ICIC Express Letters
Part B: Applications
Volume **15**, Number **5**, May **2024**

ICIC International ©2024 ISSN 2185-2766
pp. 479–486

RISK PREDICTION OF GASTRIC CANCER USING DECISION TREE J48

Navamin Suwannapool and Sanya Khruahong*

Department of Computer Science and Information Technology
Faculty of Science
Naresuan University
99 Moo 9, Thapo Sub-district, Muang District, Phitsanulok 65000, Thailand
navamins60@nu.ac.th; *Corresponding author: sanyak@nu.ac.th

Received September 2023; accepted November 2023

Abstract. *One of the most common forms of cancer in Thailand is gastric cancer. Knowing the likelihood of developing this malignancy is crucial. This work presents a risk prediction model for stomach cancer using Decision Tree J48. This study used a dataset of 1,000 patients from the Thai province of Phitsanulok who had six gastric cancer risk markers, including passive smoking, exhaustion, clubbing of the fingernails, obesity, snoring, and dry cough. The percentage split approach (70 : 30, 80 : 20, and 90 : 10) and the 5-fold and 10-fold Cross-Validation methods were used to assess the model's performance. The 10-fold Cross-Validation approach had the highest accuracy of all the methods tested, with a performance model accuracy score of 1.00, according to the data. The F1-score achieved using the other approaches was 0.977 using the 5-fold Cross-Validation method, and 0.953, 0.944, and 0.974 using the 70 : 30, 80 : 20, and 90 : 10 split methods, respectively. We used the findings from Decision Tree J48 to create a web application based on six risk indicators and test it using 30 simulation patterns. According to the findings, all 30 patterns represented 100% of the total success rate.*

Keywords: Data mining, Gastric cancer, Decision Tree J48, Prediction

1. Introduction.

Gastric cancer, originating from the stomach lining, remains a pressing health concern in Thailand, ranking as the second leading cause of cancer-related deaths in the country [1]. Despite witnessing a decline in its prevalence over the last decade, it still poses significant challenges. As per the National Cancer Institute of Thailand, it ranks fifth in commonality among Thai males and ninth among females. An alarming estimation suggests about 8,900 new cases for 2020, accounting for 6% of all cancer diagnoses. Furthermore, its high mortality rate is evident, with a projected 5,700 deaths in the same year, making up 5% of cancer deaths [2]. Several factors amplify the risk in the Thai population, including a diet rich in salt and processed meats, smoking habits, genetic predispositions, and chronic Helicobacter pylori infection. As prevention and early detection emerge as key strategies against this disease, focusing on a healthier lifestyle, accessible screening programs, and effective treatment plans becomes paramount. Consequently, integrating advanced technology to evaluate cancer risks is paramount in this fight.

Many studies have investigated cancer risk, but some of them have problems. According to some studies, an imaging signature [3] could improve prognostic forecasting and make it easier to recognize gastric cancer patients who benefit most from adjuvant treatment. Nevertheless, it is still essential to gather vast amounts of image data. Nevertheless, data mining uses advanced data analysis techniques to find patterns, correlations, and linkages in massive datasets [4,5]. Analyzing patient data to find trends and patterns that could aid with patient care and outcomes is one of the typical applications of data mining in

DOI: 10.24507/icicelb.15.05.479

healthcare. According to a report in the *Journal of the National Cancer Institute*, a recent study found that "early diagnosis of gastric cancer is associated with improved survival" [6]. More than 25,000 patients with stomach cancer were included in the trial, and data analysis revealed that early diagnoses had a much greater five-year survival rate than late diagnoses [6]. In big datasets, patterns, and associations that may not be immediately obvious to humans can be found using data mining approaches, such as Decision Tree, according to a study published in Cancer Epidemiology, Biomarkers & Prevention [7,8]. As early detection and therapy are essential for increasing the likelihood of successful treatment and lowering the invasiveness and cost of treatment, this can be especially helpful for predicting gastric cancer [9]. Decision Tree algorithms are easy to comprehend and analyze, which makes them helpful in explaining the outcomes of a prediction model to healthcare professionals and the general public [7]. They can also be visualized, which can aid in explaining the model's reasoning and how it came to its conclusions. The Decision Tree displays the choices taken at each stage of the J48 algorithm [10], making it simple to grasp and comprehend. However, if the tree grows too deep, it may be vulnerable to overfitting. Therefore, the J48 method offers pruning options that can streamline the tree and enhance its generalization capabilities to reduce this risk. Healthcare organizations frequently utilize the Decision Tree J48 to forecast and spot patterns in massive datasets. It is a well-known approach for building a Decision Tree that can categorize and forecast results based on input variables.

