EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Serum glycobiomarker mining suggested the improvement of cholangiocarcinoma detection using combined CA125 and CA242

Kodchakon Lekkoksung[1)], Atit Silsirivanit[2, 3)], Sukanya Luang[2, 3)], Prasertsri Ma-In[3)] and Sirorat Pattanapairoj*[1, 4)]

[1)]Department of Industrial Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen 40002, Thailand
[2)]Department of Biochemistry, Faculty of Medicine, Khon Kaen University, Khon Kaen 40002, Thailand
[3)]Cholangiocarcinoma Research Institute, Khon Kaen University, Khon Kaen 40002, Thailand
[4)]System Modeling for Industry Research Group, Faculty of Engineering, Khon Kaen University, Khon Kaen 40002, Thailand

Received 9 May 2024
Revised 8 July 2024
Accepted 17 July 2024

Abstract

Cholangiocarcinoma (CCA) is a malignant neoplasm originating from biliary epithelial cells. During the early stage, the patients do not show any symptoms, leading to wide and extensive spread of this disease. Nowadays, there has not been a single serum tumor marker which can be used for effective screening of the disease or classification of the patients. This study therefore aims to determine an appropriate serum marker for screening of the patients with early staged CCA by using a technique of data mining. Beginning with the C4.5 Decision tree and Logistic Regression for selection of serum markers for effective screening of the patients with CCA, the selected markers were then used for classification of the patients with CCA from non-CCA patients, and CCA from Benign Biliary Disease (BBD) by C4.5 Decision tree, Logistic Regression, Random Forest, and Artificial Neural Network. In this work, seven serum tumor markers were used, including Carbohydrate Antigen 125 (CA125), Carbohydrate Antigens 19-9 (CA19-9), Carbohydrate Antigen 242 (CA242), Carbohydrate Antigen 50 (CA50), Carbohydrate Antigen 72-4 (CA72-4), Carcinoembryonic Antigen (CEA), Cy-tokeratin-19Fragment (CYFRA 21-1). The model was used to classify the CCA and non-CCA patients and it was discovered that the serum tumor markers which could most efficiently classify the CCA patients from the non-CCA patients were the combination of CA125 and CA242 suggested by the Logistic Regression with C4.5 Decision tree as the classifier, yielding the best performance, with Sensitivity and Specificity being 75.88 % and 86.82%, respectively. In contrast, the classification of CCA patients from BBD patients was best performed by the serum tumor markers CA125 and CA72-4 suggested by C4.5 Decision tree with Logistic Regression or Random Forest as the classifier.

Keywords: Cholangiocarcinoma, Bile duct, Cancer, Tumor marker, Carbohydrate, Combined

1. Introduction

Cholangiocarcinoma (CCA) is a malignant neoplasm originating from biliary epithelial cells [1]. Today, the occurrence of the disease has continuously increased worldwide [2]. Specifically, in Northeast Thailand, the occurrence rate of the disease is the greatest, averaging 85 cases per 100,000 people [3]. Most patients do not show any symptoms during the early stages of the disease. Thus, the patients miss the opportunity to have the disease detected during the curable stage, leading to an annual fatality of approximately 14,000 deaths annually [4] and increasing every year.

In the past, some researchers tested serum tumor markers such as Cy-tokeratin-19Fragment (CYFRA 21-1) [5], Carbohydrate Antigen 242 (CA-242), Carcinoembryonic Antigen (CEA), Carbohydrate Antigen 19-9 (CA19-9) [6], Carbohydrate Antigen 125 (CA125), CEA [7] and Carbohydrate Antigens 19-9 (CA19-9), Carbohydrate Antigen 50 (CA50) [8] for screening during the early stages of the disease by using a single marker and a combination of markers. It was learned that screening by a single marker was not as efficient as using combinations of markers [9]. Nevertheless, combinations of markers render higher screening costs than a single marker.

The construction of models for classifying CCA patients by using combinations of markers has been continuously researched by using the efficiency indicators or classification accuracy, namely, Sensitivity (SEN), which is the ratio of the number of correct predictions of CCA patients to the total number of CCA patients. Specificity (SPEC), which is the ratio of the correct number of non-CCA patients to the total number of non-CCA patients [10-16] and Accuracy (ACC).

Pattanapairoj et al. [10] improved a method for diagnosing CCA by data mining. The research used a Decision tree to select serum tumor markers for the classification of CCA. The serum tumor markers selected were CCA-CEA CA19-9 ALP GGT BALP MUC5AC CCA-CA and CA-S27. Then The Artificial Neural Network was used in the accuracy test of the classification of CCA by the serum tumor markers selected by the Decision tree. After that Song et al. [11] increased accuracy in early-stage ovarian cancer detection using

*Corresponding author.
Email address: siropa@kku.ac.th
doi: 10.14456/easr.2024.53

16 cancer biomarkers, using various techniques. To choose a marker, they used Genetic Algorithms, Random Forests, t-tests, and Logistic Regression. Methods such as linear discriminant analysis, K-nearest neighbor and Logistic Regression were used for classification. These researchers highlighted the important impact of logistic regression as being efficient for selection of the biomarkers and classification of cancer. The most suitable biomarker combination for detecting ovarian cancer was a fusion of four biomarkers: HE4-ELISA, PDGF-AA, prolactin, and TTR. Next Kimawaha et al. [12] used multi-biomarkers to classify CCA patients from healthy people and other cancer patients. They selected their markers as S100A9 MUC5AC TGF- 1 Angiopoietin-2 and CA19-9 using the Decision tree. They found that S100A9 and CA10-9 yielded the best efficiency in screening CCA patients from healthy people. Later Rustam et al. [13] classified pancreatic cancer using Logistic Regression and Random Forest and using the Pancreatic cancer dataset variables, namely CA19-9, Hemoglobin, Leukocyte, Thrombosis Hematocrit. The results showed that Random Forest is more accurate than Logistic Regression in classifying pancreatic cancer. Furthermore, Mahesh et al. [14] various machine learning algorithms were evaluated in their work to determine the best one for breast cancer prediction. The algorithms that were evaluated included Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Classification and Regression Tree, Naive Bayes, and ensemble methods such as Majority-Voting, XG-Boost, and Random Forest. Majority-Voting, which was built on the top three classifiers (Logistic Regression, Support Vector Machine, and Classification and Regression Tree), was found to offer the highest accuracy of 99.3% in breast cancer classification. Later, Botlagunta et al. [15] compared nine different methods for classifying non-invasive breast cancer using machine learning algorithms based on Blood data. The methods included Logistic Regression, KNN, Decision trees, Random Forest, SVM (SVM linear, SVM radial), Gradient Boosting, and XG-boost. The results of the study revealed that the Decision tree classifier was the most accurate in identifying breast cancer metastasis. Chandrashekar et al. [16] created a decision support system (DSS) and a web-based application that can accurately classify different types of cancer using supervised machine learning algorithms. They evaluated 20 cancer exome datasets with five types of cancer included in the analysis. They employed various methods such as K-nearest Neighbor, Support Vector Machine, Decision tree, Naïve Bayes, and Random Forest to achieve the best possible accuracy in cancer classification. The results indicated that the Decision tree and Random Forest algorithms were the most accurate in classifying the different types of cancer. Additionally, four ovarian cancer biomarkers, HE4-ELISA, PDGF-AA, Prolactin and TTR, contributed significantly to this efficiency.

