Deep neural networks and conventional machine learning classifiers to analyze thoracic survival data

Cucu Ika Agustyaningrum[1], Yudi Ramdhani[2], Doni Purnama Alamsyah[3], Oda I. B. Hariyanto[4]
[1]Faculty of Engineering and Informatics, Universitas Bina Sarana Informatika, Jakarta, Indonesia
[2]Department of Information System, Faculty of Creative Technology, Satu University, Bandung, Indonesia
[3]Department of Entrepreneurship, BINUS Business School, Bina Nusantara University, Jakarta, Indonesia
[4]Department of Tourism, Batam International University, Batam, Indonesia

Article Info

Article history:

Received Aug 13, 2023
Revised Feb 20, 2024
Accepted Mar 15, 2024

Keywords:

Algorithm
Area under curve
Conventional machine learning
Deep neural network
Pythons
Toraric survival data

ABSTRACT

Lung cancer is a prevalent global health concern and most prevalent malignancy in Indonesian hospitals. Following thoracic surgery, patients were categorized into two classes: individuals who experienced mortality within a year and those who achieved survival. Despite being about socks, the dataset for the deceased category consisted of 70 data samples, while the dataset for the final group comprised 400 samples. Data calculation involves the utilization of both deep neural networks and standard machine learning algorithms. The study use the Python programming language to evaluate the algorithms, and it measures their performance using metrics such as accuracy, F1-score, precision, recall, receiver operating characteristic (ROC), and area under curve (AUC). The test results indicate that the deep neural network method achieves an accuracy of 95.56%, an F1 score of 79.24%, a precision of 91.96%, a recall of 85.52%, and an AUC of 85.52%. This study suggests that utilizing deep neural network data mining techniques, specifically with a cross-validation fold of 10, variations of six hidden layer encoder-decoder, relu, sigmoid activation function, optimizer Adam, and learning rate of 0,01, dropout rate of 0,2. Employing the synthetic minority over-sampling technique data preprocessing method, can effectively analyze thoracic patient survival data sets.

Corresponding Author:

Doni Purnama Alamsyah
Department of Entrepreneurship, BINUS Business School, Bina Nusantara University
Jakarta 11480, Indonesia
Email: doni.syah@binus.ac.id

1. INTRODUCTION

According to the Indonesian sample registration system (SRS) report from 2014, 10 diseases are responsible for most fatalities in Indonesia. Smoking cigarettes, both actively and passively, increases the risk of developing lung cancer. Air pollution and exposure to the workplace are additional risk factors [1]. Because of the high prevalence of smoking in society, lung cancer will become a public health issue in Indonesia. Lung cancer is the fourth most common cancer detected in hospitals in Indonesia [2]. Surgery, radiation, chemotherapy, immunotherapy, hormone therapy, and gene therapy can all be used to treat lung cancer. The airways are almost entirely located on the ribs. The thorax is essential in the breathing process [3]. Thoracic surgery is one of the most common treatments for lung cancer. However, thoracic surgery carries several risks and complications, including neurological disorders, infections, and life-threatening complications. Many complications occur in people with cardiovascular disease, including heart and blood vessel problems that can lead to strokes. As a result, the life expectancy following thoracic surgery is extremely low [4]. The research focus on predicting thoracic patient survival using a computer-aided diagnosis (CAD) system. The analysis of the patient's condition before and after surgery, CAD can help predict the life expectancy of patients with lung cancer. Data collection for patients with lung cancer undergoing thoracic surgery [5]. Following thoracic surgery, patients were divided into two groups: those who died within a year (die) and those who were able to survive. A total of 70 samples of data were included in the dataset for the die class and 400 samples were included in the dataset for the latter group [5]. The performance of the prediction model for the survival of thoracic patients may be affected by the class imbalance issue in the algorithm for detecting lung cancer in the thoracic dataset [6]. Because there are more false positives than true positives in the thoracic surgery dataset, there is a class imbalance. Bootstrap aggregating is a classification improvement technique that uses a random combination of classifications on training and bagging datasets to reduce variance and avoid overfitting [7]. Previously, several researchers used various algorithms to predict the survival of thoracic patients in previous studies, including the multi-class support vector machine algorithm with active learning for network traffic classification [8]. Lung cancer classification using support vector machine and neural networks[9]. The medical internet of things uses machine learning [10]. Based on unbalanced data, a comprehensive data-level analysis for cancer diagnosis is performed [11]. Previous study indicates that the deep neural network technology has not been employed thus far, with only traditional machine learning methods being utilized. The objective of this study is to compare conventional machine learning and deep neural network algorithms in order to determine the most effective approach for assessing potential online shopper intentions using the Python programming language. The aim is to obtain the most accurate results for doctors and medical services in receiving thoracic patient survival data.