Decision trees are good at forecasting diabetes and stomach cancer [11]. Many studies have employed Decision Tree J48 to expedite the diagnosis of stomach cancer, with promising results in studies like [12,13]. First, a model like this might help identify and diagnose early stomach cancer, vital for better patient outcomes. Early detection can open up more efficient treatment alternatives and boost the likelihood of successful treatment. Additionally, data analysis is made possible by data mining techniques like Decision Tree J48 [14], which offer more precise and trustworthy predictions than conventional approaches. It may result in better patient care and more educated decision-making. With better patient outcomes and a decrease in the total burden of stomach cancer, it can considerably advance the early identification, diagnosis, and prevention of this illness. Although J48 demonstrates effective performance, potential adjustments may be required to make it suitable for evaluating data from Thai hospitals, where more specialized data formats are prevalent. This implies that refining the algorithm has the potential to significantly improve the accuracy of the developed work.

Therefore, the Decision Tree J48 technique was used in this study to construct a risk prediction model for stomach cancer utilizing data mining. This approach has the potential to pinpoint disease-related risk variables, which will help in the creation of focused interventions for stomach cancer prevention. Users can forecast their probability of acquiring stomach cancer using a web application incorporating the model's uncovered risk factors.

This article is organized into five sections. It begins with an overview of the current state of stomach cancer research and relevant research papers before introducing the main topic. The subsequent section details the Decision Tree J48 procedure, explaining its steps and methodology. The third section delves into the specific research approach and techniques utilized. The fourth section is dedicated to analyzing and discussing the research results. Finally, the last section provides conclusions and outlines potential future work related to Decision Tree J48.

2. **Decision Tree J48.** This section describes the Decision Tree J48, an extension of Ross Quinlan's ID3 method based on the C4.5 algorithm [15,16]. Based on the characteristics of the data, the J48 algorithm creates a Decision Tree. In order to divide the data into subgroups depending on the chosen feature, the feature that best divides the data into

distinct groups is first chosen. This process is repeated for each subgroup until all the data is categorized or a criterion is finished. Based on the characteristics of the data, the resulting Decision Tree can be used to forecast the class of new data points. Consider a set of customer records as the data being studied. In that situation, the Decision Tree might forecast a new customer's chance of purchasing based on factors like age, income, location, and other characteristics. The features of various emergencies are considered using the Fuzzy C-Means with the Fuzzy Decision Tree [17], and the number of clusters is established, which determines the model's prediction accuracy.

The Decision Tree J48 model is the most effective at predicting disease [18]. It might help predict stomach cancer in the future. More research is required to comprehend the value of these approaches in the detection of stomach cancer completely and to maximize their application in clinical practice. Utilizing a Decision Tree for prediction has several benefits, such as simplicity, efficiency, robustness, and adaptability. Each node and branch of the decision tree directly relates to the result, simulating a tree-shaped data model with a hierarchical structure. Calculations of information gain, entropy, and gain are frequently used to choose a tree's root node [19]. The following is an explanation of attribute selection by information gain:

$$I(s_1, s_2, \ldots, s_n) = -\sum_{i=1}^{n} \frac{s_1}{s} \log_2 \frac{s_1}{s} \tag{1}$$

The following formula is used to calculate the amount of entropy or information required for classifying items within sub-trees:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{nj}}{s} I(s_{1j}, s_{2j}, \ldots, s_{nj}) \tag{2}$$

By branching on A, it is possible to achieve the encoded information.

$$Gain(A) = I(s_{1j}, s_{2j}, \ldots, s_{nj}) - E(A) \tag{3}$$

A is attribute, I for information gain, and j and n are class J and class N items, respectively, in this example.

Overall, we decided to adopt the J48 algorithm to help doctors make precise and quick diagnoses and choose the best course of therapy. Decision Tree J48 has also been used to find patterns and connections in medical data sets, which can help us comprehend the fundamental causes of conditions like gastric cancer.

3. **Research Methodology.**

This research aimed to analyze the risk factors for gastric cancer using patient data from hospitals in Phitsanulok Province. The Decision Tree J48 technique was used to create a model for predicting the risk of gastric cancer in the following steps: data collection, data preparation, model training, model evaluation, and model deployment.