In this work, we attempted to select practical serum tumor markers for efficient classification of CCA patients by using a model built from a technique of data mining. Using the C4.5 Decision tree and Logistic Regression to select the markers suitable for efficient screening of the CCA patients, the markers were then used to classify the CCA patients from the non-CCA patients and the CCA patients from the BBD patients. The model was built from a C4.5 Decision tree, Logistic Regression, Artificial Neural Network and Random Forest. The novelty of the present study is the comparison of Decision Tree and Logistics Regression as marker selectors for the classification models. Additionally, four classifiers were used and evaluated for effectiveness in classifying the CCA patients and BBD patients, and classifying CCA and Non-CCA. The previous models by Pattanapairoj et al. [10] used only C4.5 Decision Tree as the marker selector with ANN as the classifier whereas Kimawaha et al. [12] also used only Decision Tree as both the selector and classifier.

2. Materials and methods

2.1 Patients and serum tumor samples

Serum samples were obtained from the specimen bank of the Cholangiocarcinoma Research Institute, Khon Kaen University, Khon Kaen, Thailand. Informed consent was obtained from each subject, and the study protocol was approved by Khon Kaen University Ethics Committee in Human Research (HE621176). Healthy controls (HE) recruited in this study were age-sex matched individuals, presented normal fasting plasma glucose (< 100 mg/dl) and normal liver and renal functions, having annual check-up at Srinagarind Hospital, Khon Kaen University. The samples used in this study are including those from 110 HE, 23 patients with benign biliary disease (BBD), 33 patients with other gastrointestinal cancers (GI-CA), and 86 patients with CCA, totaling 252 patients. HE, BBD, and GI-CA were defined non-CCA. The data collected from the subjects were divided into two sets, namely the training set and the test set, for classification of the patients by the model. For the classification of CCA patients from non-CCA patients, the data were divided into two groups according to the method of 5-fold cross validation and the classification of CCA patients from BBD patients by the split test method (70%, 30%). This was due to the number of BBD subjects not being sufficient for the method of 5-fold cross validation.

2.2 Determination of serum tumor markers

All serum tumor markers, Carbohydrate Antigen 125 (CA125), Carbohydrate Antigens 19-9 (CA19-9), Carbohydrate Antigen 242 (CA242), Carbohydrate Antigen 50 (CA50), Carbohydrate Antigen 72-4 (CA72-4), Carcinoembryonic Antigen (CEA) and Cytokeratin-19Fragment (CYFRA21-1), were measured by enzyme-linked immunosorbent assay-based methods in the automatic MAGLUMI®800 Chemiluminescence Immunoassay (CLIA) Analyzer (Shenzhen New Industries Biomedical Engineering (SNIBE), Shenzhen, China), according to the instruction of manufacture.

2.3 Construction of classification model

In the present work, construction of the model for classification of CCA patients began by selecting serum tumor markers for classification of CCA patients from non-CCA patients and classification of CCA patients from BBD patients using C4.5 Decision tree and Logistic Regression.

C4.5 stands as an algorithm employed for constructing a decision tree classification model presented in a logical format [17]. The model constructs a decision tree based on the information gain principle [18]. The information gain indicates the capacity to classify the data using a specific marker. From the tree's top node, C4.5 systematically identifies the attribute with the utmost information gain to categorize the data in a "top-down recursive divide-and-conquer manner" [19]. Upon reaching a branch node, C4.5 repeats the process to select another attribute from the remaining set. These iterations persist until no further attributes are available. Additional insights into C4.5 can be found in [17].

Logistic regression is the most popularly used for binary data [20], such as sick/healthy and yes/no. It is also a method for data analysis where the relationship between the independent variable and the dependent variable results in only two values, 0 or 1 [21]. For this method, the data were statistically significant if the P-value < 0.05, meaning that the relationship between the dependent and independent variables was statistically significant.

Then, the selected markers were used in constructing of the model for classification of CCA patients from non-CCA patients, and classification of CCA patients from BBD patients by C4.5 Decision tree, Logistic Regression, Random Forest, and Artificial Neural Network. The results of these four techniques in the model were then compared to the most suitable serum tumor markers for classification of CCA patients.

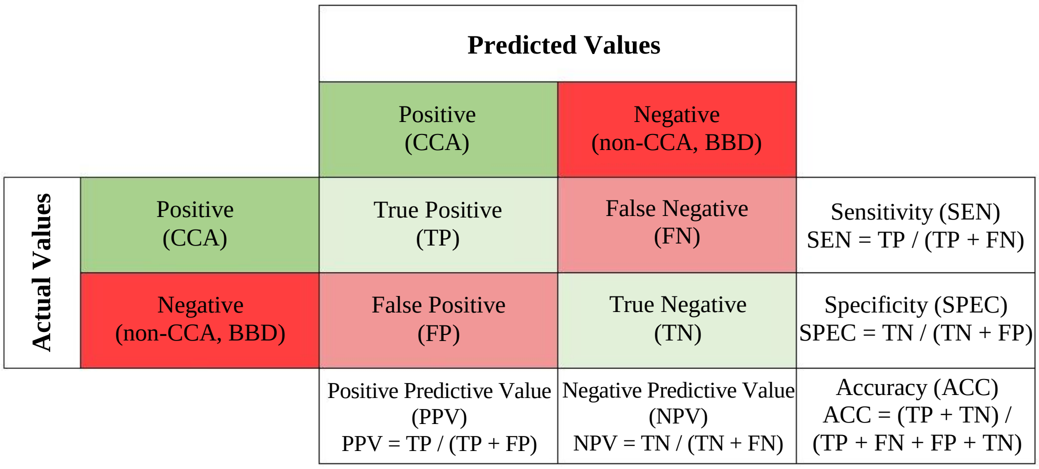
Random Forest is a technique that is extensively used for a variety of problems. It can be used for regression problems and classification problems [22]. The technique was developed from the Decision tree, with the difference being an increase in the number of trees in the Random Forest method [23]. The increased number of trees permits different data with different characteristics.

