

Implementation of SVM machine learning Algorithm to predict lung And Breast Cancer

M.Ramana Reddy

^a Assistant Professor, Dept. of ECE, CBIT, Gandipet, Hyderabad, India

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

Abstract: The proposed work will be implemented and developed. Related to learning algorithms the support vector machine is under supervised learning algorithm. To predict cancers like Breast and lung cancers so many statistical and Machine learning models are there, but out of all available models the super vector machine algorithm is best. Maximum edge of SVM hyperplane and edge prepared in 2 class test. To build an SVM classifier, you first need to define a kernel function. The predicted performance of A may vary. However, there are various studies investigating the characteristics of SVM predictions based on various factors. The proposed model is fully analyzing the predictive performance of SVMs and SVM sets and comparing training with large and small lung cancer datasets with 99.52% accuracy, roc 0.876, and major f 0.996%. SVMS and SVM set time..

Keywords: SVM, accuracy, ROC and F-measure, AI, ML

1. Introduction

Among various diseases, cancer poses a great danger to people around the world. According to the Indian Census, cancer deaths in India were high, causing an alarming number of about 806,000 cases by the 21st century. And it has the highest mortality rate. This is due to the limited potential for prevention, diagnosis and treatment of the disease. All types of cancer have been reported in the Indian population, including cancers of the skin, lungs, breasts, rectum, stomach, prostate, liver, cervix, esophagus, bladder, blood and mouth. The high incidence of this type of cancer is due to internal factors (genetic, mutational, hormonal, immunodeficiency) and external or environmental factors (dietary, productive, population growth, social, etc.). It can be both. Comparison of various cancer cases in India and around the world. In addition, attempts have been made to explain the main causes of cancer and how to prevent it. In addition, efforts have been made to predict the impact of rising cancer rates on the Indian economy.

1.1.Cancer scenario in India:

Based on the increasing trend of cancer patients over the last few decades, it is expected that there will be cancer patients in India by the end of 2015 and 2020. These aggregates were 390,809, 428,545 and 819,354 in 2004, respectively. The number of men and women with cancer continued to grow until 2009, with the numbers of men, women, and all cancer patients at 454,842, 507,990 and 96, 2,832, respectively. Similarly, 462,408 male and 517,378 female cancer patients were enrolled, bringing the total number of patients in 2010 to 979,786. Therefore, this number indicates that the number of cancer cases is gradually increasing over time. We also forecast the number of cancer patients in 2015 and 2020, respectively. There are different types of cancer in India, including lung cancer, breast cancer, stomach cancer, gallbladder cancer, cervical cancer, oral cancer, and different types of cancer.

1.2.Indian states and cancers:

The most affected states of India are Jam and Kashmir, Himachal Pradesh, Delhi, Uttarakhand, Rajasthan, Maharashtra, Jharkhand, West Bengal, Andhra Pradesh, Kerala, Tripla, Manipur. The state from this figure, cervical cancer is the second most common cancer in the women's population of Himachal Pradesh, Haryana, Rajasthan, Goa, Tamil Nadu and West Bengal, among women in Punjab and Andhra Pradesh. It is clear that it is the third most common cancer. -Pradesh and Uttar Pradesh. Breast cancer is the most common cancer in women in Himachal Pradesh, Delhi, Rajasthan, Nagaland and Goa, and the second most common cancer in women in Punjab, Maharashtra and Gujarat.

1.3.Cancer causes in India:

The causes of cancer in India are similar to those in other parts of the world. Chemical, biological and ecological identities are responsible for the uncontrolled and chaotic growth of (cancer) cells. In fact, carcinogens interact with normal cellular DNA under special conditions, initiating a series of complex multi-step processes that lead to uncontrolled cell or tumor growth (Carmaia, 1993). Cancer can be caused by intrinsic factors such as genetic variation, hormones and immune status, as well as environmental factors such as tobacco, diet, radiation and other infectious pathogens. Significant changes in cancer incidence due to lifestyle and diet have been reported (Hel bock et al, 1998). For example, Asians are 25 and 10 times less likely to develop prostate and breast

cancer. This may be due to the relatively simple lifestyle and safer sexual activity that every Asian uses compared to Western countries.

1.4.Preventive measures of cancer in India:

It is said that the preventive strategy that "prevention is better than treatment" is important for eradication. This approach presents important public health issues. The National Cancer Control Program (started in India from 1975 to 1976) led to the development of the Regional Cancer Center (RCC), which is much of the cancer department of the Medical Association. We support the purchase of telemedicine equipment. Cancer control programs have also been launched in the region, but this has not led to sustainable and productive interventions (National Cancer Control Program). Education should focus on the harmful effects of tobacco and should not be used any further.

Cancer prediction has long been a research topic in the fields of health and medicine. Cancer has started in the tissues of the lungs or breast [1]. These include obesity, lack of exercise, alcohol use, hormonal problems associated with ionization, early first menstruation, missing children, delayed pregnancy, and old age. The literature review used a variety of statistical and machine learning techniques to develop better predictive models such as logistic regression, linear discriminate analysis, decision trees, artificial intelligence, neural networks, nearest neighbors K, and SVMs. The proposed SVM method is superior to the corresponding method in the proposed problem [1-10].

According to the adage "prevention is better than cure," preventive strategies are essential to eradicate cancer. This approach poses a major public health challenge and represents a long-term, cost-effective way to fight cancer. The National Cancer Control Program (launched in India from 1975 to 1976) led to the creation of the Regional Cancer Center (RCC), which is a series of cancer departments in the medical school. We supported the purchase of telemedicine equipment. An anti-cancer program was also launched in the area, but it was not possible to establish sustainable and productive activities (the national anti-cancer program). Education should focus on the harmful effects of tobacco and discourage its use.

Cancer prediction has long been the subject of medical research. Cancer begins in the tissues of the lungs or breast [1]. These include obesity, lack of exercise, alcohol consumption, problems with hormone ionization, early first menstrual periods, missing children, late pregnancy, and old age. Literary studies can apply a variety of statistical and machine learning techniques such as logistic regression, linear discriminate analysis, decision trees, AI, neural networks, K-curtains, and SVMs to develop better predictive models. An Algorithm [1-10] because it performs better than related methods.

