Journal of Autonomous Intelligence (2024) Volume 7 Issue 3
doi: 10.32629/jai.v7i3.799

ORIGINAL RESEARCH ARTICLE

Hybrid GA-mSVM: Dimensionality reduction using hybrid genetic algorithm and modified support vector machine classifier

Ashish Kumar Rastogi[1,*], Swapnesh Taterh[1], Billakurthi Suresh Kumar[2]

[1] *Amity Institute of Information Technology, Amity University, Rajasthan 300202, India*

[2] *School of Computer Science and Engineering, Sanjay Ghodawat University, Kolhapur 222001, India*

*** Corresponding author:** Ashish Kumar Rastogi, ashish.k.rastogi24@gmail.com

ABSTRACT

The expansion of technology results in the generation of enormous amounts of data in all areas. The researchers face a difficult challenge when they attempt to categorize these high-dimensional data. A method called as feature selection is used to reduce the high-dimensionality of the data. It is possible to consider selecting features as an issue of global combinatorial optimization in the field of machine learning. This minimizes on the total number of features, gets rid of data that is insignificant, noisy, or duplicative, and ultimately achieves an acceptable level of classification accuracy. For the purpose of dimensionality reduction, a unique approach known as the Hybrid Genetic Algorithm – modified Support Vector Machine Classifier (Hybrid GA-mSVM) is presented in this study. The genetic algorithm component conducts a search using the principles underlying the evolutionary process in order to find the best feature set and after that the trimmed dataset is provided to the SVMs. The results of the experiments reveal that the proposed approach efficiently minimizes the features and achieves a better classification accuracy than other feature selection methods.

Keywords: dimensionality reduction; evolutionary algorithm; machine learning; hybrid GA-mSVM

ARTICLE INFO

Received: 29 July 2023
Accepted: 4 September 2023
Available online: 27 December 2023

1. Introduction

These days, an incredible variety of information in many forms is being collected. The advancement of information technology is accelerating at a rapid rate. Moreover, the data alone does not contain sufficient information, and there are a lot of irrelevant or inconsistent data that need to be identified. On the contrary hand, it will make it more difficult to make decisions. A method known as feature selection (FS) involves selecting the attributes that are pertinent while deleting the attributes that are irrelevant or repetitive[1–5]. When it comes to the construction of predictive models, FS selection has three primary advantages: it enhances the model's comprehensibility, it reduces the amount of time spent in training, and it increases generalization by minimizing overfitting. In the fields of data mining, pattern classification and feature integration, FS is among the most essential data preprocessing techniques.

While still preserving an acceptable level of classification accuracy, the goal of the feature selection procedure is to decrease the quantity of superfluous features[6]. The expense of measuring features can be decreased by using an effective approach for selecting features, which also improves the effectiveness and precision of categorization[7,8]. The process of selecting features is of critical

significance in many application areas, including pattern recognition, data analysis, and the extraction of information about multimedia content, healthcare data processing, Machine Learning (ML), and data mining[9–11]. In the past, feature selection on testing and training data has been accomplished through the use of a variety of techniques, including genetic algorithms, linear search algorithms, data fusion. It is necessary to have a more refined approach for selecting features in order to get a greater classification accuracy while dealing with classification issues.

According to Yin et al.[12] the method has proposed for best feature selection of GA based evolutionary algorithm. When compared to other relevant optimization-based search methodologies or evolutionary computation processes do not require any prior knowledge of the domain and it does not work on assumptions about the search space, including whether it is sequentially or nonlinearly distinguishable and discrete[13]. An additional significant benefit of EC processes is that their population-based processes can generate multiple solutions during the course of a single run. The heuristically approach that ECs take in the context of a directed stochastic search gives the impression that they could be useful in the area of attribute selection. The both techniques elaborate a double point crossover selection method and random mutation for fitness selection. The 50% selection criteria have applied and elitism-based techniques has utilized for selection of best features in each iteration.

The Genetic Algorithm (GA), is a method of feature selection that is used in evolutionary computing and is recognized for its high adaptability and efficiency. In its most basic form, it is a technique for searching that is derived from natural genomics and natural selection. It is a technique of heuristic search that is population-based and statistical, and it is an optimization technique. Iterative processes are used in a GA to manipulate one population of chromosomes in order to establish a new population by making use of genetic features such as crossover, which is the pairing of 2 individual chromosomes, and mutation (i.e., random modification of the chromosomes).

Fundamentally, the learning algorithms that are utilized in the construction of the classification algorithm can be split into two distinct types: supervised learning method and unsupervised learning algorithm. The labelled data are utilized in the supervised learning method, which leads to the development of models that are known as classifiers. The unsupervised learning process takes in data that has not been labelled and builds a model that may be used for prediction. Clustering is a good illustration of a case like this one. In feature selection methods, the supervised learning technique is frequently used to assess the efficiency of the feature subsets that are being considered for selection. In certain circumstances, the supervised learning method is the one that is implemented to judge how well the feature selection techniques perform in terms of the correctness of their classifications. Support vector machines (SVMs) are a group of closely related supervised learning techniques that can be utilized for regression analysis and classification. Recent research has shown that support vector machines (SVMs) are generally capable of providing superior results in terms of accuracy rate than other data classification techniques, such as statistical approaches, decision tree-based methodologies, and instance-based learning techniques.

