Coronary artery disease classification using support vector machines tuned via randomized search cross-validation

Kemal Akyol (kakyol@kastamonu.edu.tr)
Kastamonu University

Research Article

Keywords: Coronary artery disease, machine learning, support vector machines, hyperparameter tuning, randomized search cross-validation

Posted Date: January 13th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2459099/v1

Abstract

Coronary artery disease outstands health problem that causes high mortality in the world population. This disease brings with it fateful problems such as heart attack and heart failure in patients with cardiovascular problems. Early diagnosis of coronary artery disease is essential for the timely administration of the right treatment and reduction of mortality. Angiography is the most preferred method for CAD detection. However, the complications and costs of this method have led researchers to forage alternative methods through machine learning algorithms. By developing a machine learning model with high generalization ability, prediction errors can be minimized. Thus, these models could potentially be useful for specialist physicians in the effective detection of coronary artery disease. The main focus of this study is to perform coronary artery disease detection with improved support vector machines. k-fold cross-validation experiments were performed on the Z-Alizadeh Sani dataset to evaluate the performance of the models. According to the results obtained, support vector machines with randomized search cross-validation provided the best performance when compared to other models. 87.102% average accuracy, 91.176% average sensitivity, 90.852% average precision, 76.996% average specificity, and also 8.824% average false negative rate obtained by 5-fold cross-validation competes with the known approaches in the literature.

1. Introduction

Cardiovascular disease is the general name for many disorders relevant to heart diseases. There are a particularly coronary artery disease (CAD) [1]. CAD causes a reduced flow of blood to the heart muscles [2]. This disease is known as an important clinical disorder that occurs due to the abnormal functioning of the heart [3]. The blockage or narrowing of blood vessels in the heart causes heart disease or stroke [4]. Cardiovascular diseases, which include heart disease, cerebrovascular disease, and blood vessel disease, are one of the leading causes of death worldwide [2]. A good deal of people is affected by heart disease, especially CAD. World Health Organization reported that approximately 17.9 million people die each year and it is estimated that 32% of the deaths in the world are due to this disease (WHO, 2022). CAD-related deaths could be prevented by accurate detection and timely intervention [6].

Early diagnosis and appropriate treatment of heart disease can reduce and prevent the death rate of patients. Angiography is a common method of diagnosing abnormal narrowing of the heart vessel [7]. Accurate and early diagnosis of heart disease allows timely and accurate treatment and also reduces mortality [8]. Moreover, with timely treatment, the severity of the side effects of CAD can be reduced. Angiography is used to detect stenosis and its location in the heart vessels [9]. In order to diagnose CAD, field experts use different methods based on angiography, which is considered a definitive solution [10]. Coronary angiography [11] is costly and time-consuming. Therefore, machine learning-based systems in CAD diagnosis will help field experts.

In the literature, it is seen that different machine-learning approaches have been used for the CAD problem, so far. However, most of these studies were carried out on outdated datasets such as Cleveland, Hungarian. For example, Polat et al. first obtained new values for each attribute based on the k-nearest neighbor method and the authors performed an artificial immune recognition system with fuzzy resource allocation mechanism-based classification to detect heart disease [12]. In another study, Bahani et al. proposed a method based on a fuzzy rule-based classification [13]. Gárate-Escamila et al. performed experiments with 6 classifiers on the features which are the combination of chi-square and principal component analysis on the heart disease dataset downloaded from the UCI Machine Learning Repository [14]. Lastly, Jabbar et al. proposed a genetic algorithm and k-Nearest Neighbors-based approach to improve accuracy in diagnosing heart disease [15].

In recent years, studies based on artificial intelligence and machine learning have been very popular. Different solutions are obtained with different models by performing machine learning studies on the data obtained from patients [16]. Machine learning-based studies are often used to investigate the relationship between attributes and target classes.

This study was conducted with the motivation of applying machine learning techniques to CAD data in order to develop decision support systems that give support to specialist doctors in their fields. The main purpose of this study is to identify the most successful model for a decision support system that detects CAD at an earlier stage with minimum error. In line with this, a comprehensive comparison of the models based on three scenarios was made. The main contributions of this study are as follows:

1. This study investigates a prospering machine learning method in detecting CAD and proposes a machine learning model that detects this disease with good accuracy.
2. A hybrid model, which includes SVM optimized with Randomize Search cross-validation that has not been used in previous studies, is proposed to classify CAD.
3. The performances of the models are validated on the Z-Alizadeh Sani heart disease dataset with the 5-fold cross-validation techniques.

The remainder of the study is organized as follows: Section 2 provides detailed information on materials and methods. Section 3 introduces the experimental procedure within the data preparation and data classification steps. Section 4 presents the experimental results in detail and discusses the studies in the literature. Finally, Section 5 concludes the work with final remarks.

