Hittite Journal of Science and Engineering, 2022, 9 (2) 111-116
ISSN NUMBER: 2148-4171
DOI: 10.17350/HJSE19030000261

Hepatitis C Disease Detection Based on PCA–SVM Model

Serdar Gundogdu
Dokuz Eylul University, Department of Computer Technologies, Izmir, Turkey

ABSTRACT

Hepatitis C is a liver disease caused by infection with the hepatitis C virus (HCV), which is transmitted through the blood. The disease can lead to diseases ranging from a mild form to serious lifelong illness. Studies to detect the disease early and reduce its effect are continuing. This study proposes an effective support vector machine model supported by principal component analysis for detecting hepatitis c disease. The dataset consisted of twelve independent variables, each containing 582 samples, and these variables were used as inputs to the two classifiers, support vector machine (SVM) and artificial neural network (ANN). The accuracy, sensitivity, specificity, Matthews correlation coefficient (MCC) and KAPPA were calculated using two classification models. In addition, performance comparisons of classifiers were made for the two cases with and without principal component analysis (PCA) applied to the inputs. The highest accuracy (98.7%), sensitivity (99.1%), specificity (95.2%), MCC (92.3%) and Kappa (92.3%) in the binary class label were obtained with the SVM with PCA. In the four-class label, the highest accuracy was achieved with the same model with 95.7%. The results show that an SVM classifier model, in which PCA-reduced independent variables are applied to its inputs, may be a candidate for an accurate prediction model to predict hepatitis C disease.

Article History:
Received: 2022/01/01
Accepted: 2022/06/20
Online: 2022/06/30

Correspondence to: Serdar Gundogdu, Dokuz Eylul University, Department of Computer Technologies, Izmir, Turkey
E-Mail:serdar.gundogdu@deu.edu.tr
Tel: +90 232 433 6648

Keywords:
Artificial neural network; Dimensional reduction; Hepatitis C; Principal component analysis; Support vector machine

INTRODUCTION

Hepatitis C is liver inflammation caused by the HCV. Globally, an estimated 58 million people are infected with chronic hepatitis virus, and approximately 1.5 million new infections occur each year. World Health Organization (WHO) estimated that approximately 290 000 people died from hepatitis C in 2019, mostly cirrhosis and hepatocellular carcinoma. Antiviral drugs can cure more than 95% of people with hepatitis C infection, but access to treatment and diagnosis is low [1]. Early diagnosis is very important for hepatitis C risk groups as it will slow down the progression of the disease. Therefore, researchers are looking for ways to recognize the disease early and thus stop the progression of the disease.

Machine learning (ML) plays a very important role in medical applications. ML has recently been used for disease prediagnosis due to its ease of application and high accuracy [2-4]. It has also been used in the literature to assist in the diagnosis of hepatitis C. Ayeldeen et al. (2015) found 93.7% accuracy by applying a decision tree-based classifier as well as feature selection to investigate hepatitis C data [5]. Orczyk and Porwik (2016) studied various classification methods and obtained approximately 70% accuracy with three classifiers [6]. Bahargav et al. (2018) developed a logistic regression model to analyze 155 samples for diagnosing hepatitis C. [7]. They achieved 87.17% classification accuracy. Ahammed et al.'s study (2020), a ML based model has been proposed that can classify HCV infected patient's stages of liver [8]. They applied variable selection methods to identify significant features of HCV. They compared different methods and decided k-nearest neighbors (KNN) as the best performing ML method for the problem. KNN showed the best 94.40% accuracy than the others. Syafa'ah et al.'s study (2021) aimed to evaluate the accuracy using the algorithm classification method to detect the disease hepatitis C virus [9]. The data parameters used in this study included bilirubin (BIL), albumin (ALB), γ-glutamyl-transferase (GGT), aspartate amino-transferase (AST), choline esterase (CHE), alanine amino-transferase (ALT), creatinine (CREA), protein (PROT), Alkaline phosphatase (ALP) and cholesterol (CHOL). They achieved 95.12% classification accuracy with neural network (NN). In their study using the Egyptian patient dataset, Nandipati et al. (2020) compared the performances between dual and multi-class labels of

the same dataset [10]. They tried to find out which selected features play a key role in predicting hepatitis C virus. SVM, random forest (RF), Gaussian naive Bayes (GNB), neural network (NN), Bagging, Boosting, and KNN were the evaluated classifiers in the study. The highest accuracy is shown by random forest (54.56%) and KNN (51.06%) in binary and multi class label respectively. In addition, 12 features selected by principal component analysis showed similar performances to complete the dataset.