2.Proposed work:

2.1.The Support vector Machine

The Support Vector Machine (SVM), first introduced by Vapnik [30], demonstrates its effectiveness. There are many problems with pattern recognition [31]. They can provide better classification performance than many other classification methods. SVM classification performs binary classification. That is, it splits a set of training vectors, two different classes $(x_1, y_1), (x_2, y_2) \dots (x_m, y_m)$. Where $x_i \in R^d$ is d -Dimensional vector, which is a function space, $y_i \in \{-1, +1\}$ -class label. SVMs are created by injecting vectors into a new higher dimensional feature space represented by $\phi: R^d \rightarrow H^f$. Where $d < f$. Then use kernel functions to build the optimal hyperplane separation in the new function space.

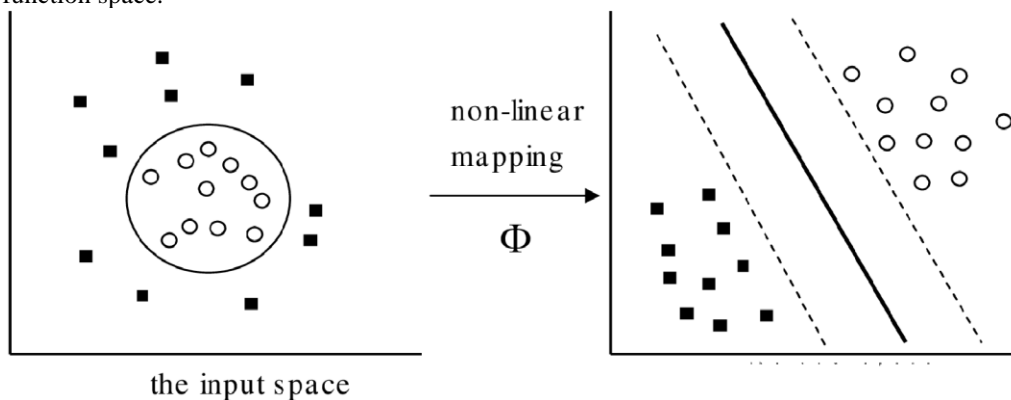


Figure1: linear kernel-based SV

Figure 1 shows a linear kernel-based SVM procedure for nonlinear displays. Enter the space into a new linear common space. In particular, all vectors on one side of the hyperplane are labeled -1, and all vectors on the other side are labeled +1. The training example closest to the hyper-plane in transformed space is called a support vector. The number of these support vectors is usually small compared to the size of the training. The set and they define the edge of the hyperplane and, therefore, the determinant. According to relevant studies, there is no formal

way to determine the best kernel function problem for a given area, but among the various kernel functions it is linear and polynomial. Kinematic base function kernels are the most widely used and compared in various fields. The Problems such as culture classification [32], gene expression classification [33], protein localization [34], speaker identification [35], compound identification [36].

$$\text{oly}(x_i, x_j) = [(x_i, x_j + 1)]^p$$

Related research shows that there is no formal way to determine the best kernel function, problems with certain domains. However, of the various kernel functions, the linear, polynomial and kinematic base function kernels are the most widely used and compared in various fields. Culture classification [32], gene expression classification [33], protein localization [34], identification dynamics [35], and identification of the splicing site [36].

2.2.Classifier ensembles

A set of classifiers that combine multiple classifiers is now considered as one. How to classify models [37–39]. This improves the classification characteristics of one classifier [26]. The concept of a set of classifiers is based on the nature of information processing by modular brains. In other words, individual functions can be subdivided into different sub processes or subtasks without mutual intervention [40]. This is an ironic principle that allows you to break a complex problem into several simpler subtasks. The task is simpler) and can be solved by a variety of training methods or algorithms.

Two commonly used methods for combining multiple classifiers are packing and lifting. In bagging, several classifiers were individually trained from different training sets using the bootstrap method [41]. Bootstrap creates k replicas of the training dataset and makes the k replicas independent. Randomly samples the specified original training dataset, but replaces them. In other words, each training example may appear to be repeated in a particular queue. Figure 1. Creating an SVM model. The SVM and SVMk training data sets either consist of breast cancer predictions or not at all. The k classifiers are then combined using the appropriate combination method, such as majority voting [42]. With improvements such as bagging, each classifier is trained to use a different training set. However, the k classifiers are not formed in parallel and independently, but in sequence. The amplification approach, which is the initial filtering impulse, was proposed by Shapir [43]. Ada Boost (or Adaptive Boosting) is currently the most commonly used accelerated learning algorithm for pattern recognition. Initially, the weight of each sample in a particular training set is the same. Use the S training set n (n < m) to train the k-class classifier as a weak training model, the k-k classifier. The trained classifier is then rated S to identify an example of this training. It cannot be classified correctly. In addition, the k + 1 classifier is trained in corrective learning. Together, this reinforces the importance of this misclassified example. This sampling procedure was repeated until a training sample K was created to build the classifier K. The inclusion of the final solution is based on the weighted value of each classifier [44].

Algorithms Used:

2.3.Support Vector Machines (Svm):

In machine learning, a support vector machine is a supervised learning model with relevant learning algorithms that analyze data for regression and classification. The figure shows the hyperplane and the maximum edges of the SVM prepared in the two test classes

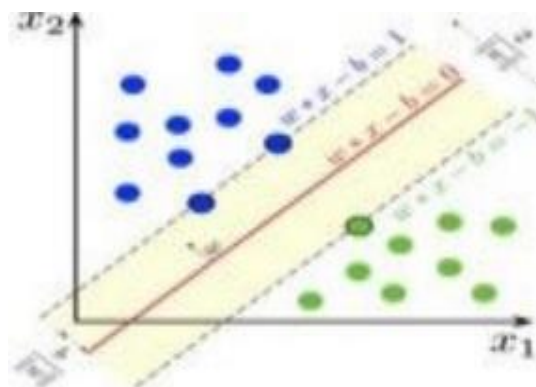


Figure.2 : SVM Hyperplane Graphs

3.Results:

3.1.Confusion Matrix Of Svm:

In the field of machine learning, especially in statistical classification problems, the confusion matrix (also known as the error matrix) [9] is a specific table layout that allows you to:Algorithm performance visualization

Each row of the matrix represents an instance of the expected class, and each column represents an instance of the actual class (or vice versa) [9]. The name comes from the fact that it's easy to see if the system is confusing the two classes (that is, one is commonly mislabeled as the other). This is a special type of contingency table, with the same group of two dimensions ("real" and "expected") and "classes" in both dimensions (each combination of dimensions and classes is a contingency table). It is a variable of).