2. Material And Methods

2.1. Dataset

The performances of the models built in this study were validated on the Z-Alizadeh Sani dataset. This dataset is a rich new dataset with no missing values for CAD which is one of the deadliest and most common types of cardiac disorders known as cardiovascular diseases [17]. This dataset includes 303 samples, 87 Normal (non-CAD) and 216 CAD patients, and comprises 55 features that identify demographic characteristics, electrocardiogram, symptom and examination, and laboratory and echo characteristics. If the patient's coronary diameter is narrowed by 50% or more, the patient is labeled "CAD", otherwise labeled "normal".

2.2. Support vector machines (SVM)

SVM which is one of the classifiers uses a hyperplane to perform the classification by using essential training tuples called support vectors. It is a machine-learning method commonly used in regression and classification [18]. Figure 2 demonstrates the working principle of the SVM classifier.

2.3. Hyperparameter Tuning

High performance is aimed at the machine learning concept by adjusting the hyperparameter values of the classifier algorithms. There are many studies [19–25] in the literature dealing with this issue. In this study, Randomize Search cross-validation and bayesian search cross-validation methods were used for the optimal combination of SVM hyperparameters.

The Randomize Search technique is a technique in which a random combination of hyperparameters is used to find the best parameter values for a model. It works with the best result, assuming that not all hyperparameters are equally important. In this type of search model, random parameter sets are considered at each iteration.

Bayesian optimization algorithm aims to find the optimal solution for the optimization of computation-intensive objective functions with a few evaluations. Therefore, it uses a spare function to carefully consider where the next objective function will be evaluated in the solution space. Bayesian search is based on this algorithm [26]. Based on the Bayesian theorem [26], a surrogate model, which is a probabilistic model of the objective function, looks for specific hyperparameters that give maximum or minimum performance [27]. It is used to predict the next combination that may yield the greatest benefit, based on the hyperparameter combination tested [28].

3. Experiments

With the motivation of the design of a new and effective machine learning model to detect CAD, this study consists of three steps: data preparation, modeling, and model testing. Figure 2 demonstrates the general block diagram of the study.

Hyparameter tuning effectiveness for SVM was performed with a series of experiments on the dataset. In the first step of this study, the dataset was normalized between 0 and 1 by applying the min-max normalization given in Eq. 1. Here, x indicates the current value, x_{min} indicates the minimum value, x_{max} indicates the maximum value and x' indicates the normalized value.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

1

In experimental studies performed with the 5-fold cross-validation technique demonstrated in Fig. 3, the dataset is divided into 5 different subsets. Four of them were used to train the classifier and the remaining one was used to test the performance of the model in each fold. In the next step, Base SVM, Randomize Search cross validation + SVM referred to as Randomize Search + SVM, and Bayesian search cross validation + SVM referred to as Bayesian Search + SVM-based models built were trained on the training sets, and a comparative analysis of these models was made on the test sets. The performances of the models were calculated by using the averages of the measurements in the test sets on each fold. Randomize Search cross-validation and bayesian search cross-validation methods were used with default parameters in the *sklearn* framework.

In this study, classification accuracy (Acc), specificity (Spe) and sensitivity (Sen) evaluation metrics, which are given between Eqs. 2 and 6, were used to compare the models. Here, TP is the True Positive, TN is the True Negative, FP is the False Positive, and the FN is the False Negative. Accuracy (Acc) measures the general classification performance of the model. Sensitivity (Sen) indicates the ratio of correctly predicted CAD samples to all CAD samples. Specificity (Spe) denotes the ratio of correctly predicted non-CAD samples to all non-CAD samples. False Negative Rate (FNR) is the ratio of the CAD samples incorrectly classified to all CAD samples. Precision (Pre) is the ratio of correctly classified CAD samples to all samples classified as CAD.

$$\mathrm{Acc} = \frac{TP + TN}{TP + FN + TN + FP}$$

2

$$\mathrm{Sen} = \frac{TP}{TP + FN}$$

3

$$\mathrm{Spe} = \frac{TN}{TN + FP}$$

4

$$\mathrm{Pre} = \frac{TP}{TP + FP}$$

5

$$\mathrm{FNR} = \frac{FN}{FN + TP}$$

6

4. Results And Discussion

This section considers the CAD classification performance of all models and also available comparisons with studies in the literature. The performance of each model is validated using metrics such as accuracy, sensitivity, specificity, precision, and false negative rate. Table 1 summarizes the experimental results of the models. The bold font in this table indicates the best performance. In experiments, SVM classifiers tuned by bayesian search and the randomized search performed better than base SVM. Specifically, SVM with Randomize Search with cross-validation has higher Acc than other models. Especially Sen and Pre are important metrics for comparing models in terms of the positive class. Sen and Pre metrics show that Randomize Search + SVM is more successful in detecting CAD when compared to other models with 91.176% and 90.852%, respectively. On the other hand, although the Spe values of the models are not very satisfying, Randomize Search + SVM is more successful in detecting control group patients without CAD. In addition, Randomize Search + SVM has 8.824% fewer false negative rates. Overall, this model achieved a 4.276% improvement over base SVM with an average accuracy of 87.102% and is 1.304% more successful than the Bayesian + SVM model. In conclusion, Randomize Search + SVM has proven its success in detecting of CAD.