The principal component analysis is a non-parametric method for extracting relevant information from complex and large data sets [11]. This method is widely used for feature extraction, dimension reduction, data visualization, and lossy data compression [12-14]. Adjusting the input properties can result in better revealing predictor-response relationships [15]. PCA attempts to extract important data from a set of interrelated variables and describe them in a new field as a set of orthogonal variables called principal components (PCs) [16] where the number of components is less than or equal to the number of original features [13]. PCs can be used as inputs to learning models instead of original variables [17].

Although SVM has proven to be a good classifier, its applications may not reach the expected level due to the complexity of time and space. The linear dependence between its different variables affects the generalization of the SVM method. PCA can effectively deal with the linear dependence between variables [18]. PCA is used to select the relevant variables by reducing the data size, and to decrease the complexity of the SVM classifier [19]. To improve the classification performance of SVM, I propose to use PCA and SVM models together. The proposed PCA-SVM approach offers two distinct advantages. The first advantage is that as the size of the model inputs decreases, the computational complexity decreases and the model run speed increases. Second, the model can avoid some defects of neural networks such as local minima and overfitting [20].

The present study compared classifier models with or without PCA to perform diagnosis prediction. PCA was applied to investigate the effect of input sizes on the performance of the classifier models.

In this paper, a principal component analysis assisted effective support vector machine model was proposed for hepatitis c detection. Feature dimensions were reduced using PCA. Then, binary and multi classifications were performed by ANN and SVM. Classifier results were compared using performance metrics.

MATERIALS AND METHODS

The data mining classification method, which is widely used in many engineering fields, is a data-based predictive model that can significantly support clinical decisions and provide a role for intelligent systems. The workflow, including the classification methods used in this study, is shown in Fig. 1.

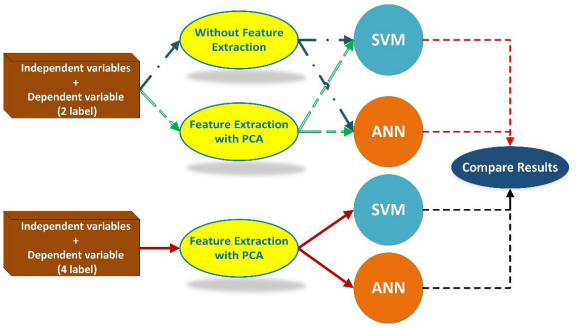

Figure 1. Workflow of the proposed work

Dataset

The dataset used in the study was obtained using the publicly available UCI machine learning repository [21]. The multivariate data type consists of 582 samples with 13 attributes, 12 of which are input and 1 is output attribute. Multiclass dataset samples consist of 4 different labels (frequencies), C0 (healthy, 526), C1 (Hepatitis, 20), C2 (Fibrosis, 12) and C3 (Cirrhosis, 24). The description of the features is shown in Table 1.

Table 1. The description of the features

No	Features	No	Features
1	Age (in years)	7	Bilirubin (BIL)
2	Sex (f, m)	8	Choline esterase (CHE)
3	Albumin (ALB)	9	Cholesterol (CHOL)
4	Alkaline phosphatase (ALP)	10	Creatinine (CREA)
5	Alanine amino-transferase (ALT)	11	Gamma-glutamyl transferase (GGT)
6	Aspartate amino-transferase (AST)	12	Protein (PROT)
13	Dependent variable; Donor (C_0), Hepatitis (C_1), Fibrosis (C_2), Cirrhosis (C_3)		

Feature Extraction Method Using PCA

PCA is one the most effective and widely used in multivariate techniques [22]. PCA is used in place of the original variables in machine learning and statistical models. It is generally adopted as a feature extraction method to reduce the dimensionality of data [23]. It reduces independent variables. The newly reduced variables are simply linear combinations of the original variables. The first principal component discriminates maximally between the variables in the sample, that is, it has a large sample variance [24].

