JOURNAL MEDICAL INFORMATICS TECHNOLOGY
medinftech.org

Optimizing Lung Cancer Prediction Using Evaluating Classification Methods and Sampling Techniques

Dika Putri Metalica[1] and Fahmi B Marasabessy[2]*

[1,2]*Faculty of Information Technology Universitas Nusa Mandiri, Jakarta*

ARTICLE HISTORY

Received: 13 March 23

Final Revision: 16 March 23

Accepted: 30 March 23

Online Publication: 31 March 23

KEYWORDS

Lung Cancer, Classification, Sampling Techniques, Gboost, Level Based

CORRESPONDING AUTHOR

14220003@nusamandiri.ac.id

DOI

10.37034/medinftech.v1i1.4

A B S T R A C T

Lung cancer is a highly malignant form of cancer and a leading cause of death worldwide. This research focuses on improving the detection and prediction of lung cancer by evaluating different classification methods and sampling techniques. The study utilizes a dataset consisting of 1000 patients and 24 attributes. The objective of this study is to compare the performance of classification methods such as Logistic Regression, AdaBoost, and GradientBoosting, along with different sampling techniques including random over-sampling, random under-sampling, and SMOTE by Level Considering for Lung Cancer prediction. The evaluation metrics used include accuracy, precision, recall, and F1-score. The experimental results reveal that Gradient Boosting (GBoost) achieves perfect accuracy, precision, recall, and F1-score values of 100% in identifying lung cancer cases within the dataset. This highlights the effectiveness of GBoost in accurately predicting lung cancer occurrence. The findings of this research aim to contribute significantly to the development of more effective diagnostic and predictive methods for lung cancer.

1. Introduction

Lung cancer is one of the deadliest diseases with an increasing incidence rate worldwide [1]. The survival rate and treatment outcomes for lung cancer patients greatly depend on the stage of cancer at the time of diagnosis. Therefore, it is crucial to develop accurate prediction models to identify the precise stages of lung cancer, thus providing guidance for appropriate clinical decision-making. The study involves a comprehensive review of relevant studies, testing classification methods, data processing techniques, and validation approaches used in creating specific prediction models for cancer stages. Algorithms such as Logistic Regression, AdaBoost and Gradient Boosting (Gboost) are discussed, with an emphasis on optimizing their performance using relevant clinical features and appropriate datasets.

The performance evaluation of the algorithms involves metrics such as accuracy, precision, recall, and F1-score, using a well-documented lung cancer dataset [2]. The results of this research will provide valuable insights into effective algorithm variants in predicting lung cancer risks, considering datasets that have three levels of risk: low, medium, and high. This will facilitate the development of more accurate prediction methods and personalization to enhance the management and treatment of lung cancer.

This research utilizes supervised data modeling in machine learning, consisting of algorithms used for comparison. The tested algorithms include Logistic Regression (LR), AdaBoost, and Gboost with sampling technique Random Under Sampling (RUS), Random Over Sampling (ROS) and SMOTE. Based on the Kaggle dataset, which consists of 1000 patient data with 23 attributes and the target level as a class [3].

2. Research Method

In this section, we will describe the dataset used and the main steps of the methodology adopted for predicting lung cancer risk. We will perform class balancing in the dataset to ensure a balanced distribution, as well as rank the features in the balanced data. Additionally, we will provide an overview of the frequency of occurrence of nominal features in relation to lung cancer classes. Furthermore, we will explain the machine learning models used and the performance metrics to be measured [4]. Figure 1 shows overview of the architecture in the prediction system.

