


# Kernel Parameter Tuning to Tweak the Performance of Classifiers for Identification of Heart Diseases


Annu Dhankhar, B. M. Institute of Engineering and Technology, Sonapat, India

Sapna Juneja, IMS Engineering College, Ghaziabad, India

Abhinav Juneja, KIET Group of Institutions, Delhi NCR, Ghaziabad, India

 <https://orcid.org/0000-0003-1984-0125>

Vikram Bali, JSS Academy of Technical Education, Noida, India

 <https://orcid.org/0000-0002-2809-8455>

## ABSTRACT

Medical data analysis is being recognized as a field of enormous research possibilities due to the fact there is a huge amount of data available and prediction in initial stage may save patient lives with timely intervention. With machine learning, a particular algorithm may be created through which any disease may be predicted well in advance on the basis of its feature sets or its symptoms can be detected. With respect to this research work, heart disease will be predicted with support vector machine that falls under the category of supervised machine learning algorithm. The main idea of this study is to focus on the significance of parameter tuning to elevate the performance of classifier. The results achieved were then compared with normal classifier SVM before tuning the parameters. Results depict that the hyperparameters tuning enhances the performance of the model. Finally, results were calculated by using various validation metrics.

## KEYWORDS

Cardiovascular Disease, Hyperparameter Tuning, Medical Analyser, Support Vector Machines

## 1. INTRODUCTION

The primary reason of higher death rate around the world is linked with Cardiovascular diseases whether it is a developed, developing or under developed country (Ayatollahi et al., 2019). In order to have a better understanding of the conditions and diseases that impact the heart and its vessels, it must be explored that along with the circulatory system of blood may have some relationship to the coronary vascular diseases (Winker et al., 2015) (Task et al., 2013). Heart is one of the vital organs of the human body structure and furnishes blood to every corner of our anatomy. If heart discontinues to function as expected, then the other body parts like brain and multiple other organs will cease to function, and this can further cause sudden death of person. Heart diseases have emerged as one of the principal factors of deaths throughout the world. In the current competitive and busy lifestyle scenario, the detected cardiovascular diseases (CVD) are escalating on a daily basis. The World Health Organization (WHO) approximates that nearly 17 million people fail to survive every year pertaining to these cardiovascular disease, primarily caused due to heart attacks and strokes (Yamashita et al.,

DOI: 10.4018/IJEHMC.20210701.oa1

This article, published as an Open Access article on April 23rd, 2021 in the gold Open Access journal, the International Journal of Information and Communication Technology Education (converted to gold Open Access January 1st, 2021), is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

2018). Overweight, obesity, hypertension, hyperglycaemia, and high cholesterol are few potential drivers that are most eventful in triggering heart related issues. Additionally, American Heart Association (Stewart et al., 2019) cited symptoms including sudden gain in weight (1–2kg per day), disordered sleep pattern, swelling of body parts, cough and amplified heart beat rate (Parasuraman et al., 2019). That is the reason it became evident to annotate the extremely governing indicators and healthy lifestyles that can contribute to CVD. Prior to diagnosis of CVD there are multiple tests performed, which includes blood pressure, ECG, blood sugar, cholesterol etc. Typically such tests are generally repeated in case the patient becomes critical in condition and patient must begin taking medical aid immediately, it often becomes critical to give priority to tests (Rumsfeld et al., 2016). There are a countless heart diseases which includes failure of heart, stroke and coronary artery diseases as well (Jatav & Sharma, 2018). With respect to medical analysis, there has been a huge demand for identifying or predicting the diseases a person is potentially suffering from, before the start of damage by that disease, with only knowing about its symptoms. Working with a few classifiers and by analysing the symptoms we can predict the actual disease with which a person can suffer in near future. The presence of innovative technologies like machine learning, we can reasonably justify the matches that coexist among the data very quickly.

### **Machine Learning Classification Methods**

Classification is that form of supervised learning which is used for assigning a class label to the incoming data sample in a machine learning model. There are a number of proposed algorithms/methods for classification which are trusted for prediction applications. The popular methods are given as under:

(a) Decision Tree (DT)

Decision Tree uses an approach to split the data into one or the other classes by applying some constraints on the sample data. The method uses a top–down algorithm, which selects a particular attribute at every stage and splits the tree based on the appropriate separator of the classes at that stage, this process iteratively continues in the sub-classes/problems that originate from the current split (Karaolis et al., 2010).

(b) Logistic Regression (LOG)

This method is quite helpful in prediction when there is a binary dependent variable to classify. Logistic regression is generally used for exploring the data. It also comprehends the association among a dependent dichotomous variable and some nominal/ordinal independent variables (Nordhausen, 2009).

(c) Naïve Bayes (NB)

This method is based on conditional probability where we predict the probability of occurrence of one event provided some other event/s has already occurred (Joachims, 2000). This is one of the very commonly used prediction methodology.

(d) Random forests (RF)

This classifier encompasses a collection of decision trees from a set of randomly chosen training data subset. The final class of test sample object is decided based on the net aggregated weight of voting from all the participating decision trees (Jatav & Sharma, 2018).

(e) K-Nearest Neighbour (KNN)

This algorithm assumes some initial categories/ classes of data as its initial state. The algorithm assigns one of the category to the incoming sample based on the most similar category for the sample. The neighbours of sample decide its category or class in the most simple words (Khourdifi & Bahaj, 2019).

(f) Support Vector Machine (SVM)

This method exhibits high performance when used for high dimensional spaces. Also the method shows its effectiveness in instances where samples available are less than the number of dimensions. The method is memory efficient as it employs only a subset of the training points for the decision function which are referred to as the support vectors (Ayatollahi et al., 2019).

(g) Neural networks (NN)

These methods can be considered as a learning system based on computing through a network of functions to realize and transform a data input received in one form to a chosen output, generally in alternative form (Lecun et al., 2015).

(h) AdaBoost

Adaboost is a popular boosting algorithm that may be applied to several machine learning classifiers to further improve their performance. It is dominantly employed with weak learning algorithms. These methods show significant accuracy in a classification problem compared to a random chance (Tu et al., 2017).