Therefore, it would be helpful to examine whether or not ECs and SVMs may be coupled in an effective manner to build a good classifier that is enabled by attribute selection. In order to complete the tasks of attribute selection and data categorization using a hybrid strategy, genetic algorithms and support vector machines have been combined as part of the hybrid technique. There are two primary steps involved in the process of this hybrid approach. In the first step, the selection of an attribute set is done using GA. After that, these characteristics are given to the SVM classifier so that it can calculate a fitness measurement for each attribute that was established during the second phase of the process. After that, these fitness values are utilized in the process of selecting the best selection of features based on GA. In addition, enhancements were made by removing unfit individuals from an already existent population. The goal was to bring the overall

fitness level of the population up to a higher average and achieve better results. As a result, the Hybrid GA-mSVM technique is proposed in this research forthe purpose of Dimensionality Reduction (DR).

The remaining portions of the paper are structured as follows: The existing study of dimensionality reduction is discussed in Section II. The proposed framework of Hybrid GA-mSVM for dimensionality reduction is discussed in Section III. The empirical findings and discussions are illustrated in Section IV. Finally, conclusion and future scope of this proposed method is discussed in Section V.

2. Literature survey

In high-dimensional input, the "curse of dimensionality" makes it impossible to record the irregularity of data points in complete data space, as stated by Li et al.[14]. In order to address this issue, a model for the identification of outliers that is dependent on variational autoencoder and the evolutionary algorithms has been suggested. This model is intended for the subspace outlier analysis of large-dimensional data. In order to perform an initial screening for anomalies, the variational autoencoder (VAE) is built using the Variational Autoencoder with Genetic Algorithm (VAGA) model that was proposed. After that, the genetic algorithm is put into action in order to explore the anomalous subspace of the outliers that were acquired by the VAE layer in order to establish a foundation for the subspace process of analyzing. After that, grouping the aberrant subspaces helps screen out the false positives, which are then recirculated to the VAE layer so that network weights can be adjusted. The comparative experimental studies that were carried out on three different public benchmark datasets revealed that the proposed VAGA model produced outlier detection performance that were highly interpretable and had good accuracy effectiveness than the methods that are currently considered to be state-of-the-art in the field.

A comparative study was carried out by Zhong et al.[15] to evaluate and discuss the efficacy of dimensionality reduction strategies in supporting general problem solving for high-dimensional sparse representation issues. For the sake of comparison and explanation, three well-known DR methods have been chosen: the Maximal information coefficient, Pearson Correlation Coefficient and the Principal Component Analysis. The findings of the experiments demonstrated that taking into account correlation alone during DR is not efficient enough to produce a suitably reduced set of issue dimensions, and so that GP with DR may operate less well than its counterpart that does not use DR. In the meantime, an innovative two-phase DR technique that takes into account both correlation and duplication has been developed. The proposed method has the potential to generate a set of finite dimensions that is more rational, which, in turn, has the potential to effectively enhance the effectiveness ofGP on HDSR situations.

Sivaranjani et al.[16] used the Machine Learning methods Support Vector Machine (SVM) & Random Forest (RF) to assist determine the possible chances of being impacted by Diabetes Related Diseases. These algorithms were able to help determine the possible chances of being affected by Diabetes Related Illness. After the data have been preprocessed, the characteristics that have an effect on the prediction are chosen using step forward as well as backward feature selection. Following the selection of particular features, the Principal Component Analysis dimensionality reduction technique is analyzed. The correctness of the prediction is found to be 83% when incorporating Random Forest, which is important when compared to the SVM which has an accuracy of 81.4%.

According to Li et al.[17] when working with high-dimensional data, it makes it harder to identify outliers in the complete data space. In this paper, an outlier identification and anomalous subspace search approach that is focused on autoencoder with genetic algorithm for high-dimensional data is suggested to ease this problem. In order to do this, the study was written. This research makes use of neural networks in order to construct an autoencoder. This autoencoder compels the mapping of high-dimensional input into a low-dimensional domain. In order to accomplish unsupervised outlier identification, the compacted data

must first be decoded and then rebuilt. The error that occurs between the source data and the recreated data must then be evaluated to a threshold value. After that, a genetic algorithm is used to seek the subsets of high-dimensional anomalies. The subdomain with the greatest degree of outlier is produced as the abnormal subspace of the identified point. This is done by encoding the subsets of the outliers and then using the revised fitness value to determine the amount of outlier of the identified subspace. The peculiar characteristics of the subspace can serve as a foundation for conducting an investigation into the factors that give rise to outliers. The results of the experiments show that the suggested technique, in contrast to previous competitive methods, is able to efficiently identify anomalies in high- dimensional data and retrieve the correct anomalous subspace of outliers.