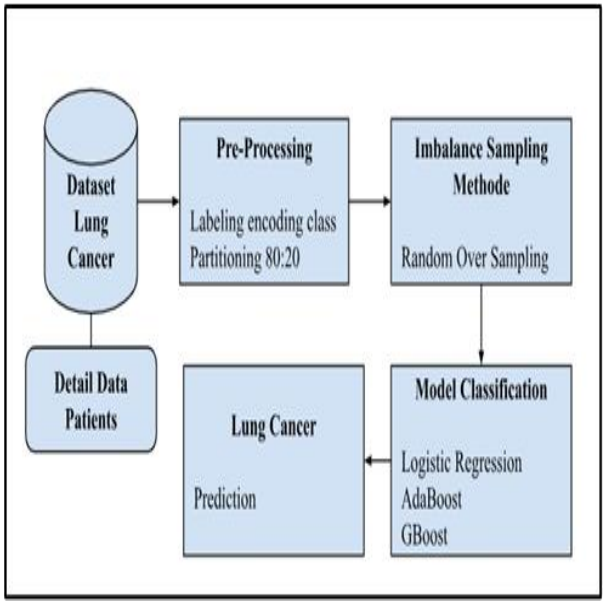

Figure 1. Block Diagram of Lung Cancer Prediction

2.1 Dataset

This research relies on a public dataset [3] with a total of 1000 patients, and all attributes (23 attributes as inputs for the ML model and 1 for the target class) are described as follows in Table 1.

Table 1. Selection Attribute Patients

No	Attribute	No	Attribute
1	Age	13	Chest Pain
2	Gender	14	Coughing of Blood
3	Air Pollution	15	Fatigue Levels
4	Alcohol use	16	Weight Loss
5	Dust Allergy	17	Shortness of Breath
6	Occupational Hazards	18	Wheezing
7	Genetic Risk	19	Swallowing Difficulty
8	Chronic Lung Disease	20	Clubbing of Finger Nails
9	Balanced Diet	21	Frequent Colds
10	Obesity	22	Dry Coughs
11	Smoking	23	Snoring
12	Passive Smoker	24	Level

2.2 Pre-Processing

In this stage, data preprocessing is performed before modeling. The data preprocessing stage involves cleaning or examining the data to identify invalid or missing values. After the data is checked using data cleaning techniques, feature selection is conducted to determine the most relevant features or attributes that will be used in the data modeling process. In the data cleaning stage, the data is examined to identify any missing values in each data point [5]

2.3 Data Partitioning

The stage of data splitting in the lung cancer prediction system involves data collection and determining the training and testing datasets. In this study, we allocate 80% of the dataset for training the model, while the remaining 20% is used as the testing dataset to evaluate the system's performance [6].

2.4. Data Modeling

In the data modeling stage, this study employs the Supervised Machine Learning method. Supervised Machine Learning is an algorithm that learns from labeled training data to assist in predicting outcomes for unseen data. In supervised learning, the machine is trained using correctly "labeled" data [7]. This can be compared to learning in the presence of a supervisor or teacher. The data is divided into training data and testing data, where the training data is used to train the machine, and the testing data is used to evaluate the machine's ability to provide accurate predictions. Supervised learning can be further divided into regression and classification methods [8].

2.5. Classification

The types of classification Machine Learning algorithms used for predicting lung cancer include Logistic Regression, AdaBoost, and Gboost with RUS, ROS and SMOTE for Technique Sampling.

a. Logistic Regression (LR)

Logistic regression is used to estimate parameters of the logistic model and is commonly used in binary classification tasks. Logistic regression estimates the probability of a binary response based on one or more independent variables. In the field of machine learning, logistic regression is widely used in various domains, including the medical field [2] Below is the logistic regression algorithm mathematical Equation (1) [9].

$$i=\text{Logistic regression } (p)=1n(p/(1-p)) \quad (1)$$

b. AdaBoost

AdaBoost is a powerful ensemble learning algorithm that can effectively leverage a small training dataset by dynamically adjusting the weights assigned to each training example. The algorithm follows a series of steps to iteratively train weak classifiers and combine them into a strong classifier [10] These steps can be summarized as follows in Algorithm 1.

Algorithm 1: AdaBoost Pseudocode [15]

Input: Let D be the dataset that includes $\{(a_1,b_1), (a_2,b_2), \ldots .. (am, bm)\}$;
Let λ be the learning (base) algorithm
Let T be the total No. of learning rounds.
Process:
D1(i) = 1/m
for time = 1, . . . , T;
ht = λ (D, Dt); weak learner is trained with Distribution Dt
∈t = PrPri~Dt [ht(ai ≠ bi)]; Error measure (entropy)

$\propto_t = \frac{1}{2} 1n \left(\frac{1-\epsilon t}{\epsilon t}\right)$; % determine the weight of h_t

