Lung cancer patients survival prediction using outlier detection and optimized XGBoost

Wirot Yotsawat, Peetiphart Suebpeng, Saroch Purisangkaha, Akarapon Poonsawad, Kanyalag Phodong
Computer Science Program, Department of Science, Faculty of Science and Technology, Phranakhon Si Ayutthaya Rajabhat University, Phranakhon Si Ayutthaya, Thailand

Article Info

Article history:

Received Jul 2, 2024
Revised Feb 14, 2025
Accepted Mar 15, 2025

Keywords:

Hyper-parameter optimization
Lung cancer
Machine learning
Outlier detection
Survival prediction
XGBoost

ABSTRACT

This research aims to improve the prediction's model for survival time of lung cancer patients by using outlier detection, hyper-parameter optimization, and machine learning technique. The research compares the performance of several methods including multilayer perceptron (MLP), decision tree (DT), linear regression (LR), Bagging, XGBoost, and random forest (RF). The dataset used for the experiment is obtained from the surveillance, epidemiology, and end result (SEER) cancer database, which contains diagnoses data from 2004 to 2015. The total number of records used is 196,031 with 22 features. 10-fold cross-validation is used for training and testing sets. The evaluation metrics are root mean square error (RMSE), mean squared error (MSE), R-squared (R^2), and mean absolute error (MAE). The results show that the lung cancer patient survival prediction model using the optimized XGBoost (O-XGBoost) model performs the best with an RMSE of 13.74 and outperforms the baseline-XGBoost model as well as other models. This research will be useful for developing a clinical decision support system for the care of lung cancer patients. Physicians can use the developed model to assess the patient's chance of survival in order to plan more effective treatment.

Corresponding Author:

Kanyalag Phodong
Computer Science Program, Department of Science, Faculty of Science and Technology
Phranakhon Si Ayutthaya Rajabhat University
Phranakhon Si Ayutthaya, Thailand
Email: kanyalagp@gmail.com

1. INTRODUCTION

In 2022, lung cancer was the second most common type of cancer and the leading cause of cancer-related deaths. It is particularly prevalent among men [1]. The primary cause of lung cancer is smoking, with smokers having a 20-30 times higher risk of developing lung cancer compared to non-smokers [2]. Cigarette smoke contains over 4,000 chemicals, including at least 69 known carcinogens and other toxins. Other causes of lung cancer include a history of respiratory diseases (such as asthma, pneumonia, tuberculosis, chronic bronchitis, and chronic obstructive pulmonary disease), exposure to workplace carcinogens, genetic factors [3], and environmental air pollution, especially prolonged exposure to fine particulate matter (PM2.5), which has been linked to an increased risk of lung cancer [4].

Predicting the survival time of cancer patients from the initial diagnosis can help patients and caregivers plan their time and resources, as well as guide the medical team's care and treatment approach [5]. However, predicting the survival time of cancer patients requires specialized expertise, posing a challenge in areas lacking sufficient specialists. Therefore, artificial intelligence (AI) tools built with machine learning techniques can address this issue. Machine learning can utilize historical cancer patient data, stored in databases, to accurately predict patient survival times [6]. This not only helps avoid unnecessary treatments, surgeries, and expenses for patients but also reduces diagnostic time and alleviates stress for doctors.

Traditional lung cancer survival prediction research has focused primarily on classification tasks, grouping patients into broad survival categories, such as survived/not survived or achieving five-year survival [7]–[12]. However, classification often lacks sufficient detail for precise survival estimates, limiting its usefulness for individualized patient care. As a result, interest has shifted toward regression models that predict continuous survival times rather than discrete categories, providing a finer and more accurate view of patient prognosis. Techniques like linear regression (LR), decision tree (DT), and random forest (RF) are frequently used to enhance model performance and reduce prediction error, especially in terms of root mean square error (RMSE) [6]. Ensemble methods that combine multiple algorithms further improve accuracy, producing models more resistant to overfitting and more closely aligned with actual survival outcomes [5], [6], [13]. Recently, large datasets like the surveillance, epidemiology, and end result (SEER) database have been utilized with traditional machine learning methods, including LR, gradient boosting machines (GBM), and support vector machines (SVM), to predict lung cancer survival times, with GBM and ensemble models generally demonstrating the highest predictive accuracy, particularly for patients with shorter survival times [6]. By incorporating key patient attributes such as age, tumor size, and stage these models often perform comparably to established statistical approaches. Although deep learning has also been explored for survival prediction, traditional machine learning methods remain reliable and interpretable, making them effective alternatives for predicting continuous survival outcomes in lung cancer patients [5].

A major challenge in using large datasets for prediction is that the collected raw data is often not ready for training machine learning algorithms. It requires several preprocessing steps, such as handling missing values, standardizing the data format, transforming data into suitable forms for the algorithms, feature selection and outliers removal. Preprocessing not only prepares the data for learning but can also enhance model performance. For instance, outliers removal [14], [15] and feature selection [16], [17] can significantly improve model performance. Neglecting these steps can result in highly inaccurate models which unsuitable for practical use. Additionally, during the learning process, model performance can be further optimized by hyper-parameter optimization [18]–[21] and using suitably ensemble methods. Ensemble methods, which develop by multiple training iterations, typically provide better performance than single techniques [11], [21]–[24].

This research presents an improvement in the performance of survival time prediction models for lung cancer patients using machine learning algorithms. The research processes include outliers removal using isolation forest (IF) [25] and auto encoding (AE) [26], hyper-parameter optimization using Bayesian methods [27], as well as model construction using various machine learning techniques. Models performance is evaluated using 10-fold cross-validation, and the effectiveness is measured by RMSE, mean squared error (MSE), R-squared (R^2), and mean absolute error (MAE). Once the model demonstrates the best performance, it will be used to develop an application for predicting the survival time of lung cancer patients.

The objectives of this research are as follows. First, to improve the performance of survival time prediction models for lung cancer patients constructed using traditional machine learning techniques under regression problems that generating continuous outcomes. Second, to compare the efficacy of individual algorithms, including LR, DT, and multilayer perceptron (MLP), with ensemble methods such as Bagging, XGBoost, and RF in predicting the survival time of lung cancer patients.

