ0, PRZ1, PRZ2}
PRE7	Pain before surgery	Nominal	{T or F}
PRE8	Hemoptysis before surgery	Nominal	{T or F}
PRE9	Dyspnoea before surgery	Nominal	{T or F}
PRE10	Cough before surgery	Nominal	{T or F}
PRE11	Weakness before surgery	Nominal	{T or F}
PRE14	T in clinical TNM - the size of the original tumor, from smallest to largest	Nominal	{OC11, OC12, OC13, OC14}
PRE17	Type 2 DM - diabetes mellitus	Nominal	{T or F}
PRE19	MI up to 6 months	Nominal	{T or F}
PRE25	PAD - peripheral arterial diseases	Nominal	{T or F}
PRE30	Smoking	Nominal	{T or F}
PRE32	Asthma	Nominal	{T or F}
AGE	Age at surgery	Numeric	{21, 87}
Risk1Y	1-year survival period - (T)rue value if died (T, F)	Nominal	{T or F}

2.2. Pre-processing Steps

The preprocessing phase in this study involved five key steps:

a. Encoding

Binary categorical variables such as “T” and “F” were converted to numerical format (1 and 0). For categorical string features such as DGN, PRE6, and PRE14, the last character of each string was extracted and label-encoded using scikit-learn’s LabelEncoder.

b. Column Renaming

To enhance interpretability, feature names were renamed (e.g., DGN → Diagnosis, Risk1Y → Death_1yr).

c. Missing Value Handling

The dataset was checked for missing (NaN) values. No missing values were found, and all 470 records were retained.

d. Normalization

Numerical attributes such as Age, FEV1, and FVC were normalized using Min-Max scaling, based on the following formula:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

This transformation rescaled the data into a [0, 1] range, allowing uniform contribution of all features during model training.

e. Class Balancing

The dataset was found to be imbalanced with 396 “survived” and 74 “deceased” cases. ADASYN (Adaptive Synthetic Sampling) was applied with $k = 5$ neighbors to synthetically generate new samples for the minority class, resulting in a balanced 1:1 class distribution.

2.3. Cross validation

To evaluate the model's generalization performance and reduce the risk of overfitting, this study employed the K-Fold Cross-Validation technique. This method partitions the dataset into K equally sized subsets (folds). The model is trained K times, each time using a different fold as the validation set and the remaining $K - 1$ folds for training.

In this study, 10-fold cross-validation was used, which is considered a standard balance between bias and variance. The process ensures that every data point is used for both training and testing, allowing for a more reliable estimate of model performance. The average performance metric (e.g., accuracy) across the folds is calculated using the following formula:

$$Accuracy_a vg = (1/K) \times \sum_{i=1}^{K} Accuracy_i \ (2)$$

Where:

- K is the number of folds (in this study, $K = 10$),
- $Accuracy_i$ is the accuracy obtained in the i^{th} fold.

This validation method is especially beneficial in medical classification problems involving small or imbalanced datasets, as it maximizes data usage while minimizing the risk of biased evaluation.

2.4. C4.5 Algorithm

The C4.5 algorithm is a decision tree algorithm developed as an enhancement of the ID3 algorithm. It is widely used in classification tasks due to its ability to generate interpretable rules and manage both numerical and categorical data. In medical applications, C4.5 is particularly advantageous because it produces decision trees that are transparent and easy to understand, making it suitable for clinical decision-making [16].

The algorithm works by recursively selecting the most informative feature at each node based on information gain ratio [17], an extension of the basic entropy-based information gain from ID3. The entropy $E(S)$ of a dataset S is calculated as:

$$E(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (3)$$

Where pi is the proportion of class i in subset S, and n is the number of classes.

The **information gain** of an attribute A is computed using:

$$Gain(S, A) = E(S) - \sum_{v=1}^{V} \frac{|S_v|}{|S|} E(S_v) \qquad (4)$$

Where Sv is the subset of S where attribute $A = v$, and V is the number of distinct values in attribute A. The attribute with the highest gain ratio is selected as the splitting criterion.