2. METHOD

2.1. Research stages

The research on thoracic surgical data involved the utilization of a model constructed through a systematic research method. This approach encompassed various steps, including dataset acquisition, preprocessing, feature selection, smoothing, modeling, and evaluation. The research commences with gathering data from the UCI machine learning repository website [12]. Subsequently, the data undergoes a transformation to generate initial data for preprocessing. Following this, features are chosen utilizing the Python programming language. The data is then balanced using the synthetic minority over-sampling technique (SMOTE) method. Finally, the data is evaluated using conventional machine learning algorithms and deep neural networks through data cross-validation. During this modeling phase, two techniques will be evaluated: traditional machine learning and deep neural networks. Prior to inputing the modeling procedure, the data is gathered using the SMOTE method, a data augmentation methodology that ensures a balanced representation of both one-year survival and mortality possibilities. After a seamless procedure, the data undergoes cross-validation, a model validation approach employed to evaluate the accuracy of the analytic outcomes. Cross-validation is a technique used to do classification on preprocessed data. It involves partitioning the data into separate training and testing sets [13]. Once the data has undergone preprocessing, feature selection, smoothing modeling, and testing, the last stage involves analyzing the data produced by the Python programming language using traditional machine learning methods and deep neural networks as shown in Figure 1.

2.2. Application of research methods

2.2.1. Dataset

The main goal of this research is to examine how research approaches can be practically applied in the setting of the thoracic surgery data set. To accomplish this goal, we painstakingly divided our study into six distinct but related steps, each of which added to the overall analytical process. The data utilized is derived from secondary sources, namely obtained from UCI machine learning. The dataset contains 470 data points, each with 17 parameters and 1 class. Its purpose is to analyze and determine the survival statistics for patients with thoracic conditions [12]. In order to overcome these challenges and offer more convenient, efficient, and precise outcomes, classification methods that exhibit superior levels of forecasting and precision can be employed. This study employed a comparison between the deep neural network algorithm and the classic machine learning algorithm to yield highly predictive and accurate findings.

2.2.2. Pre-processing

A total of 470 patients' data were gathered during this phase of the trial, and they will be examined to predict the patients' survival and mortality rates after one year. The initial step in the data preparation stage is the selection or examination of attributes that will modify the data type. The data cleaning process commences once the data selection phase is completed. During this technique, our objective is to identify any absent values.