2. LITERATURE REVIEW

Enhancing a machine learning model’s performance can be achieved through effective data preprocessing steps. Outlier detection is the process of identifying data that differs significantly from most of the data in a dataset. These outliers may result from errors in data collection or from data that is markedly different from normal data [25]. The IF algorithm is one method used for outlier detection, developed in [14]. This method relies on the principles of DT that create a large number of DT. The main concept of IF is that outliers are often isolated points that differ from most data points, while normal data points are distributed throughout the main data area [25]. Another popular method is AE techniques which employ special types of neural networks comprising an encoder part for compressing data into lower dimensions and a decoder part for reconstructing the data back to its original dimensions. AE is first trained with normal data. When new data is passed through this network, the reconstruction error is calculated. If the error exceeds a certain threshold, the data is considered an outlier because it cannot be compressed and reconstructed accurately. This principle is used to rank the degree of data outlier [26].

Optimizing hyper-parameters is another method to enhance the performance of machine learning models [18], [19]. The Bayesian method employs Bayesian probability theory to iteratively refine hyper-parameter selection for AI model construction. Beginning with a prior probability distribution, typically based on expert knowledge or assumptions about hyper-parameter ranges, the method samples values and evaluates them by training the model and assessing performance metrics like error rates or accuracy. Using Bayes' theorem, the prior distribution is updated to a posterior distribution that better reflects observed data, enhancing the probability estimates for subsequent iterations. This iterative process continues, with each cycle using updated probabilities to guide the selection of new hyper-parameter values, aiming to optimize model performance effectively. This approach significantly reduces the computational burden compared to exhaustive search methods, ensuring more efficient and effective hyper-parameter tuning [27].

In recent studies, various machine learning models have been applied to predict cancer patient survival periods using the SEER dataset. Thomgkam *et al.* [28] conducted a comparative study on models predicting breast cancer survival, analyzing 115,184 records from 2004 to 2014. They categorized survival outcomes into two classes: less than 5 years and more than 5 years. The machine learning techniques compared included naïve Bayes, partial decision tree (PART), MLP, SVM, and Bagging of the four methods. Their findings indicated that Bagging of PART had the best performance, with a sensitivity of 99.39%, specificity of 96.85%, and accuracy of 98.89%. The researchers suggested that the model and decision rules could be developed into a disease surveillance system for early risk screening by physicians. Similarly, Doppalapudi *et al.* [5] used SEER data to understand factors related to lung cancer patient survival and developed predictive models using deep learning techniques like artificial neural networks (ANN), convolutional neural networks (CNN), and recurrent neural networks (RNN). They defined three survival categories for classification: less than or equal to 6 months, more than 6 months up to 2 years, and more than 2 years, with ANN achieving the highest accuracy at 71.18%. For regression, CNN showed the best performance with an RMSE of 13.50 and an R^2 of 0.5066. Lynch *et al.* [13] proposed a linear model for predicting lung cancer patient survival, comparing it with SVM, DT, RF, GBM, and a custom ensemble. Their study, which analyzed data from 2004 to 2009 with 18 features and 10,442 records, found the custom ensemble to be the most effective, with an RMSE of 15.30.

Bartholomai and Frieboes [6] studied the prediction of lung cancer patients' survival times in months using regression and classification models based on SEER database data. They observed that while models were highly accurate for predicting short survival times (less than 6 months), their accuracy diminished for longer survival periods. Their approach utilized RF for classification and a combination of LR, GBM, and RF for regression. Model accuracy was evaluated using a confusion matrix and RMSE. The study found that RF performed best for predicting survival times of ≤6 months (RMSE 10.52) and >24 months (RMSE 20.51), while GBM was most effective for 7–24 months (RMSE 15.65). Their findings suggest that regression models are more reliable for shorter survival predictions than RMSE values might indicate. Overall, these studies highlight the effectiveness of machine learning techniques in predicting cancer patient survival and suggest further development and integration of these models into clinical practice for early risk screening and improved patient management.

3. METHOD

3.1. Research framework

This research applies the cross-industry standard process for data mining (CRISP-DM). The CRISP-DM process consists of the following steps: business understanding, data understanding, data preparation, modeling, model evaluation, and deployment [29]. In this research, the workflow is divided into seven stages, as shown in Figure 1, with the following details.

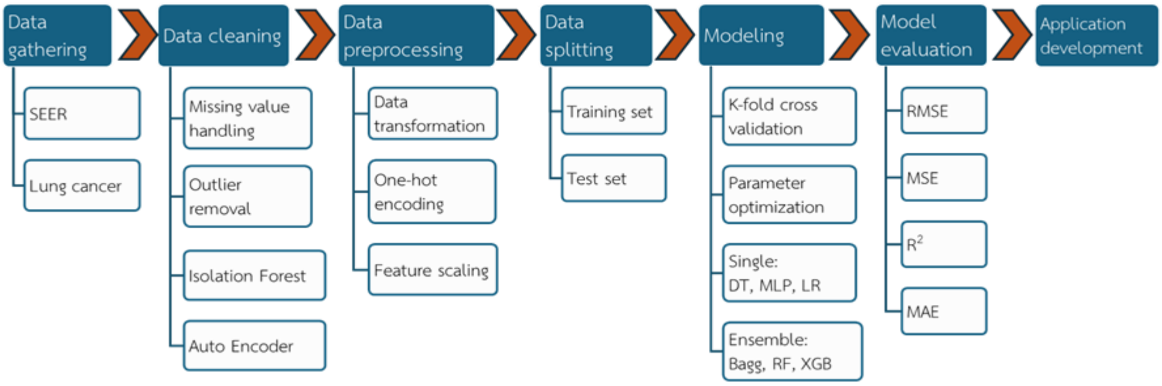


Figure 1. Research framework

3.1.1. Data gathering

The collection of lung cancer patient data from the SEER database, which is a comprehensive cancer registry and epidemiologic surveillance program of the United States. Data spanning from 1973 to 2015 is gathered using the SEER*Stat application [30]. The SEER database contains a sufficient amount of data for developing models and applications to predict survival time. From the data collection, it was found that there are a total of 721,697 records of lung cancer patients, each comprising 35 attributes.

