Semarak Current Biomedical Technology Research Journal

Journal homepage:
https://semarakilmu.my/index.php/scbtrj/index
ISSN: 3030-5616

Prediction of Hepatitis C Patient by using Support Vector Machine

Nurul Husna Md Nasri[1], Noor Hidayah Zakaria[1,*], Anis Farihan Mat Raffei[2]

[1] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
[2] Faculty of Computing, Universiti Malaysia Pahang, 26600 Pekan, Pahang, Malaysia

ARTICLE INFO

Article history:
Received 12 February 2025
Received in revised form 2 April 2025
Accepted 12 April 2025
Available online 30 June 2025

Keywords:
Hepatitis C virus; classification; Support Vector Machine; feature selection; CFS with p-value

ABSTRACT

Hepatitis is one of the most often diagnosed infectious diseases worldwide. It is difficult to identify important conclusions from the large amount of data which is accessible in the medical field. With the advancement of technology, data mining methods have established themselves as the most widely used approaches in a variety of fields. Predictive analysis is one of the areas in the medical sector. The goal of this research is to find ways to diagnose the disease using the machine learning techniques for early prediction in hepatitis C patients based on their clinical examination. The dataset was obtained from the UCI Machine Learning repository. The method for selecting features include leveraging on correlation-based feature selection (CFS) with p-value which is to find the significant independent attributes, hence improving the classifier performance. Bilirubin, Albumin, Protime, Fatigue, Malaise, Spiders, Ascites, varices and Histology were found to be the most significant independent attributes. Consequently, this research performed the prediction of hepatitis C using Support Vector Machine on the selected features, comparing the prediction of hepatitis C with and without CFS with p-value feature selection. The performance evaluation of the algorithm was evaluated using accuracy, precision and recall. Experimental result has demonstrated that CFS with p-value for Support Vector Machine were outperforming with the highest accuracy 0.9388 compared to the default SVM.

1. Introduction

The huge amount of data generated by modern molecular biology generally requires ML for accurate classification and prediction algorithms. A wide range of variables may give effects to the accuracy of classification algorithms, some of which are generic to all machine learning algorithms and therefore relevant to research in other application domains [16]. According to the World Health Organization (WHO) in 2019, about 3% of the world's population, or 120-130 million people in the world, was contaminated with hepatitis C virus, with 3-4 million new cases reported. Thus, representing one of the world's most serious public health issues, which should be addressed by efficient programme strategies for detection and care [1]. Basically, hepatitis can be diagnosed by blood testing [2]. Many of the factors should be considered if using medical diagnostics since it will

* *Corresponding author.*
E-mail address: noorhidayah.z@utm.my

https://doi.org/10.37934/scbtrj.5.1.1119

be quite challenging [3,4]. Thus, it would be helpful if the research could build the new development in terms for accuracy in the diagnosis system for detection of hepatitis, to reduce the influence of the virus and to classify the hepatitis C virus [5]. With the advancement of technologies, ML will get to investigate significant symptoms and do an analysis of the patient's condition. The study on the machine learning classifiers is done by Yarasuri *et al.,* [6] proposed and compared the most common classifiers in predicting hepatitis which were Support Vector Machine (SVM), K-Nearest Neighbour (KNN) and Artificial Neural Network (ANN). Other than that, Kumar and Vigneswari [7] introduced another five desired classifiers such as Logistic Regression, Random Forest, Decision Tree, C4.5 and MLP. Each classifier has their own strengths and weaknesses, and therefore the best classifier for any method algorithm can be determined by performance evaluation techniques, including sensitivity, specificity, and classification accuracy.