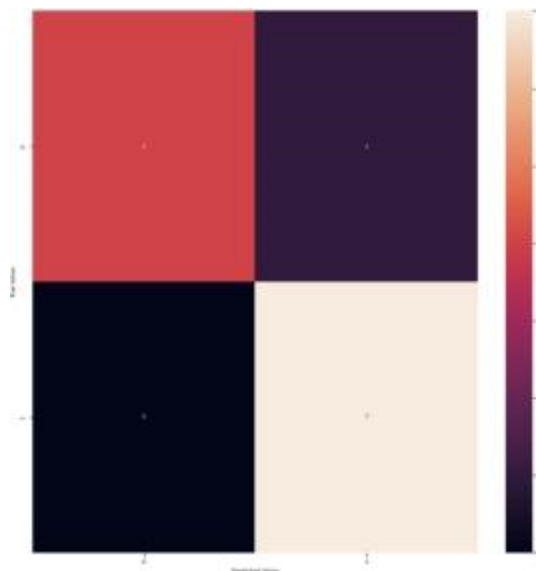


Figure .3 : Confusion matrix of SVM SVM gives an output with 100% accuracy.

4. Experimental Methodology

4.1. Experimental Procedure

The experimental procedure is based on the following steps. First, a specific data set is divided into 10-90% training and 10% testing based on a cross-validation strategy [48]. The second step is to create a multicore SVM classifier. Individual functions (linear, polynomial, RBF, etc.). In addition, a set of SVM classifiers. It is also constructed by compression and amplification to create linear, polynomial, and RBF SVM sets. Finally, the test suite is entered into a classifier created before the F test for classification accuracy, ROC, and measurement rate. In addition, the classifier learning time is also compared to analyze the complexity of the classifier learning computation. Also check if atypical features are excluded when selecting features. By using the selected dataset, you can improve the performance of the classifier compared to a non-functional classifier. Selection, in this case a genetic algorithm (GA) is used [49].

4.2. Experimental setup

Data sets: This article uses two breast cancer data sets. UCI Machine Learning Repository (available at <http://archive.ics.uci.edu/ml/>) and ACM SIGKDD Cup 2008 (available at <http://www.sigkdd.org/kddcup/index.php>). A relatively small data set of 699 data samples contained in each data sample with 11 different characteristics. In contrast, the last data set contains 10,2294 data samples. Each data sample is represented by 117 different characteristics and is considered important. The dataset for this article. The Vault project uses Weka data mining software to create various SVM classifiers. In addition to the basic features chosen to design a specific SVM classifier, other related options are based on Weka's defaults. The same approach is used to create SVM packages and extended suites. Therefore, there are three simple SVM classifiers. Linear SVM and polynomial SVM. Additionally, there are ROC and F classes for evaluating the performance of various SVM classifiers in addition to classification accuracy. It also compares the speed and time it takes to train each classifier. Pay attention to the calculation. This environment is based on a PC with an Intel i7-2600 processor at 3.40 GHz and 4 GB of RAM. SVM and SVMWeka, tuned to predict breast cancer, use Weka to perform feature selection tasks using genetic algorithms. This setting is based on the default

4.3. Experimental result

A simple SVM classifier (Figures 4 and 5) shows linear, polynomial and kernel RBFs with and without functions selected in terms of classification precision. The ROC, F, and estimated time (in seconds) for two sets of data, after performing the selection of characteristics using a genetic algorithm. The features selected from small

and large data sets are 10 and 36 respectively. As you can see performing feature selection before training SVM classifier allows feature selection. Is significantly improved performance (high classification accuracy, ROC, F, etc.). In particular, the best performances are obtained with linear GA + SVM with classification precision (96.85%), linear GA + SVM with ROC (0.967) and GA + RBFSVM (0.988) in F major. In addition, there is no significant performance difference between GA + Linear SVM and GA + RBF. SVM, In addition, the computing time for learning the SVM classifier after performing the feature selection is greatly reduced compared to the basic SVM classifier without it.