In this study, the C4.5 algorithm was implemented to construct a decision tree for classifying lung cancer patient survival based on 16 features. The tree structure highlights the most influential predictors in patient outcomes. For instance, FEV1 and Age consistently appeared as top-level nodes in the tree, indicating their strong contribution to survival prediction. This interpretability makes the model not only effective but also aligned with clinical reasoning, where understanding feature influence is as critical as prediction accuracy.

2.5. Adaptive Synthetic Sampling (ADASYN)

ADASYN (Adaptive Synthetic Sampling) is a method designed to address class imbalance in classification datasets. Class imbalance occurs when the number of data points in one class (majority class) is significantly larger than in another class (minority class), which can cause machine learning models to focus more on the majority class while ignoring the minority class [18]. ADASYN aims to enhance the representation of the minority class by generating synthetic data points, particularly in regions that are difficult to classify, allowing the model to learn in a more balanced manner [19]. The first step in ADASYN is to calculate the imbalance ratio between the number of data points in the minority and majority classes. This ratio helps determine whether oversampling is necessary. If the ratio indicates that the minority class is underrepresented, the data needs to be balanced to reduce model bias toward the majority class. Next, ADASYN calculates the number of synthetic data points required for the minority class based on this imbalance ratio, ensuring that the class distribution becomes more balanced.

The method then uses the K-Nearest Neighbor (KNN) algorithm to analyze the nearest neighbors of each minority class data point. KNN computes the distance between data points (typically using Euclidean distance) to identify underrepresented and hard-to-classify areas. These areas are given higher priority in the synthetic data generation process by assigning greater weights, ensuring that additional data is more relevant to the model's needs [20]. Synthetic data is created by combining the original minority class data points with their nearest neighbors. Technically, the new data is generated using the formula.

$$Synthetic\ Data = Original\ Data + (Neighbor - Original\ Data) \times Random\ Factor \qquad (5)$$

Where the random factor is a value between 0 and 1 to ensure variation, this process generates data that is not identical to the original but still reflects the characteristics of the minority class. With this approach, ADASYN not only increases the quantity of the minority class data but also improves its representation in hard-to-classify regions, enabling machine learning models better to understand patterns in both classes [21]. The final result is a more balanced dataset, allowing the model to produce

Vol. 6, No. 5, October 2025, Page. 3456-3467
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4724

more accurate and fair predictions. ADASYN significantly enhances performance in classification tasks involving highly imbalanced datasets.

2.6. Evaluation

Confusion Matrix is an evaluation tool used to measure the performance of machine learning models in classification tasks [22]. It provides detailed insights into the model's predictions by breaking down the number of correct and incorrect predictions for each target class. Unlike a single metric such as accuracy, the confusion matrix offers a more comprehensive understanding of how well the model performs for each class individually [23]. This makes it particularly useful for identifying the strengths and weaknesses of the model, especially in datasets with imbalanced class distributions [24].

The confusion matrix consists of four key components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive (TP) refers to the number of correctly classified positive samples, while True Negative (TN) represents the number of correctly classified negative samples. False Positive (FP) occurs when a negative sample is incorrectly classified as positive, and False Negative (FN) happens when a positive sample is misclassified as negative [25]. The following table illustrates the structure of a confusion matrix.

Table 2. Confusion Matrix Table

	Predicted Class = Yes	Predicted Class = No
Class = Yes	True Positive (TP)	False Negative (FN)
Class = No	False Positive (FP)	True Negative (TN)

Each metric provides a unique perspective on model performance:

a. Accuracy

Accuracy measures the proportion of correctly classified instances over the total number of samples. Although widely used, it can be misleading in imbalanced datasets.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

b. Precision

Precision quantifies the proportion of true positive predictions among all samples predicted as positive. It is useful in contexts where **false positives are costly**, such as medical misdiagnosis.

