JOURNAL OF SOFTWARE ENGINEERING & INTELLIGENT SYSTEMS
ISSN 2518-8739
31 December 2019, Volume 4, Issue 3, JSEIS, CAOMEI Copyright © 2016-2019
www.jseis.org

A SUPPORT VECTOR MACHINE BASED HEART DISEASE PREDICTION

TSEHAY ADMASSSU ASSEGIE

Department of Computing Technology, College of Engineering and Technology
Aksum University, Aksum, Ethiopia
Email: tsehayadmassu2006@gmail.com

ABSTRACT

Academic Editor: Sajid Khan, APCOMS, Rawalpindi.

Published: 31 December, 2019

Funding: Authors have no funds.

Article Type: Research Article

Disease classification problem can be solved by building machine learning models by training a machine to identify disease classes. The classification of disease is achieved by using machine learning algorithm like Support Vector machine (SVM). A support vector machine is an approach to machine learning in which the machine uses predefined labels from the known set to determine or predict new classes of disease which has never seen before. In this paper, we used the standard Kaggle heart disease dataset for classification of heart disease using a support vector machine learning algorithm. Finally, the accuracy of SVM is evaluated and the evaluation result shows 73.41% accuracy on heart disease classification.

Keywords: machine learning; heart disease classification; support vector machine; heart disease, prediction;

1. INTRODUCTION

The field of Computer Science is advancing and the success in machine learning is becoming vital to disease diagnosis. Disease classification is one the applications of the Computer Science field where Artificial Intelligence is used in support of the medical fields where physicians are aided with computers in disease diagnosis with better accuracy solving the problems associated with the physicians due to lack of experiences or stress which may in turn make the diagnosis difficult [1] .

The heart disease is among main reason for death around the world in both developing and developed countries although, the disease causes less mortality rate in developed nations when compared to developing countries [2]. Therefore, medical diagnosis is significant to reduce the mortality rate due to heart disease. However, diagnosis is not a simple task, due to the complexities in confirming the accuracy and precision of the results. Usually, the heart disease diagnosis is carried manually and this requires time especially in situations where there is small number of physicians than the patients waiting for the diagnosis.

As discussed in [3], heart disease causes 17 million deaths across the world. Among the many factors to the death caused by the heart disease, the major causes to this are lack of proper diagnosis and treatment to the disease. Supervised machine learning is widely applied to disease classification to simplify the diagnosis process in order to minimize the death. This paper is therefore aimed to identify the major factors making the heart disease diagnosis difficult and finally propose a solution to the identified problems.

In this study we are going to answer the following questions: 1) Can we build a machine learning model with the SVM, a supervised machine learning algorithm to classify heart disease with better accuracy? 2) What is the accuracy of SVM learning algorithm on heart disease classification? 3) How can the SVM algorithm be optimized on the heart disease classification to improve the accuracy on classification of the heart disease?

2. RELATED WORKS

In this section, the previous studies related to heart disease classification using supervised machine learning algorithm, such as SVM, Random Forest and Decision Tree is discussed. The literature reviews [4-10] are discussed the following sections. A heart disease classifier works once the heart disease features are extracted using [4] feature extraction method. Heart disease diagnosis using supervised machine learning algorithms is becoming a greater demand as a classification problem is automated using such algorithms to simplify the classification process and ensure better accuracy in classification avoiding human errors and saves time. A Support vector Machine is one of the simplest supervised machine learning algorithms used in disease classification [5]. The algorithm has better accuracy in classification compared to other supervised learning algorithms, such as Decision Tree and Random Forest.

Corresponding Author: Tsehay Admasssu Assegie
Email Address: tsehayadmassu2006@gmail.com

One of the common machine learning algorithms used in disease classification problem is the Naïve Bayes algorithm. In the study, the authors proposed a machine learning system for heart disease classification using the University of California Irvine (UCI) data repository is used to build a machine learning model for heart disease classification using the Naïve Bayes learning algorithm. The authors have also performed model evaluation on Random Forest, SVM and Naïve Bayes models. The accuracy analysis result of the models showed that Naïve Bayes model has better accuracy than the other classifiers on heart disease classification.