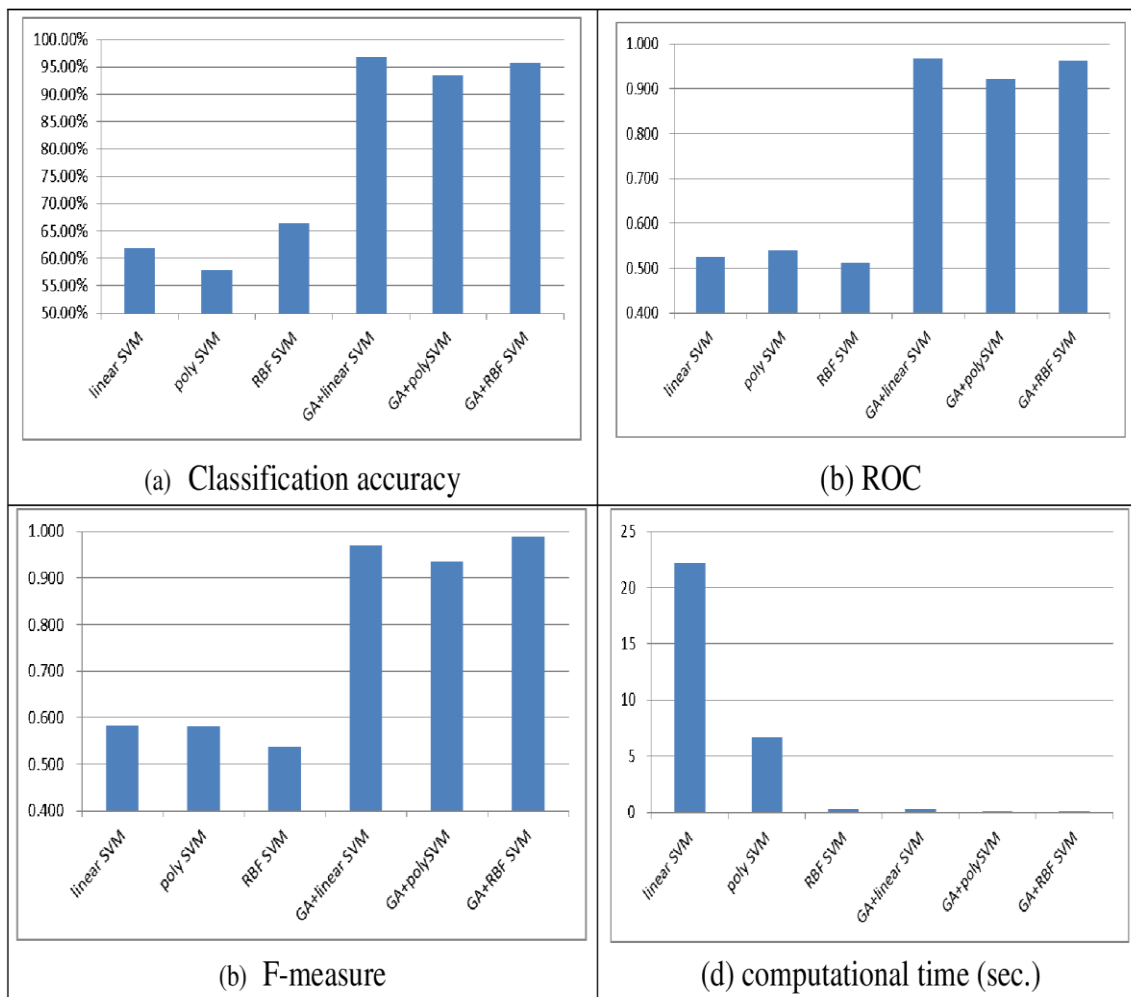


Figure 4. Performance of a single SVM classifier on a small dataset. (A) Classification accuracy, (B) ROC, (C) F measurement, (D) Calculation time (seconds)

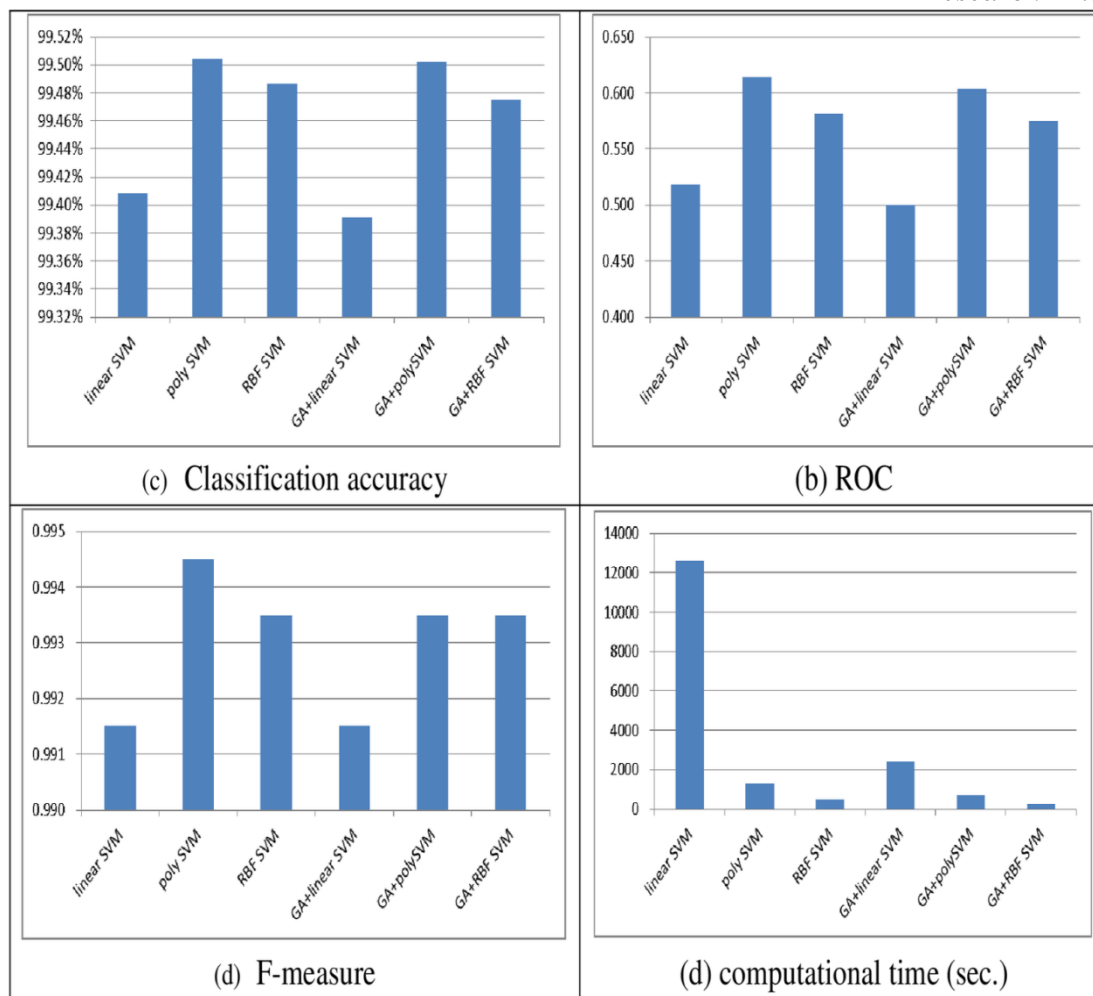


Figure .5: Performance of individual SVM classifiers on large datasets.(A) Classification accuracy, (B) ROC, (C) F-measure, (D) Computational time (sec.)

