Abstract: This work focuses on the combining a feature selection technique based on genetic algorithm and support vector machines (SVM) of medical disease classification. SVMs have been used for various applications as a powerful tool for pattern classification. We use evolutionary computation which is a subfield of artificial intelligence that involves combinatorial optimization problems. Evolutionary computation uses iterative progress, such as growth or development in a population. This population is then selected in a guided random search using parallel processing to achieve the desired end. Such processes are often inspired by biological mechanisms of evolution. The proposed genetic-SVM classifier is used to the best subset of features that can optimize the SVM classifier. It was concluded that the proposed genetic-SVM with a existing approach performs generally better and shown that our proposed method been very powerful for different data sets. The obtained results using the genetic algorithms approach show that the proposed method is able to find an appropriate feature subset and SVM classifier achieves better results than other methods.

Keywords: SVM, Genetic Algorithm, Feature Selection and Classification

I. INTRODUCTION

The classification of medical data has become an increasingly challenging problem, due to recent advances in medical mining technology. Classification of this tremendous amount of data is time consuming and utilizes excessive computational effort, which may not be appropriate for many applications. In this work, we develop an approach to optimize the support vector machine parameters which combines the merits of support vector machine (SVM) and genetic algorithm (GA). Later we compare our proposed method with the statistical approach to show its superiority in terms of computational efficiency. Feature selection is often an essential data pre-processing step prior to applying a classification models. Feature selection has been an active research area in pattern recognition, medical diagnosis, statistics, and data mining communities

The objective of feature selection is to select a subset of relevant features for building effective prediction models. By removing irrelevant and redundant features, feature selection can improve the performance of prediction models by alleviating the effect of the curse of dimensionality, enhancing the generalization performance, speed up the learning process, and improving the model interpretability. Support vector machines (SVMs) have been used for various applications as a powerful tool for pattern classification. The success of SVMs is based on (1) mapping the input space to a high-dimensional feature space, and (2) the maximization of the margin between two classes in the feature space. One of the advantages of SVMs is that we can improve generalization ability by proper selection of kernels.

This paper is structured as follows. Section 2 introduces our methodology for classification, recent developments for feature selection, Genetic Algorithm and SVMs are reviewed. Several important aspects that arise from this work are discussed in Section 3. Section 4 provides experimental results using five real-world UCI data sets are discussed. A summary of this paper can be found in Section 5 where we provide its main conclusions and address future developments.

II. METHODOLOGY

A. Feature Selection

Feature selection can be defined as a process that chooses a minimum subset of M features from the original set of N features, so that the feature space is optimally reduced according to a certain evaluation criterion. Feature selection is a fundamental problem in many different areas, especially in medical diagnosis, document classification, bioinformatics, and object recognition or in modeling of complex technological processes [3, 4, 12].

The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. Further, it is often the case that finding the correct subset of predictive features is an important problem in its own right. For example, physician may make a decision based on the selected features whether a dangerous surgery is necessary for treatment or not.
Finding the best feature subset for a given problem with \( N \) number of features requires evaluating all \( 2^N \) possible subsets. The best feature subset also depends on the predictive modeling, which will be employed to predict the future unknown values of response variables of interest. Feature selection involves minimizing the number of relevant features for maximizing the predictive power of the model.

Feature selection reduces the dimensionality of feature space, removes redundant, irrelevant, or noisy data. It brings the immediate effects for application: speeding up a data mining algorithm, improving the data quality and thereof the performance of data mining, and increasing the comprehensibility of the mining results. Feature selection algorithms may be divided into filters [10], wrappers and embedded approaches. Filters methods evaluate quality of selected features, independent from the classification algorithm, while wrapper methods require application of a classifier (which should be trained on a given feature subset) to evaluate this quality. Embedded methods perform feature selection during learning of optimal parameters (for example, neural network weights between the input and the hidden layer). According to [3], there are three main directions for feature selection: filter, wrapper[4], and embedded methods.