In this method, most of the original data is contained in the first component, while the remaining data is annotated by subsequent components in descending order. Therefore,

in terms of amount of variance, the first PC variants represent the best, and the amount of information on other PCs decreases [17].

Classification Methods

The classification method is one of the main methods of data mining and is based on a learning algorithm. The two classification methods used in the study are based on machine learning. In particular, two types of techniques have been considered: artificial neural network (NN) and support vector machines (SVM). The first considered classifier is based on a classical NN. This was achieved using the Matlab "fitcnet" command on the training dataset and all parameters were left as default [25-26]. The SVM technique used have gaussian kernel functions and the classifier was trained in Matlab using the command "fitcecoc" on the training dataset. For experiments on dataset, the original dataset is randomly divided into a training set and a test set. The dataset was divided into 60% training data and 40% test data to evaluate the performance of the classifiers while testing each model. Test data selection was done randomly, regardless of whether the dependent variable was patient or not. In future work, I plan to test the proposed model further in other diseases such as blood cancer.

Performance Metrics

In this study, the performance of the classifiers is measured using the confusion matrix. The confusion matrix evaluates the errors of the established models by revealing error metrics in solving classification problems. The confusion matrix table is shown in Fig. 2.

Confusion Matrix				
		Target Class		
		Donor	Patient	Total
Output Class	Donor	**TP**	FP	TP+FP
	Patient	FN	**TN**	FN+TN
	Total	TP+FN	FP+TN	

Figure 2. Confusion matrix

Here TN, TP, FN, and FP show the true negative, true positive, false negative, and false positive, respectively.

The performance of the proposed method was evaluated using equations (1) to (5). For better evaluation, the values of the accuracy (ACC), sensitivity (SENS), specificity (SPEC), Matthews correlation coefficient (MCC) and KAPPA criteria represent excellent classification performance.

Accuracy indicates the number of correct predictions made over all values observed by the model. Sensitivity and specificity are the proportion of true positives and true negatives that are correctly identified, respectively. Matthews correlation coefficient is the correlation coefficient of target and predicted data. MCC considers the fact that positives and negatives are not balanced [27]. A statistical indicator that compares observed and expected accuracy is the kappa coefficient.

$$ACC = \frac{TN + TP}{TN + FN + TP + FP} \quad (1)$$

$$SENS = \frac{TP}{FN + TP} \quad (2)$$

$$SPEC = \frac{TN}{FP + TN} \quad (3)$$

$$MCC = \frac{TP * TN - FP * FN}{[(TP + FP)(TP + FN)(TN + FP)(TN + FN)]^{0.5}} \quad (4)$$

$$KAPPA = \frac{2 * (TP * TN - FP * FN)}{(FP + TP) * (TN + FP) + (FN + TP) * (FN + TN)} \quad (5)$$

RESULTS AND DISCUSSION

PCA feature extraction was conducted using MATLAB codes. The PCA function was used to get principal component variances, principal component scores, and principal component coefficients [28]. Fig. 3 shows percentages of explained variances for principal components. Here, the first, second, and third principal components account for 42.1%, 34.1%, and 11.14% of the total variance, respectively. Together five components explain 95.67% of the variation in the data. Five principal components, PC1, PC2, PC3, PC4 and PC5, were chosen to best capture the characteristics of healthy and HCV-infected people.