$Dt+1(i) = \left(\frac{Dt(t)}{Zt}\right) * \begin{cases} \exp(-\propto t) \; if \; ht\,(ai) = bt \\ \exp(\propto t) \; if \; ht\,(ai) \neq bt \end{cases}$

$$\frac{Dt(t)\exp(-\propto tytht(ai)}{Zt}$$

Outcome: H(a) = sign $\sum_{t=1}^{T} \propto t\, ht(b)$

c. Gboost

Gradient boosting is a boosting-like algorithm for regression. Gradient boosting builds an additive approximation of F*(x) as a weighted sum of functions. formulas calculation it can be written using Equation (2), Equation (3), Equation (4), and Equation (5) [11].

$$F_m(\boldsymbol{x}) = F_{m-1}(x) + \rho_m h_m(x), \quad (2)$$

$$F_0(\mathbf{x}) = \arg min_{\alpha} \sum_{i=1}^{N} L(y_i, a). \quad (3)$$

$$(\rho_m, h_m(\mathbf{x})) = \arg min_{p,h} \sum_{i=1}^{N} L(y_i F_{m-1}(x_i) + \rho h(x_i)) \quad (4)$$

$$R_{mi} = \left[\frac{\partial L(y_i, F(x))}{\partial F(x)}\right] F_{(x)} = F_{m-1}(x) \quad (5)$$

d. RUS

RUS technique works by randomly eliminating data points from the majority class, either with or without replacement, until the proportion between the two classes is balanced. This approach involves randomly removing instances belonging to the majority class in the target variable until their quantity matches that of the minority class [12], formulas calculation it can be written using Equation (6).

$$x_{new} = x_i + (x\hat{}_i - x_i) * \delta \quad (6)$$

e. ROS

Random Oversampling (ROS) technique is utilized to address the challenge of imbalanced datasets. This method involves replicating instances from the minority class to achieve a balanced class distribution [13]. The process entails randomly selecting instances from the minority class and adding them to the dataset until the class distribution is leveled out.

f. SMOTE

The SMOTE algorithm finds application in various domains, such as network intrusion detection systems, where it effectively addresses the challenge of imbalanced data by mitigating the class imbalance problem. SMOTE technique synthetically increase the minority class using Equation (7) [14].

$$x_{syn} = x_i + (x_{knn} - x_i) * t \quad (7)$$

the formulas calculation it can be written using equation

2.6. Feature Analysis

Based on the observed dataset, Figure 4 shows the distribution of patients by age groups. The dataset indicates that lung cancer predominantly affects individuals between the ages of 24 to 53, with the age group of 34-38 having the highest frequency in Figure 2.

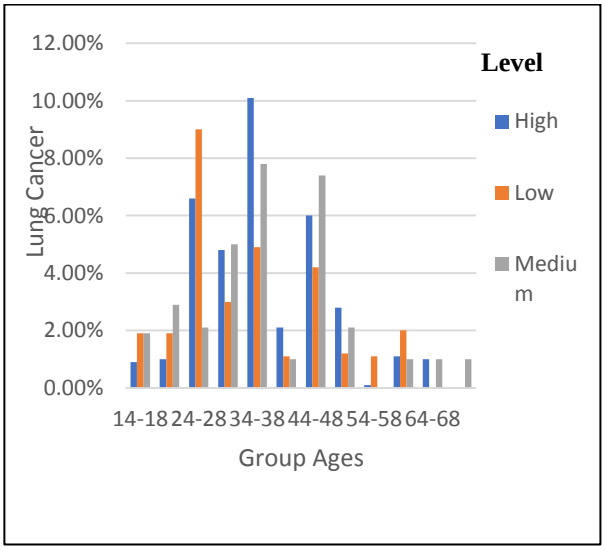

Figure 2. Distribution of Patients by Age Groups in the Balanced Data.

3. Result and Discussion

To evaluate the performance of the machine learning model, several metrics are used, including accuracy, precision, recall, and F-measure. These metrics are evaluated using a 3 x 3 confusion matrix, as seen in the accompanying Figure 5, which consists of four elements: true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [15]. Accuracy measures how well the model can predict the overall data correctly. Additionally, recall measures the model's ability to accurately identify positive cases or true positives, compared to the total number of actual positive cases. Precision is a measure of the model's quality in providing accurate predictions, while recall

is a measure of the quantity of predictions made by the model. F-measure is a harmonic value that combines precision and recall, providing an overall evaluation of the model using a single score in Figure 5 [16] the formulas calculation it can be written using Equation (8), Equation (9), Equation (10), and Equation (11).