Feature selection involves determining the highest classifier accuracy of a subset or seeking acceptable accuracy of smallest features. This study compromises between accuracy and feature numbers by the same amount of features. From this study, application of data dimensionality reduction pre-processing step is prior to the classification procedures which really improve the overall classification performance.

**B. Genetic Algorithm (GA)**

The Genetic Algorithm (GA) is an optimization and search technique based on the principles of genetics and natural selection. Generally GA are not used to find patterns, but rather to guide the learning process of data mining algorithms such as neural nets. A GA allows a population composed of many individuals [3] (basically the candidates) to evolve under specified selection rules to a state that maximizes the fitness. GA is known as a subset of evolutionary algorithms that model biological processes which is influenced by the environmental factor to solve various numerical optimization problems. GA allows a population composed of many individuals or called chromosomes to evolve under specified rules to a state that maximizes the fitness or minimizes the cost functions.

A genetic algorithm mainly composed of three operators: selection, crossover, and mutation. In selection, a good string (on the basis of fitness) is selected to breed a new generation; crossover combines good strings to generate better offspring; mutation alters a string locally to maintain genetic diversity from one generation of a population of chromosomes to the next. In each generation, the population is evaluated and tested for termination of the algorithm. If the termination criterion is not satisfied, the population is operated upon by the three GA operators and then re-evaluated. The GA cycle continues until the termination criterion is reached.

In feature selection, Genetic Algorithm (GA) is used as a random selection algorithm, capable of effectively exploring large search spaces[5], which is usually required in case of attribute selection. For instance; if the original feature set contains \( N \) number of features, the total number of competing candidate subsets to be generated is \( 2^N \), which is a huge number even for medium-sized \( N \).

Genetic Algorithm creates an initial population with a set of solutions. The evaluation process finds the best individual with the help of a fitness function. Sorting the individuals in the population brings out the top best individuals with better fitness values. The selected individuals stand as the populations for the next generation. Then the genetic operators like crossover and mutation helps in combining the different individuals and changing the feature of an individual. Each solution is called a chromosome. Again the selection operator is applied to bring out the best offspring. The procedure is repeated till the stopping condition is satisfied. In many cases the stopping criteria is the number of generations.

**C. SUPPORT VECTOR MACHINE**

The Support Vector Machine (SVM) is a supervised learning method for Data analysis, Pattern recognition, Classification and Regression analysis. It is a classification technique based on statistical learning theory [7][10]. The SVM, a promising new method for the classification of both linear and nonlinear data. A SVM performs classification by constructing an \( N \)-dimensional hyperplane that optimally separates the data into two categories. The goal of an SVM is to separate data instances into two classes using examples of each from the training data to define the separating hyperplane. The SVM method [8] provides an optimally separating hyperplane in the sense that the margin between two groups is maximized. The subset of data instances that actually define the hyperplane are called the “support vectors”, and the margin is defined as the distance between the hyperplane and the nearest support vector[11]. By maximizing this separation, it is believed that the SVM better generalizes to unseen data instances, while also mitigating the effects of noisy data or over-training. Error is minimized by maximizing the margin, and the hyperplane is defined as the center line of the separating space, creating equivalent margins for each class. Performance is most commonly evaluated as classification accuracy and/or margin width. Given two SVMs with identical classification accuracy, one would prefer to choose the SVM with a larger margin width, and vice versa. This trade-off is usually incorporated into the training of an SVM. An important generalization aspect of SVMs is that frequently not all the available training examples are used in the description and specification of the separating hyperplane. The subset of points that lie on the margin (called support vectors) are the only ones that define the hyperplane of maximum margin. This is shown in Figure 1. The algorithm constructs a maximum margin hyperplane which separates a set of positive examples from a set of negative examples. In
the case of examples not linearly separable, SVM uses a kernel functions to map the examples from input space into high dimensional feature space. Using a kernel function can solve the non-linear problem.

There are two types of SVMs, (1) Linear SVM, which separates the data points using a linear decision boundary and (2) Non-linear SVM, which separates the data points using a non-linear decision boundary.