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

c. Recall (Sensitivity or True Positive Rate)

Recall measures the model’s ability to correctly identify all actual positive cases. In medical applications, a high recall is critical to ensure that high-risk patients are not missed.

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

d. F1-Score

The F1-Score is the harmonic mean of precision and recall, offering a balanced metric when both false positives and false negatives are important.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{9}$$

This metric is particularly useful in **imbalanced datasets**, as it accounts for both types of classification errors.

2.7. Clinical Relevance of Evaluation Metrics

In medical prognosis, each evaluation metric carries distinct clinical implications. A high recall ensures that most high-risk patients (e.g., those likely to die post-surgery) are correctly identified, minimizing missed diagnoses. Precision helps avoid false alarms that may lead to unnecessary medical interventions. F1-score balances both, making it suitable when both types of errors carry serious consequences. Therefore, the chosen metrics not only validate the model's technical performance but also align with real-world healthcare needs, where decision errors can directly affect patient outcomes.

3. RESULT

This section presents the results of the predictive modeling process, following the methodological steps outlined previously. The analysis includes class balancing with ADASYN, decision tree generation using C4.5, and comparative evaluation before and after oversampling.

3.1. Class Balancing Using ADASYN

The thoracic surgery dataset initially exhibited a significant class imbalance, with 396 instances labeled “Survive” (Class 0) and only 74 labeled “Deceased” (Class 1). This imbalance posed a challenge to the classification model, which tended to favor the majority class during learning. To address this, ADASYN (Adaptive Synthetic Sampling) was applied. It synthetically generates data points for the minority class based on the density distribution and local decision boundaries.

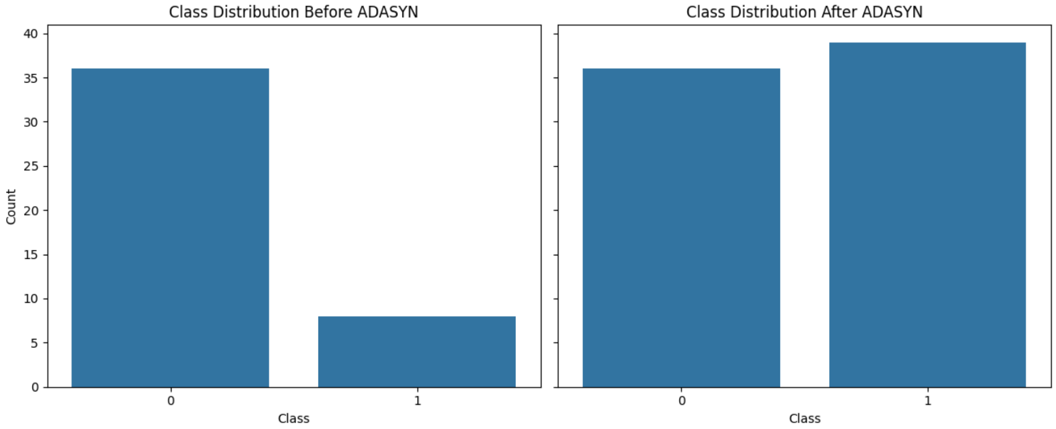

Figure 2. Class Distribution Before and After Applying ADASYN

In Figure 2, the left plot shows the original imbalanced distribution, while the right plot shows the balanced dataset after applying ADASYN. The number of "Deceased" samples is increased to match the "Survive" class, enabling the classifier to learn from both classes equally.

While ADASYN improves data balance and model fairness, it may introduce risks such as overfitting due to synthetic noise and potential distortion of the minority class distribution. Careful validation using cross-validation is necessary to ensure model generalizability.

3.2. C4.5 Algorithm Result

The initial C4.5 decision tree model was trained on the imbalanced dataset, where the “Survive” class significantly outnumbered the “Deceased” class. This class dominance caused the model to focus

Vol. 6, No. 5, October 2025, Page. 3456-3467
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4724

primarily on the majority class, reducing its ability to detect minority class cases, which are critical in clinical contexts.