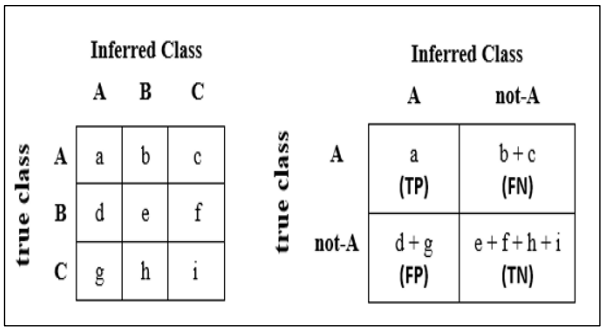

Figure 3. Illustrates an example of a 3x3 confusion matrix depicting classes A, B, and C on the left side. On the right side, there is the corresponding binary confusion matrix specifically for class A.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (8)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \quad (10)$$

$$\text{F1} - \text{Score} = 2 \text{ x } \frac{\text{Recall x Precission}}{\text{Recall} + \text{Precission}} \quad (11)$$

Table 2. Selection Accuracy Result (%)

Sampling Technique	LR	AdaBoost	GBoost
Without Imbalance Sampling	99.50	72.50	100
RUS	99.50	71.00	100
ROS	100.,00	68.50	100
SMOTE	99.50	71.00	100

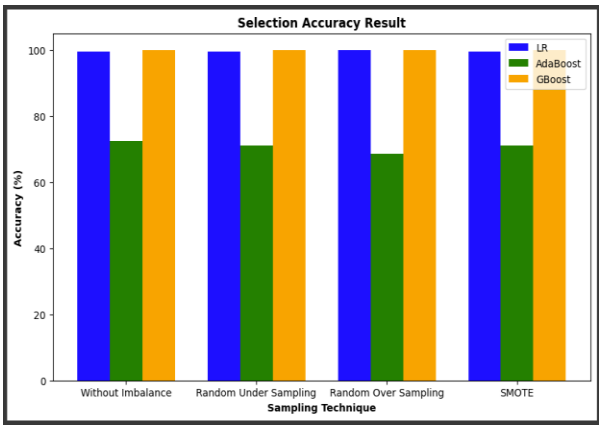

Figure 4. The Accuration of the Classification Methods for Predicting Lung Cancer.

Table 3. Selection Precision Result (%)

Sampling Technique	LR	AdaBoost	GBoost
Without Imbalance Sampling	99.51	57.82	100
RUS	99.51	77.27	100
ROS	99.51	53.82	100
SMOTE	99.51	77.27	100

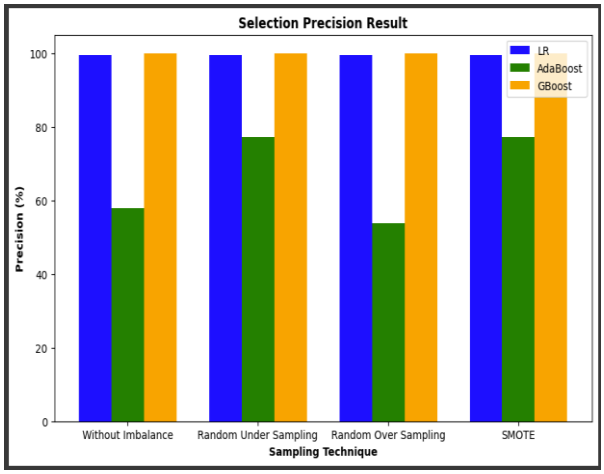

Figure 5. The Precision of the Classification Methods for Predicting Lung Cancer.