Table 1
Experimental results of the models with 5-fold cross-validation.

Model	Fold no.#	Acc	Sen	Pre	Spe	FNR
Base SVM	1	86.89	93.18	89.13	70.59	6.82
	2	83.61	90.7	86.67	66.67	9.3
	3	81.97	86.05	88.1	72.22	13.95
	4	78.33	88.37	82.61	52.94	11.63
	5	83.33	88.37	88.37	70.59	11.63
	Average	82.826	89.334	86.976	66.602	10.666
SVM with Bayesian search cross validation	1	86.89	93.18	89.13	70.59	6.82
	2	85.25	90.7	88.64	72.22	9.3
	3	88.52	90.7	92.86	83.33	9.3
	4	83.33	88.37	88.37	70.59	11.63
	5	85.0	90.7	88.64	70.59	9.3
	Average	85.798	90.73	89.528	73.464	9.27
SVM with Randomized search cross-validation	1	95.08	97.73	95.56	88.24	2.27
	2	83.61	90.7	86.67	66.67	9.3
	3	90.16	90.7	95.12	88.89	9.3
	4	83.33	86.05	90.24	76.47	13.95
	5	83.33	90.7	86.67	64.71	9.3
	Average	87.102	91.176	90.852	76.996	8.824

Figure 4 shows the confusion matrices obtained by overlaying the confusion matrices achieved by the 5-fold cross-validation strategy. As a result of 5-fold cross-validation, base SVM classified 58 correctly and 29 incorrectly out of 87 samples in the non-CAD class. This model correctly classified 193 of the 216 samples in the CAD class. Compared to Base SVM, the Bayesian + SVM model correctly classified 64 samples from the CAD class and 196 from the non-CAD class. Finally, the Randomize Search + SVM model correctly classified 67 samples from the CAD class and 197 samples from the non-CAD class compared to Base SVM.

Figure 5 shows the elapsed times for training of the models. As can be seen in this figure, while the base SVM was trained in very short time in each fold, Randomize Search was trained for less time when compared to Bayesian Search.

Also, Table 2 summarizes the comparison of the Randomize Search + SVM-based model, which provided the highest accuracy in this study, with other methods presented in the literature. Shahid and Singh hybridized emotional neural networks with the particle swarm optimization technique, which is a metaheuristic approach, and obtained an average accuracy of 88.34% on the 22 attributes with 10-fold cross-validation [29]. Hu et al. proposed a statistical model based on a finite mixture model of inverted Beta-Liouville distributions for clustering multivariate positive data and achieved an overall classification accuracy of 81.84% on the Z-Alizadeh Sani dataset [30]. Ali and Bukhari used the mutual knowledge-based DNN model, which offers better generalization performance with a significantly lower error rate, and obtained 76.92% classification accuracy on the test dataset [31]. Nasarian et al. examined the heterogeneous hybrid feature selection technique they developed in the CAD problem and also the effects of dataset balancing techniques on various classifiers. The authors reported that the XGBoost classifier offered an overall classification accuracy of 92.58% [32]. As a conclusion of this experiment, it is evidently clear that Randomize Search + SVM-based model has acceptable capability. In particular, the proposed method presented a higher success rate of 87.102% when compared to the studies of [30] and [31].

Table 2
Comparison of the different machine learning studies.

Study	Method	k-fold cross-validation (Yes/No)	Acc(%)
Shahid and Singh [29]	Hybrid PSO-EmNN coupled with feature selection	Yes	88.34
Hu et al. [30]	Var-IBLMM	No	81.84
Ali and Bukhari [31]	MI-DNN	No	76.92
Nasarian et al. [32]	SMOTE + XGBoost	No	92.58
Proposed study	**Randomize Search + SVM**	**Yes**	**87.102**
***The result of the proposed study is in bold**			

5. Conclusion

CAD is one of the most common causes of death worldwide and is recognized as a major disease in middle and older age. In particular, it is a common cardiovascular disease that causes high mortality rates. Angiography is often considered the best method for diagnosing CAD, on the other hand, it has high costs and significant side effects. In this study, the performances of the models built with different scenarios were compared in order to detect the CAD correctly. In experiments with 5-fold cross-validation techniques, the performances of the models built were found as acceptable. Statistical results demonstrate the robustness of Randomize Search CV + SVM in reliably distinguishing CAD patients from healthy ones with a 90.852% value of precision, 87.102% value of accuracy, 91.176% value of sensitivity, and 76.996% value of specificity, respectively. For early detection of CAD, Randomize Search CV + SVM model with a minimum mistake can be included in a decision support system. Finally, it is anticipated that this system can reduce the input of experts and save both cost and time during CAD diagnosis.

Declarations

Acknowledgments: The author would like to thank Arabasadi et al. for providing the Z-Alizadeh Sani dataset [17].

Competing Interests The author has no relevant financial or nonfinancial interests to disclose.