Linear SVM performs well on datasets that can be easily separated by a hyperplane into two parts. But sometimes datasets are complex and are difficult to classify using a linear kernel. Non-linear SVM classifiers can be used for such complex datasets. The concept behind non-linear SVM classifier is to transform the dataset into a high dimensional space where the data can be separated using a linear decision boundary.

A non-linear support vector machine (SVM) is amongst the most powerful pattern classification algorithms, as it can obtain maximal generalization when predicting the classification of previously unseen data compared to other nonlinear classifiers [13]. By using a kernel function, it maps the original features into higher dimensional space where it computes a hyperplane that maximizes the distance from the hyperplane to the examples in each class. Having found such a hyperplane, the SVM can then predict the classification of an unlabeled example by mapping it into the feature space and checking on which side of the separating plane the example lies.

III. THE PROPOSED GENETIC–SVM ALGORITHM
In this section, we describe the proposed Genetic-SVM system for the feature selection. The aim of this system is to select the subset of features automatically for optimizing the SVM classifier.

A. Genetic set up
The first step in GAs is to define the encoding allowing describing any potential solution as a numerical vector, we use a vector of (0 and 1) with length of 22 (the number of features) which 0 and 1 is for the omitted and selected features respectively. At first, randomly we generate 50 chromosomes as a population. We use Roulette Wheel Selection for the cross-over and also we apply Swap mutation. This operator simply changes the position of two samples at random. The probability parameter of mutation is equal 0.1. The choice of the fitness function is important because it is on this basis that the Genetic evaluates the goodness of each candidate solution for designing our SVM classification system.

B. SVM Classification with genetic Algorithm
The procedure describing the proposed SVM classification system is as follows:
Step 1. Generates randomly an initial population of size 100.
Step 2. Training SVM Classifier. SVM classifier is trained by training set with feature subset selected and variable value of parameters.
Step 3. For each chromosomes of the population, train \( \frac{n(n-1)}{2} \) SVM Classifier for computing fitness of each chromosome (subset of features).
Step 4. Select individuals from population directly based on fitness values and regenerate new individuals from old ones.
Step 5. If the maximum number of iteration is not yet reached, we proceed with the next generation operation. The termination criteria are that the max generation number reached or the fitness function value does not improve during the last 15 generations return to step 2.
Steps 6. Select the best fitness as optimal subset feature.
Steps 7. Apply the optimal feature to dataset.
Step 8. Genetic Operation. The reproduction operators selected 20% of the best chromosome.

IV. RESULT AND DISCUSSION

This section describes the experimental results obtained by applying the proposed algorithms to a variety of data sets. For experimentation, five benchmark datasets (Breast cancer, Pima Diabetes, Mammographic Mass, Dermatology and Thoracic Surgery data sets) are taken from the UCI machine learning repository[9] as shown in Table 1. In order to validate the prediction results of the comparison of the two classification (SVM, SVM + Genetic) techniques and the 10-fold crossover validation is used. The k-fold crossover validation is usually used to reduce the error resulted from random sampling in the comparison of the accuracies of a number of prediction models. The present study divided the data into 10 folds where 1 fold was for testing and 9 folds were for training for the 10-fold crossover validation.

Table 1 provides the attribute information of five datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Features</th>
<th>Instances</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wisconsin Breast cancer</td>
<td>11</td>
<td>699</td>
<td>2</td>
</tr>
<tr>
<td>Pima Diabetes</td>
<td>9</td>
<td>768</td>
<td>2</td>
</tr>
<tr>
<td>Mammographic Mass</td>
<td>5</td>
<td>961</td>
<td>2</td>
</tr>
<tr>
<td>Dermatology</td>
<td>35</td>
<td>366</td>
<td>6</td>
</tr>
<tr>
<td>Thoracic Surgery</td>
<td>17</td>
<td>470</td>
<td>2</td>
</tr>
</tbody>
</table>

A. Evaluation Methods

We have used the Weka toolkit to experiment with these five data mining algorithms [13]. The Weka is an ensemble of tools for data classification, regression, clustering, association rules, and visualization. WEKA version 3.6.9 was utilized as a data mining tool to evaluate the performance and effectiveness of the SVM and Proposed SVM-Genetic technique. This is because the WEKA program offers a well defined framework for experimenters and developers to build and evaluate their models.