3.1.2. Data cleaning

This process consists of two sub-steps: i) missing value handling: management of missing data, where it was found that 70% of the collected lung cancer patient data contained missing values that cannot be used for model building. Therefore, data with missing values was removed from the dataset. Additionally, features with missing values exceeding 70% were also excluded. ii) outliers removal: removal of abnormal data from the dataset. In this research, the effectiveness of two outlier detection techniques, IF and AE, was compared. A maximum threshold of 20% of the total data was set for outliers' removal.

3.1.3. Data preprocessing

Preparation of data prior to processing to ensure readiness for machine learning techniques includes: i) transformation using one-hot encoding, applied to nominal categorical features that cannot be processed mathematically, and ii) scaling of numeric features (numeric) through feature scaling, specifically using normalization to standardize values within the range of 0 to 1. This prevents biases towards features with larger or smaller values that could affect the final model outcomes. This study follows the data scaling procedures similar in [13] to expedite model adjustments and achieve optimal research outcomes.

3.1.4. Data splitting

Data splitting is a critical step in machine learning where the preprocessed dataset is divided into two main subsets: a training dataset comprising 80% of the data used to train the model by learning patterns and relationships, and a testing dataset comprising 20% used to evaluate the model's performance on unseen data. This approach ensures the model's ability to generalize well and make accurate predictions beyond the training data. By separating training and testing data, machine learning practitioners can assess the model's effectiveness in real-world applications, guarding against issues like overfitting or underfitting, and ensuring robust performance across different datasets.

3.1.5. Modeling

In this process, the training dataset is used to allow machine learning algorithms to learn and discover patterns collectively within the data. This involves cross-validation with 10 folds and adjusting hyper-parameters of the algorithms to find the most suitable values. The learning methods employed include single methods such as DT, LR, and MLP, as well as ensemble methods such as Bagging, XGBoost, and RF. Hyper-parameter tuning for these methods utilizes Bayesian methods to determine the optimal hyper-parameter values, as demonstrated in Table 1.

3.1.6. Model evaluation

Model evaluation involves assessing the effectiveness of the trained model using the testing dataset. This evaluation is crucial for determining how well the model generalizes to new, unseen data. Metrics such as RMSE, MSE, MAE, and R^2 are utilized. RMSE, MSE, and MAE measure the average magnitude of errors between predicted and actual values, with lower values indicating better accuracy. R^2, on the other hand, quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables; a higher R^2 value (closer to 1) suggests that the model explains a larger portion of the variance in the data. These metrics collectively provide insights into the model's predictive performance and its ability to make reliable predictions in practical applications.

3.1.7. Application development

The last stage involves transforming the highest-performing model into a sophisticated web application tailored specifically for predicting the survival duration of lung cancer patients. This application aims to provide healthcare professionals with comprehensive preliminary data essential for making informed medical assessments and treatment decisions. By leveraging advanced machine learning techniques and intuitive user interfaces, the application seeks to streamline the prediction process, ensuring accuracy and reliability in prognosis. Emphasis will be placed on integrating scalable architecture and robust backend systems to support seamless deployment and utilization in clinical settings. Ultimately, the application aims to enhance the efficiency of medical practitioners in delivering personalized care and improving patient outcomes through timely and accurate prognostic insights.

Table 1. Hyper-parameter scape for each modeling approach used to find the optimal hyper-parameter values

Algorithms	Hyper-parameters	Space of hyper-parameters
DT	ccp_alpha	[0, 10]
	criterion	[‘squared_error’, ‘friedman_mse’, ‘poisson’]
	max_depth	[1, 20]
	min_samples_leaf	[1, 100]
	min_samples_split	[0, 0.9]
	splitter	[‘best’, ‘random’]
MLP	alpha	[0, 1]
	batch_size	[1, 800]
	hidden_layer_sizes	[(100), (200), (300), (400), (123), (246), (61)]
	activation	[‘tanh’, ‘relu’, ‘identity’, ‘logistic’]
	learning_rate_init	[0, 1]
	max_iter	[10, 200]
	learning_rate	[‘constant’, ‘adaptive’, ‘invscaling’]
	momentum	[0, 1]
XGB	max_depth	[1, 30]
	gamma	[0, 1]
	learning_rate	[0, 1.0]
	min_child_weight	[1, 20]
	alpha	[0, 1]
	subsample	[0, 1]
	colsample_bytree	[0, 1]
	max_delta_step	[0, 10]
	n_estimators	[100, 800]
Bag	n_estimators	[1, 300]
	max_samples	[0.2, 1.0]
	max_features	[0, 1]
RF	n_estimators	[1, 350]
	max_depth	[1, 20]
	min_samples_leaf	[1, 100]
	min_samples_split	[0.001, 0.9]
	max_features	[1, 21]

3.2. Dataset detail

The data utilized for this research originates from lung cancer patients sourced from the SEER database of the United States of America. After undergoing comprehensive missing data handling procedures, the dataset encompasses a total of 245,034 patient records, each comprising 22 distinct attributes. These records span the period from 2004 to 2015, with survival durations ranging from 0 to 191 months. The 22 attributes include demographic information, clinical characteristics, and survival outcomes, all meticulously documented and stored within the SEER database. These details serve as data for developing predictive models aimed at forecasting survival times for lung cancer patients, thereby supporting clinical decision-making processes in healthcare settings. The details of all 22 features are presented in Table 2.

Table 2. Characteristics of lung cancer patient data from the SEER database

#	Variable name	Original name	Data type
1	AGE	Age	Nominal
2	SEX	Gender	Binary
3	RACE	Races recode (W, B, AI, API)	Nominal
4	PRIMARY_SITE	Primary Site	Nominal
5	GRADE	GRADE	Nominal
6	HIST_BROAD_GROUP	Histology Record – Broad Groupings	Nominal
7	SEQUENCE_NUMBER	Sequence Number	Binary
8	REGIONAL_NODES_POSITIVE	Regional nodes positive (1988+)	Numeric
9	REGIONAL_NODES_EXAMINED	Regional nodes examined (1988+)	Numeric
10	REASON_SURG	Reason no cancer-directed surgery	Nominal
11	RXSUMM_SURG_PRIM_SITE	RX Summ – Surg Prim Site (1998+)	Nominal
12	RXSUMM_SCOPE_REG_LN_SUR	RX Summ – Scope Reg LN Sur (2003+)	Nominal
13	DIANOSTIC_CONFIRM	Diagnostic Confirmation	Nominal
14	CSEXTENSION	CS extension (2004-2015)	Nominal
15	CSTUMOR_SIZES	CS tumor size (2004-2015)	Numeric
16	CSLYMPH_NODES	CS lymph nodes (2004-2015)	Nominal
17	SUMMARY_STAGE	Summary stage 2000 (1998-2017)	Nominal
18	T	Derived AJCC T, 6th ed (2004-2015)	Nominal
19	N	Derived AJCC N, 6th ed (2004-2015)	Nominal
20	M	Derived AJCC M, 6th ed (2004-2015)	Nominal
21	STAGE	Derived AJCC Stage Group, 6th ed (2004-2015)	Nominal
22	SURVIVAL_MONTH	survival months	Numeric