Artificial Neural Network is an algorithm developed for simulating the working of neural networks in the human brain, which can process, learn, and remember information for solving various problems. The technique allows the efficient construction of the relationship between inputs and outputs by adjusting the connection weights using single-hidden layer feedforward and training using the backpropagation algorithm [24].

A model was built to classify data using C4.5 Decision trees, Logistic Regression, Random Forest, and Artificial Neural Network. The model took combined markers from the C4.5 Decision tree and Logistic Regression as inputs and outputted the diagnostic result as either "CCA" or "non-CCA" and "CCA" or "BBD." To build the C4.5 model, the combination of markers was experimented with using binary splits of at least five instances per leaf and a confidence threshold for pruning of 0.25 as a setting. The ANN model was constructed using 20 hidden units, a sigmoid function, and 50,000 learning cycles. Based on our literature review, it was found that the ANN parameters were suitable for this study. They generally give high prediction efficiency and accuracy [10]. All models were developed using WEKA [25]. An open-source machine learning software. The model's accuracy was validated using a 5-fold cross-validation and split test of 70% and 30%, respectively.

2.4 Performance evaluation and experimental trials

The developed model must be evaluated to determine its accuracy and efficiency in evaluation or classification of the patients as desired. For this purpose, this research adopted the confusion matrix, as shown in Figure 1, to measure the performance of data classification.



		Predicted Values		
		Positive (CCA)	Negative (non-CCA, BBD)	
Actual Values	Positive (CCA)	True Positive (TP)	False Negative (FN)	Sensitivity (SEN) SEN = TP / (TP + FN)
	Negative (non-CCA, BBD)	False Positive (FP)	True Negative (TN)	Specificity (SPEC) SPEC = TN / (TN + FP)
		Positive Predictive Value (PPV) PPV = TP / (TP + FP)	Negative Predictive Value (NPV) NPV = TN / (TN + FN)	Accuricy (ACC) ACC = (TP + TN) / (TP + FN + FP + TN)

Figure 1 Confusion Matrix

From Figure 1, the confusion matrix is a 2 × 2 table, with the horizontal axis representing the actual results and the vertical axis indicating the predicted results. The names and meanings of each cell in the table are explained as follows: True Positive (TP) means a correct prediction by the model that the patients were CCA, True Negative (TN) means a correct prediction the patients were non-CCA and BBD, False Positive (FP) means a false prediction that the patients were CCA, False Negative (FN) means a false prediction that the patients were non-CCA and BBD.

The values TP, FP, TN, and FN can be used to calculate various values of the matrices to evaluate the classification performance of the model. The matrices used include Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value, and Accuracy. Each matrix can be calculated using the following equations: Sensitivity (SEN) is the ratio of the correct predictions of CCA patients to the total number of CCA patients, which can be determined as SEN = TP / (TP + FN), Specificity (SPEC) is the ratio of the number of correct predictions of non-CCA and BBD patients to the total number of non-CCA and BDD patients as SPEC = TN / (TN + FP), Positive Predictive Value (PPV) is the ratio of the correct predictions of CCA patients to the total number of CCA patients predicted by the model as PPV = TP / (TP + FP), Negative Predictive Value (NPV) is the ratio of the correct predictions non-CCA and BBD patients to the total number of non-CCA and BBD patients as NPV = TN / (TN + FN), Accuracy (ACC) is the ratio of the number of correct predictions to the total number of predictions as ACC = (TP + TN) / (TP + FN + FP + TN).

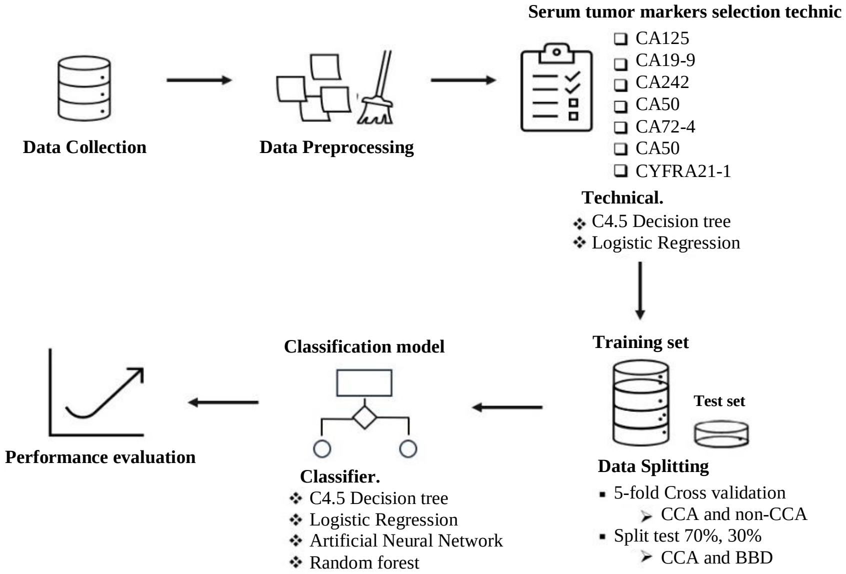


Figure 2 Process diagram of construction of the model for CCA patient classification

Figure 2 depicts the process diagram for building the model for CCA classification. Step 1 deals with CCA data collection (raw data). Step 2 involves data preprocessing by cleaning the data. This step is essential because it could influence the analysis in a negative way, such as erratic analytical results. Step 3 involves selecting serum tumor markers to classify CCA patients from non-CCA patients and CCA patients from BBD patients by C4.5 Decision tree and Logistic Regression will be able to know serum tumor markers, But Random Forest doesn't need to know. Serum tumor markers are compatible with classification instead. Step 4 deals with data splitting, which divides the data into two groups, namely Training Set and Test Set, for testing the model's efficiency. To classify the CCA patients from the non-CCA patients, division of the data, training, and testing of the data followed the method of 5-fold cross-validation, whereas classification of the CCA patients from the BBD patients followed the 70%, 30% Split Test. The data were randomly sampled for training and testing to reduce experimental bias. This process was repeated twenty times (twenty replications), and the average values were taken as the results. Step 5 deals with constructing the model for classifying the CCA patients by C4.5 Decision tree, Logistic Regression, Artificial Neural Network, and Random Forest. Finally, Step 6 is performance evaluation.