3.1. **Data collection.**

The first step is to collect data on individuals with and without gastric cancer from the hospital in Phitsanulok, Thailand; this data is non-personal and does not provide any indication of an individual. Collect a dataset of 1,000 gastric cancer patients, including relevant demographic, clinical, and laboratory data. The selection of relevant attributes is thought to cause gastric cancer. The 18 attributes of risk factors are age, gender, air pollution, alcohol use, dust allergy, genetic risk, balanced food, obesity, smoking, passive smokers, coughing of blood, fatigue, weight loss, difficulty swallowing, clubbing of fingernails, frequent cold, dry cough, or snoring. However, the data collected for this study was solely used for research purposes and focused on analyzing risk factors for gastric cancer. The information gathered does not contain personally identifiable data and is solely relevant to the risk factors associated with gastric cancer.

3.2. **Data preparation.** The collected data is clean and prepared for analysis. It includes handling missing values, transforming variables, and selecting a subset of the data for model training. As the data may include attributes not pertinent to the analysis, specific attributes, such as Patient ID, must be removed. Stored data may exhibit various irregularities, such as missing attribute data (Missing Value), lack of compelling or detailed information, and noise data (Noisy Data), which can be caused by factors like errors (Error) or outliers (Outliers) resulting from an invalid questionnaire response, faulty data entry onto the computer, or staff errors. Next, identify the pertinent attributes and transform them into a data file format with a CSV extension, followed by attribute definition. Designate the result attribute as a Label type and convert all 18 associated attributes into the desired format. Finally, transform the CSV file into an ARFF file, a compatible standardized format, and modify the result attribute's value in both the training and testing files to ensure it is numeric.

3.3. **Model training.** The training dataset is used to examine patterns and correlations suggestive of an increased risk of stomach cancer to build a Decision Tree model. This entails building the Decision Tree based on the features that are selected as relevant in order to divide the dataset into subgroups. Splitting the data into training and testing sets will help in this procedure. With 970 data points allotted for training and 30 for testing, each file needs to be explicitly separated. The first stages are examining each column for missing values and importing the file. "Replace Missing Values" is the command applied for suitable handling if any column contains missing data. Then, experiments are carried out for classifier selection and testing alternatives with several split methods, such as the percentage split (70 : 30, 80 : 20, and 90 : 10), and Cross-Validation techniques (5-fold and 10-fold). The analysis of the model ultimately identifies six key risk factors necessary for creating a prediction model that can precisely estimate the probability of stomach cancer. The main objective is to create the best model possible with accurate disease-related factor analysis.

3.4. **Model evaluation.** Once the Decision Tree model is trained, it is evaluated to determine its accuracy and performance. Evaluate the model's performance on the test set using accuracy, precision, recall, and F1-score [20]. True Positive is TP, True Negative is TN, False Negative is FN, and False Positive is FP.

Accuracy is the percentage of correct predictions the model makes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Precision is the percentage of accurate optimistic predictions the model makes.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

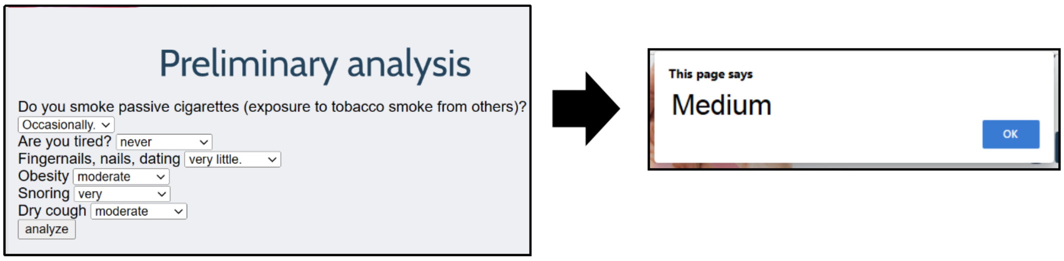
Recall is the percentage of positive cases the model correctly predicted.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

F1-score is a metric for measuring the performance of models.

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

3.5. **Model deployment.** Once the Decision Tree model has been trained and optimized, it can be deployed to predict gastric cancer in new individuals. This article developed a web application for predicting the risk of gastric cancer, as shown in Figure 1.