The generated decision tree is shown in **Figure 3**, where **FEV1** was selected as the root node, indicating its dominant influence in classification. The model then split further based on features like **Haemoptysis** and **Age**. Each node in the tree provides key information including entropy, sample count, and predicted class. The resulting tree was relatively shallow and biased, with multiple branches leading to predictions of "Survive."

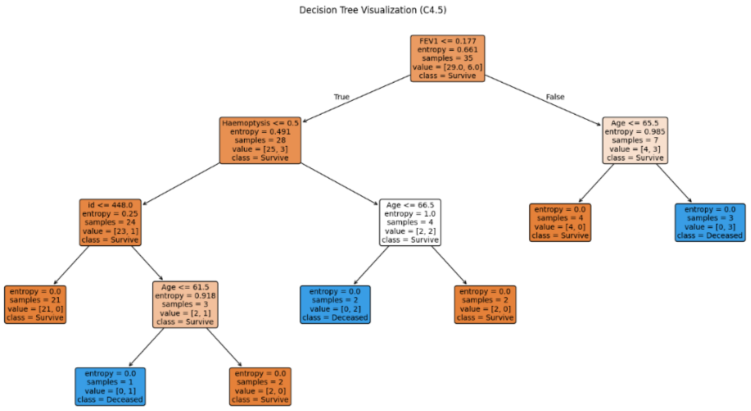

Figure 3. Decision Tree Visualization Using C4.5 Algorithm

However, due to the imbalance in class distribution, the model struggled to classify "Deceased" instances. This weakness is reflected in the confusion matrix in **Figure 4**, which shows the model's poor performance on the minority class:

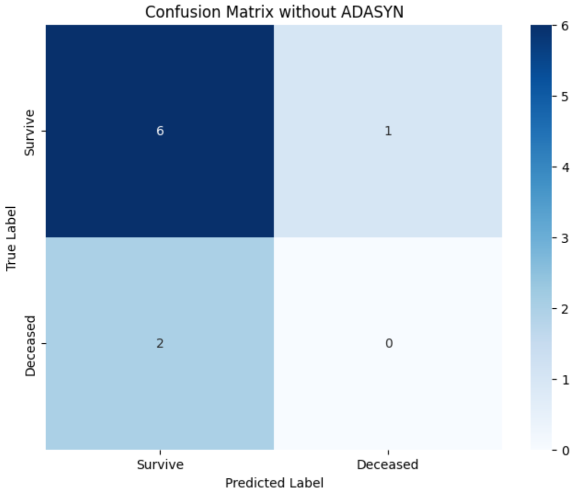

Figure 4. Confusion Matrix of C4.5 Model Without ADASYN

Based on this confusion matrix:

- True Positives (Survive): 6
- False Negatives (Survive misclassified as Deceased): 1
- False Positives (Deceased misclassified as Survive): 2
- True Negatives (Deceased): 0

The evaluation metrics for this model are:

$$Accuracy = \frac{6}{6+1+2+0} = \frac{6}{9} = 0.67 = 67\% \qquad (10)$$

$$Precision = \frac{6}{6+2} + \frac{6}{6+1} = \frac{0.75+0}{2} = 0.375 = 37.5\% \quad (11)$$

$$Recall = \frac{6}{6+1} + \frac{0}{0+2} = \frac{0.86+0}{2} = 0.43 = 43\% \quad (12)$$

$$F1 - Score = 2 * \frac{0.375*0.43}{0.375+0.43} = 0.4 = 40\% \quad (13)$$

These results clearly indicate that the model exhibited bias toward the majority class, performing poorly on minority class detection. This limitation motivated the application of the ADASYN oversampling technique to improve class representation and balance the decision process.

3.3. Decision Tree Result After ADASYN

To address the issue of class imbalance in the original dataset, this study applied the ADASYN (Adaptive Synthetic Sampling) technique prior to model training. The goal was to generate synthetic data points for the minority class ("Deceased") to ensure a more balanced learning process. Following this, the C4.5 decision tree algorithm was retrained using the newly balanced dataset.