(i) Gradient Boosting Machines (GBM)

These are also popularly referred to as boosting ensemble methods. The basic fundamental of this method is that it relies on the fact that when all the previously generated models are combined with the next feasible ideal model, results in the minimization of the resulting prediction error. The major focus is to ensure the projected outcomes corresponding to the next model in order to ensure that the error is minimized (Natekin & Knoll, 2013).

All these methods are having their own advantages in terms of accuracy, data compatibility, predictions scores based on varied prediction and classification domains. These methods find applications in all the modern day businesses and systems that are dependent on forecasting and prediction driven by past experiential data.

In the remaining sections of this paper, the kernel tuning process has been elaborated with the help of a dataset. Section 2 explores the work relevant to current work already contributed by other researchers. In Section 3 the proposed methodology for tuning the kernel parameters has been discussed, in section 4 the implementation and machine learning toolkit used in the experimental setup has been discussed and section 5 highlights the significant outcome of the research. In our work, Support Vector Classifier has been used on a heart disease data from the open dataset to predict whether person is suffering from disease or not. Kernel parameter tuning has been performed, most of the existing research work uses the linear kernel parameters or simple SVM while in the current work, kernel parameter have been changed from linear to non-linear and its impact on the accuracy score has been analysed. The accuracy of prediction significantly increases with kernel parameter tuning.

## 2. RELATED WORK

The efforts on behalf of research community to harness the best prediction capabilities of the machine learning methods are very significant. There are a lot of supervised and unsupervised learning methods which have been applied for medical diagnostics to make prediction based on symptoms. In the current survey on existing literature we have explored the contributions from author's relevant to disease prediction and mainly we have focussed on the accuracy of prediction. In (Ranga & Rohila, 2018), the authors presented an experimental work to evaluate the outcomes observed by using five different classifiers called as SVM, KNN, decision tree, random forest and artificial neural network classifiers and couple of clustering methods namely k-means clustering and EM (Expectation- Maximization) clustering. The authors (Ahmed et al., 2020) in their work demonstrated that Decision tree attained the highest accuracy at 92.2% by incorporating unique feature selection algorithm. Random Forest displayed the maximum accuracy of 90% accompanied by specific features through Relief algorithm. Additionally, Random Forest displayed the ambient accuracy of 94.9% incorporating entire feature set. In Khourdifi & Bahaj, 2019, the authors presented usefulness of the hybrid PSO and ACO approaches, the proposed best model by, Particle Swarm Optimization ACO and FCBF obtained an accuracy score of 99.6% with RF and 99.65% with KNN. Latha & Jeeva, (2019), in their work displayed that in ensemble algorithms boosting and bagging are instrumental in prediction, bagging outperforms boosting in their work. The work inferred that with the use of bagging method, there was an enhancement in accuracy by a maximum of 6.92% while boosting enhanced it by 5.92%. The highest accuracy was attained using mass voting incorporating the feature set FS2.

Arif et al., (2017) showcased their research which is done with the prime objective to observe the significant relationships between various supervised ML algorithms, also they analysed the performance and utility pertaining to disease risk prediction. They inferred that the SVM algorithm is more effectively employed in predictive analysis compared to the Naïve Bayes algorithm due to its inherent prediction accuracies. Arif et al., (2017) suggested a newer knowledge dependent system for diseases forecasting with the incorporation of noise removal, clustering, and prediction techniques. In their work, the Classification and Regression Trees (CART) technique was used to populate the fuzzification rule base to be triggered in the knowledge-based system. The experimental set up inferences that the mix of fuzzy rule-based, CART with noise elimination and clustering techniques can be productive in diseases predictions (Nilashi et al., 2017). Dahiwade et al., (2019) introduced common disease forecasting mechanism relying on machine learning algorithm which includes K nearest neighbour and convolutional neural networks in order to classify patient related data due to the fact that in present scenario, the patients' data is expanding enormously and it needs to be processed for the identification of exact diseases taking vital inputs from the patient symptoms. They found that CNN is efficient as compared to the KNN in respect to accuracy and time. Haq et al., (2018) worked on a composite predictive and intelligent system to administer the heart disease. Three feature selection algorithms namely mRMR, Relief and LASSO were employed on the common classifiers including logistic regression, naïve bayes, decision tree, k nearest neighbour, Support vector machine, Random forest and ANN. The logistic regression classifier employing 10-fold cross-validation displayed the highest accuracy of 89%, choosing the FS algorithm. Employing the 16 hidden neurons with Relief features algorithm is instrumental in attaining the highest sensitivity of the ANN classifier. Mohan et al., (2019) employed the hybrid HRFLM method by conjoining the features of Linear Method and the Random Forest. The proposed method proved to be most accurate in the forecasting of heart disease.

Masih & Ahuja, (2019) defined three in-demand data mining algorithms Decision Table, Classification and Regression Tree, and then iteratively dichotomized the 3, generated by a rule-based classifier for optimization of heart disease forecasting systems employing a sizeable patient's record. Asl et al., (2008) used HRV signal for the classification of 15 features. Features were narrowed with the help of GDA to 5 and the precision was computed to be 100% by incorporating SVM. Table 1

summarizes some of the significant work done by researchers and provided us a strong foundation to generate the initial framework for the experimental process.

### 3. PROPOSED METHODOLOGY

Machine learning (ML) algorithms are being predominantly employed in detecting and classifying the chronic diseases like cancer, heart disease, tumour, diabetes, and several others Joachims, (2000). For the proposed system, dataset has been congregated from the UCI repository. Initially, the data set has been pre-processed. Then it has been classified using SVM. It has been observed that using the concept of parameters tuning, the performance of classifier gets enhanced considerably. Finally, classifier performance has been evaluated. In the current research work, heart diseases have been identified based on the discussed process, depending on the historical dataset of patients and support vector machine classifier (Patil, 2019). The sequential steps followed for evaluation of the kernel parameters of the proposed prediction system have been presented in Fig. 1 below.