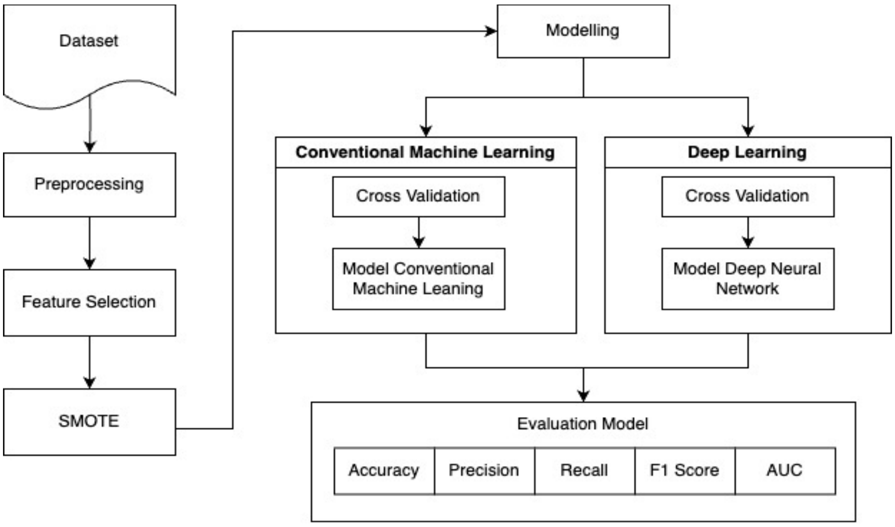

Figure 1. Research stages

2.2.3. Feature selection

During the pivotal feature selection stage, our primary goal is to identify the attributes that have the greatest influence on the dataset's dynamics. This determination is made using a thorough feature analysis, with a particular emphasis on the feature option. In accordance with our research objectives, we make the strategic decision to incorporate all attributes into our modelling process, apart from the class attribute. This deliberate choice is based on the expectation that including these attributes will significantly improve our model's performance metrics, which include accuracy, F1 score, precision, recall, and area under curve (AUC) measurement. This meticulous feature selection process is critical for fine-tuning our models and ensuring their efficacy in extracting valuable insights from the dataset.

2.2.4. Smooth

Following the critical feature selection phase, the next step in our research is to address class imbalance within the dataset. This is accomplished by utilizing the SMOTE, a method designed to equalize class distribution. Simultaneously, a critical step is taken during the modelling stage to ensure the dependability and robustness of our analysis. Cross-validation techniques are used in both deep neural network models and traditional machine learning methods to achieve this. We reduce the risk of overfitting by splitting the data into training and testing sets in a systematic manner, allowing our models to generalize effectively to previously unseen data and providing a solid foundation for our research findings.

2.2.5. Modelling

The proposed approach is used to carry out the prediction procedure at the modelling stage. The suggested method uses deep neural networks and conventional machine learning. Deep neural networks use three to eight layers and the Python programming language to assess the level of accuracy, F1 scores, precision, recall, and AUC from live thoracic patients. Conventional machine learning includes several methods, such as random forest, support vector machine, and logistics regression.

- Machine learning: machine learning is the automatic recognition of significant patterns in data. Computers can learn things from people through machine learning. The computer can learn to process the data that is supplied to it without any explicit programming. Algorithms for machine learning are used to train computers to process data [14].
- Random forest: the random forest concept involves the generation of several decision trees that are correlated, where each decision tree functions as a collection of models. Every decision tree creates class predictions, and the ultimate decision is determined by the highest yield [15]. The random forest classification method employs a decision tree approach, where attributes are randomly selected at each node to decide categorization. The decision tree utilizes the largest number of votes it receives to classify data [16], [17]. Random forest employs a voting system, namely the highest count, to merge classifiers (CARTs) that are mutually independent and originate from the same distribution. This process results in classification predictions. Decreased correlation can diminish the impact of prediction errors in random forest, which is an inherent characteristic of this algorithm [18]. Random forest formula [19]:

$$=\text{Entropy }(Y)=-\ \sum_{i} \square\ P\ (Y)\log^{2}p(Y), \qquad (1)$$

$$=\text{Entropy }(Y)-\ \sum_{V} \square\ \varepsilon\text{values}(a)\ \frac{|Y_v|}{|Y_a|}\ \text{Entropy }(Y_v).$$

Information:
Y=case set
P(c|Y) is the ratio of grades in class Y to those in class c.
Values(a)=Possible values when a is set.
Yv=subclass of Y with class v, which is related to class a.
Ya=All values that correspond to a.

- Support vector machine: because it requires specific learning objectives during training, support vector machine is an integrated (supervised) classification method [20]. The following is the support vector machine formula [21]:

$$\text{similarity} = \frac{\sum_{i=1}^{n} f(T_i, S_i)}{W_i} \qquad (2)$$

Information:
T: A new case
S: cases in storage
n: the number of attributes
I: individual attribute between 1 and n
f: TRIBUTE similarity function between case T and case S
W: weight assigned to the i-th attribute

- Logistics regression: the supervised classification algorithm includes logistic regression. This algorithm has grown in popularity in recent years, and its application has expanded significantly. This is a sigmoid curve. It is a subset of logistic regression. Next step is start with a simple linear regression formula to understand the mathematical version of the explanation [22].

$$y=b0+b1*x \qquad (3)$$

Thus, it has now been subjected to the sigmoid function, and the result is provided by the formula.

$$p=\frac{1}{1+e^{-y}} \qquad (4)$$

Now that one formula has been substituted for another to get the value of y, we have our logistic regression formula.

$$\text{logistic}(S)=b0+b1M1+b2M2+b3M3\ldots bkMk\ldots \qquad (5)$$

where S denotes the likelihood of the presence of interesting features. The predictor values are M1, M2, M3, ... Mk. The intercepts of the model are b0, b1, b2, b3, ... bk.