3. Results

3.1 Single serum tumor marker diagnosis of CCA patients and non-CCA patients

The results of classification of the CCA patients from the non-CCA patients by each individual serum tumor marker as determined by the cut-off values as per the standards of the hospital (Presented as Column 2 of Table 1) revealed that use of a single serum tumor marker yielded a high ratio of non-CCA to the total number of non-CCA (SPEC) whereas the ratio of the number of correct predicted of CCA to the total number of CCA (SEN) was low. In addition, it was found that the use of the cut-off value of single marker in classifying the patients yielded a high error in FN, leading to error in screening, denial of treatment of the patients with CCA, and hence possible deaths. Even though use of CA125 gave the highest value SEN, which was 72.71 ± 6.78, and the lowest value of FN which was 27.29 ± 6.78. However, the variance was as high as 6.78 as compared with other markers.

Table 1 Diagnostic values of single marker for diagnosis of CCA patients and non-CCA patients

Marker	Cut-Off Value	SEN	SPEC	PPV	NPV	ACC	FP	FN
CA125	25 U/ml	72.71 ± 6.78	90.06 ± 5.09	80.23 ± 8.75	86.59±2.79	84.15 ±3.60	9.94 ±5.09	27.29 ± 6.78
CA19-9	37 U/ml	45.67 ± 1.53	94.07 ± 1.68	80.71 ± 3.54	77. 12 ±0.26	77.56 ±0.66	5.93 ±1.68	54.33 ± 1.53
CA242	20 U/ml	38.38 ± 3.56	93.19 ± 1.75	76.15 ± 2.66	74.58 ± 0.52	73.95 ±3.19	7.76 ±4.39	61.51 ± 2.22
CA50	25 U/ml	58.79 ± 2.49	86.81 ± 1.43	71.53 ± 1.86	80.42 ± 0.66	76.88 ±2.88	13.82 ± 3.53	41.18 ± 2.09
CA72-4	6 U/ml	59.35 ± 4.45	86.81 ± 2.13	71.80 ± 1.95	80.69 ± 1.31	77.02 ±3.28	13.76 ± 3.79	40.47 ± 4.37
CEA	5 ng/ml	45.31 ± 2.92	92.11 ± 7.23	82.49 ± 3.97	77.06 ± 0.58	76.86 ±3.27	6.6.5 ±4.33	54.78 ± 3.34
CYFRA21-1	3.3 ng/ml	58.20 ± 1.53	95.65 ± 2.41	89.55 ± 2.06	81.98 ± 1.07	82.32 ±3.55	4.79 ±3.76	41.78 ± 2.65

SEN = Sensitivity; SPEC = Specificity; PPV = Positive Predictive Value; NPV = Negative Predictive Value; ACC = Accuracy; FP = False Positive; FN = False Negative

3.2 Combination of serum tumor markers using the C4.5 Decision tree method and logistic regression method

CA125, CA50, CA72-4, and CEA which were selected by C4.5 Decision tree technique were used as markers and inputs to Classification Model 1 whereas CA125 and CA242 markers selected by the Logistic Regression method were inputs to Classification Model 2. The two models used a C4.5 Decision tree, Logistic Regression, Artificial Neural Network, and Random Forest in classifying the CCA patients from the non-CCA patients with the results being compared in Figure 3.