Preliminary analysis

Do you smoke passive cigarettes (exposure to tobacco smoke from others)? Occasionally.

Are you tired? never

Fingernails, nails, dating very little.

Obesity moderate

Snoring very

Dry cough moderate

analyze

This page says

Medium

OK

Figure 1. The interface of the web application

4. **Result and Discussion.** The study utilized a sample of 1,000 patients from a hospital to develop a model best suited for identifying the risk factors associated with gastric cancer. The model's effectiveness was then evaluated using the Decision Tree J48 technique. Patterns are created a tree-like structure, with each branch representing a different decision or prediction, as depicted in Figure 2.

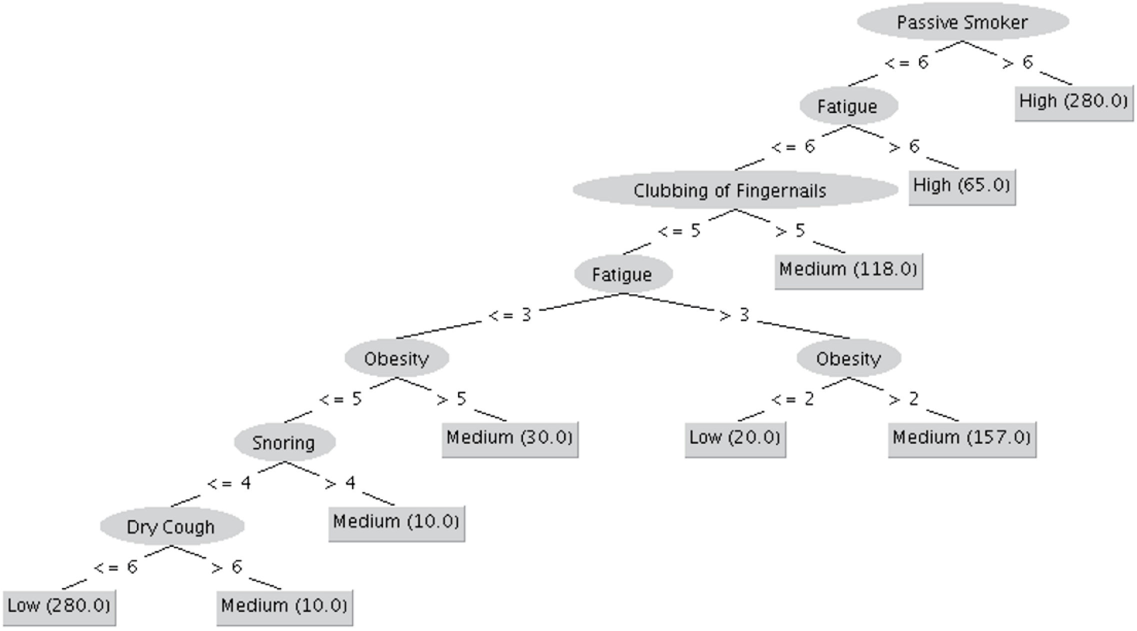


Figure 2. The result of a Decision Tree

Figure 2 describes the operation of the tree model for every nine outcomes of the prediction model that will occur in each case of the Passive Smoker node, which can use this result to develop an application.

1) Passive Smoker > 6
The forecast result is High.

2) Passive Smoker <= 6 and Fatigue > 6
The forecast result is High.

3) Passive Smoker <= 6 and Fatigue <= 6 and clubbing of fingernails > 5
The forecast result is Medium.

4) Passive Smoker <= 6 and Fatigue <= 6 and clubbing of fingernails <= 5 and Fatigue > 3 and obesity > 2
The forecast result is Medium.

5) Passive Smoker <= 6 and Fatigue <= 6 and clubbing of fingernails <= 5 and Fatigue > 3 and obesity <= 2
The forecast result is Low.

6) Passive Smoker <= 6 and Fatigue <= 6 and clubbing of fingernails <= 5 and Fatigue <= 3 and obesity > 5

The forecast result is Medium.

7) Passive Smoker <= 6 and Fatigue <= 6 and clubbing of fingernails <= 5 and Fatigue <= 3 and Obesity <= 5 and Snoring > 4

The forecast result is Medium.

8) Passive Smoker <= 6 and Fatigue <= 6 and clubbing of fingernails <= 5 and Fatigue <= 3 and Obesity <= 5 and Snoring <= 4 and Dry Cough > 6

The forecast result is Medium.