Comparison of training time to RBF SVM training. The longest calculation time is the shortest and the SVM polygon is the second longest. This is necessary for the linear SVM classifier. For large data sets, selecting objects doesn't always create an SVM. Classifiers work better than classifiers without feature selection. Especially for classification accuracy (99.50%), poly SVM and GA + poly SVM provide the best performance. Poly SVM ROC (0.614) and F major (0.994). Similar to the results obtained with the SVM classifier, it works the same in terms of F classification and measurement accuracy with or without feature selection based on small data sets, polynomials, and RBF kernel functions. Specifically, the difference in performance was 0.02% and 0.001 for F. Accuracy of the baseline classification, but considering ROC as a score metric, poly SVM far outperformed other SVM classifiers. Comparison of the calculation time of the SVM classifier training shows that the linear SVM classifier without function selection requires maximum calculation time. It will take time to learn how to select objects from other classifiers, SVMpoly and RBF. SVM is about half the size of the unselected features. However, from our point of view, the 11 and 20 minutes spent on GA + poly SVM are not much different. In particular, the numbers released by Poly SVM when this new forecast model improves performance in terms of classification accuracy, ROC, and F.

In short, GA + RBF SVM and poly SVM have improved performance in terms of classification accuracy, making them suitable for both small and large data sets. The ROC and F major do not require significant training time from the evaluators.

4.4.Collection of SVM classifiers

Figure 6-9 shows the performance of the linear, poly, and RBF SVM classifier result sets. Gi the simple SVM classification (When using two datasets for a small dataset such as Figure 4 (see Figure 4) a) -4 (c), classification accuracy perspective, ROC, mean F, and time calculations Specific characteristics are selected from) Devices using SVM sequences GA performs better than the SVM suite that does not select features. In particular, the GA + RBF SVM set using the extended method was the most efficient in terms of classification accuracy (98.28%),

while the GA + linear SVM and GA + poly SVM sets using the packaging method were another set of classifiers. .. is better than (0.98). GA + SVM linear, the maximum measurement speed F (0.966) is obtained by the expansion method and the pressure method. Generating an SVM dataset from a data set that has been trimmed since the implementation of GA can significantly reduce compute time. In particular, less is needed to build an RBFS VM. The linear results and training times using the original dataset are similar to the previous results (see Figure 4 (d)).

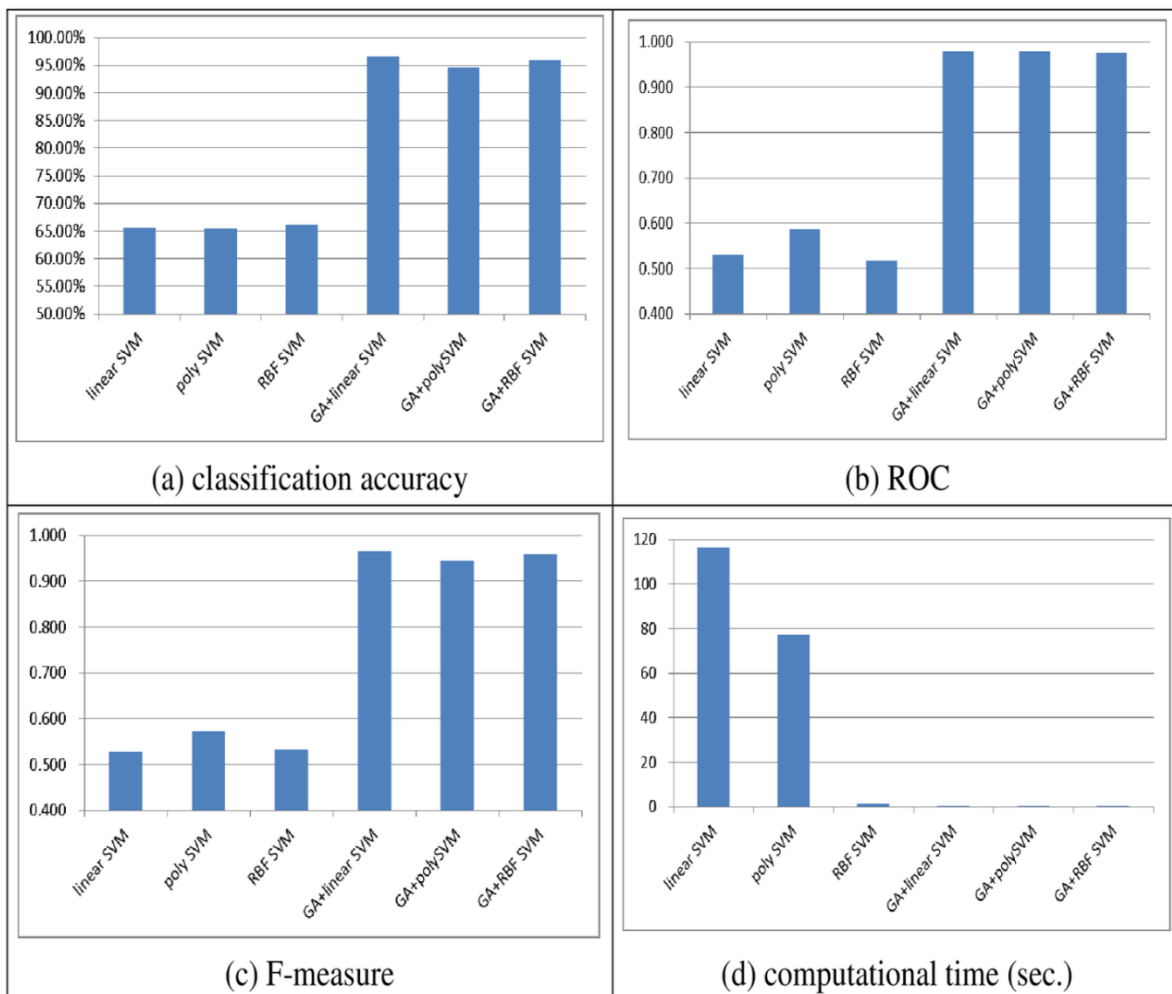


Figure 6. SVM classifier set performance based on reduced dataset bagging. (A) ClassificationAccuracy, (B) ROC, (C) F measure, (D) Calculation time (seconds)