Note: Further study details can be found on the website https://seer.cancer.gov

4. RESULTS AND DISCUSSION

4.1. Results of outliers removal

The results after removing outliers data from the dataset using the IF and AE methods, each with a removal rate of 20%, show that the initial dataset, which contained 245,034 records, was reduced to 196,031 records using both methods. This reduction in data will be used for subsequent experimental steps. Despite the two datasets being equal in size, they differ significantly in content because the IF and AE methods employ distinct processes to identify and remove outliers. The IF method detects anomalies based on how isolated the data points are in a feature space, while AE uses neural networks to encode data and identify deviations from the norm. Therefore, the datasets resulting from these methods, although numerically equivalent, vary in the specific data they contain due to these methodological differences. The details of the comparison between the initial dataset and the datasets after outliers removal is presented in Table 3.

Table 3. Comparison the results of the data before and after removing outliers

Algorithms	Original instants	Good instants	Anomaly instants	% Removal
Isolation forest	245,034	196,031	49,003	20.0
Auto encoding	245,034	196,031	49,003	20.0

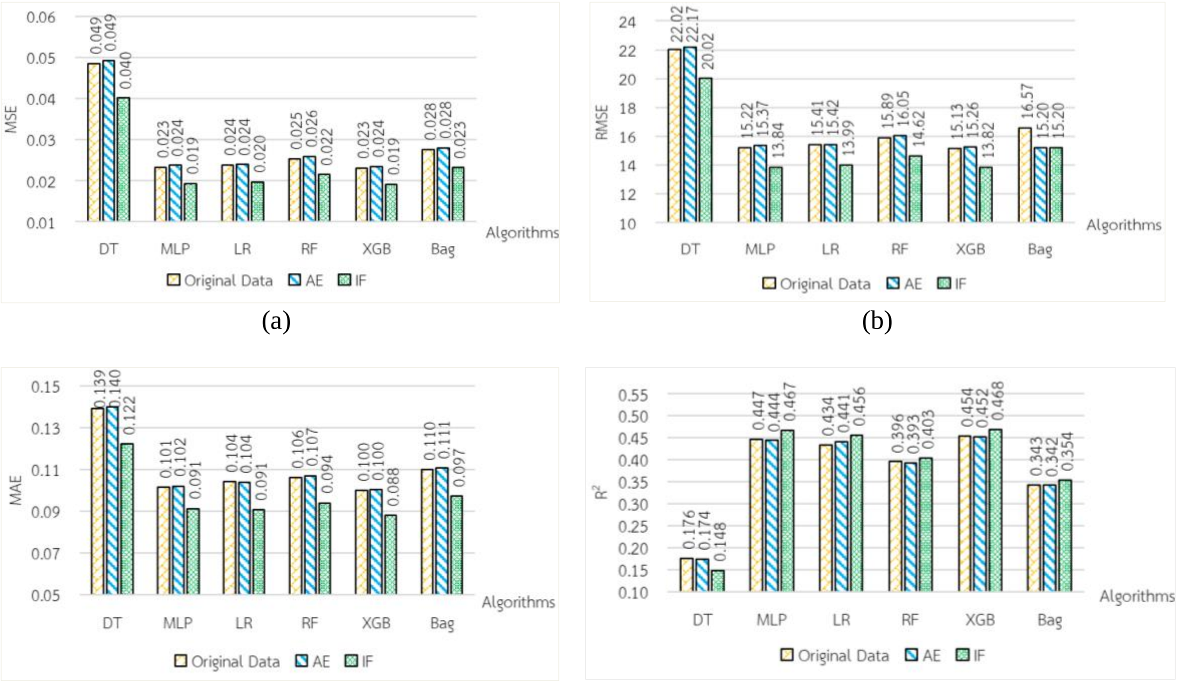
When training the algorithms with both datasets, it was observed that the models exhibited improved results and predictive performance compared to those trained with data that retained outliers. The reduced dataset size also contributed to faster model creation. A detailed comparison of MSE, RMSE, MAE, and R^2 values revealed that the dataset processed with the IF method yielded superior results across all metrics, except for the R^2 value in the DT algorithm. Additionally, when the XGBoost algorithm was trained with the IF-processed dataset, it demonstrated the highest predictive performance among all models and datasets. This indicates that the IF method not only enhances prediction accuracy but also streamlines the modeling process by effectively identifying and removing outliers. The comprehensive comparison and performance details are illustrated in Figure 2.

From Figure 2, when considering the overall picture, the experimental results can be explained as follows: i) removing outliers using the IF method sufficiently reduces MSE as shown in Figure 2(a), RMSE in Figure 2(b), and MAE in Figure 2(c) across all methodological steps when compared to both datasets: the dataset without outliers removal and the dataset with outliers removal using the AE method. ii) outliers removal with the AE method, however, slightly decreases the overall performance of the models, evidenced by higher MSE as shown in Figure 2(a), RMSE in Figure 2(b), and MAE in Figure 2(c) values, as well as decreased R^2 in Figure 2(d) values when compared to the dataset without outliers removal. iii) although outliers removal with the IF method leads to decreased MSE as shown in Figure 2(a), RMSE in Figure 2(b), and MAE in Figure 2(c) values in DT, the R^2 value decreases as well as shown in Figure 2(d). This is due to the DT method's initial hyper-parameters, which do not specify a maximum depth and leaf nodes, resulting in an unnecessary proliferation of leaf nodes during training. Consequently, the final learned results show an inaccurate average of each record mixed within the excessive leaf nodes. iv) despite the decreased R^2 as shown in Figure 2(d) value in DT after outliers removal with the IF method, other methods show an improvement in R2 values in Figure 2(d).