Raia et al.[18] aims to find a solution for the problem of Model Order Reduction (MOR) for permanent motor synchronous machines (PMSM) by employing artificial neural networks and ML approaches for the purpose of reducing the dimensionality of the data. The neural networks are trained with the use of data obtained from a number of electromagnetic finite element analyses (FEA), which were carried out according to the requirements given by the means of data dimensionality reduction. The workflow that is suggested to develop the PMSM MOR begins with the creation of the data, continues on to the post-processing of the data, and then concludes with the training of the model and the clinical validation of the model. In this study, a process known as data dimension reduction is carried out in order to improve the efficiency of the computing while simultaneously preserving the accuracy of the model. Comparisons are made between various data reduction strategies with regard to the amount of computational resources they require and how simple they are to implement. The findings that were obtained are evaluated to the results that were achieved using FEA in an effort to find the optimal solution for constructing the dynamic model. The resultant read-only memory (ROM) is included into a real-time control indicate a strong in order to evaluate the performance of the machine. The accuracy of the model as well as its applicability are demonstrated by a comparison analysis using simulated dataand empirical studies.

The research conducted by Yaswanthram and Sabarish[19] investigates the effect that dimensionality reduction has on the effectiveness and precision of machine learning techniques used for face recognition. The investigation is carried out through the utilization of a number of distinct techniques, such as RF, SVM, K-nearest neighbour and Logistic Regression. When executed without Principal Component Analysis (PCA), Logistic Regression accomplished an accuracy score of 0.97 within a time of 5.74 seconds; when executed with PCA, Logistic Regression accomplished an accuracy score of 0.93 within a time of 0.15 seconds. Based on the findings of the analysis, Logistic Regression provides better results in terms of accuracy and time. While there is a substantial difference in the amount of time it takes to compute (about 20 times), there is not much of a variation in the accuracy.

Islam et al.[20] provided two distinct methods namely Term Presence Count (TPC) and Term Presence Ratio (TPR) to eliminate the redundant and irrelevant features in both negatively and positively tagged articles with almost equal distribution. Both of these methods may be found in the referenced article. For the purpose of sentiment classification utilizing the film critic dataset, the following four machine learning-based classification techniques were applied: Logistic Regression (LR), Support Vector Machine, Random Forest (RF) and Naïve Bayes (NB). At last, the classifiers are assessed according to their accuracy, specificity, recall, and mean F- measure values. According to the findings of the experiments, the feature dimension can be lowered by around 83 % using the proposed strategy while maintaining or even boosting classification accuracy.

The goal of the work by Siddique et al.[21] is to incorporate the Deep Autoencoder as well as Neighborhood Components Analysis (NCA) dimensionality reduction methodologies in Matlab and to examine the implementation of these techniques on nine different datasets taken from the UCI ML repository.

In the first place, the dimensions of these datasets have been cut down to 50% of their initial dimensions through the application of the Deep Autoencoder as well as NCA dimensionality methodologies. This was achieved by choosing and retrieving the features or attributes that are the most pertinent and suitable. After that, every dataset is categorized using the classification methods of K-Nearest Neighbors (KNN), SVM and Extended Nearest Neighbors (ENN). Matlab is used as the primary platform for the development of all classification techniques. Within every classification, the ratio of training data to test data is consistently set at 90%:10%. During the classification process, the variation in accuracies is noticed and studied. This is done in order to determine the degree of compatibility amongst each technique for reducing dimensionality and every classifier, as well as to assess the effectiveness of every classifier using every dataset.

Using the publicly released Cardiotocography (CTG) dataset from the California State University and Irvine ML Repository, Reddy et al.[22] analyses two of the prevalent methods for reducing dimensionality namely Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) on four popular ML methodologies, namely Decision Tree Induction, SVM, NB and RF Classifier. The findings of the experiments demonstrate that PCA performs better than LDA in all of the measures. Additionally, applying PCA and LDA does not have a significant impact on the performance of the classifiers that were studied. In order to conduct additional research on PCA and LDA's effectiveness. The experimentation is conducted on datasets from both diabetic retinopathy (DR) and the intrusion detection system (IDS). The findings of the experiments demonstrate that machine learning algorithms that use PCA give superior results whenever the dimension of the datasets is large. It has been discovered that machine learning techniques that do not use dimensionality reduction get superior results when the dimensionality of the datasets is minimal.

According to Cutura et al.[23], Dimensionality Reduction, sometimes known as DR, is a well-known method that is frequently used in the fields of Machine Learning and Visual analytics to examine high-dimensional data. The methodology has been shown to be effective through extensive empirical testing in locating previously hidden structures in data sets. It was by chance that the artistic beauty of the DR procedure was uncovered when observing the outcomes of the intermediate optimizing steps of DR techniques. It has opted to look at DR from the perspective of generative art rather than from the perspective of their technical application aspects, and it will apply DR techniques to make artwork. This decision was motivated by passion for the beauty. In specific, the optimization technique was used to produce images by attempting to draw each initial stage of the optimization procedure with some opaqueness over the prior transitional result. This was done by placing every intermediate stage of the optimization procedure over the prior intermediate result. In order to apply DR to images while still preserving some of the facial attributes of the subject, a neural-network model is employed for face landmark identification. This results in abstracted facial representations. Another potential input is a neural-network model.