9) Passive Smoker <= 6 and Fatigue <= 6 and clubbing of fingernails <= 5 and Fatigue <= 3 and Obesity <= 5 and Snoring <= 4 and Dry Cough <= 6

The forecast result is Low.

Furthermore, the researchers evaluated the accuracy of the model developed with the Decision Tree J48 technique for predicting the risk of stomach cancer using the Cross-Validation Test and Percentage Split methods. The accuracy was measured to assess the effectiveness of the model.

Table 1 provides information on the performance of different predictive models for evaluating the risk of gastric cancer. The models were evaluated using percentage split (70 : 30, 80 : 20, and 90 : 10), 5-fold, and 10-fold Cross-Validation techniques. The performance metrics used to assess the models were accuracy, precision, recall, and F1-score. The *accuracy* of the models ranged from 0.954 to 1, with the 10-fold Cross-Validation achieving the highest accuracy of 1. This model indicates that the model's predictions were accurate, and the actual outcomes were predicted with a high degree of certainty. Next, the *precision* of the models ranged from 0.945 to 1, with the 10-fold Cross-Validation model achieving the highest precision of 1. This model means that the model was able to predict the occurrence of gastric cancer with high precision. The *recall* of the models ranged from 0.937 to 1, with the 10-fold Cross-Validation model achieving the highest recall of 1. This model indicates that the model identified individuals with gastric cancer correctly. Lastly, the *F1-score* of the models ranged from 0.944 to 1, with the 10-fold Cross-Validation model achieving the highest F1-score of 1. This score represents the harmonic mean of the precision and recall, indicating that the model could predict gastric cancer with high accuracy and precision. Overall, the models demonstrated high performance in predicting the risk of gastric cancer, with the 10-fold Cross-Validation model achieving the highest accuracy, precision, recall, and F1-score.

Table 1. Performance of each model

Model evaluation	**Percentage split**			**Cross-Validation Test**	
	70 : 30	**80 : 20**	**90 : 10**	**5-fold**	**10-fold**
Accuracy	0.975	0.954	0.982	0.987	1
Precision	0.945	0.952	0.966	0.974	1
Recall	0.962	0.937	0.983	0.981	1
F1-score	0.953	0.944	0.974	0.977	1

Finally, an analysis of the risk of stomach cancer is produced by this research. Developing a web application with six risk indicators included a risk assessment question. These questions address things like dry cough, snoring, clubbing of the fingernails, exhaustion, and passive smoking. Following processing, three risk levels – high, medium, and low – are shown for the results. After evaluating the predictions for 30 simulation patterns, the results showed that they had a perfect success rate – all 30 patterns had been correctly anticipated, or a 100% success rate.

5. **Conclusion and Future Work.** This study discovered that the Decision Tree model trained using the Decision Tree J48 algorithm was highly influential in predicting the risk of gastric cancer in a large dataset of individuals. The model exhibited high accuracy and precision and was able to identify critical risk factors and forecast the likelihood of gastric cancer in new individuals. To create a predictive model for assessing the likelihood of stomach cancer, we utilized data mining techniques and specific factors from 1,000 patients at a Hospital in Phitsanulok. We evaluated the model's effectiveness using the Decision Tree J48 technique and standard data mining procedures. Our results identified six significant predictors of gastric cancer from the model: passive smoking, fatigue, clubbing of fingernails, obesity, snoring, and dry cough. Five methods were employed to assess the model's performance in this research: Percentage Split (70 : 30, 80 : 20, and 90 : 10), 5-fold Cross-Validation, and 10-fold Cross-Validation. The results indicated that the model performed best using the 10-fold Cross-Validation method, achieving 100% accuracy. Lastly, the Decision Tree J48 enabled us to create a web application that employs six risk factors to assess 30 simulation patterns. Our analysis revealed that all 30 patterns achieved a 100% success rate, indicating flawless performance.

However, the study also had some limitations when interpreting the results. One limitation was the potential for overfitting, as the specific characteristics of the training data may have overly influenced the Decision Tree model. Further experimentation could be conducted using different datasets or algorithms to validate the results. In addition, the study did not consider the impact of certain factors, such as diet and lifestyle, on the risk of gastric cancer. Further research could investigate the role of these factors and how they may be incorporated into the model to improve its predictive accuracy. Overall, the study demonstrates the potential of Decision Tree models for predicting gastric cancer and highlights the need for further research and experimentation to validate and improve the model.