Based on the experimental findings, it was evident that the dataset treated with the IF outliers removal method consistently exhibited superior performance compared to both the dataset processed with the AE method and the original dataset without outliers removal. Consequently, the IF-processed dataset was selected for further model development stages. Specifically, across the six models evaluated (MLP, LR, DT, XGBoost, and Bagging) it was noted that five of these models demonstrated enhanced predictive capabilities when trained on the IF-processed dataset. This enhancement was reflected in improved metrics such as reduced RMSE, MAE, and MSE, as well as higher R^2 values, indicating better model accuracy and robustness. These results underscore the effectiveness of the IF method in enhancing data quality by effectively identifying and removing outliers, thereby contributing to improved model performance across various machine learning algorithms.

4.2. The results of hyper-parameters' optimization

In Table 4, the hyper-parameter values obtained through the Bayesian search method, using the dataset processed with IF outliers removal, were used to build models. The models were evaluated using K-folds cross-validation with K set to 10. This resulted in average performance metrics across all 10 folds for each of the 6 methods, as shown in Table 5.



(a) (b)

(c) (d)

Figure 2. Comparison the results of the models trained on data before and after outlier removal on (a) MSE value, (b) RMSE value, (c) MAE value, and (d) R^2 value

Table 4. Results of hyper-parameters optimization using Bayesian search

Algorithms	Hyper-parameter	Best parameters
DT	ccp_alpha	0.00339225
	criterion	squared_error
	max_depth	8
	min_samples_leaf	61
	min_samples_split	0.860951952
	splitter	random
XGB	max_depth	24
	gamma	0.173941856
	learning_rate	0.05028551
	min_child_weight	9.006166167
	alpha	0.26153907
	subsample	0.700633533
	colsample_bytree	0.403669676
	max_delta_step	6
	n_estimators	480
MLP	alpha	0.203240288
	batch_size	798
	hidden_layer_sizes	200
	activation	relu
	learning_rate_init	0.000542794
	max_iter	82
	learning_rate	constant
	momentum	0.637558948
Bagging	n_estimators	296
	max_samples	0.695150026
	max_features	0.207177187
RF	n_estimators	114
	max_depth	12
	min_samples_leaf	82
	min_samples_split	0.003641991
	max_features	16

In Table 5, it is evident that after adjusting the hyper-parameters to optimal values, the DT, MLP, LR, RF, Bagging, and optimized XGBoost (O-XGBoost) models all showed varying degrees of improvement. Notably, the O-XGBoost model emerged as the top performer across all evaluation metrics (MSE, RMSE, MAE, and R^2) indicating its robustness and superior predictive capabilities. Following XGBoost, the MLP, Bagging, RF, LR, and DT models sequentially exhibited commendable performance improvements. Specifically, the O-XGBoost model achieved impressive metrics with RMSE at 13.74, R^2 of 0.4747, MAE of 0.0876, and MSE of 0.0190. Comparatively, when compared to the Baseline-XGBoost model trained on the initial dataset, the XGBoost model refined through the data mining process showcased sufficiently enhanced performance. These findings underscore the effectiveness of hyper-parameter tuning and data preprocessing techniques in optimizing model accuracy and reliability across diverse machine learning algorithms.

Table 5. Results of the models after adjusting hyper-parameters to appropriate values

Models	MSE	RMSE	MAE	R^2
DT	0.0233	15.2200	0.1007	0.3554
MLP	0.0193	13.8600	0.0880	0.4656
LR	0.0197	13.9900	0.0906	0.4555
RF	0.0196	13.9600	0.0899	0.4579
Bagging	0.0195	13.9100	0.0908	0.4610
O-XGBoost	0.0190	13.7400	0.0876	0.4747

4.3. Results discussion

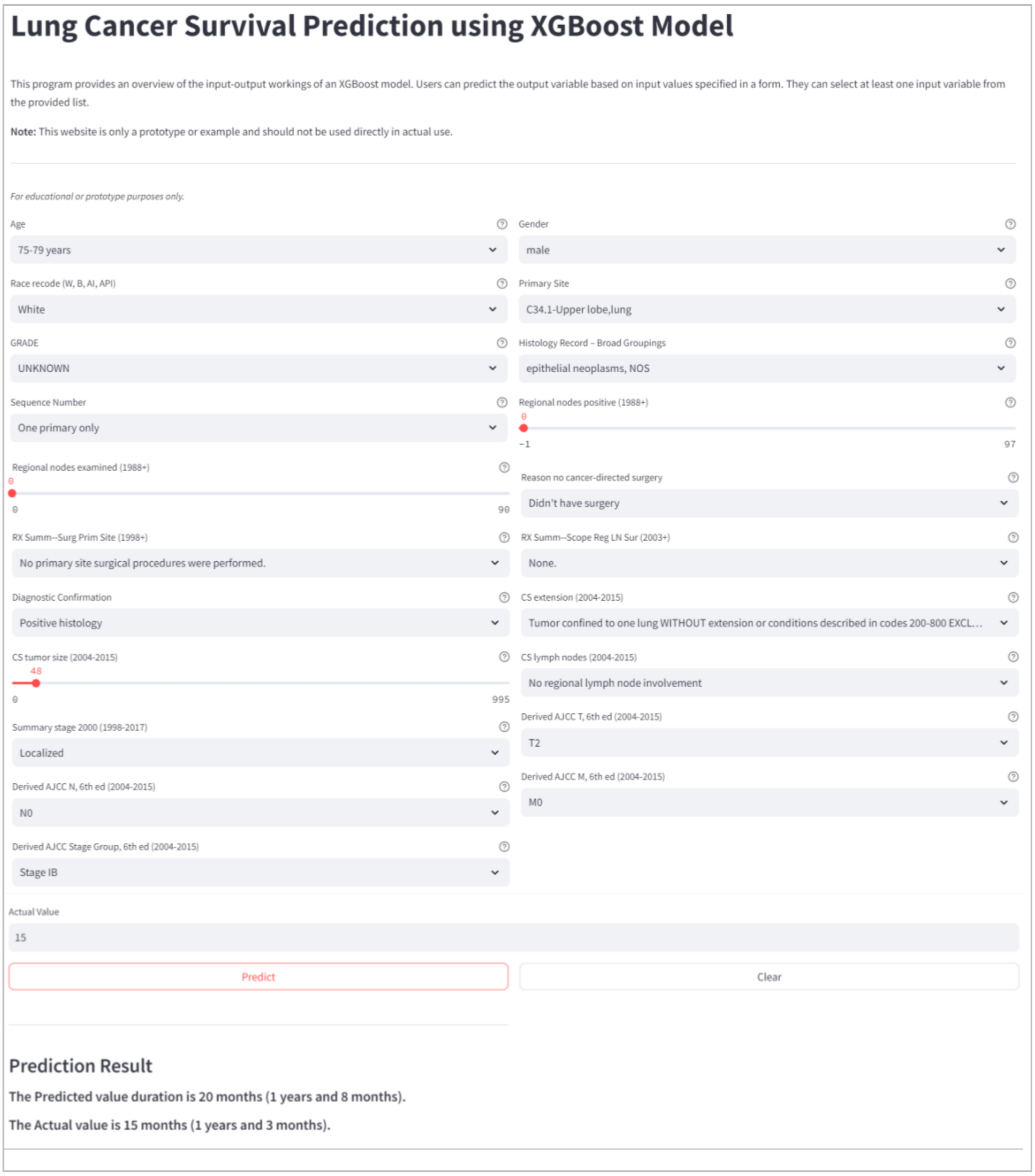
In Table 6, the comparative analysis between the O-XGBoost and Baseline-XGBoost models shows that the O-XGBoost model outperforms the baseline across all evaluation metrics. The O-XGBoost model achieved a lower MSE at 0.0190 compared to 0.0231, and a lower RMSE at 13.74 compared to 15.13, indicating more precise and accurate predictions. Additionally, the MAE is reduced from 0.0999 to 0.0876, further confirming the model’s improved accuracy. The R² value increased from 0.4537 to 0.4747, demonstrating that the optimized model explains a greater proportion of variance in the data. Overall, the O-XGBoost model provides more reliable predictions for lung cancer patient survival times. This aligns with the research in [18], [19], which used hyper-parameter optimization to enhance the performance of XGBoost.