- Deep neural network: the deep neural network is a complex artificial neural network consisting of multiple layers. A deep neural network often consists of more than three layers, including an input layer, hidden layers, and an output layer. This architecture classifies it as a multilayer perceptron (MLP) with multiple layers. The depth is attributed to the multitude of layers. Deep learning refers to the process of learning in deep neural network [23]. Deep neural network belongs to the category of neural networks. Deep neural network consists of multiple hidden units that are interconnected across layers, but not inside each individual unit within a layer. This approach employs supervised training and shares a comparable structure with an artificial neural network. The deep learning approach can optimize numerous parameters for speech recognition. deep neural network possesses the ability to accurately identify and interpret speech, exhibiting enhanced efficiency in understanding various languages and dialects [24].
- Confusion matrix: a particularly helpful tool for examining bias in the recognition of tuples of various types is the confusion matrix [25]. This technique makes use of a matrix array with positive and negative classes [26]. The values of accuracy, precision, recall, and error rate will be obtained during the evaluation stage using the confusion matrix. The accuracy ratio is the number of correctly identified cases divided by the total number of cases. The following formula can be used to calculate accuracy:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

The proportion of cases with a true positive result is referred to as precision.

$$\text{Precision}=\frac{TP}{TP+FP} \quad (7)$$

The proportion of correctly identified positive cases is referred to as recall.

$$\text{Recall}=\frac{TP}{TP+FN} \quad (8)$$

Information:
The number of positive cases is designated as positive.
The number of negative cases is considered positive.
The number of negative cases is classified as such.
The number of positive case fields is considered negative.

2.2.6. Evaluation

During the evaluation phase, we conduct a comprehensive assessment of the model's performance using Python. This involves examining important metrics such accuracy, F1 score, precision, recall, and success or mistake rates. This study utilizes two separate algorithms: conventional machine learning and deep neural networks, enabling us to completely evaluate the usefulness and efficiency of the model.

2.3. Method of collecting data

Primary data and secondary data are the two categories of data sources that are available for utilization in the process of data collection. Primary data are the ones that are obtained directly, whereas secondary data are the ones that are obtained from other researchers who have already carried out an investigation that is comparable. To complete this investigation, researchers looked at secondary data. Thoracic surgery data set results from UCI machine learning were used for the research. These results included 470 records with a total of 17 attributes and 1 class attribute [27], as stated in Table 1. Table 2 contains categories of user behavior analysis.

Table 1. Description of the attributes of the survival dataset of patients with thorax

No	Attributes	Data type	Description
1	DGN	Category	Diagnosis: specific combination of ICD-10 codes for primary and secondary tumors and multiples if present (DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1).
2	PRE4	Numeric	Forced vital capacity - FVC
3	PRE5	Numeric	Volume that has been exhaled at the end of the first second of a forced expiration, FEV
4	PRE6	Category	Performance status, Zubrod scale (PRZ2, PRZ1, PRZ0)
5	PRE7	Category	Pain before surgery (T, F)
6	PRE8	Category	Hemoptysis before surgery (T, F)
7	PRE9	Category	Dyspnea before surgery (T, F)
8	PRE10	Category	Cough before surgery (T, F)
9	PRE11	Category	Weakness before surgery (T, F)
10	PRE14	Category	T in clinical TNM: original tumor size, from OC11 (smallest) to OC14 (largest) (OC11, OC14, OC12, OC13)
11	PRE17	Category	Type 2 DM, diabetes mellitus (T, F)
12	PRE19	Category	MI for up to six months (T, F)
13	PRE25	Category	PAD: peripheral arterial disease (T, F)
14	PRE30	Category	Smoking (T, F)
15	PRE32	Category	Asthma (T,F)
16	AGE	Numeric	Age at surgery (numeric)
17	Risk1Y	Category	1 year survival period - (T) true value if dead (T,F)