#### 3.1 Heart Attack Prediction Data Set

Dataset for Heart Attack Prediction is taken from UCI repository for heart disease dataset which consists of various independent variables or features for prediction of the disease. As per medical industry insights, if a nation is able to manage and control the key factors like smoking, cholesterol and diabetes, the patients suffering from heart diseases can be significantly dropped close to 15 percentage (Karaolis et al., 2010).

The selected dataset comprises of a set of 76 attributes, generally the most of the researchers in their published work have considered a similar subset of 14 of them as shown in Table 2. It includes 303 patient records with each row consisting of a single patient record. Vital parameters including sex, age, trestbps, cp, cholesterol, fasting blood sugar, restecg, maximum heart rate achieved, exercise induced angine, slope, oldpeak, number of major vessels (0-3) coloured by fluoroscopy are used for prediction of a number, which is either 1 or 0 on the basis of the outcome where '1' signifies that the person has a heart attack and '0' infers that the Person is safe from heart attacks. This can be further used for prediction of the possibility that a person can suffer from heart attack.

#### 3.2 Importing The Data Set

This model has been implemented through Scikit-learn library using Python language (Varoquaux et al., 2015). First, the pandas module has been loaded and the comma separated values are read i.e. csv through read\_csv() method and that csv file is translated into a pandas DataFrame. A sample of some records is invoked using the head() method is shown in fig 2. Below.

#### 3.3 Data Exploration

Data exploration is a crucial step in every Machine Learning problem. In the above code, if some argument has been used i.e. integer number inside the head () method, to show first five records from the data set. Tail () method was invoked to show the last records that depend on the argument inside it (fig 3. below).

In order to get some information about the datatype of every attribute; this data is important for conversion of data type. For this dtypes() method has been used, it displays all the columns available and their corresponding data types also.

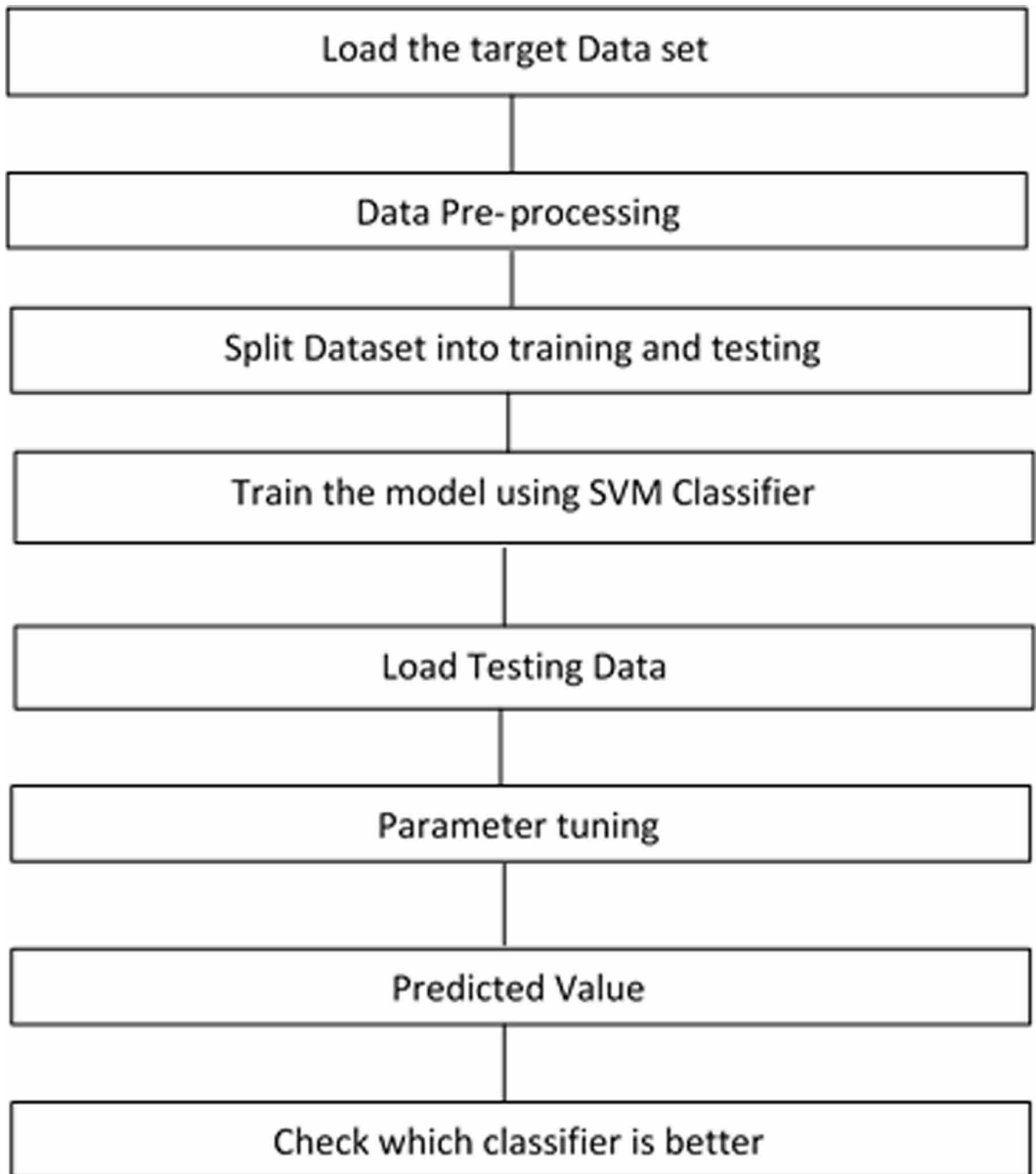
##### 3.3.1 Pre-Processing Data

In earlier sections, it has been illustrated most of the attributes in the data set are of integer datatype, and all algorithms of Scikit-learn library work only on numeric data type, there is so no need to change their data types. If the attributes present in the data set belong to string datatype then we first