The dataset consists of many disease measurements and some of them may not be related to hepatitis C virus disease. Consequently, it might affect the prediction accuracy of hepatitis C virus disease and make the model algorithm learn based on irrelevant features [8]. Hence, during feature extraction, a new reduced collection of attributes is generated by applying multidimensional space mapping into less dimensional space. This results in the transformation of the initial feature space into more simplified and informative newer feature space [9]. Using this method, attributes are grouped into a new reduced collection of features [10,14]. This research also considers on the importance feature selection process, where this method selects a group of important features and eliminates irrelevant, noisy, and repetitive features for the data representation to become simpler and more concise [11,12]. This method is useful in reducing the cost of diagnosis by focusing on the most important attributes and also to show an improvement of accuracy. Furthermore, the extraction of important features plays a key role in the prediction and classification of hepatitis C virus disease. Afterwards it can improve the performances of the machine learning algorithm with high accuracy resulting in time and effort reduction. The purpose of this paper is to use the dataset in diagnosing hepatitis C using Support Vector Machine to measure performances of the classification in terms of accuracy, sensitivity and specificity. Before building up the classifier, the CFS with p-value feature selection conducted and performed for classification algorithm.

2. Methodology

The objectives of this research is to compare the result obtained from Support Vector Machine (SVM) with and without CFS with p-value feature selection in order to predict and classify hepatitis patient. Therefore, by using hepatitis dataset, a new feature selection method (CFS with p-value) obtained from a descriptive analysis method is proposed to eliminate an irrelevant attribute. The significant attributes are given to inputs of classifiers SVM in the classification stage [15]. The performance of SVM classifiers in diagnosing the hepatitis disease is estimated using classification accuracy, precision and recall respectively. To identify the best result of accuracy, different kernel of the classifier is used and then the classifier is implemented. The hyperparameter tuning is performed to get the best performances of the algorithms. The result is compared between of cross validation and train/test split. After that, an analysis based on the evaluations and the validation step that is best in classifying hepatitis is identified. The flow of research methodology can be seen in Figure 1 below.