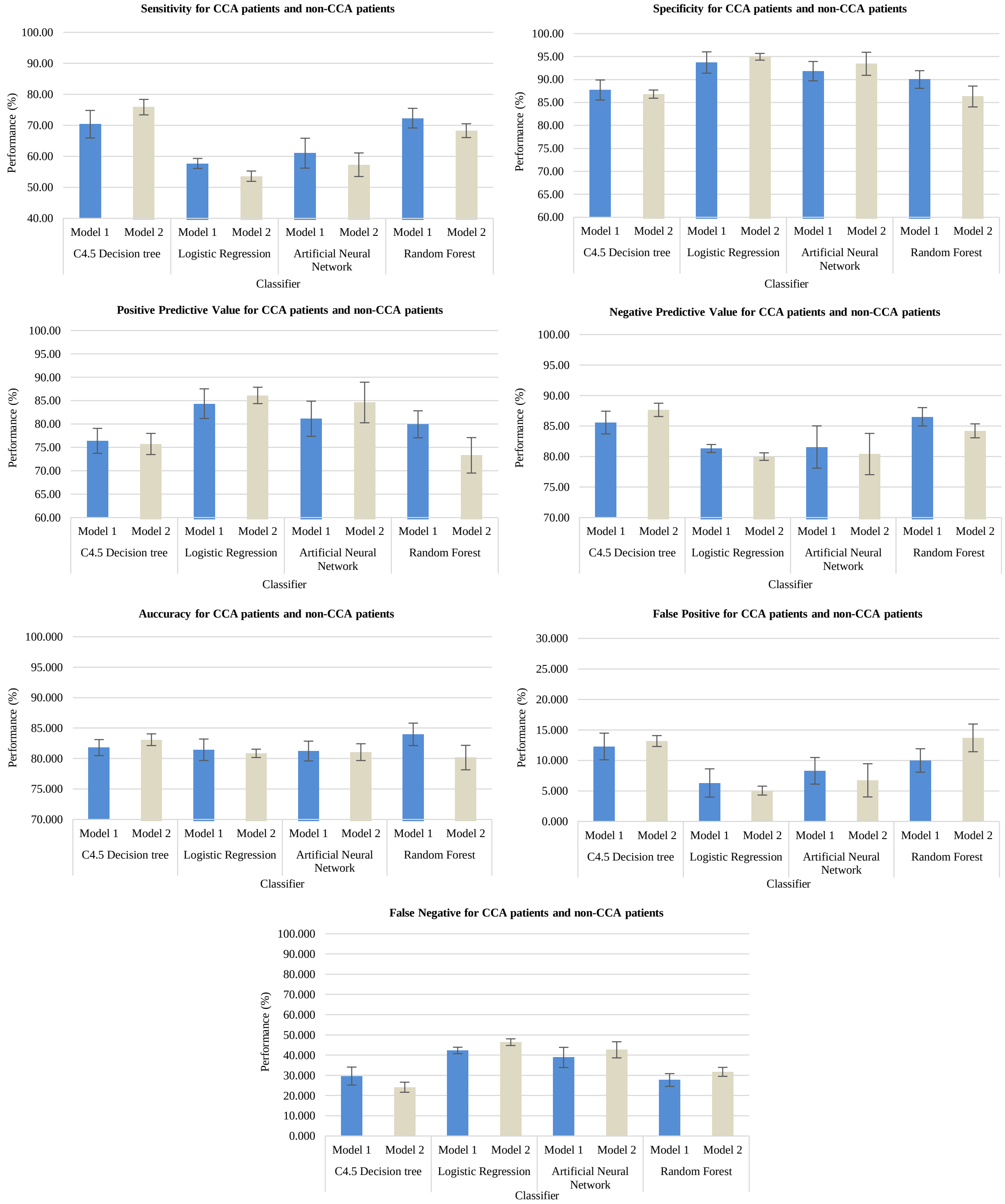


Figure 3 Diagnostic values of combined-multiple markers suggested by C4.5 Decision tree and Logistic Regression for CCA patients and non-CCA patients

3.3 The serum tumor markers have the highest diagnostic Performance for identifying patients with CCA and non-CCA

In Figure 3, the results of CCA and non-CCA classification using Model 1 and Model 2 which employed different markers and 4 different classifiers previously discussed revealed that the model employing Logistic Regression, Artificial Neural Network, and Random Forest yielded good results when CA125, CA50, CA72-4, and CEA were used. In contrast, the model using C4.5 Decision tree yielded good results when CA125 and CA242 were used. This could be since use of too many input variables with the C4.5 Decision tree technique might create complexity data, leading to overfitting [26]. On the other hand, the ANN technique is suitable for learning and finding the relationship among the data. The higher the number of input variables, the better the ANN technique learns and finds the complex relationship of data [17]. Similarly, for Logistic Regression, a high number of input variables yields more accurate output variables [21]. Lastly, the Random Forest technique which make uses of several decision trees with different features, hence a greater diversity of features. The higher number of input variables results in more diverse features for each decision tree, and thus avoiding overfitting [22]. When the performance of the models in classifying the CCA and non-CCA patients by using different markers and classifiers, it was found that the model which had the highest efficiency was which one that used two markers (CA125 and CA242) suggested by Logistic Regression with C4.5 Decision tree as the classifier. The model yielded the highest ratio of the number of correct predictions of CCA patients to the total number of CCA patients SEN (75.88 ± 2.49) with the lowest FN (24.12 ± 2.49) and the highest accuracy (ACC) of 87.65 ± 1.09. The model yielded the lowest number of patients with treatment opportunity loss.

3.4 Singer serum tumor marker diagnosis of CCA patients and BBD patients

The results of classification of the CCA patients from the BBD patients by each individual serum tumor marker as determined by the cut-off values as per the standards of the hospital (presented as Column 2 of Table 2) revealed that. Similar results to the classification of CCA patients from the non-CCA patients.

Table 2 Diagnostic values of single marker for diagnosis of CCA patients and BBD patients

Marker	Cut-Off Value	SEN	SPEC	PPV	NPV	ACC	FP	FN
CA125	25 U/ml	66.37 ± 6.75	81.43±13.19	67.05 ± 6.19	73.10 ± 6.59	82.49±10.14	18.57±13.19	33.63 ± 6.75
CA19-9	37 U/ml	56.81 ± 6.62	52.86±13.98	50.37 ± 9.68	55.07 ± 8.24	59.40 ± 8.33	47.14±13.98	43.19 ± 6.62
CA242	20 U/ml	40.95 ± 6.51	81.43 ± 8.16	53.18 ± 4.12	59.06 ± 6.01	71.81±12.00	18.57 ± 8.16	59.05 ± 6.51
CA50	25 U/ml	55.61 ± 7.49	52.86±13.98	49.31 ± 9.51	54.14 ± 8.34	58.39 ± 9.11	47.14±13.98	44.84 ± 7.49
CA72-4	6 U/ml	40.74 ± 7.55	85.71±11.35	54.29 ± 5.00	69.86 ± 6.70	78.67±14.80	14.29±11.35	59.26 ± 7.55
CEA	5 ng/ml	54.98 ± 6.04	72.86±13.83	57.49 ± 5.90	62.98 ± 7.15	71.97±11.46	27.14±13.83	45.02 ± 6.04
CYFRA21-1	3.3 ng/ml	60.09 ± 7.69	92.86±11.82	66.14 ± 4.91	74.77 ± 5.60	92.88±11.48	7.14 ± 11.82	39.91 ± 7.69

SEN = Sensitivity; SPEC = Specificity; PPV = Positive Predictive Value; NPV = Negative Predictive Value; ACC = Accuracy; FP = False Positive; FN = False Negative

3.5 Combination of serum tumor markers using the C4.5 Decision tree method and logistic regression method

CA125 and CA72-4 were the markers which were selected by C4.5 Decision tree technique and were used as inputs to Classification Model 3 whereas CA125 Markers selected by the Logistic Regression method were inputs to Classification Model 4. The two models used C4.5 Decision tree, Logistic Regression, Artificial Neural Network, and Random Forest in classifying the CCA patients from the BBD patients with the results being compared in Figure 4.