Author Contributions Kemal Akyol: Methodology, Software, Writing - original draft, review & editing.

Ethical approval This research did not require ethics approval.

Consent to Participate A requirement for informed consent was waived for this study because of the anonymous nature of the data.

Consent for Publication Not applicable.

Data Availability Statement This study uses public dataset available at:

https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani

References

1. Libby, P., Ridker, P.M., Hansson, G.K.: Progress and challenges in translating the biology of atherosclerosis. Nature 2011 473:7347. 473, 317–325 (2011). https://doi.org/10.1038/nature10146
2. Global atlas on cardiovascular disease prevention and control / edited by: Shanthi Mendis ... [et al.], https://apps.who.int/iris/handle/10665/44701
3. Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., Yang, G.-Z.: Deep Learning for Health Informatics. IEEE J Biomed Health Inform. 21, 4–21 (2017). https://doi.org/10.1109/JBHI.2016.2636665
4. Prabhakaran, D., Jeemon, P., Roy, A.: Cardiovascular Diseases in India: Current Epidemiology and Future Directions. Circulation. 133, 1605–1620 (2016). https://doi.org/10.1161/CIRCULATIONAHA.114.008729
5. Cardiovascular diseases, https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
6. Abdar, M., Książek, W., Acharya, U.R., Tan, R.S., Makarenkov, V., Pławiak, P.: A new machine learning technique for an accurate diagnosis of coronary artery disease. Comput Methods Programs Biomed. 179, 104992 (2019). https://doi.org/10.1016/J.CMPB.2019.104992

7. Alizadehsani, R., Habibi, J., Sani, Z.A., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Khozeimeh, F., Alizadeh-Sani, F.: Diagnosing Coronary Artery Disease via Data Mining Algorithms by Considering Laboratory and Echocardiography Features. Res Cardiovasc Med. 2, 133 (2013). https://doi.org/10.5812/CARDIOVASCMED.10888
8. Ashish, L., V, S.K., Yeligeti, S.: Ischemic heart disease detection using support vector Machine and extreme gradient boosting method. Mater Today Proc. (2021). https://doi.org/10.1016/J.MATPR.2021.01.715
9. Alizadehsani, R., Habibi, J., Hosseini, M.J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Bahadorian, B., Sani, Z.A.: A data mining approach for diagnosis of coronary artery disease. Comput Methods Programs Biomed. 111, 52–61 (2013). https://doi.org/10.1016/j.cmpb.2013.03.004
10. Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., Yarifard, A.A.: Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Comput Methods Programs Biomed. 141, 19–26 (2017). https://doi.org/10.1016/j.cmpb.2017.01.004
11. Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P.M., Khozeimeh, F., Nahavandi, S., Sarrafzadegan, N., Acharya, U.R.: Machine learning-based coronary artery disease diagnosis: A comprehensive review. Comput Biol Med. 111, (2019). https://doi.org/10.1016/J.COMPBIOMED.2019.103346
12. Polat, K., Şahan, S., Güneş, S.: Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing. Expert Syst Appl. 32, 625–631 (2007). https://doi.org/10.1016/J.ESWA.2006.01.027
13. Bahani, K., Moujabbir, M., Ramdani, M.: An accurate fuzzy rule-based classification systems for heart disease diagnosis. Sci Afr. 14, e01019 (2021). https://doi.org/10.1016/J.SCIAF.2021.E01019
14. Gárate-Escamila, A.K., Hajjam El Hassani, A., Andrès, E.: Classification models for heart disease prediction using feature selection and PCA. Inform Med Unlocked. 19, 100330 (2020). https://doi.org/10.1016/J.IMU.2020.100330
15. jabbar, M.A., Deekshatulu, B.L., Chandra, P.: Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm. Procedia Technology. 10, 85–94 (2013). https://doi.org/10.1016/J.PROTCY.2013.12.340
16. Condie, T., Mineiro, P., Polyzotis, N., Weimer, M.: Machine learning on big data. Proc Int Conf Data Eng. 1242–1244 (2013). https://doi.org/10.1109/ICDE.2013.6544913
17. Alizadehsani, R., Habibi, J., Hosseini, M.J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Bahadorian, B., Sani, Z.A.: A data mining approach for diagnosis of coronary artery disease. Comput Methods Programs Biomed. 111, 52–61 (2013). https://doi.org/10.1016/j.cmpb.2013.03.004
18. Huang, S., Nianguang, C.A.I., Penzuti Pacheco, P., Narandes, S., Wang, Y., Wayne, X.U.: Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics. 15, 41–51 (2018). https://doi.org/10.21873/CGP.20063