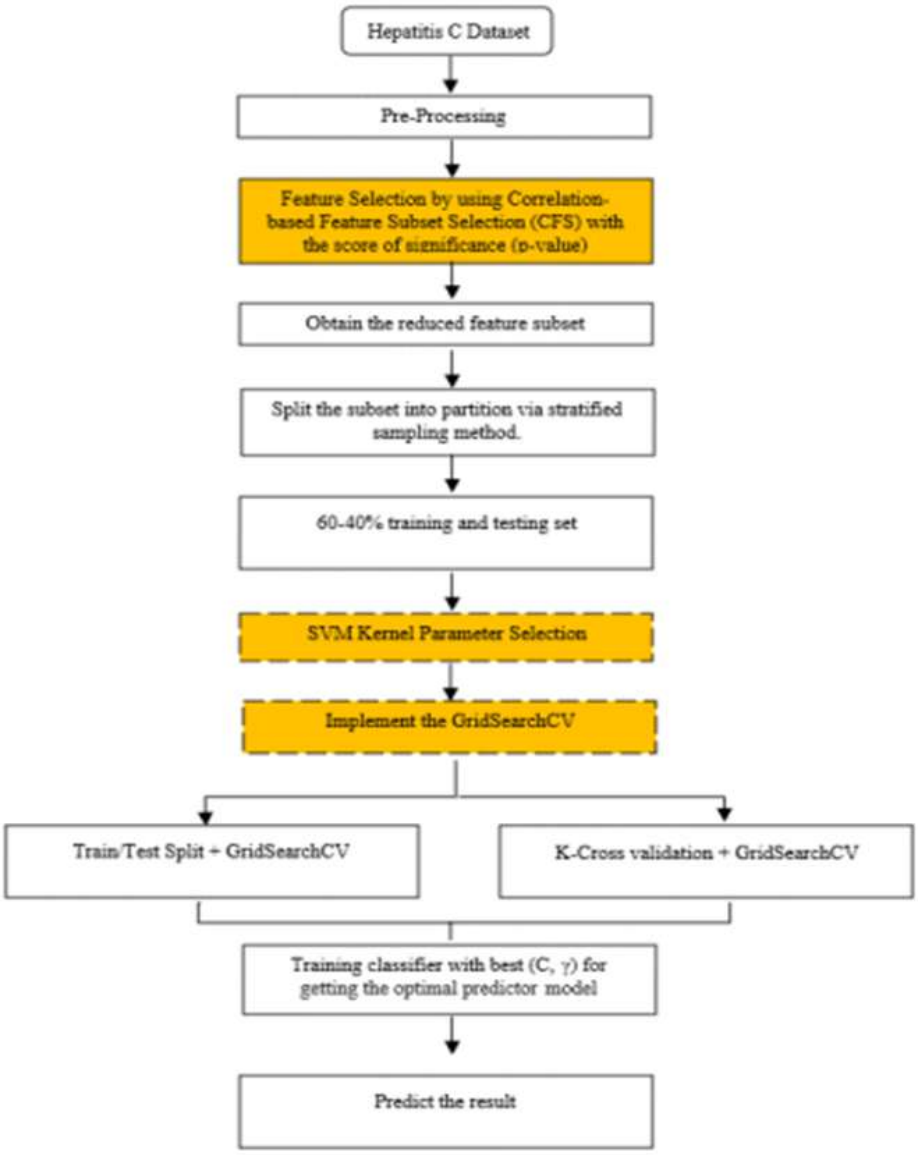

Fig. 1. Flow of the research methodology

2.1 Dataset

The dataset included in this research for HCV prediction is taken from the UCI Machine Learning Repository and contains samples from 155 instances [13]. This dataset is the records of various medical tests carried out on patients which could determine whether the patients are suffering hepatitis disease or not. The dataset consisted of a total of 19 features, which may be classified into two categories which are quantitative and qualitative features. The total of quantitative features is equivalent to 6, whereas the number of qualitative features is equivalent to 13. The details of the dataset are given in Table 1.

Table 1
Hepatitis dataset [13]

No	Attribute	Values
1	Age	10, 20, 30, 40, 50, 60,70, 80
2	Sex	Male, Female
3	Steroid	No, Yes
4	Antivirals	No, Yes
5	Fatigue	No, Yes
6	Malaise	No, Yes
7	Anorexia	No, Yes
8	Liver big	No, Yes
9	Liver firm	No, Yes
10	Spleen palpable	No, Yes
11	Spiders	No, Yes
12	Ascites	No, Yes
13	Varices	No, Yes
14	Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
15	Alk phosphate	33, 80, 120, 160, 200, 250
16	SGOT	13, 100, 200, 300, 400, 500
17	Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
18	Protime	10, 20, 30, 40, 50, 60, 70, 80, 90
19	Histology	No, Yes

2.2 Phase 1: Data Pre-Processing

The dataset may have some missing value or irrelevant attributes which affected in obtaining the best results for the classification accuracies. The activities involved in data pre- processing is checking the missing values, outliers and data normalization. Missing values can be handled by removing replacing instances with an average, frequency, maximum or minimum. Outliers are data values that deviate significantly from the majority of observations in a dataset. Data normalization is the process to uniform the instance values from different ranges into the range 0 to 1 by generating new values while maintaining the general distribution and ratios of the given dataset.

2.3 Phase 2: Feature Selection

Feature selection plays an important role in building the classifier models by selecting only significant attributes. Considering all the FS methods available, Correlation based Feature Selection with the p-value method is chosen to assist the FS in this research for selecting the most significant attributes. Figure 2 shows the methodology of the chosen feature selection.