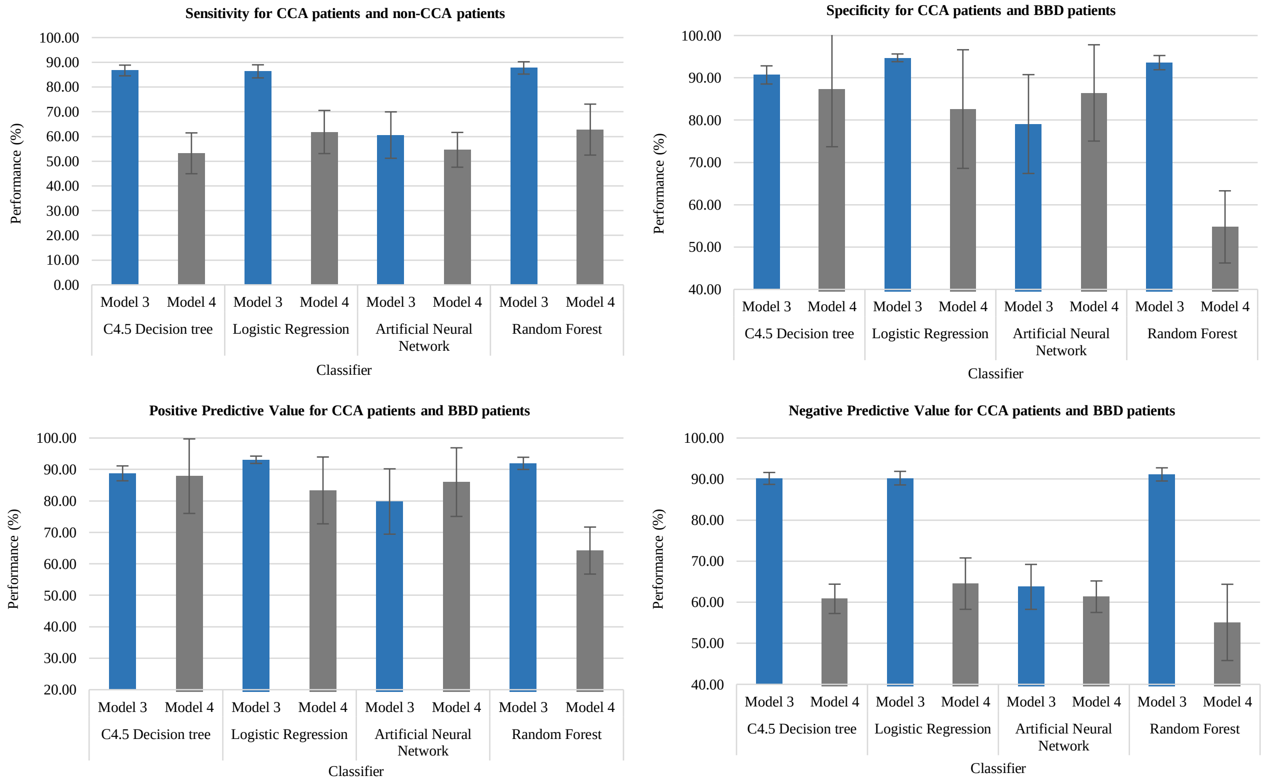


Figure 4 Diagnostic values of combined-multiple markers suggested by C4.5 Decision tree and Logistic Regression for CCA patients and BBD patients

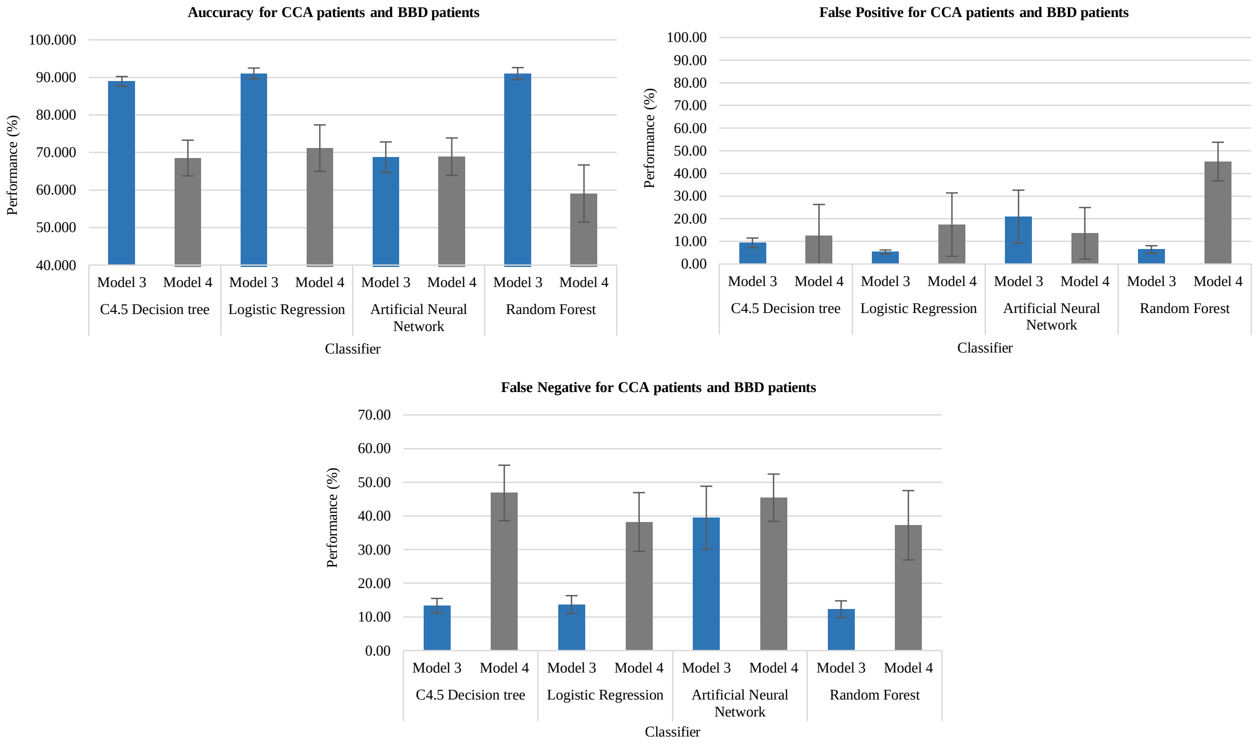


Figure 4 (continued) Diagnostic values of combined-multiple markers suggested by C4.5 Decision tree and Logistic Regression for CCA patients and BBD patients

3.6. The serum tumor markers have the highest diagnostic Performance for identifying patients with CCA and BBD

In Figure 4, the results of the classification of CCA patients from the BBD patients by Model 3 and Model 4 using different markers and Classifier 4, as in the case of the classification of CCA patients from the non-CCA patients disclosed that the model which performed the best in classifying the CCA patients from the BBD patients was Model 3 which used the two markers (CA125 and CA72-4) suggested by C4.5 Decision tree, and Logistic Regression and Random Forest as classifiers. The model yielded the best classification performance with statistically insignificant results at 0.05.