19. P, P.K., V, M.A.B., Nair, G.G.: An efficient classification framework for breast cancer using hyper parameter tuned Random Decision Forest Classifier and Bayesian Optimization. Biomed Signal Process Control. 68, 102682 (2021). https://doi.org/10.1016/J.BSPC.2021.102682
20. Valarmathi, R., Sheela, T.: Heart disease prediction using hyper parameter optimization (HPO) tuning. Biomed Signal Process Control. 70, 103033 (2021). https://doi.org/10.1016/J.BSPC.2021.103033
21. Nayak, J., Naik, B., Dash, P.B., Souri, A., Shanmuganathan, V.: Hyper-parameter tuned light gradient boosting machine using memetic firefly algorithm for hand gesture recognition. Appl Soft Comput. 107, 107478 (2021). https://doi.org/10.1016/J.ASOC.2021.107478
22. Kousalya, K., Saranya, T.: Improved the detection and classification of breast cancer using hyper parameter tuning. Mater Today Proc. (2021). https://doi.org/10.1016/J.MATPR.2021.03.707
23. Kalita, D.J., Singh, V.P., Kumar, V.: A dynamic framework for tuning SVM hyper parameters based on Moth-Flame Optimization and knowledge-based-search. Expert Syst Appl. 168, 114139 (2021). https://doi.org/10.1016/J.ESWA.2020.114139
24. Srinivas, P., Katarya, R.: hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost. Biomed Signal Process Control. 73, 103456 (2022). https://doi.org/10.1016/J.BSPC.2021.103456
25. Xia, Y., Liu, C., Li, Y.Y., Liu, N.: A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Syst Appl. 78, 225–241 (2017). https://doi.org/10.1016/J.ESWA.2017.02.017
26. Elsayad, A.M., Nassef, A.M., Al-Dhaifallah, M.: Bayesian optimization of multiclass SVM for efficient diagnosis of erythemato-squamous diseases. Biomed Signal Process Control. 71, 103223 (2022). https://doi.org/10.1016/J.BSPC.2021.103223
27. Cho, H.U., Nam, Y., Choi, E.J., Choi, Y.J., Kim, H., Bae, S., Moon, J.W.: Comparative analysis of the optimized ANN, SVM, and tree ensemble models using Bayesian optimization for predicting GSHP COP. Journal of Building Engineering. 44, 103411 (2021). https://doi.org/10.1016/J.JOBE.2021.103411
28. Li, H., Lin, Z., An, Z., Zuo, S., Zhu, W., Zhang, Z., Mu, Y., Cao, L., Daniel, J., García, P.: Automatic electrocardiogram detection and classification using bidirectional long short-term memory network improved by Bayesian optimization. Biomed Signal Process Control. 73, 103424 (2022). https://doi.org/10.1016/J.BSPC.2021.103424
29. Shahid, A.H., Singh, M.P.: A Novel Approach for Coronary Artery Disease Diagnosis using Hybrid Particle Swarm Optimization based Emotional Neural Network. Biocybern Biomed Eng. 40, 1568–1585 (2020). https://doi.org/10.1016/j.bbe.2020.09.005
30. Hu, C., Fan, W., Du, J.X., Bouguila, N.: A novel statistical approach for clustering positive data based on finite inverted Beta-Liouville mixture models. Neurocomputing. 333, 110–123 (2019). https://doi.org/10.1016/J.NEUCOM.2018.12.066
31. Ali, L., Bukhari, S.A.C.: An Approach Based on Mutually Informed Neural Networks to Optimize the Generalization Capabilities of Decision Support Systems Developed for Heart Failure Prediction.

IRBM. 42, 345–352 (2021). https://doi.org/10.1016/J.IRBM.2020.04.003

32. Nasarian, E., Abdar, M., Fahami, M.A., Alizadehsani, R., Hussain, S., Basiri, M.E., Zomorodi-Moghadam, M., Zhou, X., Pławiak, P., Acharya, U.R., Tan, R.S., Sarrafzadegan, N.: Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach. Pattern Recognit Lett. 133, 33–40 (2020). https://doi.org/10.1016/j.patrec.2020.02.010