Table 2. Numerical attributes and categories of user behavior analysis

No	Nama atribute	Min. Value	Max. Value	SD
1	PRE4	1.44	6.30	0.87
2	PRE5	0.96	86.30	11.77
3	AGE	21.00	87.00	8.71

3. RESULTS AND DISCUSSION

Research data on the survival of thoracic patients is secondary data, where the total data is 470 with 17 attributes and 1 class. In a previous study that discussed the survival of thoracic patients conducted by Zięba *et al.* [5] with focused on boosted support vector machine to prediction of the post-operative life expectancy in the lung cancer patients. This study aims to determine the best results using the boosted support vector machine algorithm to test each predicted classification for the survival of thoracic patients. In a study, the experimental process used two methods of class formation, namely the minority class and the majority class. The study's conclusions consist of calculations derived from both qualitative and quantitative processing methods, which are based on the proposed model. All available datasets were utilized for this investigation. This study employs deep neural networks and conventional machine learning approaches to predict data sets through trials and testing. The datasets utilized in this experiment have undergone preprocessing, feature selection, smoothing, and modeling using Python in Google Collaboratory.

3.1. Pre-processing step validation

During the validation process of survival data for thoracic patients, we obtained a set of noteworthy findings during the course of our study, which involved meticulous data preprocessing, as shown in Table 3. When comparing the performance of conventional machine learning algorithms, the random forest algorithm stood out, boasting an impressive accuracy rate of 92 percent, an F1 score of 91.06%, a precision of 92.94%, and a recall rate of 88.17%. Surprisingly, these results outperformed those of the support vector machine algorithm and another regression logistics, albeit by only 1-2 percent.

Table 3. Preprocessing value results of a conventional machine learning algorithm

Model	Accuracy (%)	F1 Score (%)	Precision (%)	Recall (%)
Random forest	92	91.06	92.94	88.17
Support vector machine	88	81.67	79.06	85.87
Logistic regression	87	-	-	-

3.2. Conventional machine learning algorithm model

This study employed various conventional machine learning methods, such as random forest, support vector machine, and logistic regression. This work use conventional machine learning methods to compute accuracy, F1 score, precision, and recall metrics for analyzing survival data in thoracic patients. Table 4 presents the test results of many conventional machine learning techniques that have been employed, facilitating a comparison of the accuracy, F1 score, precision, and recall values. When the conventional machine learning algorithm is applied to the survival data of thoracic patients, the random forest algorithm is generated. This random forest algorithm has outstanding performance metrics, including a 92% accuracy rate, a 91.06% F1 score, a precision level of 92.94%, and a recall rate of 88.17%. Notably, these performance metrics outperform the support vector machine algorithm and another regression logistics.

Table 4. The results of a comparison of thorax conventional machine learning algorithm

Model	Accuracy (%)	F1 score (%)	Precision (%)	Recall (%)
Random forest	92	91.06	92.94	88.17
Support vector machine	88	81.67	79.06	85.87
Logistic regression	87	-	-	-

3.3. Deep neural network algorithm

Several techniques are used in this study, beginning with standard machine learning algorithms and progressing to a deep neural network algorithm with a variable number of hidden layers ranging from 3 to 8. We conducted extensive testing with multiple iterations of the deep neural network method, resulting in a thorough comparison of various performance metrics. Accuracy, F1 score, precision, recall, and AUC values are among the metrics presented in Table 5 for a thorough analysis and evaluation of our models' effectiveness. The test results, as presented in Table 5, demonstrate that the optimization of the deep neural network algorithm using different variations of 6 hidden layer decoders (encoder, activation function parameter sigmoid, optimizer Adam), with a learning rate of 0.01 and dropout rate of 0.2, yields higher values compared to the deep neural network algorithm with alternative variations.