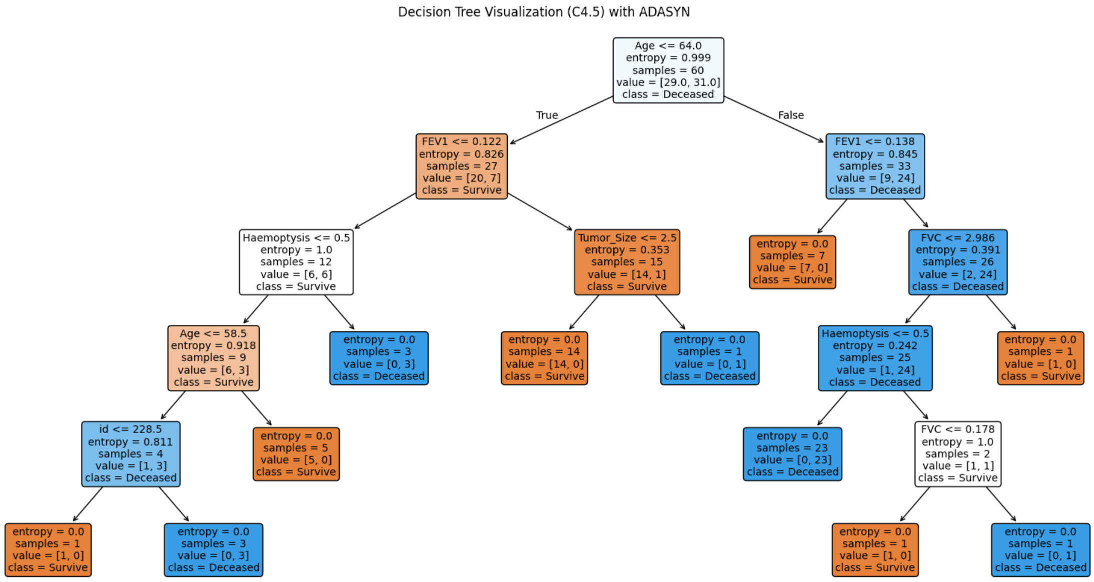
Decision Tree Visualization (C4.5) with ADASYN
Age <= 64.0 entropy = 0.999 samples = 60 value = [29.0, 31.0] class = Deceased
True
False
FEV1 <= 0.122 entropy = 0.826 samples = 27 value = [20, 7] class = Survive
FEV1 <= 0.138 entropy = 0.845 samples = 33 value = [9, 24] class = Deceased
Haemoptysis <= 0.5 entropy = 1.0 samples = 12 value = [6, 6] class = Survive
Tumor_Size <= 2.5 entropy = 0.353 samples = 15 value = [14, 1] class = Survive
entropy = 0.0 samples = 7 value = [7, 0] class = Survive
FVC <= 2.986 entropy = 0.391 samples = 26 value = [2, 24] class = Deceased
Age <= 58.5 entropy = 0.918 samples = 9 value = [6, 3] class = Survive
entropy = 0.0 samples = 3 value = [0, 3] class = Deceased
entropy = 0.0 samples = 14 value = [14, 0] class = Survive
entropy = 0.0 samples = 1 value = [0, 1] class = Deceased
Haemoptysis <= 0.5 entropy = 0.242 samples = 25 value = [1, 24] class = Deceased
entropy = 0.0 samples = 1 value = [1, 0] class = Survive
id <= 228.5 entropy = 0.811 samples = 4 value = [1, 3] class = Deceased
entropy = 0.0 samples = 5 value = [5, 0] class = Survive
entropy = 0.0 samples = 23 value = [0, 23] class = Deceased
FVC <= 0.178 entropy = 1.0 samples = 2 value = [1, 1] class = Survive
entropy = 0.0 samples = 1 value = [1, 0] class = Survive
entropy = 0.0 samples = 3 value = [0, 3] class = Deceased
entropy = 0.0 samples = 1 value = [1, 0] class = Survive
entropy = 0.0 samples = 1 value = [0, 1] class = Deceased