Table 6. Comparison of XGBoost model, Baseline-XGBoost and other models

Models	Instances	Attributes	Survival time (months)	MSE	RMSE	MAE	R^2
Custom ensemble [13]	10,442	19	0-72	-	15.30	-	-
RF [6]	10,442	25	0-6	-	10.52	-	-
GBM [6]	10,442	25	7-24	-	15.65	-	-
RF [6]	10,442	25	>24	-	20.51	-	-
GBM [5]	702,411	15	0-60	0.023	15.30	-	0.3664
RF [5]	702,411	15	0-60	0.022	14.87	-	0.4015
CNN [5]	702,411	15	0-60	0.018	13.50	-	0.5066
Baseline-XGBoost	196,031	22	0-191	0.023	15.13	0.099	0.4537
O-XGBoost	196,031	22	0-191	0.019	13.74	0.087	0.4747

The findings of this research align with previous studies that have demonstrated the efficacy of machine learning models in predicting cancer patient survival times using SEER data. Similar to the study [28] on breast cancer, and similarly to the study [5], [13] research on lung cancer, our study confirms the superior performance of ensemble methods and advanced algorithms. Notably, our use of outlier detection and hyper-parameter optimization has further enhanced model accuracy, with XGBoost outperforming other models, achieving RMSE of 13.74. This result surpasses the performance metrics reported by Bartholomai and Frieboes [6], who highlighted the variability in accuracy across different survival periods. The success of XGBoost in our study underscores its potential for clinical application, providing physicians with a robust tool for assessing patient prognosis. This aligns with the suggestion by Thomgkam *et al.* [28] and other researchers that predictive models can sufficiently aid in early risk screening and the formulation of more effective treatment plans. Our research contributes to the ongoing development of clinical decision support systems, enhancing the precision and reliability of survival predictions for lung cancer patients.

O-XGBoost is a powerful tool for survival analysis, while CNN can also be effective, particularly with image or spatial data. CNN excel at extracting essential features from complex datasets, making them suitable for applications like medical imaging and geographic data analysis. However, CNN generally require larger datasets and greater computational resources for effective training [31]. In contrast, O-XGBoost is typically more interpretable, handling tabular data efficiently and often requiring fewer resources. The choice between O-XGBoost and CNN ultimately depends on the problem, data characteristics, and required interpretability. In Table 6, it's also noted that this experimental comparison with previous studies is based on differing datasets, so no definitive conclusion can be made about which model is superior. In this study, we also created a web application that predicts the survival time of lung cancer patients using the O-XGBoost models, as illustrated in Figure 3.


Lung Cancer Survival Prediction using XGBoost Model

This program provides an overview of the input-output workings of an XGBoost model. Users can predict the output variable based on input values specified in a form. They can select at least one input variable from the provided list.

Note: This website is only a prototype or example and should not be used directly in actual use.

For educational or prototype purposes only.

Age: 75-79 years

Gender: male

Race recode (W, B, AI, API): White

Primary Site: C34.1-Upper lobe,lung

GRADE: UNKNOWN

Histology Record – Broad Groupings: epithelial neoplasms, NOS

Sequence Number: One primary only

Regional nodes positive (1988+): 0 (-1 to 97)

Regional nodes examined (1988+): 0 (0 to 90)

Reason no cancer-directed surgery: Didn't have surgery

RX Summ--Surg Prim Site (1998+): No primary site surgical procedures were performed.

RX Summ--Scope Reg LN Sur (2003+): None.

Diagnostic Confirmation: Positive histology

CS extension (2004-2015): Tumor confined to one lung WITHOUT extension or conditions described in codes 200-800 EXCL…

CS tumor size (2004-2015): 48 (0 to 995)

CS lymph nodes (2004-2015): No regional lymph node involvement

Summary stage 2000 (1998-2017): Localized

Derived AJCC T, 6th ed (2004-2015): T2

Derived AJCC N, 6th ed (2004-2015): N0

Derived AJCC M, 6th ed (2004-2015): M0

Derived AJCC Stage Group, 6th ed (2004-2015): Stage IB

Actual Value: 15

Prediction Result

The Predicted value duration is 20 months (1 years and 8 months).

The Actual value is 15 months (1 years and 3 months).