3.4. Comparison model

The results of each algorithm are meticulously documented in Table 6, revealing a compelling pattern of performance. Notably, deep neural network methods are significantly more effective for research purposes when compared to traditional machine learning algorithms. In line with previous studies, it has been demonstrated that deep neural networks improve the capabilities of intrusion detection systems and optimize classification algorithms [28], [29]. A thorough examination of the test results, including an examination of both the confusion matrix and the receiver operating characteristic (ROC) curve, elucidates the specific conditions under which this superiority is most apparent. The deep neural network algorithm was optimized, with variations of 6 hidden layers in a decoder-encoder architecture, an activation function parameter set to sigmoid, the Adam optimizer with a learning rate of 0.01, and a dropout rate of 0.2. This configuration produces impressive performance metrics, including an accuracy rate of 95.56%, an F1 score of 79.24%, a precision level of 91.96%, a recall rate of 85.52%, and an AUC of 85.52%. In stark contrast, the conventional machine learning model, specifically the random forest algorithm, achieves an accuracy rate of 92.00%, an F1 score of 91.06%, a precision of 92.94%, and a recall of 88.17%. These figures show a noticeable average difference in accuracy of 3.56%, an F1 score difference of 11.82%, a precision difference of 0.98%, and a recall difference of 2.65%, highlighting the significant performance gains offered by the deep neural network approach.

Table 5. Results of the deep neural network algorithm's accuracy, precision, recall, F1-score, and AUC scores

Layers	Activation function	Optimizer	Learn rate	Accuracy (%)	F1 score (%)	Precision (%)	Recall (%)	AUC (%)
3 Hidden Decod-Encod	Sigmoid	RMSProop	0.001	93.95	65.11	96.79	74.13	74.13
3 Hidden Encod-Decod	Sigmoid	Adagrad	0.1	93.14	58.53	96.39	70.68	70.68
4 Hidden Decod-Encod	Sigmoid	Adam	0.01	94.35	70.83	92.11	78.85	78.85
4 Hidden Encod-Decod	Sigmoid	RMSProop	0.01	92.33	55.81	89.22	70.23	70.23
5 Hidden Decod-Encod	Sigmoid	Adam	0.1	94.75	73.46	92.58	80.57	80.57
5 Hidden Encod-Decod	Sigmoid	Adam	0.1	94.75	72.34	94.61	79.08	79.08
6 Hidden Decod-Encod	Sigmoid	Adam	0.01	95.56	77.55	95.30	82.53	82.53
6 Hidden Encod-Decod	Sigmoid	Adam	0.1	95.56	79.24	91.96	85.52	85.52
7 Hidden Decod-Encod	Sigmoid	Adam	0.01	95.56	76.59	97.60	81.03	81.03
7 Hidden Encod-Decod	Sigmoid	Adagrad	0.1	92.74	57.14	92.53	70.46	70.46
8 Hidden Decod-Encod	Sigmoid	Adam	0.01	95.56	76.59	97.60	81.03	81.03
8 Hidden Encod-Decod	Sigmoid	RMSProop	0.01	92.33	53.65	92.01	68.73	68.73

Table 6. Testing conventional machine learning algorithms and deep neural networks

Layers	Activation function	Optimizer	Learning rate	Accuracy (%)	F1 score (%)	Precision (%)	Recall (%)	AUC (%)
6 hidden encod-decod	Sigmoid	Adam	0.01	95.56	79.24	91.96	85.52	85.52
	Random forest			92.00	91.06	92.94	88.17	0.00

4. CONCLUSION

Based on real-time data collected from patients with thoracic conditions, the crucial step in studying survival data for thoracic patients is a pre-processing procedure that includes data filtering and cleansing, feature selection, and SMOTE. The deep neural network data mining technique can be a valuable tool for analyzing the survival data set of thoracic patients using a 10-fold cross-validation approach. The deep neural network's six hidden layers can be modified with different variations, such as decoder-encoder, relu, and sigmoid activation functions, optimizer Adam, a learning rate of 0.01, and a dropout rate of 0.2. By implementing these modifications, the network can achieve an accuracy of 95.56%, precision of 91.96%, recall of 85.52%, F1 score of 79.24%, and an accuracy of 85.52%. When comparing this value to conventional machine learning algorithms, it becomes evident how superior it is. Based on this rationale, deep neural network algorithms can produce superior outcomes in terms of accuracy, precision, recall, F1 score, and AUC compared to machine learning approaches and related research. There are no definitive guidelines dictating the learning rate, node, and dropout, although reducing these factors often enhances the accuracy of the network. An increase in learning speed, node density, and dropout rates will lead to a decrease in network accuracy, hence prolonging the process.