Figure 5. Decision Tree Visualization Using C4.5 Algorithm and ADASYN

As shown in Figure 5, the decision tree structure changed substantially after balancing. The root node shifted to Age ≤ 64, indicating that Age became the most decisive feature in the new training context. This is consistent with clinical knowledge, as age is a well-established factor influencing postoperative outcomes in lung cancer patients. Additional splits were made using Haemoptysis ≤ 0.5, FEV1 ≤ 0.122, and Tumor_Size ≤ 2.5, suggesting these features also play a significant role in differentiating between "Survive" and "Deceased" classes.

Each node in the tree displays entropy, number of samples, and class distribution, with several leaf nodes showing entropy = 0, indicating pure classification results. The overall tree is deeper and more balanced than before ADASYN, demonstrating the model's improved ability to recognize patterns from both classes.

The improved performance of the model is shown in the **confusion matrix in Figure 6**, which evaluates the model on the balanced dataset:

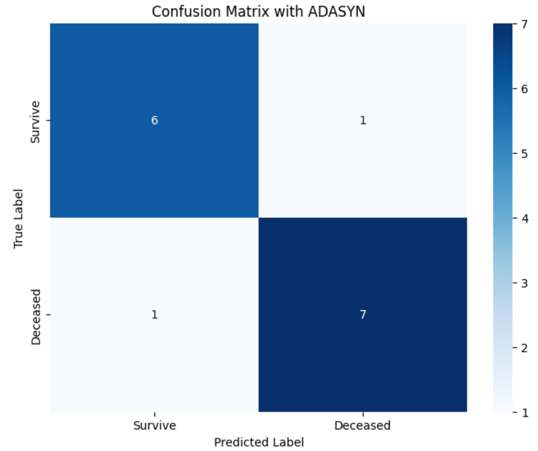

Figure 6. Confusion Matrix of C4.5 Model With ADASYN

From the matrix, we calculate the following performance metrics:

$$Accuracy = \frac{6+7}{6+1+1+7} = \frac{13}{15} = 0.87 = 87\% (14)$$

$$Precision = \frac{6}{6+1} + \frac{7}{7+1} = \frac{0.86+0.88}{2} = 0.87 = 87\% \qquad (15)$$

$$Recall = \frac{6}{6+1} + \frac{7}{7+1} = \frac{0.86+0.88}{2} = 0.87 = 87\% \ (16)$$

$$F1 - Score = 2 * \frac{0.87*0.87}{0.87+0.87} = 0.87 = 87\% (17)$$

3.4. Discussion of Misclassifications

Despite the overall improvement, two misclassifications remain:

- **1 false positive**: a “Deceased” patient was classified as “Survive”
- **1 false negative**: a “Survive” patient was classified as “Deceased”

From a clinical standpoint, **false negatives are more critical**, as they could result in missed intervention opportunities for patients at risk of death. However, the reduction in false negatives from **2 to 1** after ADASYN demonstrates meaningful progress in reducing this risk. Additionally, the new tree structure aligns more closely with **domain-relevant features** (Age, FEV1), suggesting improved model interpretability and potential for medical application.

4. CONCLUSION

This study aimed to develop a predictive model for the one-year survival of lung cancer patients following thoracic surgery, utilizing the C4.5 decision tree algorithm in combination with ADASYN to address data imbalance. The results demonstrated that this combined approach significantly improved the model’s performance across multiple evaluation metrics, including accuracy, precision, recall, and F1-score.

The most notable improvement was observed after applying ADASYN, where the model’s F1-score increased from 40% to 87%. This improvement highlights the importance of addressing data imbalance in medical datasets, particularly when the minority class (i.e., deceased patients) represents the more clinically critical group. The confusion matrices also showed a reduction in false negatives, which is essential in healthcare scenarios where failing to identify high-risk patients can lead to serious consequences.