A supervised learning algorithm has significant importance in medical diagnosis, object recognition such as handwritten digits recognition, face recognition and so on [6-7]. The Support Vector Machine (SVM) is a supervised machine learning algorithm widely applied to several applications in recent days like bioinformatics, text categorization, object detection, big data analysis.

One of the common machine learning algorithms used in disease classification problem is the Naïve Bayes algorithm [7]. In the study, the authors proposed a machine learning system for heart disease classification using the University of California Irvine (UCI) data repository was used for testing the algorithm. The accuracy test on Random Forest, Naïve Bayes, SVM and Decision Tree model showed that Naïve Bayes model has better accuracy than the other classifiers on heart disease classification.

In [8], a comparative study on heart disease classification algorithms was conducted and the accuracy of the Decision Tree, linear SVM and quadratic SVM was reported in the study. Linear SVM has better accuracy compared to other classification algorithms with an accuracy score of 65.3% and the quadratic SVM performed well next to the linear SVM. The Decision Tree has least accuracy compared to all of the classifiers the SVM, ensemble space discriminant methods.

One of the significant contributions of machine learning field in medicine is disease diagnosis. The application of this field in identification of feature form large data repositories of disease is vital to the identification of disease classes. Recently, many researchers have applied machine learning models on the classification and clustering larger medical datasets to assist the disease diagnosis process. Some of the disease with complex features includes heart disease and cancer [9]. In [10], heart disease prediction system using KNN (K-Nearest Neighbor) was proposed. The study showcased that features used in classification of disease are important to achieve better accuracy in classification using machine learning algorithms like the KNN.

Another study on heart disease classification is proposed in [11] using the deep neural network algorithm for classification. The deep neural network has an accuracy of 83.67% and accuracy is important metric to evaluate the classification performance and the relevance of prediction in diagnosis of heart disease.

Diagnosis of heart disease, specially the coronary heart disease is difficult and complicated [12]. The prediction of the disease can be modelled using machine learning algorithm like the multi-layer perception (MLP-Artificial Neural Network). The complexity in diagnosis of this type is that many diseases have similar features with the coronary heart disease and this makes it difficult to easily identified and diagnosed early as possible before the disease causes death. The model proposed by the authors has good accuracy although; the accuracy was affected by the occurrences of over fitting. A comparative study on the performance of machine learning models, SVM, Logistic Regression and Artificial Neural Network (ANN) shows that Logistic Regression having better performance than the SVM and ANN models in the prediction of heart disease [13,14].

Disease diagnosis is the process of making decision on unknown event from known set of medical records and the physician's experiences [15]. The known sets of medical records from clinics, hospitals or medical centers in general can be provided to a machine to model intelligent systems for decision making support and classification of disease. But, the machine learning systems vary in accuracy of prediction and classifications in the same manner medical specialists have different experience on threating disease. The main factor affecting the accuracy of machine learning algorithms, are the amount of data presented to the machine to learn from and the quality of the data. The larger the dataset and the more features are presented to the machine in training, the better accuracy in prediction of the learning algorithm on a given new and unknown cases.

3. RESEARCH METHODOLOGY

To build a machine learning system using the support vector machine (SVM) algorithm for diagnosis of heart disease, the Kaggle heart disease dataset is used. We have used python for implementing and testing the accuracy of the classifier on heart disease classification.

2.1 Dataset description

The Kaggle heart disease dataset is used to build machine learning model using the SVM. This dataset consists of 303 samples of heart disease patients and non-heart disease patients. The dataset has 138 non-heart disease patients' features and 168 heart disease patients' features. The 75% of the dataset, which is 227 of the dataset samples, are used for training and 76 dataset samples are used in testing. The dataset is visualized as shown in Figure 1.