Dewangan et al.[24] recommended a technique that utilizes three different machine learning techniques in order to recognize the Code Smells. The two datasets at the class level and the two datasets at the function level were used for this reason. In this body of research work, a technique called PCA, which is based on the reduction of dimensionality, is utilized to choose several components from every dataset. Principal component analysis oriented Logistic regression (PCA-LR), Principal component analysis oriented Random forest (PCA-RF), Principal component analysis oriented K-nearest neighbour (PCA-KNN), and Principal component analysis oriented Decision tree (PCA-DT) are the four PCA-based ML techniques algorithms that are utilised. The accuracy of the model is checked using a process known as 10-fold cross-validation. Throughout the course of this investigation, it was discovered that the PCA LR model provides the highest accuracy, 99.97%, for the data class dataset. On the other hand, the PCA KNN model provided the lowest accuracy, 77.38%, for the lengthy parameter list dataset.

Chen and Omote[25] describe a strategy for protecting personal information that makes advantage of

dimensionality reduction and is difficult to undo while preserving the extensive value of the data. The key premise of the method is that it is possible to perform accurate data analysis while maintaining the user's privacy by integrating dimensionality reduction techniques with the insertion of noise. In addition, the efficiency and safety of the approach that has been proposed are analyzed, which demonstrates the usefulness of the method that has been suggested.

In the context of classification, Fournier and Aloise[26] demonstrate that PCA is still an applicable method for dimensionality reduction. In order to achieve this goal, the effectiveness of PCA is assessed by contrasting it with that of Isomap, a deep autoencoder, and a VAE. Experiments were carried out on the CIFAR, MNIST and Fashion-MNIST image datasets, all of which are very popular and widely utilized. In order to successfully project the facts into a low-dimensional space, a variety of dimensionality reduction strategies were applied to every dataset in their own unique manner. After that, a k-NN classifier was developed on every projection by doing a cross-validated probabilistic approach across the whole number of neighbours. It was interesting to learn from the studies that k-NN obtained equivalent accuracy on PCA, and that the projections produced by both autoencoders offered a large enough dimension. Nevertheless, the time required for PCA computation was approximately two orders of magnitude less than that required by its neural network equivalents.

An enhanced PCA technique known as EW-PCA was proposed by Yumeng and colleagues[27] by incorporating the idea of similarity measure from cognitive science and integrating it with the idea of the entropy weight method. initially, an information gain threshold is established for the purpose of feature screening. Subsequently, the idea of weighted estimated value is presented in order to enhance the procedure of data centralization. In conclusion, the entropy weight is implemented to enhance the primary component in order to maximize the effectiveness of the dimensionality reduction procedure. In order to make predictions and conduct research based on the analyzed data set, the KNN and SVM algorithms are utilized. The modified EW-PCA algorithm has a greater effect on dimension reduction, and it has a higher accuracy in prediction than the conventional PCA technique.

Using a variety of machine learning methods, including KNN, AdaBoost, SVM, LR, RF, Artificial Neural Networks, and NB, Varunram et al.[28] built several IDS. The Synthetic Minority Oversampling Method is implemented in order to rectify the class imbalance that existed in the dataset and was brought about by the inherent characteristics of network analysis. After that, the dataset is simplified by applying three distinct methods for reducing dimensionality including principal component analysis, t-sne, and unsupervised multidimensional scaling. The three datasets that are produced by these methods are then utilized by the algorithms for machine learning to construct the intrusion detection system. The algorithms will perform binary classification to categorize the future attacks between normal activities and DDoS Attacks.

3. Research methodology

The framework of the hybrid GA-mSVM for dimensionality reduction is illustrated in **Figure 1**. While mSVMs are utilized by the induction algorithm, the evaluation for candidate attribute sets is carried out with the aid of hybrid GA. After going through data preprocessing, the data set is then examined for any missing values that may have been present all along. The first step is to randomly initialize a population of 20 chromosomes that will later be used to indicate the attribute subsets. After that, the size of the data collection is reduce based on these attribute sets. The mSVM classifier is then used to do a ten-fold cross validation on the data before it is sent on. Following this procedure, the degree of precision in chromosomal classification that was achieved is connected as an indicator of the individual's level of fitness. The hybrid model has successfully completed one generation as a result of this approach.

Following each iteration, the algorithm would then examine two criteria for terminating the process. In

the first step, the process of evolution can be interrupted if convergence is reached, which is the condition that occurs when all of the chromosomes in the population have the same levels of fitness. This indicates that the ideal combination of qualities has been reached at this point. Convergence is not, however, guaranteed in every situation that arises. In this particular scenario, the algorithm must be terminated at some stage. Before the beginning of the process, the user decides on the utmost number of generations that the technique is allowed to complete before the process is terminated. The user makes a choice on this particular parameter, which serves as the basis for the second criterion. If the maximal number of generations passes without convergence being obtained, the algorithmwill terminate.