When compared to prior studies using the same dataset, such as the work of Setyadi et al. (2020) with Genetic Algorithm + Naive Bayes (accuracy: 85.31%) and Prasetio & Susanti (2019) with KNN + AdaBoost (accuracy: 85.11%), the proposed method (C4.5 + ADASYN) achieved a higher accuracy of 87%. This demonstrates that the use of interpretable models like decision trees can still outperform more complex ensemble methods when paired with appropriate preprocessing techniques.

From the perspective of computer science and machine learning, this research contributes to the ongoing efforts in handling imbalanced classification problems—a major challenge in data mining and predictive analytics. The implementation of ADASYN within the clinical prediction pipeline reflects the importance of adaptive sampling algorithms in improving the robustness and fairness of classifiers.

Furthermore, the use of C4.5 adds value through its interpretability. Unlike black-box models such as deep learning or random forests, decision trees offer transparent logic that is essential in medical applications, where decision support systems must be explainable to practitioners. This aligns with the increasing emphasis on interpretable AI in both research and applied systems.

In summary, the findings of this study emphasize the dual importance of data preprocessing (balancing) and model interpretability in building effective and trustworthy predictive systems. The approach not only improved technical performance but also aligned with real-world clinical needs, making it suitable for further development as part of decision support tools in thoracic oncology.

ACKNOWLEDGEMENT

This research was supported by Lambung Mangkurat University and Gazi University, which provided essential resources and backing.

REFERENCES

[1] R. Patra, *Prediction of lung cancer using machine learning classifier*, vol. 1235 CCIS. Springer Singapore, 2020. doi: 10.1007/978-981-15-6648-6_11.

[2] I. Jabin and M. M. Rahman, “Predicting lung cancer survivability : A machine learning regression model Predicting lung cancer survivability : A machine learning regression model,” vol. 11, no. May, pp. 68–81, 2021.

[3] R. Rami-Porta, C. Wittekind, and P. Goldstraw, “Complete Resection in Lung Cancer Surgery: From Definition to Validation and Beyond,” *J. Thorac. Oncol.*, vol. 15, no. 12, pp. 1815–1818, 2020, doi: 10.1016/j.jtho.2020.09.006.

[4] N. Maleki, Y. Zeinali, and S. T. A. Niaki, “A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection,” *Expert Syst. Appl.*, vol. 164, p. 113981, Feb. 2021, doi: 10.1016/j.eswa.2020.113981.

[5] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable ai: A review of machine learning interpretability methods,” *Entropy*, vol. 23, no. 1, pp. 1–45, 2021, doi: 10.3390/e23010018.

[6] L. Indi, P. Aji, U. A. Yogyakarta, A. Sunyoto, and U. A. Yogyakarta, “An Implementation of C4 . 5 Classification Algorithm to Analyze Student ' s Performance,” pp. 5–9, 2021.

[7] A. Mailana, A. A. Putra, S. Hidayat, and A. Wibowo, “Comparison of C4.5 Algorithm and Support Vector Machine in Predicting the Student Graduation Timeliness,” *J. Online Inform.*, vol. 6, no. 1, p. 11, Jun. 2021, doi: 10.15575/join.v6i1.608.

[8] R. A. Saputra *et al.*, “Detecting Alzheimer’s Disease by the Decision Tree Methods Based on Particle Swarm Optimization,” *J. Phys. Conf. Ser.*, vol. 1641, no. 1, pp. 61–67, 2020, doi: 10.1088/1742-6596/1641/1/012025.

[9] A. Helisa, H. Saragih, I. Budiman, F. Indriani, D. Kartini, and T. H. Saragih, “Prediction of Post-Operative Survival Expectancy in Thoracic Lung Cancer Surgery Using Extreme Learning Machine and SMOTE,” *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 9, no. 2, pp. 239–249, 2023, doi: 10.26555/jiteki.v9i2.25973.