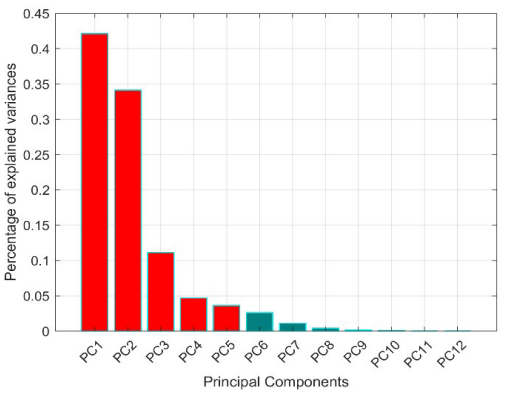

Figure 3. Percentages of explained variances for Principal Components

In the study, 12 independent variables and one dependent variable were used. The dependent variable contains

healthy (C0), and types of patients (hepatitis (C1), fibrosis (C2), cirrhosis (C3)) with a 4-label class. In the first part of the study, the classification process was performed as binary classification as a donor and HCV infected patient. The most significant three features in classifying donor and HCV infected patients were identified using the multiple linear regression (MLR). These features, in order of importance, were (i) AST (aspartate amino-transferase), (ii) GGT (gamma-glutamyl transferase) and (iii) BIL (bilirubin). In the second part of the study, the classification process was carried out with a 4-label classification as the donor, hepatitis, fibrosis, and cirrhosis disease.

Performance analysis for models is evaluated using hold out (with 60% training and 40% testing) on training data among the 582 data samples, 350 were used as training datasets and the remaining 232 are the testing datasets. The testing result of NN and SVM models without/with PCA are shown in Fig. 4a, b, c, and d where the overall accuracy and the confusion matrix are shown. The confusion matrix shows the number correctly classified with respect to the two classes. The overall accuracy of SVM models with PCA (Fig. 4d) is 98.7 % which is higher compared the ANN model with PCA result. In its confusion matrix, out from the 210 testing samples in the donor class, 209 were correctly predicted, and 1 was predicted wrongly as patient with an error percentage of 0.48%. Furthermore, 20 of the data samples under patient class were correctly predicted and only 2 was predicted wrongly as donor with an error percentage of 9.09%

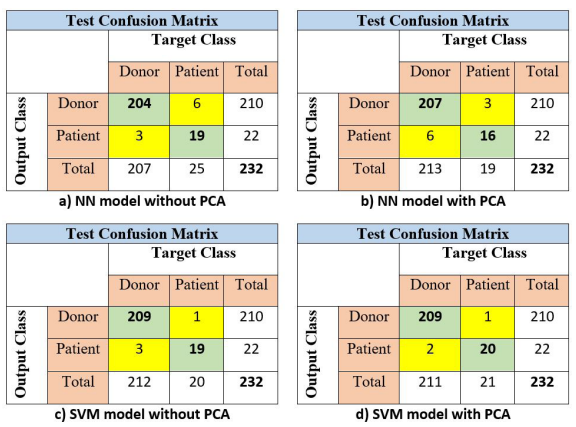

Figure 4. Results of classifier models with/without PCA

Performance metrics derived from the confusion matrix are shown in Table 2. The differences between trials from Table 2 are very significant. For example, for the test trials, the MCC of the NN model (without PCA) is 78.9%, and the MCC of the PCA-NN model is only 76.2%; while for the proposed PCA-SVM model, MCC reaches 92.3%, which is higher than the individual SVM, implying that the PCA-based variable selection has a significant impact on improving SVM performance. The accuracy, sensitivity, specificity, MCC and KAPPA values for binary classification of the best model (SVM with PCA) of this study are 98.7%, 99.1%, 95.2%, 92.3% and 92,31% respectively.