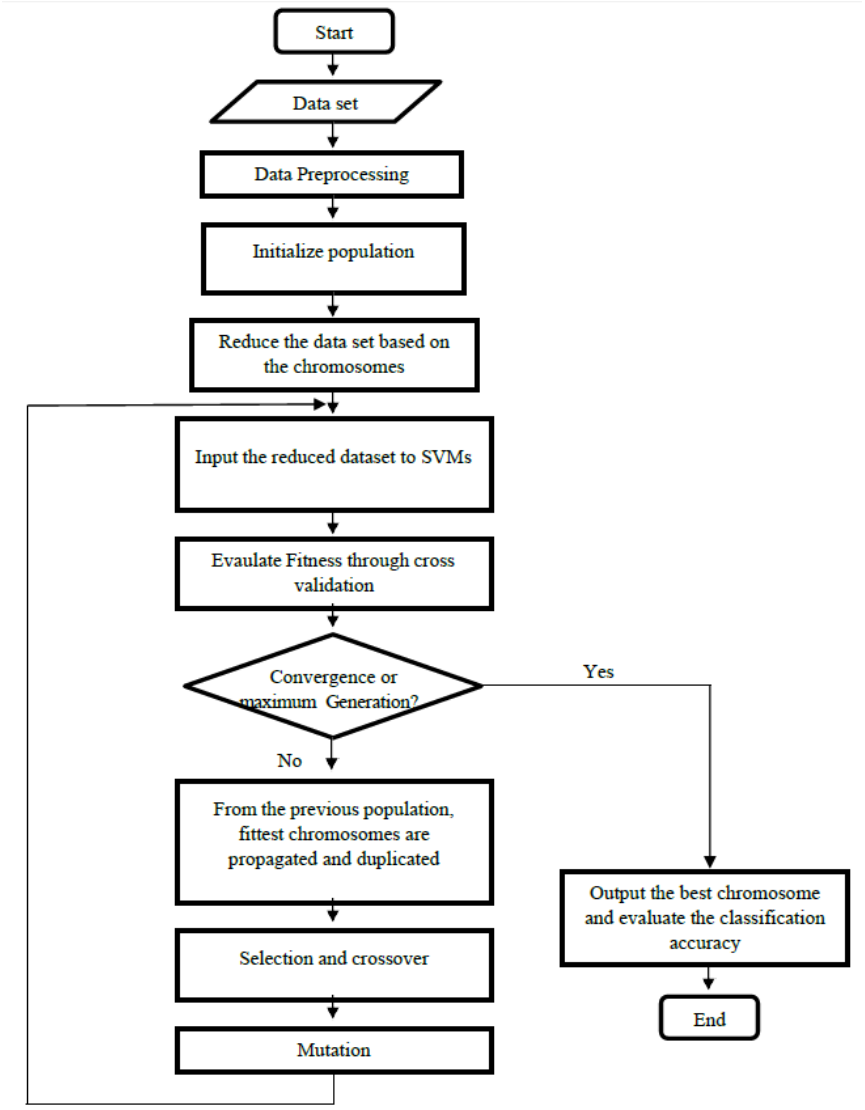

Figure 1. The flowchart of proposed hybrid GA-mSVM for dimensionality reduction.

In the event that these two criteria are not satisfied, the population of chromosomes will subsequently be subjected to the processes of selection, crossover, and mutation that are specified by the hybrid GA. This leads to the emergence of a population that has evolved to a new state. Reducing the data sets depending on the feature sets that it depicts and running them through mSVMs is a process that is repeated in order to determine the fitness levels of the chromosomes that will be used in the subsequent generation. After then, the hybrid GA process is repeated in an iterative manner until either of the requirements is satisfied. At the end of the process, the set of characteristics that is best is the one that has the maximum classification accuracy while also having the fewest number of attributes.

4. Result and discussion

In this study, a heterogeneous experimental setup has been done to evaluate the proposed work. The open-source environment has been used with JDK 1.8 with NetBeans 8.0. The 2.7 GHz processor has been used with 8 GB RAM.

The proposed GA-SVM hybrid is validated with 4 benchmark datasets that are obtained from the UCI Machine Learning Repository. The proposed model relies on n-fold cross validation to acquire the classification accuracy. The data is first randomized to ensure that the different classes of data tuples are evenly spread out to obtain an unbiased validation of the hybrid. Cross-validation then divides the data set into n different portions and the average classification accuracy for the n iterations is acquired. **Table 1** lists the parameter settings in the hybrid GA-mSVM that are applied to the 4 data sets such as KDDCUP99, NSLKDD, ISCX and BOTNET, we given details in **Table 2**. These parameters were chosen based on some groundwork experiments. In-depth analysis and trials have to be carried out to find the optimal set of parameters.

Table 1. Parameters of hybrid GA-mSVM algorithm.

Parameters	Value
Chromosome Length	No. of features
Population size	20
Maximum number of iterations	50
Probability of crossover	0.9
Probability of mutation	0.01
Number of folds of cross-validation	10
Crossover type	Uniform
Mutation type	Uniform

The level of difficulty of the problems posed by the data sets is illustrated in **Table 2**.

Table 2. Complexity of problem of 4 datasets.

Dataset	Features	No. of class	No. of instances	Percentage of majority of class
KDD	10	4	5000	32%
NSL-KDD	10	5	5000	57%
ISCX	12	2	1500	80%
BOTNET	10	2	2500	82%

The proposed hybrid GA-mSVM was tested by means of the benchmark data sets, and the classification accuracies were determined after a total of 50 iterations across all of the data sets. The 10-fold cross validation approach was utilized in order to determine the classification accuracies. **Table 3** provides a summary of the highest, lowest, and average levels of accuracy for each of the data sets.

Table 3. Classification Results of Hybrid GA-mSVM.