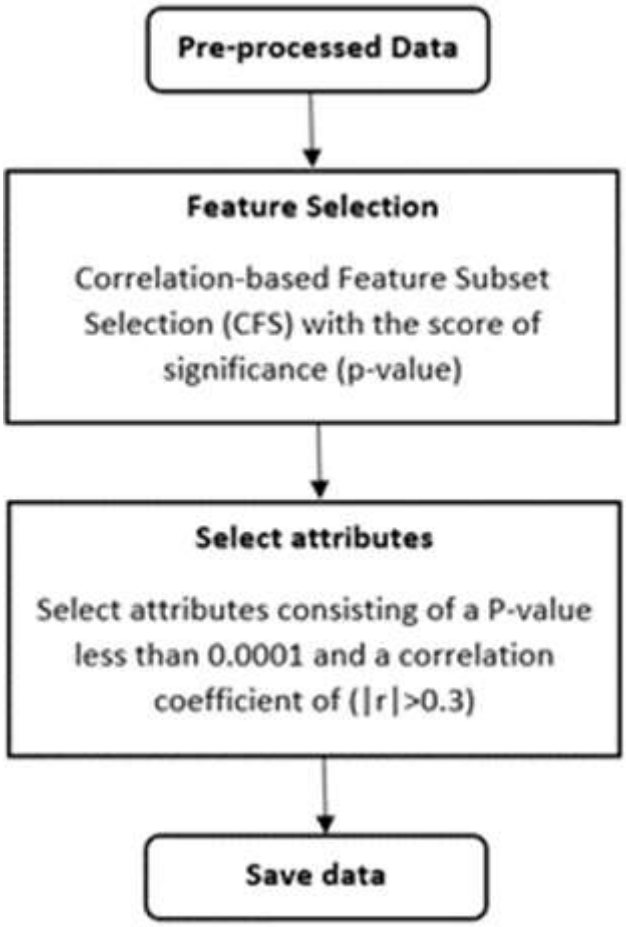

Fig. 2. Feature selection methodology

2.4 Phase 3: Machine Learning Algorithm

The proposed learning algorithm in the diagnosis of hepatitis C virus is Support Vector Machine (SVM) based on predetermined data for given patients as per shown in Figure 3. The dataset was loaded and split into training and testing sets. The splitting was 60% of training and 40% of testing. The training sets were then used to train the classifier, and the testing sets were used to conduct the prediction. Before applying SVM to a specific task, the selection of all these parameters is necessary. It is because the performance of the model somehow depends on the kernel functions. In order to get the most accurate predictions, GridSearchCV is used to find the best hyperparameter for each of the model. The performance of algorithms was evaluated using Repeated Stratified K-Cross Validator and Training and Testing.