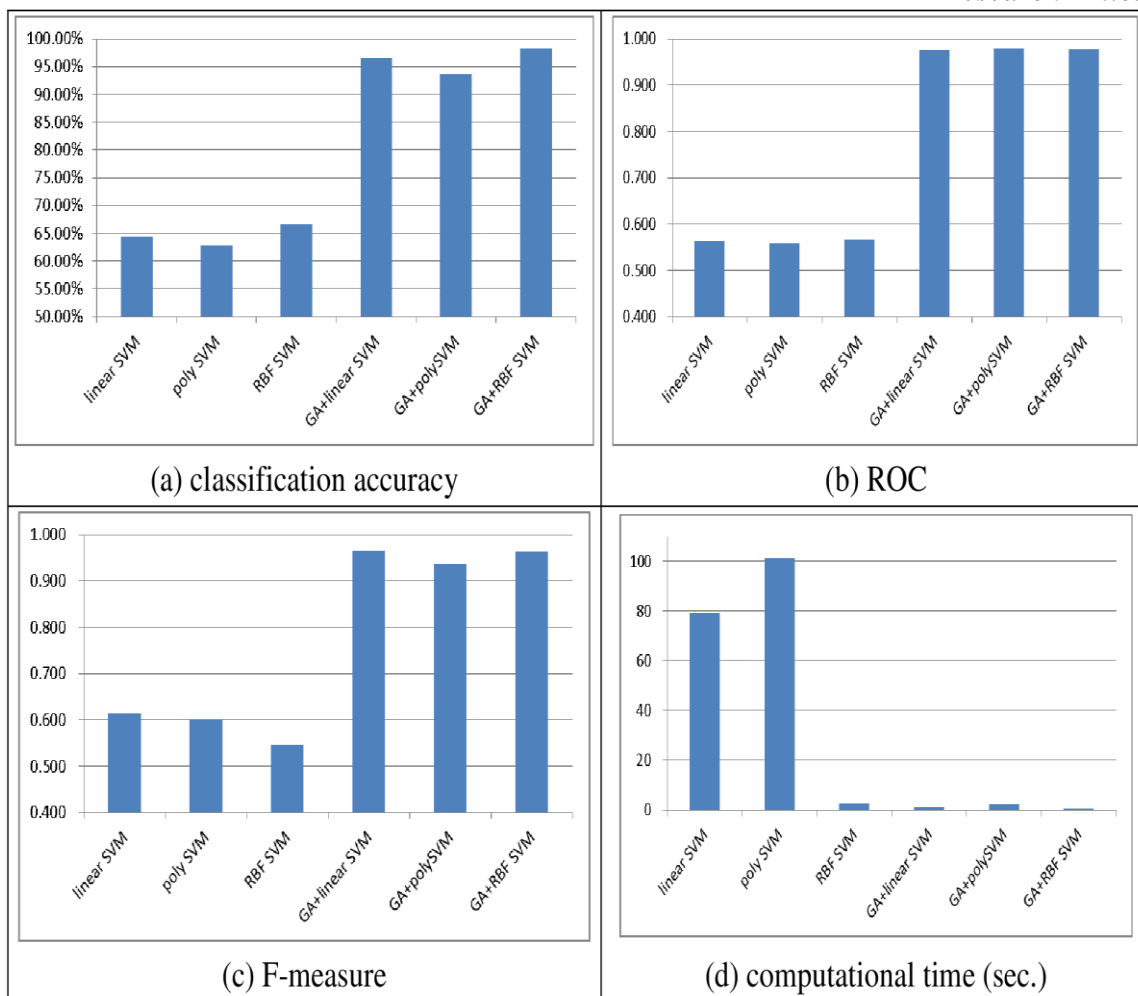


Figure 7. Amplification-based SVM classifier performance adds to small datasets. (A) Classification accuracy, (B) ROC, (C) F-measure, (D) Computational time (sec.)

For large datasets, performing feature selection does not make SVM any better than no feature selection. Specifically, classification accuracy (99.52%), ROC (0.876), and F were measured using the SVM RBF ensemble and enhancement method, ensemble and enhancement method. Expansion of linear SVM GA + and SVM RBF ensemble and improvement method. (0.995). Do. SVM kits using the amplification method require more training than SVM kits using the packaging method. However, RBF SVMs take less time than other SVMs.

5. Discussion

There is no best classifier of performance across all rating scales. Table 2 shows the three main classifiers based on classification accuracy, ROC and size comparison. Usually you will find that FSVM suites perform better than SVM alone. These results are consistent with those of related studies (Kittler et al., 1998). Update kits can be considered the best classifiers for various indicators. However, for large datasets, only the extended RBF + SVM set is included. Three lists in three different sizes.