Figures

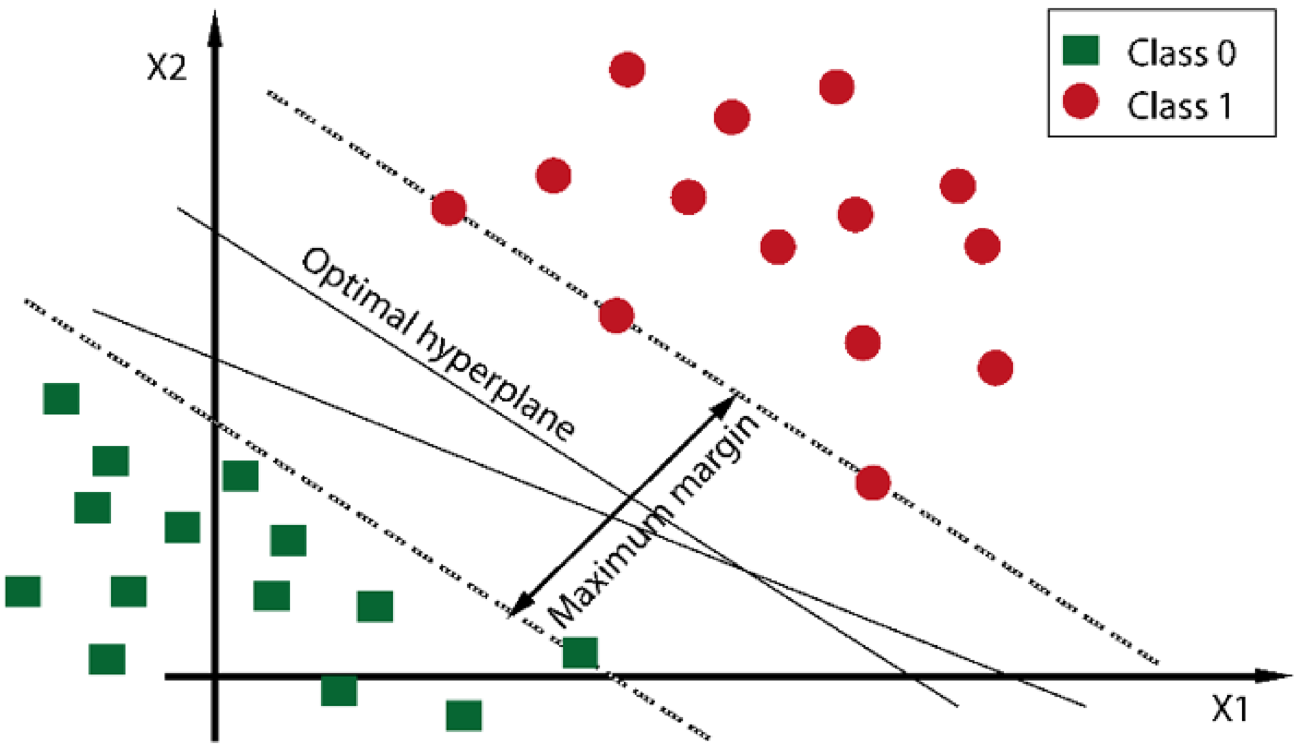

Figure 1

Schematic of a Linear SVM for two-dimensional data.

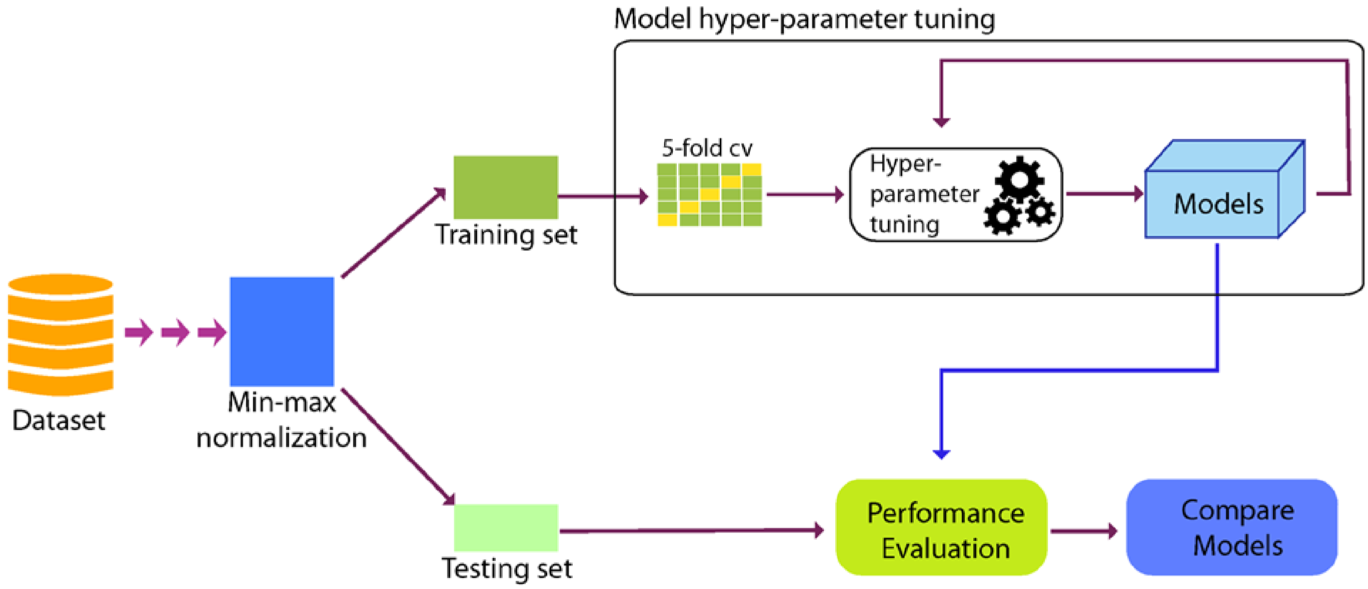

Figure 2

General block diagram of the proposed study.