4. Discussion

Diagnosis of CCA using serum tumor markers still poses some problems since the markers used are for diagnosing cancers in general. The markers are not specifically for CCA diagnosis. Thus, using a single marker is inefficient for screening CCA. This study therefore employed the techniques in data mining to determine combinations of markers which could efficiently classify CCA patients from the non-CCA patients and the BBD patients. Use of the Decision Tree and Logistic Regression to select the markers for efficient screening of the CCA patients is deemed suitable C4.5 Decision tree is easily adaptable to actual application with the result presented in the form of a decision tree which is readily understandable. Furthermore the technique is widely used in medical research [10-12]. Logistic Regression was adopted due to its suitability for binary data such as sick/healthy or yes/no decision [21]. The technique can identify factors influencing predicted results by using the statistical P-value to determine statistical significance of the independent variables on the dependent variables.

Table 3 Classification Performance of the C4.5 Decision tree for CCA patients and non- CCA patients

Input Variables	**Classification Performance of the C4.5 Decision tree**						
	SEN	**SPEC**	**PPV**	**NPV**	**ACC**	**FP**	**FN**
CA125, CA242	75.88 ± 2.49	86.82 ± 0.89	75.74± 2.26	87.6± 1.09	83.09± 0.96	13.18 ± 0.89	24.12 ± 2.49

SEN = Sensitivity; SPEC = Specificity; PPV = Positive Predictive Value; NPV = Negative Predictive Value; ACC = Accuracy; FP = False Positive; FN = False Negative

Table 3 depicts the results of the classification of the CCA patients from the non-CCA patients of Model 2, which used two markers in combination (CA125 and CA242) suggested by Logistic Regression and C4.5 Decision tree as the classifier. The model predicted the classification of the CCA patients from the non-CCA patients with SEN and SPEC being 75.88 ± 2.49 and 86.82 ± 0.89, respectively. Figure 2 shows the structure of the classification scheme. The structure of the classification scheme is simple and convenient for physicians to use in the diagnosing of the CCA patients**.**

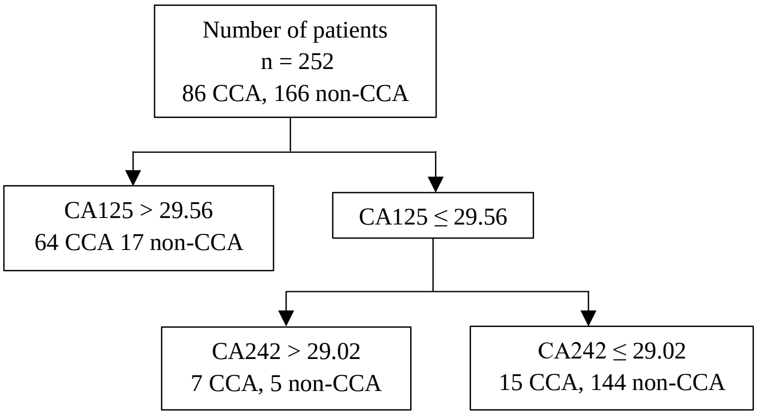


Figure 5 Structure of Classification Scheme using C4.5 Decision tree

As illustrated in Figure 5, the classification of the CCA patients from the non-CCA patients begins by considering the serum tumor markers CA125. If CA125 > 29.56, the model yields 64 CCA patients and 17 non-CCA patients. If CA125 ≤ 29.56, consider using CA242 in combination with CA125. In this case, if CA125 ≤ 29.56 and CA242 ≤ 29.02, the model gives 15 CCA patients and 144 non-CCA patients. If CA125 ≤ 29.56 and CA242 > 29.02, the model gives 7 CCA patients and 5 non-CCA patients.

Table 4 Classification Performance of the Logistic Regression and Random Forest for CCA patients and BBD patients

Input Variables	**Classification Performance of the Logistic Regression**						
	SEN	**SPEC**	**PPV**	**NPV**	**ACC**	**FP**	**FN**
CA125, CA72-4	86.34 ± 2.65	94.73 ± 0.91	93.09 ± 1.16	90.23 ± 1.66	91.04 ±1.44	5.27 ± 0.91	13.66 + 2.65
	Classification Performance of the Random Forest						
	SEN	**SPEC**	**PPV**	**NPV**	**ACC**	**FP**	**FN**
	87.72 ± 2.49	93.59 ± 1.68	91.95 ± 1.93	91.14 ± 1.61	91.02± 1.59	6.41 ± 1.68	12.28 ± 2.49

SEN = Sensitivity; SPEC = Specificity; PPV = Positive Predictive Value; NPV = Negative Predictive Value; ACC = Accuracy; FP = False Positive; FN = False Negative

Table 4 shows the results of the classification of CCA patients from the BBD patients. In this case, Model 3 which used a combination of CA125 and CA72-4 suggested by C4.5 Decision tree and Logistic Regression and Random Forest as the classifiers. The model yielded the best classification performance with statistically insignificant results at the level of 0.05.

Results from research by several researchers discovered that use of several serum tumor markers (combined analysis) could increase the efficiency of the classification of CCA patients. Algorithms have been used in selecting serum tumor markers and modeling for the diagnosis of CCA patients [10, 12]. In their research, Pattanapairoj et al. [10] in the classification of CCA patients from non-CCA patients. They found SEN and SPEC to be 98.71% and 96.94%, respectively. Kimawaha et al. [12] used CA19-9 and S100A9 to classify CCA patients from healthy individuals. In this case, SEN and SPEC were 93% and 80 %, respectively. In the present study, combined use of CA125 and CA242 suggested by Logistic Regression with C4.5 Decision tree as the classifier yielded the best performance (the highest SEN value of 75.88 %) in classifying CCA patients from non-CCA patients. It is also worth noting that the SPEC was 86.82 % in this case. In the classification of CCA patients from BBD patients, CA125 and CA72-4, suggested by C4.5 Decision tree with Logistic Regression or Random Forest as the classifiers. Gave similarly the best performance with SEN and SPEC being 86.34 % and 94.73 %, respectively for Logistic Regression, and 87.72 % and 93.59 %, respectively for Random Forest. The differences between these respective results were not statistically significant. It is worth noting that combined analysis was also applied in the diagnosis of other types of cancer since the technique is more effective than using a single serum tumor marker, hence enhancing treatment opportunity and cure from the ailment [6, 10-13].