REFERENCES

[1] E. Dimakakou, H. J. Johnston, G. Streftaris, and J. W. Cherrie, “Exposure to environmental and occupational particulate air pollution as a potential contributor to neurodegeneration and diabetes: A systematic review of epidemiological research,” *International Journal of Environmental Research and Public Health*, vol. 15, no. 8. 2018, doi: 10.3390/ijerph15081704.

[2] M. G. Sholih, D. A. Perwitasari, R. Hendriani, H. Sukandar, and M. I. Barliana, “Risk factors of lung cancer in Indonesia: A qualitative study,” *Journal of Advanced Pharmacy Education and Research*, vol. 9, no. 2, pp. 40–41, 2019.

[3] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning from class-imbalanced data: Review of methods and applications,” *Expert Syst Appl*, vol. 73, pp. 220–239, 2017, doi: 10.1016/j.eswa.2016.12.035.

[4] J. A. ALzubi, B. Bharathikannan, S. Tanwar, R. Manikandan, A. Khanna, and C. Thaventhiran, “Boosted neural network ensemble classification for lung cancer disease diagnosis,” *Applied Soft Computing Journal*, vol. 80, pp. 579–591, 2019, doi: 10.1016/j.asoc.2019.04.031.
[5] M. Zięba, J. M. Tomczak, M. Lubicz, and J. Świątek, “Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients,” *Applied Soft Computing Journal*, vol. 14, pp. 99–108, 2014, doi: 10.1016/j.asoc.2013.07.016.
[6] H. Kaur, H. S. Pannu, and A. K. Malhi, “A systematic review on imbalanced data challenges in machine learning: Applications and solutions,” *ACM Computing Surveys*, vol. 52, no. 4, 2019, doi: 10.1145/3343440.
[7] L. Yu, R. Zhou, L. Tang, and R. Chen, “A DBN-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data,” *Applied Soft Computing Journal*, vol. 69, pp. 192–202, 2018, doi: 10.1016/j.asoc.2018.04.049.
[8] S. Dong, “Multi class SVM algorithm with active learning for network traffic classification,” *Expert Systems with Applications*, vol. 176, 2021, doi: 10.1016/j.eswa.2021.114885.
[9] P. Nanglia, S. Kumar, A. N. Mahajan, P. Singh, and D. Rathee, “A hybrid algorithm for lung cancer classification using SVM and Neural Networks,” *ICT Express*, vol. 7, no. 3, pp. 335–341, 2021, doi: 10.1016/j.icte.2020.06.007.
[10] K. Pradhan and P. Chawla, “Medical internet of things using machine learning algorithms for lung cancer detection,” *Journal of Management Analytics*, vol. 7, no. 4, pp. 591–623, 2020, doi: 10.1080/23270012.2020.1811789.
[11] S. Fotouhi, S. Asadi, and M. W. Kattan, “A comprehensive data level analysis for cancer diagnosis on imbalanced data,” *Journal of Biomedical Informatics*, vol. 90, 2019, doi: 10.1016/j.jbi.2018.12.003.
[12] M. Lubicz, K. Pawelczyk, A. Rzechonek, and J. Kolodziej, “Thoracic surgery data,” *UCI Machine Learning Repository*, 2013. Accessed: Apr. 05, 2023. [Online]. Available: http://archive.ics.uci.edu/dataset/277/thoracic+surgery+data
[13] L. Akter and M. M. Islam, “Hepatocellular carcinoma patient’s survival prediction using oversampling and machine learning techniques,” in *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, IEEE, 2021, pp. 445–450, doi: 10.1109/ICREST51555.2021.9331108.
[14] I. H. Sarker, “Machine learning: Algorithms, real-world applications and research directions,” *SN Computer Science*, vol. 2, no. 3, 2021, doi: 10.1007/s42979-021-00592-x.
[15] M. R. Kabir, F. B. Ashraf, and R. Ajwad, “Analysis of different predicting model for online shoppers’ purchase intention from empirical data,” *2019 22nd International Conference on Computer and Information Technology, 5* 2019, doi: 10.1109/ICCIT48885.2019.9038521.
[16] A. R. Panhalkar and D. D. Doye, “Optimization of decision trees using modified African buffalo algorithm,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, 2022, doi: 10.1016/j.jksuci.2021.01.011.
[17] C. S. Lee, P. Y. S. Cheang, and M. Moslehpour, “Predictive analytics in business analytics: decision tree,” *Advances in Decision Sciences*, vol. 26, no. 1, 2022, doi: 10.47654/V26Y2022I1P1-30.