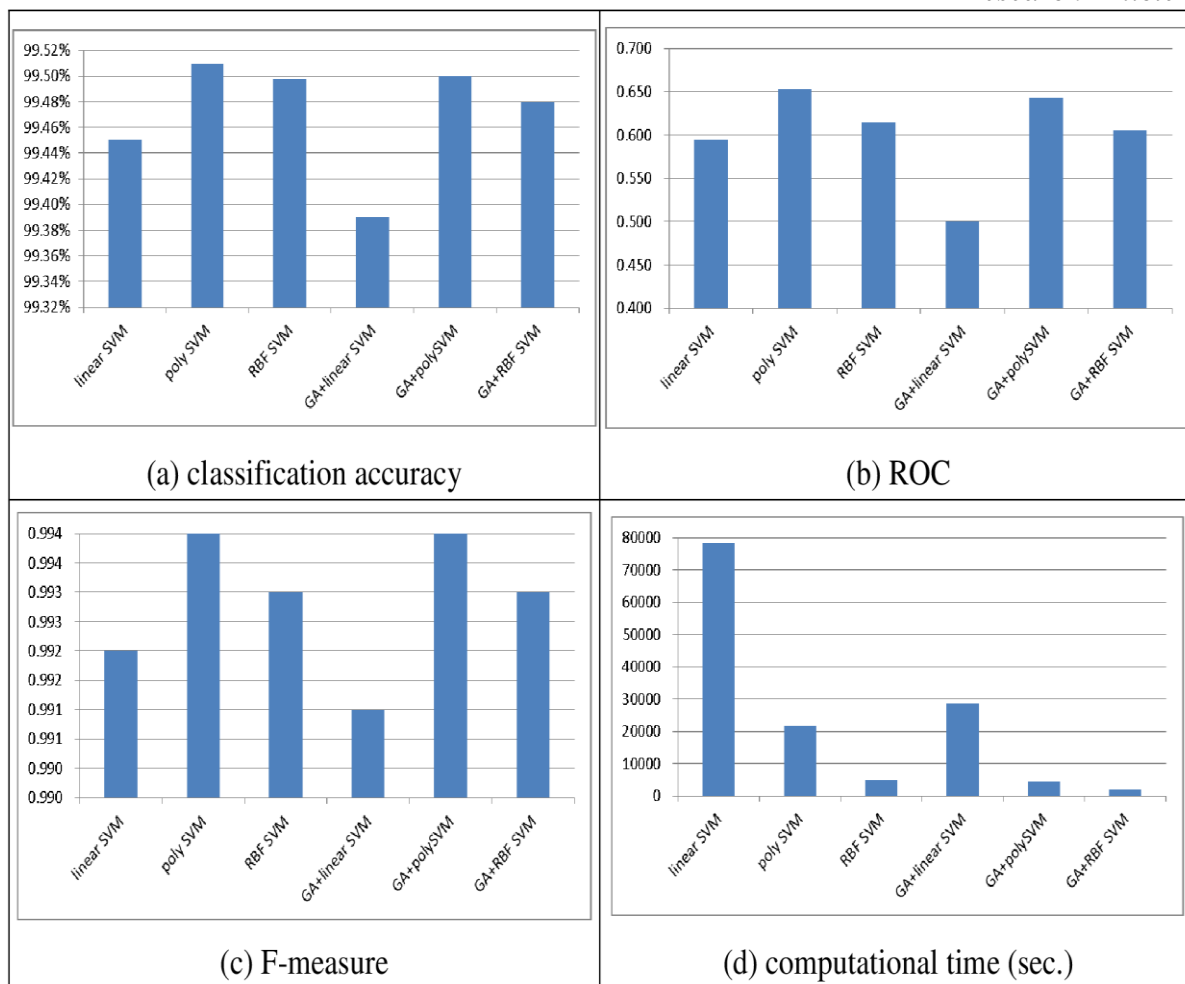


Figure 8. The performance of baggage-based SVM classifiers covers large data sets. (A) Classification accuracy, (B) ROC, (C) F-measure, (D) Computational time (sec.)

Comparing the calculated time to the training time of the linear SVM GA +, the usage of the package is similar to the usage of the SVM GA + RBF set. A small dataset is used (for example, 0.57 vs 0.5). Getting the best classifier for a large dataset takes about 301 hours with a loaded RBF SVM build. This is required to assemble an SVM with cabin 724 and an SVM 724 bag with bagging (65 hours). Therefore, when working with large datasets, both predictive characteristics and training time classifications are considered at the same time. SVMGA + RBF set is recommended for expansion. In fact, this is the accuracy of the classification, and the ROC and F environments are provided at 99.41%, 0.875, and 0.994, respectively. It also takes about 186 hours. The small cab was installed by SVMRBF. However, the cloud platform is used in the same way as the map reduction calculation implemented in Hadoop (available at the following URL): <https://hadoop.apache.org/>, you can always reduce the computing load. In this case, the boot-tuned RBF SVM is ideal for predicting milk.model. That is, these two datasets contain a limited number of characteristics and a large amount of data. These results are as simple as milk or up to 11 pairs. It is 699 for small datasets, compared to 102,294 for large datasets. This is two additional

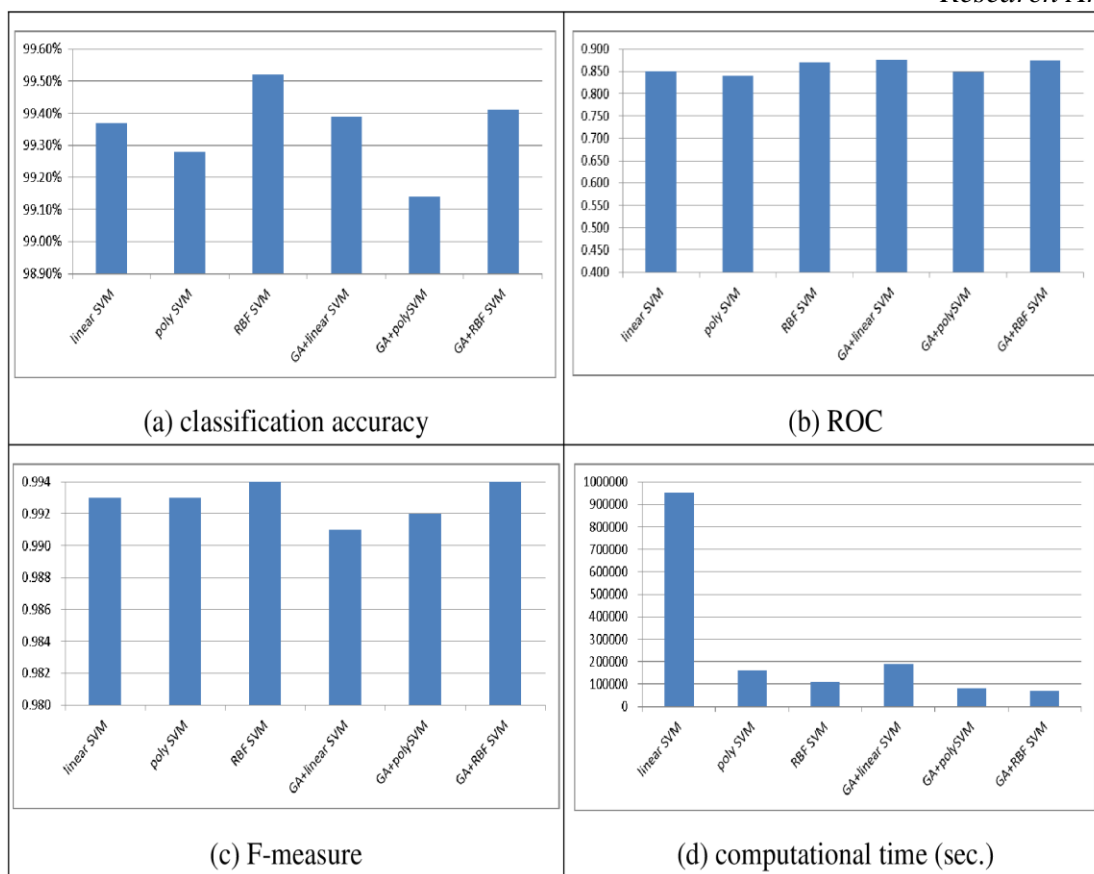


Figure 9. The performance of enhancement-based SVM classifiers covers large data sets.. (A) Classification accuracy, (B) ROC, (C) F-measure, (D) Computational time (sec.)