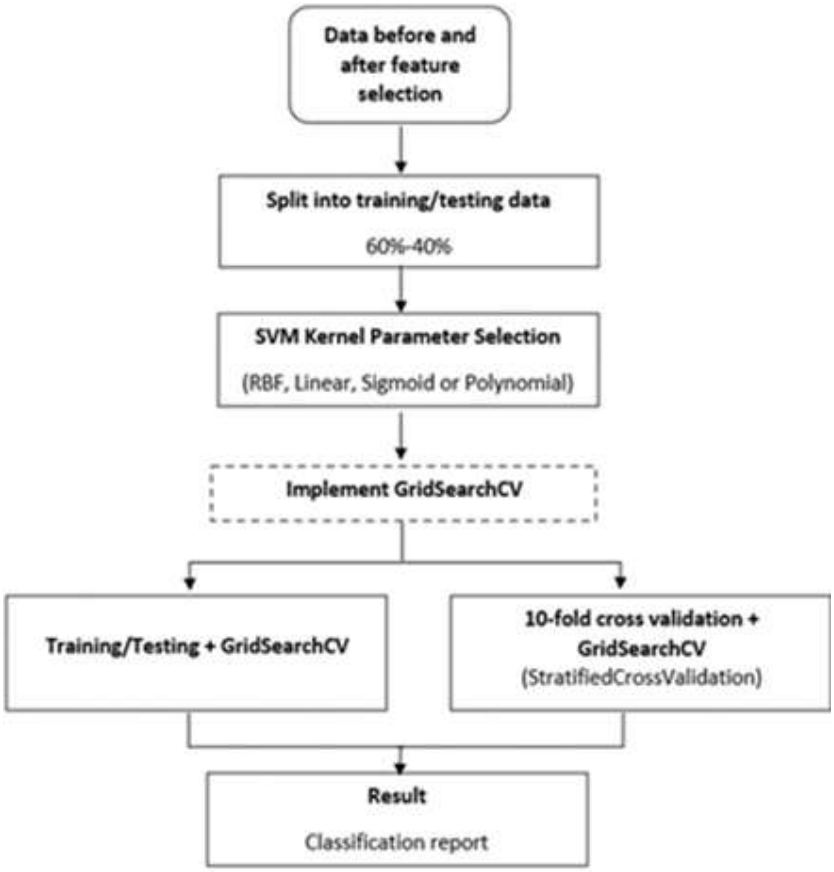

Fig. 3. Classification methodology

2.5 Phase 4: Result, Analysis and Discussion

The result of the performance classification from the model with and without applying feature selection are obtained and compared. The SVM classifier were analyzed using a confusion matrix which includes accuracy, precision and recall. It will describe more on the performance of the classification model and at the same time provides information about the actual and predicted values of the dataset. From the confusion matrix, there will be four outcomes which are True positive (TP), True negative (TN), False positive (FP) and False negative (FN) as shown in Table 2.

Table 2
Confusion matrix of hepatitis dataset

	Predicted: Normal	Predicted: Patient
Actual: Normal	True negative (TN)	False positive (FP)
Actual: Patient	False negative (FN)	True positive (TP)

The performance evaluation techniques in terms of accuracy, precision and recall were calculated using evaluation metric as shown in Eq. (1) to (3) below:

$$Accuracy = (TP + TN)/(TP + FP + FN + TN) \times 100 \quad (1)$$

$$Precision = TP/(TP + FN) \times 100 \quad (2)$$

$$Recall = TN/(TN + FP) \times 100 \quad (3)$$

3. Results

The result of correlation coefficient and p-value is shown in Table 3 below. If the correlation coefficient r was close to -1 or 1 and the p-value was less than the standard 5% (P <0.0001), it was declared as statistically significant. This research managed to identify 9 attributes as selected attributes which it can be entered to both algorithms for training and testing. All the 9 attributes were consisting of a P-value less than 0.0001 and a correlation coefficient of (|r|>0.3, which is considered acceptable.

Table 3
Result of correlation feature selection with score of significance

Dataset	FS algorithm	Size	Significant attributes
Hepatitis C	CFS with p-value	9	Bilirubin, albumin, protime, fatigue, malaise, spiders, ascites, varices and histology

Applying the baseline, this research had the point of references or benchmark to consider whether the result of proposed study's performance had improved or not. The WEKA Explorer environment was used to calculate the baseline performance with Zero Rule Algorithm. Based on the result from WEKA explorer, the correctly classified instances or accuracy which obtained was 79.53%. The four kernels from SVM have been used to find which giving the best accuracy which be implemented in this research. As the result, this research used the RBF kernel as the selected kernel due to the highest accuracy and have possibility affect the performance of the machine learning model referring to the Figure 4.