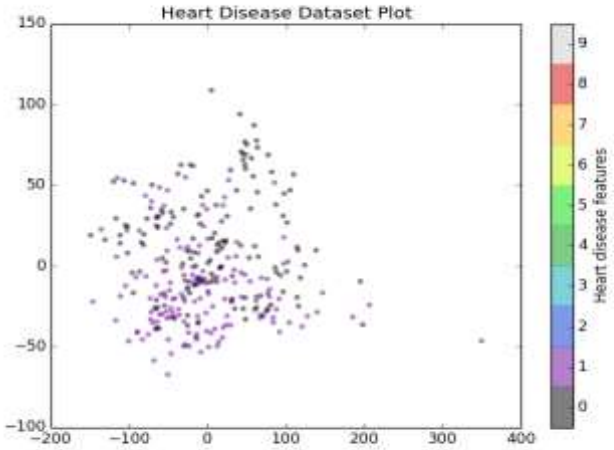

Figure. 1 Heart disease dataset features plot

Table. 1 Feature description of the kaggle heart disease dataset

Instance	Feature name	Feature code	Description
1	Age	age	Age in years

2	Sex	sex	male=1
			female=0

3	chest pain	cp	atypical angina=1 typical angina=2 asymptomatic=3 non-angina pain=4
4	calcium scan	thal	normal=3 fixed defect=6 reversible defect=7
5	Fasting blood sugar	fbs	normal=0 having ST_T=1
6	Slope of the peak exercise ST segment	slope	up sloping=1 flat=2 down sloping=3

7	Resting electrocardiographic results	restecg	normal=0
			having ST_T=1
			hypertrophy=2

8	target	target	normal=0
			patient=1

4. RESULTS AND FINDINGS

The experiment was conducted on heart disease data repository of the standard Kaggle using the support vector machine classifier for training the machine. In this data repository, features are extracted. The data repository consists of 303 heart disease positive sets and heart disease negative sets. Among the 303 samples of the dataset, 138 are heart disease negative and 165 are heart disease positive (heart disease patients). The 75 % of the dataset is used for training and 25 % is used in testing the SVM classifier.

4.1 Accuracy Analysis

The accuracy of SVM on the heart disease dataset is evaluated using the python sklearn metric library which is used to evaluate accuracy of classification algorithm. Table 2 and Figures 2-3 show the accuracy of the SVM.

Table. 2 Accuracy of SVM on heart disease diagnosis

Random test on the SVM model	Accuracy in %
1	88.15
2	64.47
3	73.68
4	67.1
5	73.68