Hybrid GA-mSVM	Iris	Heart disease	Hepatitis	Breast EW
Max accuracy	97.22%	84.21%	88.53%	90.25%
Min accuracy	95.36%	82.49%	83.41%	85.13%
Mean accuracy	96.29%	83.35%	85.97%	87.69%
Standard deviation	0.56	1.45	1.73	1.8
Best feature set	1, 2, 3	2, 4, 7, 8, 11, 12, 13	2, 5, 6, 8, 9, 11, 13, 15, 17, 18, 19	2, 3, 4, 6, 9, 10, 11, 14, 15, 17, 19, 18, 21, 23, 25, 27, 29, 30
Total number of feature set	4	14	20	30

The Iris dataset has very high classification accuracies. One possible reason for this is that the three classes are completely unique from one another. It would appear that the data set on heart disease had the least accurate measurements among the four different data sets. It has been discovered that the heart disease data set has many classes; as a result, the classification problem has been given an additional layer of complexity. As a result, the robustness of the hybrid GA-mSVM in the multi-class domain is demonstrated by a mean accuracy of 83.35% forheart disease data set.

The standard deviations of the classification accuracies were also obtained. The consistent performance of the classifier is demonstrated by the relatively minimal standard deviations that are displayed. Table 3 also provides an illustration of the attribute set that provides the best fit for the data sets. These are the features that were chosen based on their correlation to achieving the highest possible classification accuracy for every data set. It is feasible to acquire extremely positive classification accuracy while still working with a smaller data set if the eliminated features are not considered in the analysis.

Initially, we started the execution of GA with ten iterations, but it generated a redundant feature vector after stopping GA. On the other hand, we have given 100 as the maximum iterations, and then it creates the best feature vector but requires high computation. In the entire experimental analysis, we find that 50 is the optimal iteration that takes an average time for execution while generating an effective feature vector.

In order to evaluate how well the further feature selection component functions, the hybrid GA-mSVM model was put through its paces against the SVM, the mSVM, and the GA-SVM. In this scenario, using a conventional SVM indicates that there was no feature selection performed on the data sets. In a similar fashion, the best and mean classification accuracies that were achieved were based on the results of 50 iterations and 10-fold cross validation. **Table 4** displays the tabulated findings of the study.

Based on the data stated in the table, the conclusion can be obtained that it is convenience for good classification to get rid of duplicate attributes. After feature selection has been completed, there is always a noticeable rise in the overall classification precision. This demonstrates that the accuracy of the data sets has improved because the classification algorithm is now able to accurately categorize a greater percentage of the test data despite having a smaller data set to work with. After the feature selection process is complete, there is also an improvement shown in the population's best accuracy. The additional features, some of which were redundant, might have led the classifier misled, which would have resulted in the incorrect classification of the test data. A smaller number of features would mean that future data could be compiled more quickly. In this scenario, just the most relevant attributes would need to be taken into consideration. The one and only disadvantage of performing feature selection is the additional time it necessitates to select the features that make up the most desirable profile. In addition, since GA is based on randomness, it does not necessarily guarantee that the same optimal collection of features will be chosen in each and every iteration.

Table 4. Comparative analysis of accuracy between SVM, mSVM, GA-SVM and hybrid GA-mSVM.

Data sets	Mean accuracy of SVM	Best accuracy of SVM	Mean accuracy of mSVM	Best accuracy of mSVM	Mean accuracy of GA-mSVM	Best accuracy of GA-mSVM	Mean accuracy of hybrid GA-mSVM	Best Accuracy of Hybrid GA-mSVM	Best set of features
Iris	93.34%	95.24%	93.99%	94.13%	94.13%	96.13%	96.29%	97.22%	1,2,3
Heart Disease	80.46%	82.57%	81.22%	83.67%	82.24%	83.47%	83.35%	84.21%	2,4,7,8,11,12,13
Hepatitis	85.49%	88.31%	85.90%	89.00%	83.55%	87.39%	85.97%	88.53%	2,5,6,8,9,11,13,15,17,18,19
Breast-EW	85.25%	87.46%	88.02%	88.05%	86.38%	89.12%	87.69%	90.25%	2,3,4,6,9,10,11,14,15,17,19,18,21,23,25,27,29,30

Consider the below **Figure 2**, which illustrates the comparative analysis of accuracy between conventional SVM, modified SVM, GA-SVM proposed GA-mSVM. It is observed that the accuracy rate of proposed GA-mSVM is better as compared to other models.