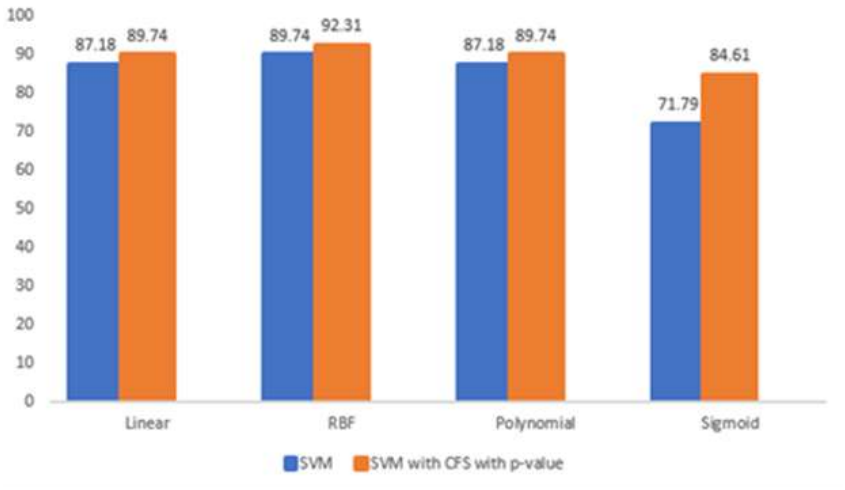

Fig. 4. Comparison accuracy from all kernel

As the result in Table 4 shown below, SVM shows the difference in performance throughout the two experiments between with and without applying feature selection. From both experiments, the implementation of 10-fold cross validation performs the best in terms of accuracy compared to Training/Testing split. The result of SVM applying CFS with p-value and implementation of 10-fold cross validation has an accuracy of 93.88% compared to implementation of training/testing split which resulting of 93.54%. The third rank falls on SVM implementation of 10-fold cross validation which is 92.75% followed with the default SVM which implementing training/testing split with only 92.47%.

Table 4
Classification report

	SVM		CFS with p-value + SVM	
	training:75% and testing: 25% + GridSearchCV	10-fold K-Cross validation + GridSearchCV (Average)	training:75% and testing: 25% + GridSearchCV	10-fold K-Cross validation + GridSearchCV (Average)
Accuracy (%)	92.47	92.75	93.54	93.88
Precision (%)	88.88	74.64	85.00	88.88
Recall (%)	76.19	73.32	85.00	69.02

Table 5 shows the comparison with previous research. The previous method has been developed by Yarasuri *et al.,* [7] which scored 0.8958 on the accuracy of the default SVM. It is lower than the accuracy which performed by the proposed parameter which is GridSearchCV which scored 0.9275 for default SVM. By implement the feature selection, the result of accuracy of the models become the highest among the other experiments which contributes to 0.9388. The GridSearchCV and Feature Selection have helped a lot and bring benefit to the medical field as it able to find which algorithm is the best in analysis hepatitis prediction by using medical dataset.

Table 5
Comparison with previous research

Classifier	Accuracy
Default SVM [7]	0.8958
This study (Default SVM + Grid SearchCV)	0.9275
This study (CFS with p-value_SVM +GridSearchCv)	0.9388

4. Conclusions

In this research, new medical diagnostics method, CFS with p-value_SVM for the purpose of resolving the hepatitis diagnosis issue were proposed. The algorithm may help the doctors to predict whether the patient suffering hepatitis C or otherwise. The hepatitis dataset was successfully undergoing preprocessing step by checking the missing value, outliers and data normalization. Then, Bilirubin, Albumin, Protime, Fatigue, Malaise, Spiders, Ascites, varices and Histology were found to be the most significant independent attributes. This research has also shown the best kernel parameter which should be applied and by using GridSearchCV for hyperparameter tuning purposes manage to save the researcher time and improved the accuracy. Referring to the analyze result, it is proven that the CFS with p-value SVM by implementation of 10-fold cross validation is the best technique in predicting and classifying the hepatitis compared to other experiments. For the further improvement of studies, experimenting the algorithm by using bagging or boosting or by implementing other improved feature selection.

Acknowledgement
The authors would like to thank Universiti Teknologi Malaysia and Universiti Malaysia Pahang for supporting this collaborative research in the present work. This work was supported by the Universiti Teknologi Malaysia under Grant vot no. Q.J130000.3851.19J73.

References
[1] Nandipati, Satish CR, Chew XinYing, and Khaw Khai Wah. "Hepatitis C virus (HCV) prediction by machine learning techniques." *Applications of modelling and simulation* 4 (2020): 89-100.