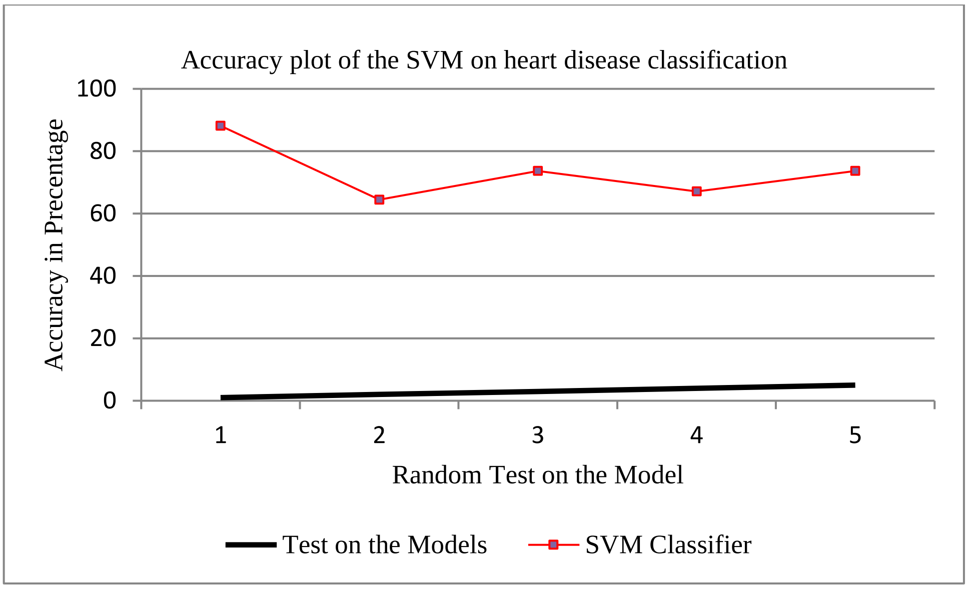

Figure. 2 Accuracy plot of SVM on heart disease classification

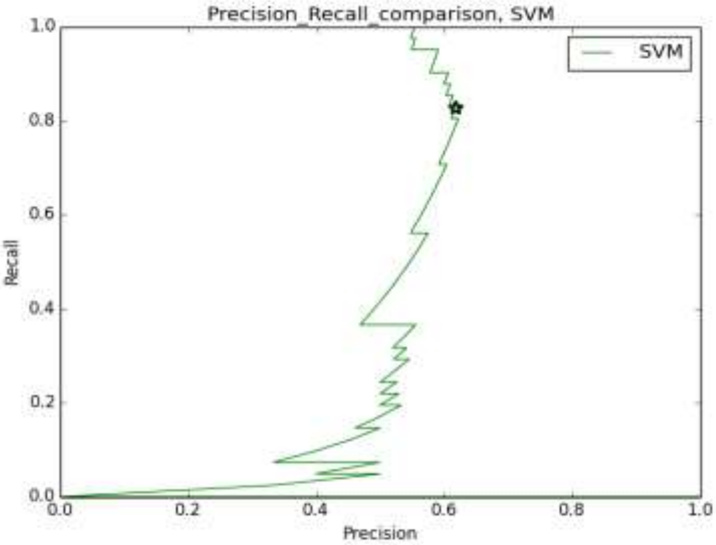

Figure. 3 SVM recall precision curve on heart disease classification

As illustrated in Figure 3, we can see that SVM performs better at around the middle (approximately precision=0.4), this shows that SVM has better performance at the middle precision requirements.

4.2 Confusion analysis

The confusion matrix is used to evaluate the prediction of the SVM model. The confusion matrix shows the miss-classifications (the false positive and false negatives) and corrects classification (the true positive and true negative) of the SVM algorithm on classification of the heart disease as illustrated in Table 3.

Table.3 SVM confusion matrix

Random test on the model	Predictions of the SVM model			
	TP	TN	FP	FN
1	33	26	5	12
2	33	22	10	11
3	36	25	7	8
4	31	27	9	9
5	29	28	8	11

Figure 4, demonstrates the confusion matrix for SVM on diabetes disease classification on five random tests on the model. Considering the test 1, we have noted that the correct prediction that is true positive (TP) and true negative (TN) outclasses the incorrect prediction that is false positive (FP) and false negative (FN). When we look at the numerical values of correct prediction the total is 59 (TP+TN=33+26=59) and three are 17 incorrect predictions by the model (FP+FN=5+12=17). The accuracy for the test 1 is calculated using the following formula:

$$\text{SVM model Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} * 100 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . (1)$$

Substituting the values, we get the following result:

$$\text{SVM model Accuracy} = \frac{33 + 26}{33 + 26 + 5 + 12} * 100 = 77.63\%$$

The confusion matrix of SVM on heart disease classification is shown in figure 4. As shown in the figure in all tests on the SVM algorithm the TP and TN or the true class is greater than the false classes or the FP and FN values for each test on the algorithm and this shows that SVM model is best suited for heart disease classification and the model can assist physicians' in the heart disease diagnosis process.