Table 1. Brief summary of inspiration from the earlier research in the domain of disease identification

S.No.	Author and Year	Technique	Objective and outcome
1.	Ahmed et al., (2020)	Univariate feature selection Relief algorithm	Results displayed that Decision tree achieved accuracy of 92.2%. The Random Forest using features from Relief algorithm delivered an accuracy of 90%.
2.	Khourdifi & Bahaj, (2019)	PSO and ACO approaches	The proposed best model by incorporating Particle Swarm Optimization ACO, and FCBF obtained an accuracy score of 99.6% with RF and 99.65% with KNN
3.	Khourdifi & Bahaj, (2019)	Majority voting and bagging, stacking and boosting.	The authors have used ensemble algorithms, boosting and majority voting and bagging, stacking with which they inference that accuracy increased by 6.92% using bagging, while with boosting it improvised by 5.92%.
4.	Uddin et al., (2019)	SVM and Random Forest	They observed that the SVM algorithm is preferred over Naïve Bayes algorithm by the researchers. However, the best accuracy is obtained by Random Forest (RF) algorithm comparatively.
5.	Dahiwade et al., (2019)	CNN and KNN algorithm	They found that CNN is efficient compared to K-Nearest Neighbour with respect to prediction accuracy and processing time.
6.	Mohan et al., (2019)	HRFLM	They used hybrid HRFLM method by combining the characteristics of Linear Method and Random Forest method.
7.	Kalaiyarasi & Suguna, (2019)	SVM parameter tuning	Classification method was used for diabetic patient data analysis and the kernel parameters were tuned to obtain a significant change in the prediction accuracy.
8.	Ranga & Rohila, (2018)	K-means and EM clustering	The researchers have compared the results obtained by using five different classifiers namely Support vector machine, decision tree, random forest and ANN, KNN clustering techniques
9.	Haq et al., (2018)	Relief, mRMR, and LASSO	They proposed a composite intelligent predictive model for forecasting of heart disease. Using 16 hidden neurons with Relief features algo helpful to obtain best sensitivity of ANN classifier.
10.	Chen et al., (2017)	Nearest Neighbour, Elbow Method	Hyper parameter tuning for two parameters bandwidth and trade-off has been achieved using a two-step process involving the nearest neighbour and elbow method. The method achieves faster training and better generalizability.
11.	Nilashi et al., (2017)	Clustering, noise removal, and prediction techniques	The researchers in their experimental work evidenced that a mix of CART and fuzzy based rule and clustering methodology can be instrumental in forecasting of disease.
12.	Sentelle et al.,(2016)	SVM path	The work presents a path-following algorithm which handles semidefinite kernels without the need of method to identify singular matrices.
13.	Chaurasia, (2013)	CART and ID3 (Iterative Dichotomized) and decision table (DT)	They developed the heart disease prediction models using rule based classifier.
14.	Asl et al., (2008)	GDA and HRV	Precision was computed to be 100% by using SVM. HRV signal is used to classify features and these features can be reduced to only five using the GDA.

Figure 1. Proposed Architecture for detecting Heart disease



need to convert it in to numeric format. LabelEncoder() method in Scikit-learn library can be used for converting string data type into numeric data type.

### 3.3.2 Training Set and Test Set

Now in the next step i.e. training of parameters, all the parameter are needed to be checked, and identify that parameter which is helpful to give better output in training process (Karaolis et al., 2010). The output of training model is used with test data to identify the result. First data is segregated into the features and target variables. To select rows and column in reverse order, we may use negative number, for instance use -1. In array notation, all the selection mechanism for row will go on left

Table 2. Various Attributes of the patient data

S.No	Attribute	Description	Values
I.	Age	Age of the person	No particular range
II.	Sex	Gender (Binary value)	Male=1,Female=0
III.	Cp	Chest pain type	Asymptomatic=4,Non-angina pain=3,Atypical angina=2,Typical angina=1
IV.	Trestbps	Blood pressure count during resting(mmHg)	No particular range
V.	Chol	Amount of Serum cholesterol (mg/dl)	No particular range
VI.	Fbs	Fasting blood sugar	False=0,true=1
VII.	Restecg	Resting electrocardiographic results	Hypertrophy=2,Abnormality=1,Normal=0
VIII.	Thalach	Maximum heart rate achieved	No fixed value
IX.	Exang	Angine induced due to exercise	No=0 and Yes=1
X.	Oldpeak	ST depression persuaded through exercise relative to rest.	No fixed value
XI.	Slope	Slope of the peak exercise ST segment	Downsloping=3,Flat=2,Upsloping=1
XII.	Ca	Count of major vessels colored by Fluoroscopy	No fixed value
XIII.	Thal	3 values	Reversible defect=7, Fixed defect=6, Nomal=3
XIV.	Target	Result (Target variable)	Presence=1; Absence=0

Figure 2. Display Records using Head Method

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 3. Display records using Tail method

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0



hand side of comma. For column notation, it will go on the right hand side. If blank notation is used then it will add all the columns and rows. The `iloc` method of pandas has been used for the purpose of executing this process. Here data is split up and stored in variables X and Y. The variable X contains all the rows and all the labels but it excludes last label that is the target. The target attribute is present in variable Y with all its rows. Further, the data frame is fragmented into a training set and a testing set. While dividing the data, 80 percent has been kept for the purpose of training while the remaining 20 percent is kept for the purpose of testing. But these numbers are not fixed and this ratio will also depend on the type of data. We may vary the ratio for different experiments based on the type of data.

### 3.4 Choice of Parameters

Provided some random dataset, it is not defined in advance that which kernel may prove to be the best one. For a problem related to the linear separable dataset, linear kernel is acceptable. It does not make sense to employ the linear classifier when the data is not linearly separable, it is suggested to use an RBF kernel in such situation. Linear problems are resolved using the linear support vector machines while for non-linear problems, RBF kernel may be employed. The Support vector decision region corresponding to the RBF kernel is a linear decision area. Creating non-linear permutations of the features to strengthen the sample dataset and upgrade it to a corresponding feature space bearing higher dimensional attributes is the main task of an RBF kernel. In this space, a linear decision boundary may be used to separate the classes. A hyper parameter search is performed and different kernels are compared with others. There are several evaluation metrics including as accuracy, F1 and ROC auc etc. to evaluate the performance and infer the fitness of a particular kernel for a problem.