Table 2. Comparison of the classification accuracy, ROC, and F-measure of the top 3 classifiers.

Small scale data set:

NO	Classification accuracy	S.ROC	F-measure
1	GA+RBF SVM ensembles (boosting) (98.28%)	GA+linear/poly SVM ensembles (bagging) (0.98)	GA+RBF SVM (0.988)
2	GA+linear SVM (96.85%)	GA+poly SVM ensembles (boosting) (0.979)	GA+linear SVM ensembles (bagging/boosting) (0.966)
3	GA+linear SVM ensembles (bagging/boosting) (96.57%)	GA+RBF SVM ensembles (boosting) (0.977)	GA+RBF SVM ensembles (boosting) (0.963)
Large scale data set			
1	RBF SVM ensembles (boosting) (99.52%)	GA+linear SVM ensembles (boosting) (0.876)	RBF SVM ensembles (boosting) (0.995)
2	Poly SVM ensembles (bagging) (99.51%)	GA+RBF SVM ensembles (boosting) (0.875)	Poly SVM; poly SVM ensembles (bagging); GA+poly SVM ensembles (bagging); GA+RBF SVM ensembles (boosting) (0.994)
3	Poly SVM; GA+poly SVM; RBF SVM ensembles (bagging); GA+poly SVM ensembles (bagging) (99.50%)	RBF SVM ensembles (boosting) (0.869)	

6. Conclusion And Future Enhancement

If you carefully read the above discussion in this article, you can clearly see that the number of cancer patients in India is increasing every year. The different factors involved in the development of cancer are discussed and must be controlled to eradicate them. India is a growing country and plays an important role in the development of the world. This issue therefore deserves special attention. Governments and NGOs are to launch various programs to sensitize the Indian population. Indians should be aware of these facts because diet and lifestyle are important factors in controlling the spread of cancer. In short, cancer undermines a country's growing economy, which can be saved with the right treatment. In light of these facts, it is very important to root out this turmoil. Let us pray for a better future for this country, necessary for the development of the whole world. In this paper, the performance of a single SVM classifier and a set of SVM classifiers provided information on the use of various kernel functions and various combinations of breast cancer predictions. In addition, two sets of data with different scales are used for comparison.

In addition, the classification accuracy, ROC, F measurements, and training calculation time are compared. Various classifiers. These specific experimental parameters have never been shown or tested. The results provide a comprehensive view of predictive performance when using SVM and SVM together. And the best predictive model can be defined as the basic classifier for the future. I am learning. Most SVM suites are slightly better than simple SVM classifiers. In particular, a small genetic algorithm (GA) is used for the selection of characteristics. Scaled datasets can dramatically improve SVM classifiers and SVM datasets. Outperforms the performance of the same classifier without feature selection. Among them are GA + linear. The first two are the SVM bagged kit and the amplified SVM GA + RBF kit. The difference between a predictive model and its performance is not important. However, for large data sets, a prediction model based on RBFSVM is used. However, acceleration-based SVM sets generally take longer to train than simple SVM classifiers or wrapped SVM sets. In practice, there are two possible solutions to reduce the computation time. First select a characteristic to reduce the size of the dataset. In this case, an assembly based on the GA + SVM amplifier outperforms many other classifiers. The second is expanding to the cloud and directly creating SVM sets. Arcene test details and classification of SVMs and single SVMs in the Micromass SVM Linear Poly dataset Poly SVM RBF MArchene Single 0.8850.890.56 Bagging 0.90.880.560.89 Rise 0.8950.56 Micromass Single dataset 0.7860.6480.105 Bagging 0.7720 0.63 90.105 increase 0.7690. 6240.105 SVM and SVM tuned for breast cancer prediction. In this case, it is not necessary to make a function selection while learning the storage. You can save even more time.

References

1. Abegunde D, Mathers C, Adam T, Ortegón M, Strong K(2007) The burden and costs of chronic diseases in low income and middle-income countries. *The Lancet* 370,1929-38 Alabaster O (1972) *Colorectal Cancer: Epidemiology Risks and prevention*. JP Lippincott, Philadelphia.
2. Ali I, Rahis-ud-din, Saleem k, Aboul-Enein HY, RatherMA (2011) *Social Aspects of Cancer Genesis*. *CancerTherapy* 8, 6-14
3. Hachesu, P.R., Mofitian, N., Dehghani, M., Soltani,T.S.: Analyzing a lung cancer patient dataset with the focus on predicting survival rate one year after thoracic surgery. *Asian Pacific J. Cancer Prevention:APJCP* 18(6), 1531 (2017)
4. Kadir, T., Gleeson, F.: Lung cancer prediction using machine learning and advanced imaging techniques. *Transl. Lung Cancer Res.* 7(3), 304 (2018)
5. Kourou, K., Exarchos, T.P., Exarchos, K.P.,Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. *Comput. Struc. Biotechnol. J.* 13, 8–17 (2015)
6. Lynch, C.M., et al.: Prediction of lung cancer patient survival via supervised machine learning classification techniques. *Int. J. Med. Inform.* 108,1–8 (2017)
7. Murty, N.R., Babu, M.P.: A critical study of classification algorithms for lung cancer disease detection and diagnosis. *Int. J. Comput. Intell. Res.*13(5), 1041–1048 (2017)
8. Shanthi, S., Rajkumar, N.: Lung cancer prediction using stochastic diffusion search (sds) based feature selection and machine learning methods. *NeuralProcess. Lett.* 1, 1–14 (2020)
9. Sidey-Gibbons, J.A., Sidey-Gibbons, C.J.: *Machine Learning in medicine: a practical introduction*. *BMC Med. Res. Methodol.* 19(1), 64 (2019)