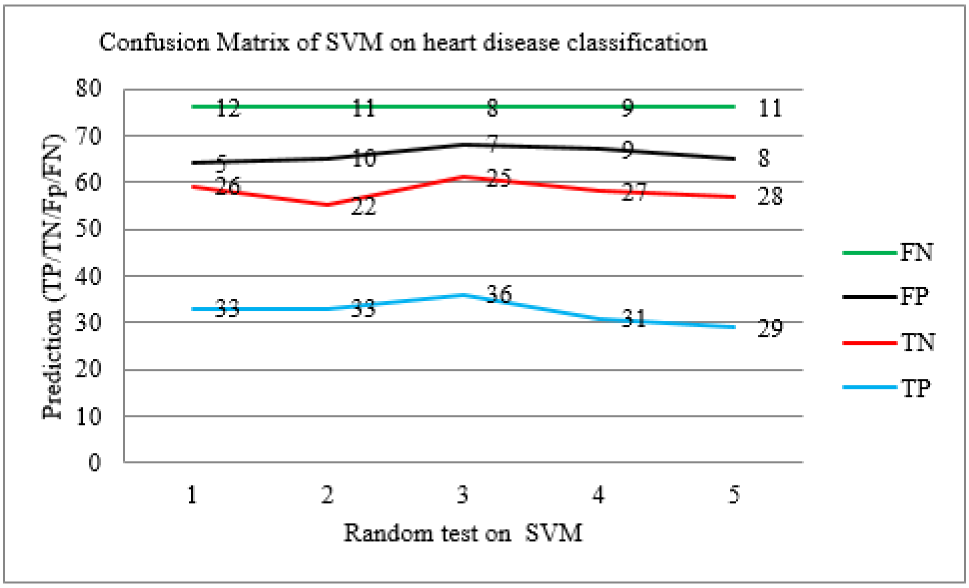

Figure. 4 Confusion matrix of SVM on heart disease diagnosis

5. CONCLUSION

In this paper, we have proposed a Support Vector Machine learning model to solve heart disease classification problem. In building the model the machine was trained with a Kaggle data repository consisting of 303 samples of which 138 are heart disease negative and 165 are heart disease positive (heart disease patients). Some of the features used during in training the model are age, slope, and sex and chest pain. Finally, we have analysed the performance of the model using different performance metrics such as confusion matrix, recall, precision and accuracy of the model on prediction of the heart disease. The accuracy analysis result shows that an average accuracy of 73.41% is achieved on classification of heart disease using the SVM model.

ACKNOWLEDGEMENT

Special thanks to Aksum University, Aksum, Ethiopia for supporting this research.

REFERENCES

1. *Heart Attack*, International Journal of Electrical and Computer Engineering (IJECE) Vol. 8, No. 2, April 2018.
2. T. R. Stella Mary, Shoney Sebastian, *Predicting heart ailment in patients with varying number of features using data mining techniques*, International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 4, August 2019.
3. Wan Hajarul Asikin Wan Zunaidi, RD Rohmat Saedudin, Zuraini Ali Shah, Shahreen Kasim, Choon Sen Seah, Maman Abdurohman, *Performances Analysis of Heart Disease Dataset using Different Data Mining Classification*, International Journal on Advanced Science Engineering and Information Technology, Vol.8, No. 6, 2018.
4. Nikhil Gawande , Alka Barhatte , *Heart Diseases Classification using Convolutional Neural Network*, Proceedings of the 2nd International Conference on Communication and Electronics Systems (ICCES), IEEE, 2017.
5. Aufzalina Mohd Yusof, Nor Azura Md. Ghani, Khairul Asri Mohd Ghani, Khairul Izan Mohd Ghani, *A predictive model for prediction of heart surgery procedure*, Indonesian Journal of Electrical Engineering and Computer Science Vol. 15, No. 3, September 2019.
6. Tsehay Admassu Assegie, Pramod Sekharan Nair, *Handwritten digits recognition with decision tree classification: a machine learning approach*, International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 5, October 2019.
7. R. Chitra, Dr.V. Seenivasagam, *Heart Disease Prediction System Using Supervised Learning Classifier*, Bonfring International Journal of Software Engineering and Soft Computing, Vol. 3, No. 1, March 2013.
8. Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun, *A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms*, Hindawi Mobile Information Systems Volume 2018.
9. Simge EKIZ, Pakize ERDOGMUS, *Comparative Study of Heart Disease Classification*, 978-1-5386-0440-3/17, IEEE, 2017.
10. Gaurav Meena, Pradeep Singh Chauhan, Ravi Raj Choudhary, *Empirical Study on Classification of Heart Disease Dataset-its Prediction and Mining*, International Conference on Current Trends in Computer, Electrical, Electronics and Communication (ICCTCEEC), 2017.
11. Theresa Princy. R, J. Thomas, *Human Heart Disease Prediction System using Data Mining Techniques*, International Conference on Circuit, Power and Computing Technologies, IEEE, 2016.
12. Kathleen H. Miaoa, b, Julia H. Miao, *Coronary Heart Disease Diagnosis using Deep Neural Networks*, International Journal of Advanced Computer Science and Applications, Vol. 9, No. 10, 2018.
13. Wiharto Wiharto, Esti Suryani, Vicka Cahyawati, *The methods of duo output neural network ensemble for prediction of coronary heart disease*, Indonesian Journal of Electrical Engineering and Informatics (IJEEI) Vol. 7, No. 1, March 2019.
14. Divyansh Khanna, Rohan Sahu, Veeky Baths, Bharat Deshpande, *Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease*, International Journal of Machine Learning and Computing, Vol. 5, No. 5, October 2015.
15. Ketkee Mankar, Dr. A. D Gawande, *Disease diagnosis, a review*, International Journal of Computer Engineering and Applications, Volume IX, Issue II, February 2015.

AUTHORS PROFILE