[18] X. Geng, S. Wang, A. Ullah, G. Wu, and H. Wang, “Prediction of hardenability curves for non-boron steels via a combined machine learning model,” *Materials*, vol. 15, no. 9, 2022, doi: 10.3390/ma15093127.
[19] J. Zhou, Y. Dai, M. Tao, M. Khandelwal, M. Zhao, and Q. Li, “Estimating the mean cutting force of conical picks using random forest with salp swarm algorithm,” *Results in Engineering*, vol. 17, 2023, doi: 10.1016/j.rineng.2023.100892.
[20] R. Rastogi and S. Sharma, “Fast Laplacian twin support vector machine with active learning for pattern classification,” *Applied Soft Computing Journal*, vol. 74, 2019, doi: 10.1016/j.asoc.2018.10.042.
[21] S. K. Sharma and X. Hoque, “Sentiment predictions using support vector machines for odd-even formula in Delhi,” *International Journal of Intelligent Systems and Applications*, vol. 9, no. 7, 2017, doi: 10.5815/ijisa.2017.07.07.
[22] K. Shah, H. Patel, D. Sanghvi, and M. Shah, “A comparative analysis of logistic regression, random forest and KNN models for the text classification,” *Augmented Human Research*, vol. 5, no. 1, 2020, doi: 10.1007/s41133-020-00032-0.
[23] S. Watanabe and H. Nishimori, “Fall lecture note on statistical learning theory,” in *Lecture note for Tokyo Institute of Technology*, 2016.
[24] S. A. Mohamed, M. A. Abdou, and A. A. Elsayed, “Residual information flow for neural machine translation,” *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3220691.
[25] J. Xu, Y. Zhang, and D. Miao, “Three-way confusion matrix for classification: A measure driven view,” *Information Sciences*, vol. 507, 2020, doi: 10.1016/j.ins.2019.06.064.
[26] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning and Data Mining*, New York: Springer Science + Business Media, 2017.
[27] R. S. Wahono, “A systematic literature review of software defect prediction: research trends, datasets, methods and frameworks,” *Journal of Software Engineering*, vol. 1, no. 1, pp. 1–16, 2015.
[28] K. Farhana, M. Rahman, and M. T. Ahmed, “An intrusion detection system for packet and flow based networks using deep neural network approach,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 5, pp. 5514-5525, 2020, doi: 10.11591/ijece.v10i5.pp5514-5525.
[29] M. Al-Smadi, M. Hammad, Q. B. Baker, and A. Sa’ad, “A transfer learning with deep neural network approach for diabetic retinopathy classification,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 4, pp. 3492-3501, 2021, doi: 10.11591/ijece.v11i4.pp3492-3501.

BIOGRAPHIES OF AUTHORS

Cucu Ika Agustyaningrum is a faculty member and researcher at Universitas Bina Sarana Informatika at the Faculty of Engineering and Informatics. Focusing research on data mining, data science, and computer science. She can be contacted at email: cucu.cuk@bsi.ac.id.

Yudi Ramdhani is faculty member and researcher from Department of Information Systems at the Faculty of Creative Technology, Satu University. His research focuses on data analysis using data mining and decision support systems. He can be contacted at email: yudi.ramdhani@univ.satu.ac.id.

Doni Purnama Alamsyah is faculty member from Bina Nusantara University, teaching at Department of Entrepreneurship in Customer Behavior study. He received a Doctorate degree from Padjadjaran University in the field of Management Science. Currently has an interest and research focus on consumer behavior and is very open to conducting research collaboration. He can be contacted at email: doni.syah@binus.ac.id

Oda I. B. Hariyanto is a Professor of Culture and Food History, currently a Faculty Member and Head of Department in Department of Tourism at Batam International University. She has an interest in studying behavior and culture related to food in Indonesia. She can be contacted at email: oda@uib.ac.id.