5. Conclusions

In the present work, mathematical models were constructed for the purpose of classifying CCA patients from non-CCA patients, and classification of CCA patients from BBD patients. Data mining techniques were employed in selecting serum tumor markers and classifiers. The total number of four models were studied, two models for the classification of CCA patients from non-CCA patients and the other two models for CCA patients from BBD patients. The study found that the use of a single serum tumor marker was less effective than combined serum tumor markers. This result agreed with the results of other researchers. For classification of CCA patients from non-CCA patients, it was discovered that CA125 and CA242 suggested by Logistic Regression with C4.5 Decision tree as the classifier, yielded the best performance, whereas classification of CCA patients from BBD patients was best performed by the serum tumor markers CA125 and CA72-4, suggested by C4.5 Decision tree with Logistic Regression or Random Forest as the classifier.

6. Acknowledgements

This work was supported or funded by the Research System Modeling for Industry (SMI) program, which is part of the Department of Industrial Engineering in the Faculty of Engineering at Khon Kaen University, Thailand. Moreover, the work was also supported by the Graduate School, and the Faculty of Engineering of Khon Kaen University and Faculty of Medicine (Grant Number IN63273), Khon Kaen University, Thailand. The authors would like to express their gratitude to Assoc. Prof. Dr. Somnuk Theerakulpisut for proofreading and suggestions for manuscript preparation.

7. References

[1] Goral V. Cholangiocarcinoma: new insights. Asian Pac J Cancer Prev. 2017;18(6):1469-73.

[2] Chaiteerakij R, Pan-ngum W, Poovorawan K, Soonthornworasiri N, Treeprasertsuk S, Phaosawasdi K. Characteristics and outcomes of cholangiocarcinoma by region in Thailand: a nationwide study. World J Gastroenterol. 2017;23(39):7160-7.

[3] Khan SA, Tavolari S, Brandi G. Cholangiocarcinoma: epidemiology and risk factors. Liver Int. 2019;39(S1):19-31.

[4] Seeherunwong A, Chaiear N, Khuntikeo N, Ekpanyaskul C. The proportion of occupationally related cholangiocarcinoma: a tertiary hospital study in Northeastern Thailand. Cancers (Basel). 2022;14(10):2386.

[5] Uenishi T, Yamazaki O, Tanaka H, Takemura S, Yamamoto T, Tanaka S, et al. Serum cytokeratin 19 fragment (CYFRA21-1) as a prognostic factor in intrahepatic cholangiocarcinoma. Ann Surg Oncol. 2008;15(2):583-9.

[6] Zhang Y, Yang J, Li H, Wu Y, Zhang H, Chen W. Tumor markers CA19-9, CA242 and CEA in the diagnosis of pancreatic cancer: a meta-analysis. Int J Clin Exp Med. 2015;8(7):11683-91

[7] Qiu Y, He J, Chen X, Huang P, Hu K, Yan H. The diagnostic value of five serum tumor markers for patients with cholangiocarcinoma. Clin Chim Acta. 2018;480:186-92.

[8] Luang S, Teeravirote K, Saentaweesuk W, Ma-In P, Silsirivanit A. Carbohydrate antigen 50: values for diagnosis and prognostic prediction of intrahepatic cholangiocarcinoma. Medicina. 2020;56(11):616.

[9] Wongkham S, Silsirivanit A. State of serum markers for detection of cholangiocarcinoma. Asian Pac J Cancer Prev. 2012;13(Suppl):17-27.

[10] Pattanapairoj S, Silsirivanit A, Muisuk K, Seubwai W, Cha'on U, Vaeteewoottacharn K, et al. Improve discrimination power of serum markers for diagnosis of cholangiocarcinoma using data mining-based approach. Clin Biochem. 2015;48(10-11):668-73.

[11] Song HJ, Yang ES, Kim JD, Park CY, Kyung MS, Kim YS. Best serum biomarker combination for ovarian cancer classification. Biomed Eng Online. 2018;17(S2):152.

[12] Kimawaha P, Jusakul A, Junsawang P, Thanan R, Titapun A, Khuntikeo N, et al. Establishment of a potential serum biomarker panel for the diagnosis and prognosis of cholangiocarcinoma using decision tree algorithms. Diagnostics (Basel). 2021;11(4):589.

[13] Rustam Z, Zhafarina F, Saragih GS, Hartini S. Pancreatic cancer classification using logistic regression and random forest. IAES Int J Artif Intell. 2021;10(2):476-81.

[14] Mahesh TR, Vinoth Kumar V, Dhilip Kumar V, Geman O, Margala M, Guduri M. The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification. Healthcare Analytics. 2023;4:100247.

[15] Botlagunta M, Botlagunta MD, Myneni MB, Lakshmi D, Nayyar A, Gullapalli JS, et al. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. Sci Rep. 2023;13(1):485.

[16] Chandrashekar K, Setlur AS, Sabhapathi CA, Raiker SS, Singh S, Niranjan V. Decision support system and web-application using supervised machine learning algorithms for easy cancer classifications. Cancer Inform. 2023;22:1-18.

[17] Bishop CM. Pattern recognition and machine learning. 2nd ed. New York: Springer; 2006.

[18] Smith DR. Top-down synthesis of divide-and-conquer algorithms. Artif Intell. 1985;27(1):43-96.

[19] Salzberg SL. C4.5: programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Mach Learn. 1994;16(3):235-40.

[20] Cramer JS. The origins of logistic regression. Tinbergen Institute Working Paper No. 2002-119/4. Amsterdam: Tinbergen Institute; 2002.

[21] Lu K. Logistic regression in biomedical study. 2022 International Conference on Biotechnology, Life Science and Medical Engineering (BLSME 2022); 2022 Jan 22-23; Jeju, South Korea. p. 589-95.

[22] Ahmad I, Basheri M, Iqbal MJ, Rahim A. Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. IEEE Access. 2018;6:33789-95.

[23] Wu Z, Lin W, Zhang Z, Wen A, Lin L. An ensemble random forest algorithm for insurance big data analysis. 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC); 2017 Jul 21-24; Guangzhou, China. New York: IEEE; 2017. p. 531-6.

[24] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533-6.

[25] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter. 2009;11(1):10-8.

[26] Rim P, Liu E. Optimizing the C4.5 decision tree algorithm using MSD-Splitting. IJACSA. 2020;11(10):41-7.