### 3.5 Parameter Tuning For SVM Classifier

For the purpose of disease prediction in healthcare industry the most popular technique is supervised machine learning. In the current work we have taken Support Vector Machine (SVM) algorithm to classify the absence and presence of disease. It can be used for both linear and nonlinear data. SVM takes data points and generates the two dimension hyper plane that separates these data points into different classes. The line that is used for the separation is referred to as the decision boundary. SVM determines the optimal decision boundary. During the current experimental work, we have used parameters tuning as a technique for analysing the performance of the SVM classifier.

## 4. RESULTS

The data has been split into two blocks, wherein the first block being training and other the testing block. Now SVM is needed to be trained. Scikit-Learn contains many libraries and built-in classes for different algorithms. In the current work SVC (support vector classifier) class in the Scikit-Learn's SVM library has been used. This class contains multiple parameters but in our case only one of the parameter is being used which is of the kernel type. For the purpose of predictions, `predict` method of the `svc` class has been used. The confusion matrix is a table that gives idea about the performance of our model. As we are using the supervised learning technique, we have both training as well as testing data available for analysing the outcome of the predictions given by our trained model. Based on these predictions, the confusion matrix gives a representation of all True-Positives, True-Negatives, False-Positives and False-Negatives as predicted by the model. The meaning of these outcomes as give as under:

1. TP expanded as true-positive means we predicted positive and it's true.
2. TN expanded as true-negative, means it was predicted negative and it's true.
3. FP expanded as false-positive means it was predicted positive and it's false (Type I error)
4. FN expanded as false-negative, means we predicted negative and its false (Type II error).

Confusion matrix exhibits information about the predicted and actual values of a classification system. It is used for summarizing the performance of the classification algorithm. It is evident to note that the false-negative predictions may result to be dangerous prediction in the domain of medical sciences. According to our confusion matrix, true positive value is 1 and true negative values are 31 (fig. 4) and the precision may be calculated dividing the total number of true positive values by the sum of true positives and false positives meaning that true positive / predicted positive which is 1.00. Further sensitivity or recall may be calculated by dividing total positives with the sum of total positives and false negatives meaning true positive /actual positive, which is 0.03. If we calculate precision and recall score perfectly than it will generate a perfect F-score. Accuracy is the measure of how much we have correctly measured among all the predictions made on the sample dataset under test. It may also be inferred to as the ratio of total correct predictions to the total number of predictions. Accuracy obtained without the model being tuned for kernel is computed to be .5245. The confusion matrix and performance data generated for prediction without tuning the kernel is shown (in fig. 4) below.

Performance metrics (fig 5.) may be analysed for the purpose of getting information on adjudging the best performing kernel.

Figure 4. Performance of Classification algorithm

	precision	recall	f1-score	support
0	1.00	0.03	0.06	30
1	0.52	1.00	0.68	31
micro avg	0.52	0.52	0.52	61
macro avg	0.76	0.52	0.37	61
weighted avg	0.75	0.52	0.38	61

Next confusion matrix will be generated for the condition where the kernel mode is linear, in order to explore, if there is any difference on accuracy score when the kernel mode will get changed. As illustrated in fig. 6, accuracy score increases when kernel mode gets changed to the linear mode. The performance of kernel in linear mode is given in fig.7.

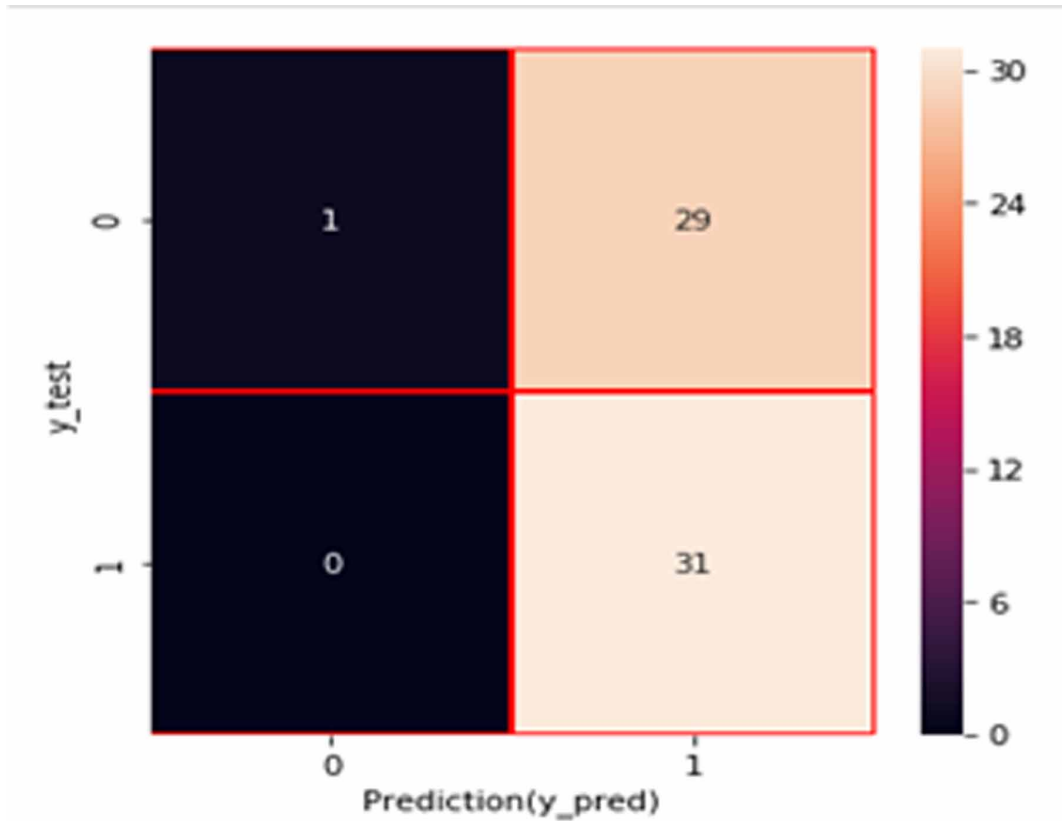
The above description shows the effect of kernel parameter variations on the classification behaviour of the SVM classifier. There is a considerable increase in the accuracy of the model when we modulate the kernel parameters and this motivates the incorporation of such methods in the future intelligent machines. Table 3 below shows a comparative analysis of our model with another similar model. There is a significant improvement in the proposed model in terms of outcome without and with tuning of kernel.