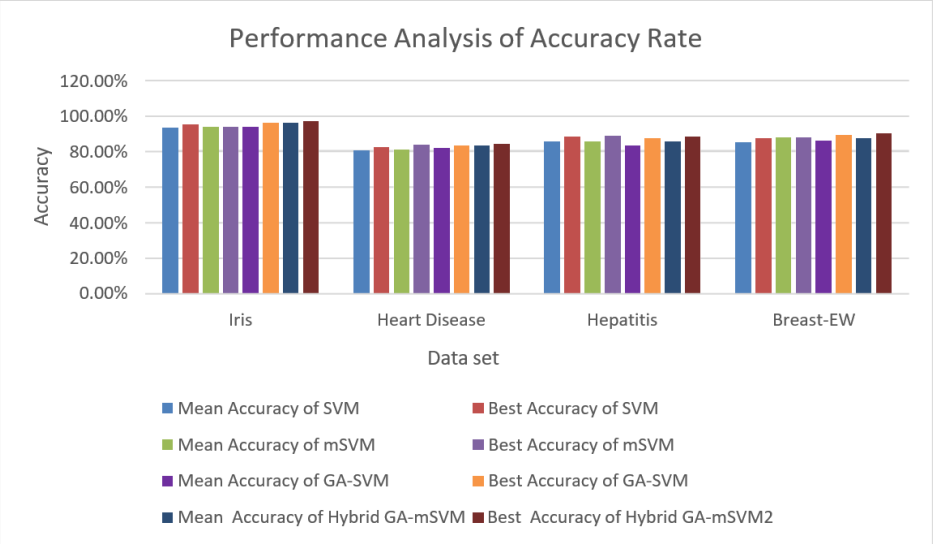

Figure 2. Performance analysis of accuracy rate between SVM, mSVM, GA-SVM and hybrid GA-mSVM.

The capacity of a classifier to make accurate classifications is the factor that should be given the most weight. As a consequence of this, the objective of the subsequent stage of development is to enhance the classification capabilities of the hybrid GA-mSVM. There are primarily two ways that have been identified as having a high probability of improving the categorization accuracy of the suggested method. Firstly, the overall fitness of the population as a whole can be improved in the aim of increasing the probability of future generations having child chromosomes that include improved genetic information. Removing the unfit chromosomes in a population and substituting them with better chromosomes could result in an improvement in the population's overall level of fitness. By performing this extra step, it is not possible to predict that the optimal state will be reached in a shorter amount of time, but it is also possible to increase the algorithm's overall level of effectiveness. Secondly, modifying the simulation's design parameters is still another technique that can be utilized to enhance the reliability of the classification. On the other hand, this approach is extremely laborious and fraught with a certain amount of risk. This is the case because to the fact that the settings of some parameters could be more adaptable to different types of data sets, which can result in improved accuracy for certain data sets while decreasing accuracy for others. In addition, there is no certain method for fine-tuning the parameters. The process of tuning must be accomplished through a combination

of trial and error as well as significant testing.

5. Conclusion and future scope

In this research, a hybrid evolutionary algorithm for feature selection and machine mSVM for classification has been developed for attribute selection in the context of data mining. The research has described three phases for dimensionality reduction using feature extraction and selection methods. In the first phase, GA generates the best subset using evolutionary techniques, while mSVM is used to classify entire test data. The selected features are used as training rules during the machine learning classifiers. This approach is very effective when dealing with ensemble and hybrid machine learning methods. The above approach is also very effective when it deals with unstructured data for selecting essential feature sets. Applying hybrid feature selection methods on large text, numeric and categorial datasets and evaluating with machine learning and deep learning algorithms will be the future task for this system.

Author contributions

Conceptualisation, AKR and ST; methodology, AKR; validation, BSK and AKR; formal analysis, ST and BSK; investigation, ST and BSK; resources, AKR; data curation, BSK and AKR; writing—original draft preparation, AKR; validation, ST and BSK; writing—review and editing, AKR; visualisation, ST and AKR; supervision, ST and BSK. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Rajeswari S, Josephine MS, Jeyabalaraja V. Dimension reduction: A PSO-PCNN optimization approach for attribute selection in high-dimensional medical database. In: Proceedings of the 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI); 21–22 September 2017; Chennai, India. pp. 2306–2309.
2. Siddique MAB, Sakib S, Rahman MA. Performance analysis of deep autoencoder and NCA dimensionality reduction techniques with KNN, ENN and SVM classifiers; In: Proceedings of the 2019 2nd International Conference on Innovation in Engineering and Technology (ICIET); 23–24 December 2019; Dhaka, Bangladesh; pp. 1–6.
3. Obeidat I, Eleisah W, Magableh K. Dimensionality reduction and supervised learning for intrusion detection. In: Proceedings of the 2022 8th International Conference on Information Management (ICIM); 25–27 March 2022; Cambridge, United Kingdom. pp. 86–91.
4. Othman AA, Hasan TM, Hasoon SO. Impact of dimensionality reduction on the accuracy of data classification. In: Proceedings of the 2020 3rd International Conference on Engineering Technology and its Applications (IICETA); 6–7 September 2020; Najaf, Iraq. pp. 128–133.
5. Stiawan D, Arifin MAS, Rejito J, et al. A dimensionality reduction approach for machine learning based IoT botnet detection. In: Proceedings of the 2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI); 20–21 October 2021; Semarang, Indonesia. pp. 26–30.
6. Wang G, Lauri F, El Hassani AH. A study of dimensionality reduction's influence on heart disease prediction. In: Proceedings of the 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA); 12–14 July 2021; Chania Crete, Greece. pp. 1–6.
7. Tulapurkar H, Banerjee B, Mohan BK. Effective and efficient dimensionality reduction of hyperspectral image using CNN and LSTM network. In: Proceedings of the 2020 IEEE India Geoscience and Remote Sensing Symposium (InGARSS); 1–4 December 2020; Ahmedabad, India. pp. 213–216.
8. AlSaeed H, Hewahi N, Ksantini R. Dimension reduction techniques for image classification. In: Proceedings of the 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT); 20–21 November 2022; Sakheer, Bahrain. pp. 358–365.
9. Geng Y, Cai S, Qin S, et al. An efficient network traffic classification method based on combined feature dimensionality reduction. In: Proceedings of the 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C); 6–10 December 2021; Hainan, China. pp. 407–414.
10. Wan Y, Li T, Wang P, et al. Robust and efficient classification for underground metal target using dimensionality