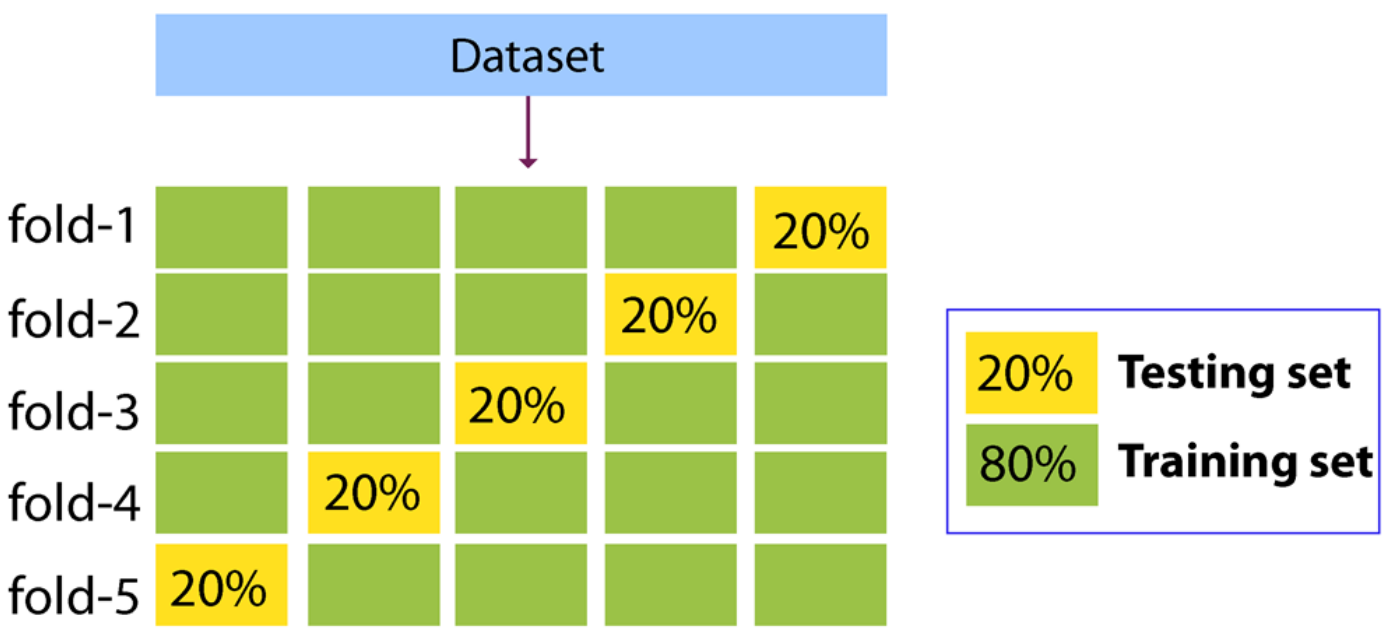

Figure 3

5-fold cross-validation technique

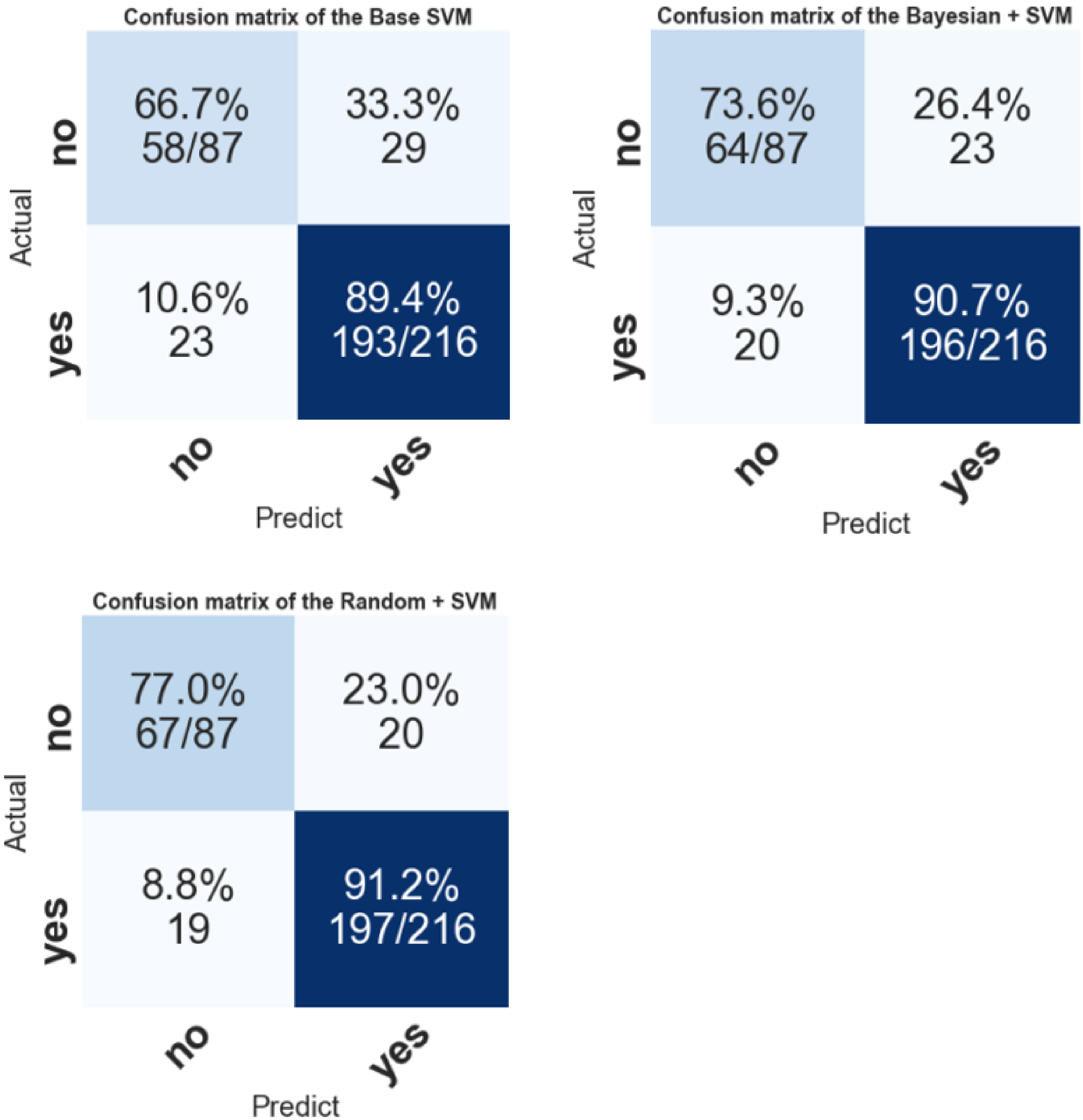

Figure 4