REFERENCES

[1] F. M. Johnston and M. Beckman, Updates on management of gastric cancer, *Current Oncology Reports*, vol.21, pp.1-9, 2019.

[2] National Cancer Institute, *Cancer Statistics in Thailand*, 2020.

[3] Y. Jiang et al., Development and validation of a deep learning CT signature to predict survival and chemotherapy benefit in gastric cancer: A multicenter, retrospective study, *Annals of Surgery*, vol.274, no.6, pp.e1153-e1161, 2021.

[4] S. Wang, J. Cao and P. S. Yu, Deep learning for spatio-temporal data mining: A survey, *IEEE Transactions on Knowledge and Data Engineering*, vol.34, pp.3681-3700, 2019.

[5] L. Liu and M. T. Özsu, *Encyclopedia of Database Systems*, Springer New York, 2019.

[6] Y. Lin, J. Chen, J. Li, S. Li, Y. Wang and L. Liu, The impact of early diagnosis on gastric cancer survival: A population-based study, *Journal of the National Cancer Institute*, vol.113, no.4, 2021.

[7] Y. Guo, J. Sun, X. Chen and L. Li, Predictive modeling of gastric cancer using data mining techniques: A systematic review, *Cancer Epidemiology, Biomarkers & Prevention*, vol.27, no.4, 2018.

[8] J. Polpinij, K. Namee and B. Luaphol, Bug reports identification using multiclassification method, *Science, Engineering Health Studies*, 22020009, 2022.

[9] National Cancer Institute, *Stomach (Gastric) Cancer*, 2020.

[10] E. B. B. Palad, M. J. F. Burden, C. R. D. Torre and R. B. C. Uy, Performance evaluation of decision tree classification algorithms using fraud datasets, *Bulletin of Electrical Engineering and Informatics*, vol.9, no.6, pp.2518-2525, 2020.

[11] A. M. Posonia, S. Vigneshwari and D. J. Rani, Machine learning based diabetes prediction using Decision Tree J48, *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pp.498-502, 2020.

[12] C. Neto, M. Brito, V. Lopes, H. Peixoto, A. Abelha and J. Machado, Application of data mining for the prediction of mortality and occurrence of complications for gastric cancer patients, *Entropy*, vol.21, no.12, 1163, 2019.

[13] M. R. Afrash, M. Shanbehzadeh and H. Kazemi-Arpanahi, Design and development of an intelligent system for predicting 5-year survival in gastric cancer, *Clinical Medicine Insights: Oncology*, vol.16, 11795549221116833, 2022.

[14] N. Bhargava, G. Sharma, R. Bhargava and M. Mathuria, Decision tree analysis on J48 algorithm for data mining, *Proc. of International Journal of Advanced Research in Computer Science Software Engineering*, vol.3, no.6, 2013.

[15] M. F. Faruque, Asaduzzaman and I. H. Sarker, Performance analysis of machine learning techniques to predict diabetes mellitus, *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox'sBazar, Bangladesh, DOI: 10.1109/ECACE.2019.8679365, 2019.

[16] L. M. Crivei, V.-S. Ionescu and G. Czibula, An analysis of supervised learning methods for predicting students' performance in academic environments, *ICIC Express Letters*, vol.13, no.3, pp.181-189, 2019.

[17] Y. Hao, M. Tian, Y. Wang and M. Huang, Demand forecasting for rush repair spare parts of power equipment using fuzzy C-means clustering and the fuzzy decision tree, *International Journal of Innovative Computing, Information and Control*, vol.19, no.4, pp.1007-1021, 2023.

[18] H. Peixoto et al., Predicting postoperative complications for gastric cancer patients using data mining, *Prof. of Intelligent Technologies for Interactive Entertainment: The 10th EAI International Conference (INTETAIN 2018)*, Guimarães, Portugal, pp.37-46, 2019.

[19] S. Moro, P. Cortez and P. Rita, A data-driven approach to predict the success of bank telemarketing, *Decision Support Systems*, vol.62, pp.22-31, 2014.

[20] M. Raihan et al., Risk prediction of ischemic heart disease using artificial neural network, *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox'sBazar, Bangladesh, DOI: 10.1109/ECACE.2019.8679362, 2019.