Table 2. Performances results of classifier models

	Accuracy	Sensitivity	Specificity	MCC	Kappa
NN (without PCA)	96.1	98.6	76.0	78.9	0.7870
NN (with PCA)	96.1	97.2	84.2	76.2	0.7593
SVM (without PCA)	98.3	98.6	95.0	89.7	0.8953
SVM (with PCA)	98.7	99.1	95.2	92.3	0.9231

The confusion matrix tables for multiclass ANN and SVM classifiers (with PCA) are constructed as shown in Fig. 5 and 6, respectively. The results show that the PCA-SVM achieves 95.7% classification accuracy with the testing data as indicated in the confusion matrix of Fig. 6. In the 4-label classification using the PCA-SVM model, the prediction accuracy of fibrosis patients was 33.3%, while the prediction accuracy of cirrhosis patients was 88.9%.

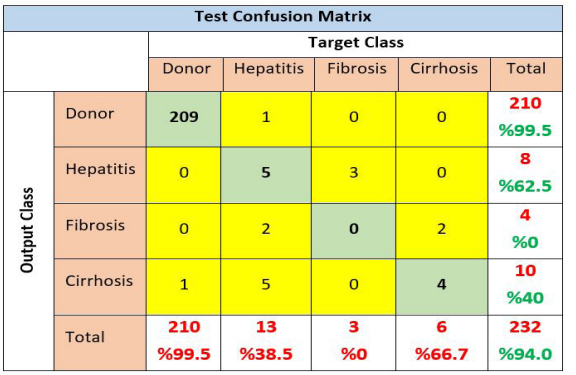

Figure 5. Confusion matrix for multiclass ANN model with PCA

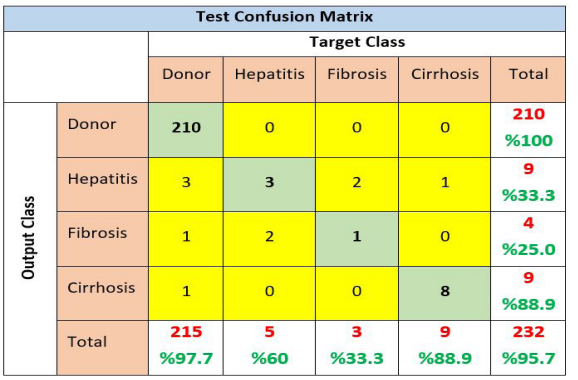

Figure 6. Confusion matrix for multiclass SVM model with PCA

Performance metrics comparison showed SVM with PCA to be a better classifier for this dataset. The highest accuracy was found with SVM (95.7%, 98.7%) in multi and binary class label respectively. The performance score of the binary class shows better that the multiclass label. In the 4-label classification using the PCA-SVM model, prediction accuracies were low for fibrosis (33.3%) and hepatitis (60%).

Since the multi-classification problem for disease diagnosis is more difficult than binary classification, future studies will try to increase the efficiency of the proposed model for multi-classification.

The approach followed in this study is similar in terms of the input variables used by Syafa'ah et al. (2021) for hepatitis C disease classification. Except for age and sex, the selected independent variables are common. While NN and SVM were used for machine learning in this study, they used four classification methods: KNN, naive Bayesian, NN and RF. Syafa'ah et al. (2021) obtained the most accurate classification using NN with 95.12%. In this study, the highest accuracy value was reached with SVM with 98.7. This achievement was achieved on large test data (40%) and size reduction with PCA.

Experimental results show that the proposed SVM model with PCA outperforms the NN (with/without PCA) and a single SVM in terms of classification accuracy, sensitivity, specificity, MCC, and kappa. As stated in [20], the proposed PCA-SVM model has some advantages over individual support vector machines and NN, such as it needs less parameters than NN, overcomes some shortcomings such as overfitting and local minima, and reduces model input space with PCA (speeds SVM learning).

The data set consists of a total of 582 samples collected from 526 donor and 56 patient individuals. The main limitation of this study was the small sample size of the participants, especially patient individuals. A joint study with doctors has been planned to reach more samples.

The study had several disadvantages. The main disadvantage of feature extraction methods such as PCA is that new components are more difficult to interpret. Moreover, it can be more difficult to determine which features contribute the most to the disease prediction. Second, all the datasets here are from the same clinical center. In the future, it will be necessary to conduct a study with larger sample size.