Similar approach has been used for evaluating the performance of the breast cancer data set available in the open repository available at [www.kaggle.com](http://www.kaggle.com) (*Breast Cancer Prediction Dataset Dataset Created for "AI for Social Good: Women Coders" Bootcamp*). There is an increase in the accuracy of the predictions in the breast cancer patient dataset also with the incorporation of the kernel parameter tuning. It has been observed that the prediction accuracy increased from 85% to 87% in this dataset.

## 5. CONCLUSION AND FUTURE SCOPE

To conclude, the significance of fine tuning parameters for forecasting the heart disease in support vector machine was discussed and analysed in the current research work. Data was pre-processed before

Figure 5. Performance metrics



modelling and then used for testing the prediction. To support previous statement, Prediction analysis result suggests the definitive progress in accordance with relevant study of these kernel parameters. According to our results, SVM algorithm performs better using tuning of kernel parameter. It shows that when the kernel parameter have been modified, it results in increase in the performance adding a significant impact to the result. The Goal of this study was to focus on the importance of parameter tuning to elevate the performance of classifier and we also compared the result with normal classifier

Figure 6. Performance of classification Algorithm when kernel is in linear mode

```
[[22  8]
 [ 5 26]]
```

	precision	recall	f1-score	support
0	0.81	0.73	0.77	30
1	0.76	0.84	0.80	31
micro avg	0.79	0.79	0.79	61
macro avg	0.79	0.79	0.79	61
weighted avg	0.79	0.79	0.79	61

Figure 7. Performance metrics with kernel in linear mode

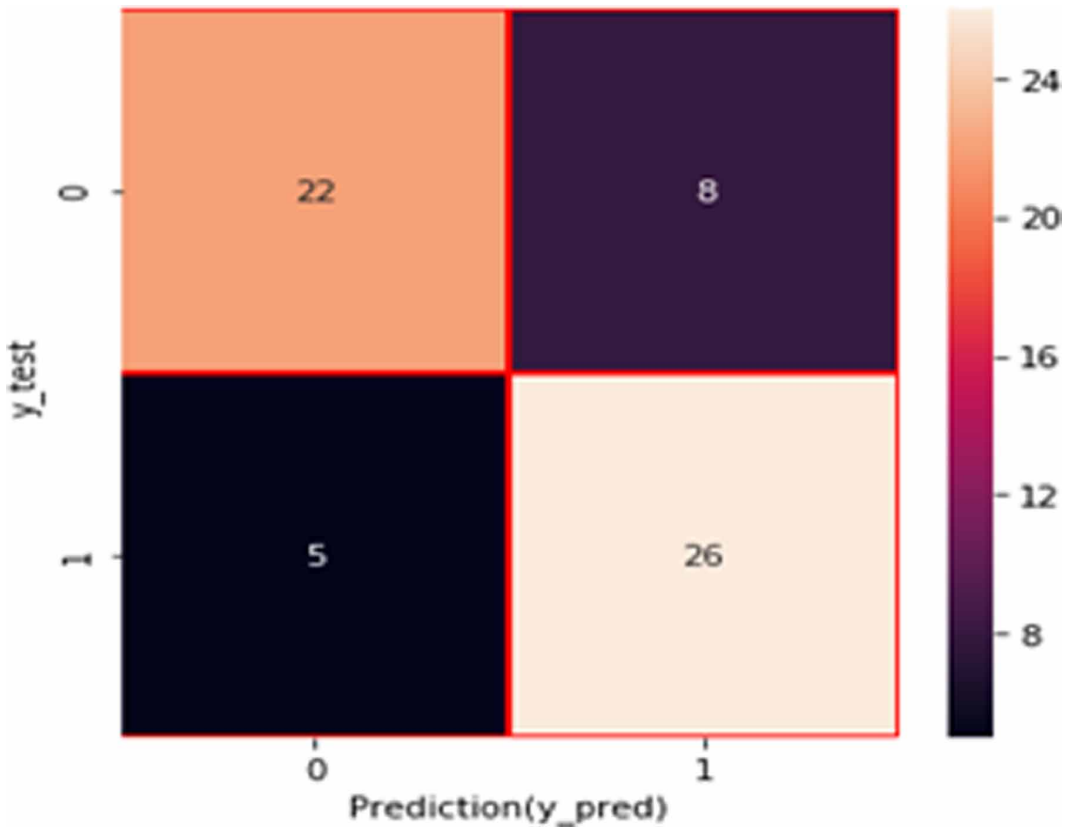


Table 3. Comparison of proposed model with a similar approach

Authors	Basic approach and Algorithm	Accuracy (%)		Deviations from the proposed work
		Before tuning parameters	After tuning parameters	
Kalaiyarasi & Suguna, (2019)	SVM	72.40	73.44	Genetic algorithms have been applied to optimize the parameters values
Goel & Srivastava, (2016)	SVM	58.21	74.37	The work has been implemented using the Matlab tool and varies in certain aspects of Kernel description
Nanda et al., (2018)	SVM	Classification error 0.17	Classification error 0.1	The work does not considers the implementation of non-parallel SVM
Amami et al. (2015)	SVM	46 (polynomial)	51.6(RBF)	The work has limitation is finding out the optimal kernel values for making an optimized prediction
Proposed Model	SVM	52	78	

SVM before tuning the parameters. Training and Testing data using SVM provided the direction to choose the correct classifier to supervise this model. Results show that the hyper parameters tuning increases the performance of this model. Along with all this, various validation metrics were also used to calculate the final results. After modification of kernel parameter the classification accuracy score is 78%, when parameter were not altered, the accuracy score was 52%. Therefore, there is significant improvement in prediction efficiency with the proposed methodology. Therefore the performance of SVM classifier is highly dependent on the kernel parameter. Overall, it's a blend of data, python, kernel tuning, classifiers and few other important parameters to get desired result.