Figure 3. A web application of lung cancer survival prediction using O-XGBoost model

5. CONCLUSION

This research, an ensemble model is developed to enhance the prediction of lung cancer patient survival times by incorporating both outlier detection and hyper-parameter optimization techniques. The ensemble approach combines multiple training iterations of base algorithm which each trained through multiple iterations to optimize their performance. For outlier detection, the IF algorithm is employed to identify and manage data points that deviate from the norm, ensuring that the model is trained on clean and relevant data. Additionally, Bayesian optimization is used to fine-tune hyper-parameters, efficiently refining the model parameters through iterative updates based on performance metrics. The dataset for this study is sourced from the SEER cancer database, including 196,031 records with 22 attributes collected between 2004 and 2015. The research utilizes 10-fold cross-validation for comprehensive model training and testing to ensure robust performance evaluation. The ensemble model's effectiveness is measured using metrics such as RMSE, MSE, R², and MAE. Results indicate that the XGBoost algorithm achieves the highest accuracy with an RMSE of 13.74, surpassing other models and the baseline-XGBoost. This research underscores the effectiveness of utilizing the ensemble methods with outlier detection and Bayesian optimization to improve prediction performance, providing sufficient contributions to the development of clinical decision support systems that can enhance treatment planning and patient care. Future work could explore more complex algorithms, thorough feature selection, or feature weighting techniques to enhance the model's performance.

FUNDING INFORMATION

The researchers would like to thank the Research and Development Institute, Phranakhon Si Ayutthaya Rajabhat University, for their research funding support (financial code: Z03301010642), which greatly contributed to the successful completion of this study. Special thanks are also extended to the experts for their evaluation of the research proposal and their valuable recommendations for improving the research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Wirot Yotsawat	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Peetiphart Suebpeng		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Saroch Purisangkaha					✓		✓			✓				
Akarapon Poonsawad					✓		✓			✓				
Kanyalag Phodong	✓	✓		✓	✓	✓	✓			✓		✓	✓	

C	:	**C**onceptualization	I	:	**I**nvestigation	Vi	:	**Vi**sualization
M	:	**M**ethodology	R	:	**R**esources	Su	:	**Su**pervision
So	:	**So**ftware	D	:	**D**ata Curation	P	:	**P**roject administration
Va	:	**Va**lidation	O	:	Writing - **O**riginal Draft	Fu	:	**Fu**nding acquisition
Fo	:	**Fo**rmal analysis	E	:	Writing - Review & **E**diting			

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data supporting the findings of this study are publicly accessible through the SEER*Stat application [30], available on the official website: https://seer.cancer.gov.

REFERENCES

[1] B. S. Chhikara and K. Parang, "Global cancer statistics 2022: the trends projection analysis," *Chemical Biology Letters*, vol. 10, no. 1, pp. 1–16, 2022.

[2] J. D. Minna, J. A. Roth, and A. F. Gazdar, "Focus on lung cancer," *Cancer Cell*, vol. 1, no. 1, pp. 49–52, 2002, doi: 10.1016/S1535-6108(02)00027-2.

[3] M. B. Schabath and M. L. Cote, "Cancer progress and priorities: lung cancer," *Cancer Epidemiology Biomarkers and Prevention*, vol. 28, no. 10, pp. 1563–1579, 2019, doi: 10.1158/1055-9965.EPI-19-0221.

[4] B. Kosanpipat *et al.*, "Impact of PM2.5 exposure on mortality and tumor recurrence in resectable non-small cell lung carcinoma," *Scientific reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-76696-0.

[5] S. Doppalapudi, R. G. Qiu, and Y. Badr, "Lung cancer survival period prediction and understanding: deep learning approaches," *International Journal of Medical Informatics*, vol. 148, 2021, doi: 10.1016/j.ijmedinf.2020.104371.

[6] J. A. Bartholomai and H. B. Frieboes, "Lung cancer survival prediction via machine learning regression, classification, and statistical techniques," in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 632–637, 2018, doi: 10.1109/ISSPIT.2018.8642753.

[7] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, “Lung cancer survival prediction using ensemble data mining on SEER data,” *Scientific Programming*, vol. 20, no. 1, pp. 29–42, 2012, doi: 10.1155/2012/920245.

[8] T. Kadir and F. Gleeson, “Lung cancer prediction using machine learning and advanced imaging techniques,” *Translational Lung Cancer Research*, vol. 7, no. 3, pp. 304–312, 2018, doi: 10.21037/tlcr.2018.05.15.

[9] Y. H. Lai, W. N. Chen, T. C. Hsu, C. Lin, Y. Tsao, and S. Wu, “Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning,” *Scientific Reports*, vol. 10, no. 1, 2020, doi: 10.1038/s41598-020-61588-w.

[10] R. Patra, “Prediction of lung cancer using machine learning classifier,” in *Communications in Computer and Information Science*, Singapore: Springer, pp. 132–142, 2020, doi: 10.1007/978-981-15-6648-6_11.

[11] S. P. Venkatesh and L. Raamesh, “Predicting lung cancer survivability: a machine learning ensemble method on SEER data,” *International Journal of Cancer Research & Therapy*, vol. 8, no. 4, pp. 148–154, 2023, doi: 10.33140/ijcrt.08.04.03.

[12] S. Makubhai, G. R. Pathak, and P. R. Chandre, “Comparative analysis of explainable artificial intelligence models for predicting lung cancer using diverse datasets,” *IAES International Journal of Artificial Intelligence*, vol. 13, no. 2, pp. 1978–1989, 2024, doi: 10.11591/ijai.v13.i2.pp1980-1991.

[13] C. M. Lynch *et al.*, “Prediction of lung cancer patient survival via supervised machine learning classification techniques,” *International Journal of Medical Informatics*, vol. 108, pp. 1–8, 2017, doi: 10.1016/j.ijmedinf.2017.09.013.

[14] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation forest,” in *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008, doi: 10.1109/ICDM.2008.17.

[15] J. Jia *et al.*, “Predictive model for totally implanted venous access ports-related long-term complications in patients with lung cancer,” *Oncology Letters*, vol. 28, no. 1, 2024, doi: 10.3892/ol.2024.14459.

[16] B. C. Ross, “Mutual information between discrete and continuous data sets,” *PLoS ONE*, vol. 9, no. 2, 2014, doi: 10.1371/journal.pone.0087357.