CONCLUSION

Hepatitis C, a liver disease caused by virus, can have a mild course or lead to many different diseases that last a lifetime. If the disease is not treated, it can reach serious dimensions that will impair human health. For these reasons, it is vital to detect the disease early. The application of machine learning classification algorithms for disease prediction is still an emerging field. The diseases can be detected by extracting the features representing the problem from the laboratory and demographic values. In this study, a technique combining SVM and PCA is proposed to construct the classification model and retain the optimal feature subset. PCA based method was used to reduce the dimension of the features. The classifier models' performance is evaluated based on the confusion matrices and their related metrics: accuracy, sensitivity, specificity, MCC, as well as the kappa value. In binary (healthy/unhealthy) classification, the highest level of accuracy obtained in the testing done using PCA was 98.7% achieved with the SVM technique and 98.3% with the NN. The performance of the model is also evaluated by kappa and the maximum kappa of 92.31% is achieved by SVM classifier and the minimum kappa of 75.93% is achieved by NN classifier. This study achieved significant success in predicting hepatitis C disease by using the PCA-SVM model. The performance of this suggested approach is high when compared to other existing methods. With the created model, it is thought that the diagnosis of hepatitis C disease can be facilitated and medical decision support can be provided to the doctor. The results obtained from the study will contribute to the existing literature on hepatitis C and other diseases detection and provide a perspective for future studies. In future studies, the proposed model will be tested on a larger data set and different diseases.

CONFLICT OF INTEREST

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. Hepatitis.https://www.who.int/news-room/fact-sheets/detail/hepatitis-c (accessed Nov. 20,2021)
2. Rayan Z, Alfonse M, Salem ABM. Machine Learning Approaches in Smart Health. Procedia Computer Science 154 (2019) 361–368. https://doi.org/10.1016/J.PROCS.2019.06.052
3. Gündoğdu S. Improving breast cancer prediction using a pattern recognition network with optimal feature subsets. Croatian medical journal 62 (2021) 480–487. https://doi.org/10.3325/cmj.2021.62.480
4. Senturk ZK. Early diagnosis of Parkinson's disease using machine learning algorithms. Medical Hypotheses 138 (2020) 109603. https://doi.org/10.1016/J.MEHY.2020.109603
5. Ayeldeen H, Shaker O, Ayeldeen G, Anwar KM. Prediction of liver fibrosis stages by machine learning model: A decision tree approach, Paper presented at Third World Conference on Complex Systems (WCCS), Marrakech, Morocco, 23-25 November, IEEE, pp. 1-6, 2015. https://doi.org/10.1109/ICOCS.2015.7483212
6. Orczyk T, Porwik P. Liver fibrosis diagnosis support system using machine learning methods, in: Chaki,R., Cortesi, A., Saeed, K., Chaki, N. (Eds) Advanced Computing and Systems for Security. Advances in Intelligent Systems and Computing. Springer, New Delhi, pp.111-121, 2016. https://doi.org/10.1007/978-81-322-2650-5_8
7. Bhargav, KS, Thota DSSB, Kumari TD, Vikas B. Application of Machine Learning Classification Algorithms on Hepatitis Dataset. International Journal of Applied Engineering Research 13 (2018) 12732–12737.
8. Ahammed K, Satu MS, Khan MI, Whaiduzzaman M. Predicting Infectious State of Hepatitis C Virus Affected Patient's Applying

Machine Learning Methods, Paper presented at 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 5-7 June, IEEE, pp. 1371-1374, 2020. http://doi: 10.1109/TENSYMP50017.2020.9230464.