Table 4. Selection Recall Result (%)

Sampling Technique	LR	AdaBoost	GBoost
Without Imbalance Sampling	99.50	72.50	100
RUS	99.50	71.00	100
ROS	100.00	68.50	100
SMOTE	99.50	71.00	100

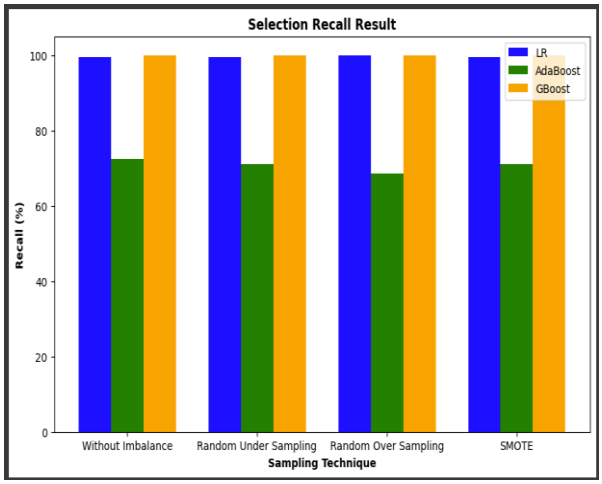

Figure 6. The Recall of the Classification Methods for Predicting Lung Cancer.

Table 5. Selection F1 Score Result (%)

Sampling Technique	LR	AdaBoost	GBoost
Without Imbalance Sampling	99.50	62.93	100
RUS	99.50	65.47	100
ROS	100.00	58.49	100
SMOTE	99.50	65.47	100

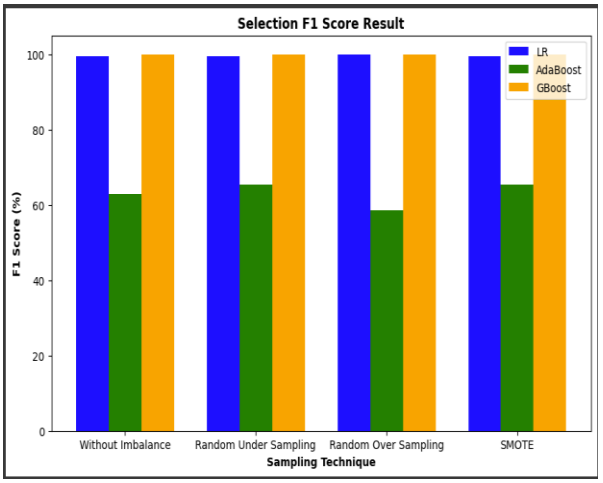

Figure 7. The F1-Score of the Classification Methods for Predicting Lung Cancer.

Table 6. Comparison Selection with Related Research (%)

Model	Accuracy	Recall	Precision	F1-Score
Rikta [17]	98.76	98.79	98.76	98.76
Yadaf [18]	87.00	98.00	88.00	93.00
Proposed Method	100.00	100.00	100.00	100.00

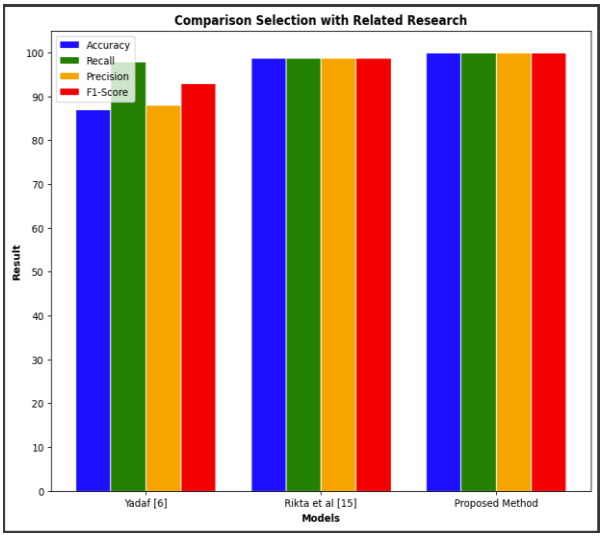

Figure 8. Comparison Selection with Existing Study.

4. Conclusion

Based on the research findings presented in Table VI, it can be concluded that Gboost, when combined with all sampling techniques, achieved the highest performance in predicting lung cancer occurrence. This indicates that Gboost is a powerful algorithm for the task of predicting lung cancer using machine learning on the Kaggle dataset with 1000 data patients. The results of the study can be show in Table II-V that Gboost consistently achieved perfect scores in accuracy, precision, recall, and f1-score, which were 100% for all sampling techniques. This demonstrates the reliability and effectiveness of Gboost in classifying lung cancer cases with high accuracy. This highlights the potential of Gboost as a reliable and accurate tool for lung cancer prediction. It should be emphasized that the choice of the optimal algorithm is influenced by several factors, including specific task demands, dataset size and properties, and computational constraints. Therefore, further analysis and evaluation may be necessary to determine the most suitable algorithm for specific scenarios.