There is still scope of improvement in sensitivity, specificity, F1 score and accuracy score. In future we will consider more parameters that affects the performance or we will create some hybrid model to improve its accuracy score.

The obtained classification accuracy after modifications of the kernel width is 84% which is the highest value reported in the literature that adopted the support vector machines approach for the Cleveland heart disease dataset.

## REFERENCES

- Ahmed, H., Younis, E. M. G., Hendawi, A., & Ali, A. A. (2020). Heart disease identification from patients' social posts, machine learning solution on Spark. *Future Generation Computer Systems, 111*, 714–722. doi:10.1016/j.future.2019.09.056
- Amami, R., Ben Ayed, D., & Ellouze, N. (2015). *Practical Selection of SVM Supervised Parameters with Different Feature Representations for Vowel Recognition*. Advance online publication. doi:10.4156/jdcta.vol7.issue9.50
- Arif, M., Alam, U., Roy, N., Holmes, S., Gangopadhyay, A., & Galik, E. (2017). *AutoCogniSys: IoT Assisted Context-Aware Automatic Cognitive Health Assessment*. Academic Press.
- Asl, B. M., Setarehdan, S. K., & Mohebbi, M. (2008). Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. *Artificial Intelligence in Medicine, 44*(1), 51–64. doi:10.1016/j.artmed.2008.04.007 PMID:18585905
- Ayatollahi, H., Gholamhosseini, L., & Salehi, M. (2019). Predicting coronary artery disease: A comparison between two data mining algorithms. *BMC Public Health, 19*(1), 1–9. doi:10.1186/s12889-019-6721-5 PMID:31035958
- Breast Cancer Prediction Dataset Dataset created for "AI for Social Good: Women Coders' Bootcamp"*. (n.d.). University of Wisconsin Hospitals, Madison.
- Chaurasia, V. (2013). Early Prediction of Heart Diseases Using Data Mining Techniques. *Caribbean Journal of Science and Technology, 1*, 208–217.
- Chen, G., Florero-Salinas, W., & Li, D. (2017). Simple, fast and accurate hyper-parameter tuning in Gaussian-kernel SVM. *Proceedings of the International Joint Conference on Neural Networks, 348–355*. doi:10.1109/IJCNN.2017.7965875
- Dahiwade, D., Patle, G., & Meshram, E. (2019). Designing disease prediction model using machine learning approach. *Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019, Iccmc*, 1211–1215. doi:10.1109/ICCMC.2019.8819782
- Goel, A., & Srivastava, S. K. (2016). Role of kernel parameters in performance evaluation of SVM. *Proceedings - 2016 2nd International Conference on Computational Intelligence and Communication Technology, CICT 2016*, 166–169. doi:10.1109/CICT.2016.40
- Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., Sun, R., & García-Magarinó, I. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems, 2018*, 1–21. Advance online publication. doi:10.1155/2018/3860146
- Jatav, S., & Sharma, V. (2018). An Algorithm for Predictive Data Mining Approach in Medical Diagnosis. *International Journal of Computer Science and Information Technologies, 10*(1), 11–20. doi:10.5121/ijcsit.2018.10102
- Joachims, T. (2000). Transductive inference for text classification using support vector machines. *Proceedings of the 20th International Conference on Machine Learning*.
- Kalaiyarasi, P., & Suguna, J. (2019). The significance of fine tuning parameters in supervised machine learning techniques for diabetic disease prediction. *International Journal of Advanced Science and Technology, 28*(17), 364–375.
- Karaolis, M. A., Moutiris, J. A., Hadjipanayi, D., & Pattichis, C. S. (2010). Assessment of the risk factors of coronary heart events based on data mining with decision trees. *IEEE Transactions on Information Technology in Biomedicine, 14*(3), 559–566. doi:10.1109/TITB.2009.2038906 PMID:20071264
- Khourdifi, Y., & Bahaj, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International Journal of Intelligent Engineering and Systems, 12*(1), 242–252. doi:10.22266/ijies2019.0228.24
- Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked, 16*(July), 100203. doi:10.1016/j.imu.2019.100203

- Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi:10.1038/nature14539 PMID:26017442
- Masih, N., & Ahuja, S. (2019). Prediction of Heart Diseases Using Data Mining Techniques. *International Journal of Big Data and Analytics in Healthcare*, *3*(2), 1–9. doi:10.4018/IJBDAH.2018070101
- Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 81542–81554. doi:10.1109/ACCESS.2019.2923707
- Nanda, M. A., Seminar, K. B., Nandika, D., & Maddu, A. (2018). A comparison study of kernel functions in the support vector machine and its application for termite detection. *Information (Switzerland)*, *9*(1), 5. Advance online publication. doi:10.3390/info9010005
- Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neuroinformatics*, *7*(DEC). Advance online publication. doi:10.3389/fnbot.2013.00021 PMID:24409142
- Nilashi, M., Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. *Computers & Chemical Engineering*, *106*, 212–223. doi:10.1016/j.compchemeng.2017.06.011
- Nordhausen, K. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition by Trevor Hastie, Robert Tibshirani, Jerome Friedman. *International Statistical Review*, *77*(3), 482–482. doi:10.1111/j.1751-5823.2009.00095\_18.x
- Parasuraman, S. K., Loudon, B. L., Lowery, C., Cameron, D., Singh, S., Schwarz, K., Gollop, N. D., Rudd, A., McKiddie, F., Phillips, J. J., Prasad, S. K., Wilson, A. M., Sen-Chowdhry, S., Clark, A., Vassiliou, V. S., Dawson, D. K., & Frenneaux, M. P. (2019). Diastolic ventricular interaction in heart failure with preserved ejection fraction. *Journal of the American Heart Association*, *8*(7), 1–11. doi:10.1161/JAHA.118.010114 PMID:30922153
- Patil, M. (2019). Prediction and Analysis of Heart Disease Using SVM Algorithm. *International Journal for Research in Applied Science and Engineering Technology*, *7*(1), 856–859. doi:10.22214/ijraset.2019.1137
- Ranga, V., & Rohila, D. (2018). Parametric analysis of heart attack prediction using machine learning techniques. *International Journal of Grid and Distributed Computing*, *11*(4), 37–48. doi:10.14257/ijgdc.2018.11.4.04
- Rumsfeld, J. S., Joynt, K. E., & Maddox, T. M. (2016). Big data analytics to improve cardiovascular care: Promise and challenges. *Nature Reviews. Cardiology*, *13*(6), 350–359. doi:10.1038/nrcardio.2016.42 PMID:27009423
- Sentelle, C. G., Anagnostopoulos, G. C., & Georgiopoulos, M. (2016). A Simple Method for Solving the SVM Regularization Path for Semidefinite Kernels. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(4), 709–722. doi:10.1109/TNNLS.2015.2427333 PMID:26011894
- Stewart, R. A. H., Held, C., Krug-Gourley, S., Waterworth, D., Stebbins, A., Chiswell, K., Hagstrom, E., Armstrong, P. W., Wallentin, L., & White, H. (2019). Cardiovascular and lifestyle risk factors and cognitive function in patients with stable coronary heart disease. *Journal of the American Heart Association*, *8*(7). Advance online publication. doi:10.1161/JAHA.118.010641 PMID:30897999
- Task, A., Members, F., Montalescot, G., France, C., Sechtem, U., Germany, C., Germany, S. A., Uk, C. A., Poland, A. B., France, T. C., Di, C., Uk, M., Germany, A. K. G., France, J. H., Germany, N. M., Opie, L. H., Africa, S., Prescott, E., Sabate, M., & Denmark, K. et al. (2013). Guidelines on the management of stable coronary artery disease. *British Journal of Cardiac Nursing*, *8*(11), 519–520. doi:10.12968/bjca.2013.8.11.519
- Tu, C., Liu, H., & Xu, B. (2017). AdaBoost typical Algorithm and its application research. *MATEC Web of Conferences*, *139*. doi:10.1051/mateconf/201713900222
- Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 1–16. doi:10.1186/s12911-019-1004-8 PMID:31864346
- Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn. *GetMobile: Mobile Computing and Communications*, *19*(1), 29–33. doi:10.1145/2786984.2786995

Winker, M. A., Ferris, L. E., Aggarwal, R., Barbour, V., Callahan, M., Cress, P. E., Habibzadeh, F., Jawad, F., Kumar, R., Laine, C., Lang, T., McGee, G., Momen, H., Rohrich, R. J., Ruiz, M. del C., Sahni, P., & Wager, E. (2015). Promoting global health: The world association of medical editors position on editors' responsibility. *The International Journal of Occupational and Environmental Medicine*, 6(3), 125–127. doi:10.15171/ijoem.2015.638 PMID:26174989

Yamashita, R., Nishio, M., Kinh, R., Do, G., & Togashi, K. (2018). *Convolutional neural networks : an overview and application in radiology*. Academic Press.

*Annu Dhankhar is working as an assistant professor in Deptt of CSE at B.M.Institute of Engineering and Technology, Sonapat. She is having interest in artificial intelligence and machine learning based projects and applications.*

*Sapna Juneja is working as a Professor in Deptt of CSE of CSE at IMS Engineering College, Ghaziabad. Earlier she was working as Professor in Department of B.M. Institute of Engineering and Technology Sonapat. She has published several papers in good and reputed journals. She is editor of books on latest technological developments. She has a total teaching experience of 16 years.*

*Abhinav Juneja is working as a Professor in Deptt of IT, KIET Group of Institutions, Ghaziabad, Earlier he was working as Professor in Deptt of CSE at B.M. Institute of Engineering and Technology Sonapat. He has published several papers in good and reputed journals. He is editor of books on latest technological developments. He has a total teaching experience of 19 years.*

*Vikram Bali is Professor and Head-Computer Science and Engineering Department at JSS Academy of Technical Education, Noida, India. He had graduated from REC, Kurukshetra – B.Tech (CSE), Post Graduation from NITTTR, Chandigarh – M.E (CSE) and doctorate (Ph.D) from Banasthali Vidyapith, Rajasthan. He has more than 20 years of rich academic experience. He has published more than 50 research papers in International Journals/Conferences and edited Books. He has authored Five text books. He has published Patent on Smart Dustbin- Sanitation & Solid-Liquid Waste Management. He is on editorial and on the review panel of many International Journals. He is life time member of IEEE, Indian Society for Technical Education (ISTE), Computer Society of India (CSI) and Institution of Engineers (IE). He was Awarded Green Thinker Z-Distinguished Educator Award 2018 for remarkable contribution in the field of Computer Science and Engineering at 3rd International Convention on Interdisciplinary Research for Sustainable Development (IRSD) at Confederation of Indian Industry (CII), Chandigarh. He has also attended Faculty Enablement programme organised by Infosys and NASSCOM. He has been the member of board of studies of different Indian Universities and member of organizing committee for various National and International Seminars/Conferences. He is working on three sponsored research projects funded by TEQIP-3 and Unnat Bharat Abhiyaan. He has written books on Fundamental of “Cyber Security and Laws”, “Software Engineering” and “Operating System”. He is reviewer to many International Journals of repute like Inderscience and IGI Global. His research interest includes Software Engineering, Cyber Security, Automata Theory, CBSS and ERP.*