[17] P. Liu, K. Jin, Y. Jiao, M. He, and S. Fei, “Prediction of second primary lung cancer patient’s survivability based on improved eigenvector centrality-based feature selection,” *IEEE Access*, vol. 9, pp. 55663–55672, 2021, doi: 10.1109/ACCESS.2021.3063944.

[18] X. Gong, B. Zheng, G. Xu, H. Chen, and C. Chen, “Application of machine learning approaches to predict the 5-year survival status of patients with esophageal cancer,” *Journal of Thoracic Disease*, vol. 13, no. 11, pp. 6240–6251, 2021, doi: 10.21037/jtd-21-1107.

[19] A. S. Azar *et al.*, “Application of machine learning techniques for predicting survival in ovarian cancer,” *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, 2022, doi: 10.1186/s12911-022-02087-y.

[20] L. Yang and A. Shami, “On hyperparameter optimization of machine learning algorithms: theory and practice,” *Neurocomputing*, vol. 415, pp. 295–316, 2020, doi: 10.1016/j.neucom.2020.07.061.

[21] W. Yotsawat, P. Wattuya, and A. Srivihok, “Improved credit scoring model using XGBoost with Bayesian hyper-parameter optimization,” *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, pp. 5477–5487, 2021, doi: 10.11591/ijece.v11i6.pp5477-5487.

[22] M. Mamun, A. Farjana, M. Al Mamun, and M. S. Ahammed, “Lung cancer prediction model using ensemble learning techniques and a systematic review analysis,” in *2022 IEEE World AI IoT Congress (AIIoT)*, 2022, pp. 187–193, doi: 10.1109/AIIoT54504.2022.9817326.

[23] A. Safiyari and R. Javidan, “Predicting lung cancer survivability using ensemble learning methods,” in *2017 Intelligent Systems Conference (IntelliSys)*, 2017, pp. 684–688, doi: 10.1109/IntelliSys.2017.8324368.

[24] N. Marwah, P. Aggarwal, and R. Kaur, “Lung cancer survivability prediction with recursive feature elimination using random forest and ensemble classifiers,” in *2022 2nd International Conference on Computing and Machine Intelligence (ICMI)*, 2022, pp. 1–5, doi: 10.1109/ICMI55296.2022.9873658.

[25] Z. Cheng, C. Zou, and J. Dong, “Outlier detection using isolation forest and local outlier factor,” in *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, 2019, pp. 161–168, doi: 10.1145/3338840.3355641.

[26] A. Abhaya and B. K. Patra, “An efficient method for autoencoder based outlier detection,” *Expert Systems with Applications*, vol. 213, 2023, doi: 10.1016/j.eswa.2022.118904.

[27] J. Wu *et al.*, “Hyperparameter optimization for machine learning models based on bayesian optimization,” *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, 2019, doi: 10.11989/JEST.1674-862X.80904120.

[28] J. Thomgkam, V. Sukmak, and P. Klangnok, “Application of machine learning techniques to predict breast cancer survival,” in *Multi-disciplinary Trends in Artificial Intelligence*, 2021, pp. 141–151, doi: 10.1007/978-3-030-80253-0_13.

[29] C. Schröer, F. Kruse, and J. M. Gómez, “A systematic literature review on applying CRISP-DM process model,” *Procedia Computer Science*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[30] B. F. Hankey, L. A. Ries, and B. K. Edwards, “The surveillance, epidemiology, and end results program: A national resource,” *Cancer Epidemiology Biomarkers and Prevention*, vol. 8, no. 12, pp. 1117–1121, 1999.

[31] L. Alzubaidi *et al.*, “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” *Journal of Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00444-8.

BIOGRAPHIES OF AUTHORS

Wirot Yotsawat received the Ph.D. and M.Sc. degree in computer science from the Faculty of Science, Kasetsart University, Thailand in 2021 and 2014, respectively. He also received his B.Sc. (computer science) from the School of Informatics, Walailak University, Thailand in 2007. He is currently a lecturer at Computer Science Program in Phranakhon Si Ayutthaya Rajabhat University, Thailand. His research includes machine learning, artificial intelligence, and internet of things (IoT). He can be contacted at email: ywirot@aru.ac.th.

Peetiphart Suebpeng received the B.Sc. degree in computer science from the Faculty of Science and Technology, Phranakhon Si Ayutthaya Rajabhat University, Thailand in 2023. He is currently an AI engineer at StoreMesh company, Thailand. His research includes machine learning, image processing, computer vision, and deep learning. He can be contacted at email: pwlnwza@gmail.com.

Saroch Purisangkaha holds a Master of Science (M.Sc.) in software engineering from the School of Information Technology, Sripatum University, Thailand in 2008. He also received his B.Sc. (computer science) from the Faculty of Science and Technology, Phranakhon Si Ayutthaya Rajabhat University, Thailand in 2000. He is currently an assistant professor at Computer Science Program in Phranakhon Si Ayutthaya Rajabhat University, Thailand. His research areas of interest include computer network, software engineering, and software testing. He can be contacted at email: ipreds@aru.ac.th.

Akarapon Poonsawad received the Ph.D. degree in computer education from the Faculty of Technical Education, King Mongkut's University of Technology North Bangkok (KMUTNB), Thailand in 2022. He received the M.Sc. in applied statistics and information technology from the School of Applied Statistics, National Institute of Development Administration, Thailand in 2012. He also received his B.Sc. (computer science) from the Faculty of Science, Srinakharinwirot University, Thailand in 2007. He is currently a lecturer at Computer Science Program in Phranakhon Si Ayutthaya Rajabhat University, Thailand. His research includes computer education, gamification for learning, and information system management. He can be contacted at email: pakarapon@aru.ac.th.

Kanyalag Phodong graduated with bachelor's degree in computer science from Naresuan University, Thailand in 2007. She obtained the M.Sc. and Ph.D. degree in computer science at the Department of Computer Science, Thammasat University, Thailand in 2011 and 2021, respectively. Currently, she is a lecturer at the Program of Computer Science, Faculty of Science and Technology, Phranakhon Si Ayutthaya Rajabhat University, Thailand. Her researches are in fields of artificial intelligence, natural language processing, learning analytics, and information system. She can be contacted at email: kanyalagp@gmail.com.