References

[1] Singh, G.A.P. and Gupta, P.K., (2019). Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans. Neural Computing and Applications, [online] 31(10), pp.6863–6877. https://doi.org/10.1007/s00521-018-3518-x.

[2] S. Bharati, P. Podder, R. Mondal, A. Mahmood, and M. Raihan-Al-Masud, (2020). Comparative Performance Analysis of Different Classification Algorithm for the Purpose of Prediction of Lung Cancer, vol. 941. Springer International Publishing. doi: 10.1007/978-3-030-16660-1_44.

[3] Guslovesmath, 2022. “Lung Cancer Prediction (ML)”. https://www.kaggle.com/code/guslovesmath/lung-cancer-prediction-ml/input.

[4] E. Dritsas and M. Trigka, (2022). “Lung Cancer Risk Prediction with Machine Learning Models,” Big Data Cogn. Comput., vol. 6, no. 4, doi: 10.3390/bdcc6040139

[5] Md, Abdul Quadir et al. 2023. “Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease.” Biomedicines 11(2). doi: 10.3390/biomedicines11020581

[6] D. Yadav, (2022). “Lung Cancer Prediction Using Supervised Ml Algorithms,” Int. Res. J. Mod. Eng. Technol. Sci., no. 10, pp. 293–298, doi: 10.56726/irjmets30472

[7] D. Bansal, R. Chhikara, K. Khanna, and P. Gupta, (2018). “Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia,” Procedia Comput. Sci., vol. 132, pp. 1497–1502, doi: 10.1016/j.procs.2018.05.102.

[8] L.J.Muhammad, E. A.Algehyne, S. S. Usman, A. Ahmad, C. Chakraborty, and I.A.Mohammed, “Supervised Machine Learning Models for Prediction of COVID- 19.pdf.” 2021, https://doi.org/10.1007/s42979-020-00394-7.

[9] Ibrahim and A. Abdulazeez, (2021). “The Role of Machine Learning Algorithms for Diagnosing Diseases,” J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 10–19, doi: 10.38094/jastt20179.

[10] A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), Gorakhpur, India, 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.

[11] Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz. 2021. 54 Artificial Intelligence Review A Comparative Analysis of Gradient Boosting Algorithms. Springer Netherlands. doi:10.1007/s10462-020-09896-5

[12] Geetha, R. et al. 2019. “Cervical Cancer Identification with Synthetic Minority Oversampling Technique and PCA Analysis Using Random Forest Classifier.” Journal of Medical Systems 43(9). doi: 10.1007/s10916-019-1402-6.

[13] A. Priya, S. Garg, and N. P. Tigga, (2020). “ScienceDirect ScienceDirect Predicting Anxiety , Depression and Stress in Modern Life using Predicting Anxiety , Depression and Stress in Modern Life using Machine Learning Algorithms Machine Learning Algorithms,” Procedia Comput. Sci., vol. 167, no. 2019, pp. 1258–1267, doi: 10.1016/j.procs.2020.03.442.

[14] Ding, Yi, Hongyang Zhu, Ruyun Chen, and Ronghui Li. 2022. "An Efficient AdaBoost Algorithm with the Multiple Thresholds Classification." Applied Sciences (Switzerland) 12(12). doi: 10.3390/app12125872

[15] S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," Behav. Processes, vol. 148, pp. 56–62, 2018, doi: 10.1016/j.beproc.2018.01.004.

[16] Ramakrishna, Mahesh Thyluru et al. 2023. "Homogeneous Adaboost Ensemble Machine Learning Algorithms with Reduced Entropy on Balanced Data." Entropy 25(2). doi:10.3390/e25020245

[17] Rikta, Sarreha Tasmin et al. 2023. "XML-GBM Lung: An Explainable Machine Learning-Based Application for the Diagnosis of Lung Cancer." Journal of Pathology Informatics 14(March). doi: 10.1016/j.jpi.2023.100307

[18] Wongvorachan, Tarid, Surina He, and Okan Bulut. 2023. "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining." Information (Switzerland) 14(1). doi: 10.3390/info14010054