The performance of a chosen classifier is validated based on error rate and computation time. The classification accuracy is predicted in terms of Sensitivity and Specificity. The computation time is noted for two classifier is taken in to account. The evaluation parameters are the specificity, sensitivity, and overall accuracy of five UCI data sets are presented in Table 2 and Table 3. Accuracies and error rates of proposed and existing SVM are shown in figure 3 and figure 4.

B. Evaluation Metrics

Three classical evaluation metrics of Precision, Recall and F-score are used to evaluate the efficiency of the proposed method. The three metrics are traditionally defined for a binary classification task with positive and negative classes. Precision is the proportion of positive predictions that are correct, and recall is the proportion of positive samples that are correctly predicted positive. That is:

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

\[
\text{F - score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

- True positive (TP) = number of positive samples correctly predicted.
- False negative (FN) = number of positive samples wrongly predicted.
- False positive (FP) = number of negative samples wrongly predicted as positive.
- True negative (TN) = number of negative samples correctly predicted.

Table 2: Performance of different data sets SVM

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>ROC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wisconsin Breast cancer</td>
<td>0.967</td>
<td>0.037</td>
<td>0.967</td>
<td>0.967</td>
<td>0.967</td>
<td>0.965</td>
</tr>
<tr>
<td>Dermatology</td>
<td>0.954</td>
<td>0.009</td>
<td>0.954</td>
<td>0.954</td>
<td>0.954</td>
<td>0.984</td>
</tr>
<tr>
<td>Thoracic Surgery Data</td>
<td>0.849</td>
<td>0.851</td>
<td>0.724</td>
<td>0.849</td>
<td>0.782</td>
<td>0.499</td>
</tr>
<tr>
<td>Mammographic Mass</td>
<td>0.803</td>
<td>0.199</td>
<td>0.805</td>
<td>0.801</td>
<td>0.801</td>
<td>0.801</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.774</td>
<td>0.34</td>
<td>0.766</td>
<td>0.774</td>
<td>0.762</td>
<td>0.728</td>
</tr>
</tbody>
</table>
### Table 3: Performance of different data sets SVM with Genetic

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>ROC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wisconsin Breast Cancer</td>
<td>0.97</td>
<td>0.034</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.968</td>
</tr>
<tr>
<td>Dermatology</td>
<td>0.967</td>
<td>0.007</td>
<td>0.967</td>
<td>0.967</td>
<td>0.967</td>
<td>0.988</td>
</tr>
<tr>
<td>Thoracic Surgery</td>
<td>0.849</td>
<td>0.851</td>
<td>0.724</td>
<td>0.849</td>
<td>0.782</td>
<td>0.499</td>
</tr>
<tr>
<td>Mammographic Mass</td>
<td>0.776</td>
<td>0.223</td>
<td>0.781</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.773</td>
<td>0.335</td>
<td>0.769</td>
<td>0.773</td>
<td>0.77</td>
<td>0.73</td>
</tr>
</tbody>
</table>

**Figure 2:** compares the genetic-SVM approach with plain SVM classifier.

**Figure 3:** compares the genetic-SVM approach with plain SVM classifier.

### V. CONCLUSION

This paper proposed a Genetic algorithm based optimization algorithm, which can optimize the parameter values for SVM, and obtain the optimal subset of features. A comparison of the proposed algorithm results with existing SVM approach demonstrates that the proposed method improves the classification accuracy rates. The GA-SVM method was applied to remove insignificant features and effectively find the best parameter values.

The goal of this paper is to design Support Vector Machine and Binary Coded Genetic Algorithm were analyzed to find the classification accuracy and runtime for various kernel functions such as Polynomial and Radical Basic function are used. Feature Selection algorithm is used to improve the classification accuracy of classifier with respect to medical datasets. The results show that the classification accuracy of GA-SVM is the highest of SVM algorithm.
Therefore, our proposed approach can be used for many classification areas in medical diagnosis for prediction of diseases.

REFERENCES