9. Syafa'ah L, Zulfatman Z, Pakaya I, Lestandy M. Comparison of Machine Learning Classification Methods in Hepatitis C Virus. Jurnal Online Informatika 6(2021) 73–78. https://doi.org/10.15575/JOIN.V6I1.719
10. Nandipati SC, XinYing C, Wah KK. Hepatitis C Virus (HCV) Prediction by Machine Learning Techniques. Applications of Modelling and Simulation 4 (2020) 89–100.
11. Shlens J. A tutorial on principal component analysis. arXiv preprint 2014. https://doi.org/10.48550/arXiv.1404.1100
12. Abdi H, Williams LJ. Principal component analysis, Wiley Interdisciplinary Reviews: Computational Statistics 2 (2010) 433-459. https://doi.org/10.1002/wics.101
13. Jolliffe I. Principal component analysis, in Lovric M. (Ed.).International Encyclopedia of Statistical Science. Springer Series in Statistics. Springer, Berlin/Heidelberg, pp.1094-1096, 2011. . https://doi.org/10.1007/978-3-642-04898-2_455
14. Kuhn M, Johnson K. Applied predictive modeling, first ed. Springer, New York, 2013.
15. Kuhn M, Johnson K. Feature engineering and selection: A practical approach for predictive models, first ed. CRC Press, London, 2019.
16. Hotelling H. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24 (1933) 417-441.
17. Shafizadeh-Moghadam H. Fully component selection: An efficient combination of feature selection and principal component analysis to increase model performance. Expert Systems with Applications 186 (2021) 115678. https://doi.org/10.1016/J.ESWA.2021.115678
18. Vinodhini G, Chandrasekaran RM. Sentiment Mining Using SVM-Based Hybrid Classification Model, in: Krishnan G, Anitha, R, Lekshmi R, Kumar, M, Bonato A, Graña M (Eds). Advances in Intelligent Systems and Computing. Paper presented at Computational Intelligence, Cyber Security and Computational Models, Coimbatore, India, 19-21 December. Springer, New Delhi, 246, pp.155-162. 2014. https://doi.org/10.1007/978-81-322-1680-3_18
19. Liang N, Tuo Y, Deng Y, He T. PCA-based SVM classification for simulated ice floes in front of sluice gates. Polar Science (2022) 100839. https://doi.org/10.1016/j.polar.2022.100839.
20. Yanqing Z, Shiwei Z, Junfeng Y, Lei W. Predicting corporate financial distress by PCA-based support vector machines, Paper presented at 2010 International Conference on Networking and Information Technology, Manila, 11-12 June, IEEE, pp. 373-376, 2010. https://doi.org/10.1109/ICNIT.2010.5508491
21. Dua D, Graff C. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science 2019. (accessed Nov. 15, 2021).
22. Jolliffe I. A 50-year personal journey through time with principal component analysis. Journal of Multivariate Analysis 188 (2022) 104820. https://doi.org/10.1016/J.JMVA.2021.104820
23. Uddin MP, Mamun MA, Hossain MA. PCA-based Feature Reduction for Hyperspectral Remote Sensing Image Classification. IETE Technical Review 38 (2021) 377–396. https://doi.org/10.1080/02564602.2020.1740615
24. Schreiber JB. Issues and recommendations for exploratory factor analysis and principal component analysis. Research in Social and Administrative Pharmacy 17 (2021) 1004–1011. https://doi.org/10.1016/J.SAPHARM.2020.07.027
25. Nocedal J, Wright SJ. Numerical Optimization, second ed. Springer, Berlin/Heidelberg, Germany, 2006.
26. Cirillo A, Laudante G, Pirozzi S. Tactile sensor data interpretation for estimation of wire features. Electronics (Switzerland) 10(2021) 1458. https://doi.org/10.3390/ELECTRONICS10121458
27. Toropova AP, Toropov AA, Benfenati E. Semi-correlations as a tool to model for skin sensitization. Food and Chemical Toxicology 157 (2021) 112580. https://doi.org/10.1016/j.fct.2021.112580
28. Lim JY, Nam JS, Shin H, Park J, Song HI, Kang M, Lim K, Choi Y. Identification of Newly Emerging Influenza Viruses by Detecting the Virally Infected Cells Based on Surface Enhanced Raman Spectroscopy and Principal Component Analysis. Analytical Chemistry 91 (2019) 5677–5684. https://doi.org/10.1021/acs.analchem.8b05533