[10] M. Koziarski, “Radial-Based Undersampling for imbalanced data classification,” *Pattern Recognit.*, vol. 102, 2020, doi: 10.1016/j.patcog.2020.107262.

[11] B. Xue *et al.*, "Use of Machine Learning to Develop and Evaluate Models Using Preoperative and Intraoperative Data to Identify Risks of Postoperative Complications," *JAMA Netw. Open*, vol. 4, no. 3, p. e212240, Mar. 2021, doi: 10.1001/jamanetworkopen.2021.2240.

[12] G. Kovács, "Smote-variants: A python implementation of 85 minority oversampling techniques," *Neurocomputing*, vol. 366, pp. 352–354, 2019, doi: 10.1016/j.neucom.2019.06.100.

[13] Y. Cao, X. Zhao, Z. Zhou, Y. Chen, X. Liu, and Y. Lang, "MIAC: Mutual-Information Classifier with ADASYN for Imbalanced Classification," *2018 Int. Conf. Secur. Pattern Anal. Cybern.*, pp. 494–498, 2018.

[14] C. Kaope and Y. Pristyanto, "The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 22, no. 2, pp. 227–238, 2023, doi: 10.30812/matrik.v22i2.2515.

[15] A. Alam, D. A. F. Alana, and C. Juliane, "Comparison Of The C.45 And Naive Bayes Algorithms To Predict Diabetes," *Sinkron*, vol. 8, no. 4, pp. 2641–2650, 2023, doi: 10.33395/sinkron.v8i4.12998.

[16] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/j.gltp.2022.04.020.

[17] Y. Fakir, M. Azalmad, and R. Elaychi, "Study of The ID3 and C4.5 Learning Algorithms," *J. Med. Informatics Decis. Mak.*, vol. 1, no. 2, pp. 29–43, Apr. 2020, doi: 10.14302/issn.2641-5526.jmid-20-3302.

[18] N. G. Ramadhan, "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus," *Sci. J. Informatics*, vol. 8, no. 2, pp. 276–282, 2021, doi: 10.15294/sji.v8i2.32484.

[19] A. Alhudhaif, "A Novel Multi-class Imbalanced EEG Signals Classification Based on the Adaptive Synthetic Sampling (ADASYN) approach," *PeerJ Comput. Sci.*, vol. 7, pp. 1–15, 2021, doi: 10.7717/PEERJ-CS.523.

[20] T. M. Khan, S. Xu, Z. G. Khan, and M. U. Chishti, "Implementing multilabeling, ADASYN, and relieff techniques for classification of breast cancer diagnostic through machine learning: Efficient computer-aided diagnostic system," *J. Healthc. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/5577636.

[21] L. A. Sevastianov and E. Y. Shchetinin, "On methods for improving the accuracy of multi-class classification on imbalanced data," *CEUR Workshop Proc.*, vol. 2639, pp. 70–82, 2020.

[22] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "Multi-label Classifier Performance Evaluation with Confusion Matrix," pp. 01–14, 2020, doi: 10.5121/csit.2020.100801.

[23] N. E. Ramli, Z. R. Yahya, and N. A. Said, "Confusion Matrix as Performance Measure for Corner Detectors," *J. Adv. Res. Appl. Sci. Eng. Technol.*, vol. 29, no. 1, pp. 256–265, 2022, doi: 10.37934/araset.29.1.256265.

[24] S. Napi, T. Hamonangan Saragih, D. Turianto Nugrahadi, D. Kartini, and F. Abadi, "Implementation of Monarch Butterfly Optimization for Feature Selection in Coronary Artery Disease Classification Using Gradient Boosting Decision Tree," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 5, no. 4, pp. 314–323, 2023.

[25] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using EfficientNets – a first step towards preventing skin cancer," *Neurosci. Informatics*, vol. 2, no. 4, p. 100034, 2022, doi: 10.1016/j.neuri.2021.100034.