reduction and machine learning. IEEE Access 2021; 9: 7384–7401. doi: 10.1109/ACCESS.2021.3049308
11. Arowolo MO, Adebiyi MO, Adebiyi AA, Okesola OJ. A hybrid heuristic dimensionality reduction methods for classifying malaria vector gene expression data. IEEE Access 2020; 8: 182422–182430. doi: 10.1109/ACCESS.2020.3029234
12. Yin J, Wang Y, Hu J. A new dimensionality reduction algorithm for hyperspectral image using evolutionary strategy. IEEE Transactions on Industrial Informatics 2012; 8(4): 935–943. doi: 10.1109/TII.2012.2205397
13. Moni V, Mattipalli M, Badar AQH. Machine learning classification techniques to predict directional change of energy prices using high dimensionality reduction. In: Proceedings of the 2022 International Conference on Computer Science and Software Engineering (CSASE); 15–17 March 2022; Duhok, Iraq. pp. 247–252.
14. Li J, Zhang J, Wang J, et al. VAGA: Towards accurate and interpretable outlier detection based on variational auto-encoder and genetic algorithm for high-dimensional data. In: Proceedings of the 2021 IEEE International Conference on Big Data (Big Data); 15–18 December 2021; Orlando, FL, USA. pp. 5956–5958.
15. Zhong L, Zhong J, Lu C. A comparative analysis of dimensionality reduction methods for genetic programming to solve high-dimensional symbolic regression problems. In: Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 17–20 October 2021; Melbourne, Australia. pp. 476–483.
16. Sivaranjani S, Ananya S, Aravinth J, Karthika R. Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In: Proceedings of the 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS); 19–20 March 2021; Coimbatore, India. pp. 141–146.
17. Li J, Zhang J, Wang J, et al. Outlier detection and abnormal subspace search based on autoencoder and genetic algorithm for high-dimensional data. In: Proceedings of the 2021 7th International Conference on Computer and Communications (ICCC); 10–13 December 2021; Chengdu, China. pp. 1510–1514.
18. Raia MR, Ruba M, Nemes RO, Martis C. Artificial neural network and data dimensionality reduction based on machine learning methods for PMSM model order reduction. IEEE Access 2021; 9: 102345–102354. doi: 10.1109/ACCESS.2021.3095668
19. Yaswanthram P, Sabarish BA. Face recognition using machine learning models—Comparative analysis and impact of dimensionality reduction. In: Proceedings of the 2022 IEEE Fourth International Conference on Advances in Electronics, Computers and Communications (ICAECC); 10–11 January 2022; Bengaluru, India. pp. 1–4.
20. Islam M, Anjum A, Ahsan T, Wang L. Dimensionality reduction for sentiment classification using machine learning classifiers. In: Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI); 6–9 December 2019; Xiamen, China. pp. 3097–3103.
21. Zhang Y, Jia Z, Ge H, Wang J. Novel SVM based SMOTE integrated LPP dimensionality reduction method for imbalanced samples fault diagnosis. In: Proceedings of the 2021 CAA Symposium on Fault Detection, Supervision, and Safety for Technical Processes (SAFEPROCESS); 17–18 December 2021; Chengdu, China. pp. 1–5.
22. Reddy GT, Reddy MPK, Lakshmanna K, et al. Analysis of dimensionality reduction techniques on big data. IEEE Access 2020; 8: 54776–54788. doi: 10.1109/ACCESS.2020.2980942
23. Cutura R, Angerbauer K, Heyen F, et al. DaRt: Generative art using dimensionality reduction algorithms. In: Proceedings of the 2021 IEEE VIS Arts Program (VISAP); 24–29 October 2021; New Orleans, LA, USA. pp. 59–72.
24. Dewangan S, Rao RS, Yadav PS. Dimensionally reduction based machine learning approaches for code smells detection. In: Proceedings of the 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP); 21–23 July 2022; Hyderabad, India. pp. 1–4.
25. Chen Z, Omote K. A privacy preserving scheme with dimensionality reduction for distributed machine learning. In: Proceedings of the 2021 16th Asia Joint Conference on Information Security (AsiaJCIS); 19–20 August 2021; Seoul, Korea. pp. 45–50.
26. Q. Fournier and D. Aloise. Empirical comparison between autoencoders and traditional dimensionality reduction methods. In: Proceedings of the 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE); 3–5 June 2019; Sardinia, Italy. pp. 211–214.
27. Yumeng C, Yinglan F. Research on PCA Data Dimension Reduction Algorithm Based on Entropy Weight Method. In: Proceedings of the 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI); 23–25 October 2020; Taiyuan, China. pp. 392–396.
28. Varunram TN, Shivaprasad MB, Aishwarya KH, et al. Analysis of different dimensionality reduction techniques and machine learning algorithms for an intrusion detection system. In: Proceedings of the 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA); 17–19 December 2021; Arad, Romania. pp. 237–242.