Overlapped confusion matrices of the models

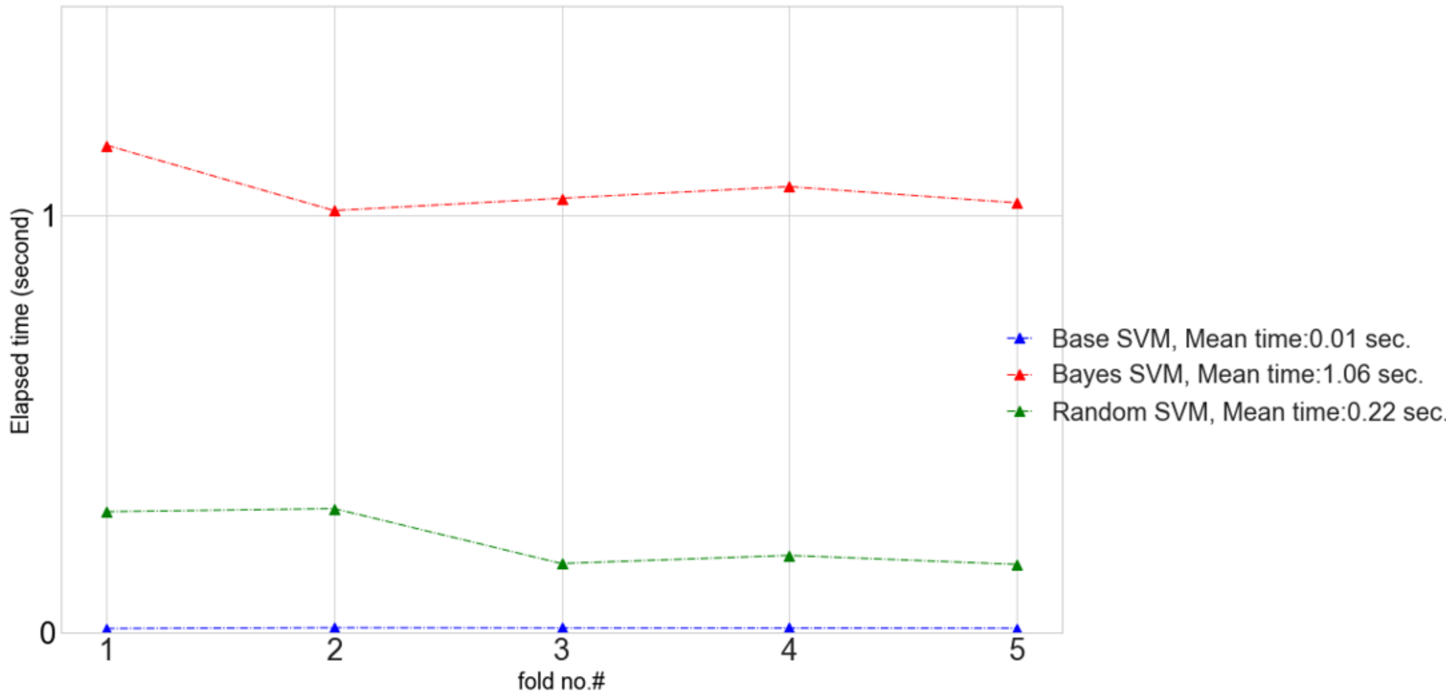

Figure 5

Elapsed times for training of the models.