[2] Bhattacharya, Renuka, and Margaret C. Shuhart. "Hepatitis C and alcohol: interactions, outcomes, and implications." *Journal of clinical gastroenterology* 36, no. 3 (2003): 242-252. https://doi.org/10.1097/00004836-200303000-00012
[3] Gupta, Ekta, Meenu Bajpai, and Aashish Choudhary. "Hepatitis C virus: Screening, diagnosis, and interpretation of laboratory assays." *Asian journal of transfusion science* 8, no. 1 (2014): 19-25. https://doi.org/10.4103/0973-6247.126683
[4] Ahammed, Khair, Md Shahriare Satu, Md Imran Khan, and Md Whaiduzzaman. "Predicting infectious state of hepatitis c virus affected patient's applying machine learning methods." In *2020 IEEE Region 10 Symposium (TENSYMP)*, pp. 1371-1374. IEEE, 2020. https://doi.org/10.1109/TENSYMP50017.2020.9230464
[5] Janardhanan, Padmavathi, and Fathima Sabika. "Effectiveness of support vector machines in medical data mining." *Journal of communications software and systems* 11, no. 1 (2015): 25-30. https://doi.org/10.24138/jcomss.v11i1.114
[6] Yarasuri, Vedha Krishna, Gowtham Kishore Indukuri, and Aswathy K. Nair. "Prediction of hepatitis disease using machine learning technique." In *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 265-269. IEEE, 2019. https://doi.org/10.1109/I-SMAC47947.2019.9032585
[7] Kumar, N. Komal, and D. Vigneswari. "Hepatitis-infectious disease prediction using classification algorithms." *Research Journal of Pharmacy and Technology* 12, no. 8 (2019): 3720-3725. https://doi.org/10.5958/0974-360X.2019.00636.X
[8] Shaikh, Raheel. "Feature selection techniques in machine learning with python." *Towards data science* 28 (2018).
[9] Plaza, Antonio, Pablo Martínez, Javier Plaza, and Rosa Pérez. "Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations." *IEEE Transactions on Geoscience and remote sensing* 43, no. 3 (2005): 466-479. https://doi.org/10.1109/TGRS.2004.841417
[10] Çalişir, Duygu, and Esin Dogantekin. "A new intelligent hepatitis diagnosis system: PCA–LSSVM." *Expert Systems with Applications* 38, no. 8 (2011): 10705-10708. https://doi.org/10.1016/j.eswa.2011.01.014
[11] Uzer, Mustafa Serter, Nihat Yilmaz, and Onur Inan. "Feature selection method based on artificial bee colony algorithm and support vector machines for medical datasets classification." *The Scientific World Journal* 2013, no. 1 (2013): 419187. https://doi.org/10.1155/2013/419187
[12] Janardhanan, Padmavathi, and Fathima Sabika. "Effectiveness of support vector machines in medical data mining." *Journal of communications software and systems* 11, no. 1 (2015): 25-30. https://doi.org/10.24138/jcomss.v11i1.114
[13] Chen, Hui-Ling, Da-You Liu, Bo Yang, Jie Liu, and Gang Wang. "A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis." *Expert systems with applications* 38, no. 9 (2011): 11796-11803. https://doi.org/10.1016/j.eswa.2011.03.066
[14] Tan, Kay Chen, Eu Jin Teoh, Qiang Yu, and K. C. Goh. "A hybrid evolutionary algorithm for attribute selection in data mining." *Expert Systems with Applications* 36, no. 4 (2009): 8616-8630. https://doi.org/10.1016/j.eswa.2008.10.013
[15] Uzer, Mustafa Serter, Nihat Yilmaz, and Onur Inan. "Feature selection method based on artificial bee colony algorithm and support vector machines for medical datasets classification." *The Scientific World Journal* 2013, no. 1 (2013): 419187. https://doi.org/10.1155/2013/419187
[16] Nancy, P., V. Sudha, and R. Akiladevi. "Analysis of feature selection and classification algorithms on hepatitis data." *Int. J Advanced Res Comp. Eng. Technol* 6, no. 1 (2017): 19-23.