CHAPTER V
CLASSIFICATION OF ELECTROCARDIOGRAM SIGNALS WITH SUPPORT VECTOR MACHINE

5.1. INTRODUCTION

In the previous chapter, the complete process of preprocessing and feature extraction is discussed. Once this process is completed, classification can be carried out. In this chapter, Support Vector Machine (SVM) is used for classification.

Numerous algorithms have been introduced for the recognition and classification of ECG signal. Some of them use time and some use frequency domain for depiction. Based on that many specific attributes are defined, allowing the recognition between the beats belonging to different pathological classes. The ECG waveforms may be different for the same patient to such extent that they are unlike each other and at the same time alike for different types of beats [96]. Artificial Neural Network (ANN) [69-71] and fuzzy-based techniques were also employed to exploit their natural ability in pattern recognition task for successful classification of ECG beats [3].

In this chapter, the approach to ECG beat classification presented thorough experimental exploration of the SVM capabilities for ECG classification. Further the performances of the SVM approach in terms of classification accuracy are evaluated: 1) by selecting the best discriminating features from the whole considered feature space and 2) by solving the model selection issue. The extraction process is implemented through AR Modeling framework (Refer Section 4.1.3) that exploits a criterion intrinsically related to SVM classifier properties. This framework is formulated in such a way that it also solves the model selection issue, i.e., to estimate the best values of the SVM classifier parameters, which are the regularization and kernel parameters.
5.2. SUPPORT VECTOR MACHINE

SVM is one of the most popular learning techniques depending on the statistical learning theory, which uses the structural risk minimization inductive principle with the aim to obtain a good generalization from narrow data sets. SVM produces black box models in the way that they do not have the capability to describe, in a clear form, the process by means of which the exit takes place.

To overcome this drawback, the theory created by either neural network or SVM [98] could be transferred into a more logical representation; these conversion methods are known as rule extraction algorithms.

SVM looks for the global hyper plane and avoids over fitting. SVM is a significant machine learning technique which is based on artificial intelligence. The exact working principle of SVM can be clearly understood from the figure 5.1.

![Figure 5.1: Principle of SVM](image)

The mapping of the input-output functions from a set of labeled (disease type) training data set is produced by the SVM supervised learning technique. SVMs in a high dimensional feature space use a theory space of linear functions which are trained with a learning approach from optimization theory that implements a learning bias derived from statistical learning theory.
SVM is a new technique for training classifiers depending on various functions such as polynomial functions, radial basis functions, neural networks etc. In Support Vector machines, the classifier is generated using a hyper-linear separating plane. SVM offers a most excellent solution for issues that cannot be linearly divided in the input space. The problem is resolved by making a non-linear transformation of the original input space into a high dimensional feature space, where an optimal separating hyper plane is found. A maximal margin classifier considering the training data is obtained when the separating planes are optimal.

The support vectors are the points which are at the margin and the solution depends only on these data points. This is the unique feature of this technique. Linear SVM can be extended to nonlinear SVM if a feature space uses a group of nonlinear basis function. The data points can be separated linearly in the feature space which are very high dimensional.

Figure 5.2: A Separating Hyper Plane in the Feature Space Corresponding to a Non-Linear Boundary in the Input Space
The figure 5.2 contains a separating hyperplane in the feature space corresponding to a non-linear boundary in the input space.

A significant feature of the SVM is that this transformation need not be implemented to determine the separating hyperplane in the possibly very-high dimensional feature space, a kernel representation can be used for the purpose of determining the separating hyperplane, where the solution evaluated at the support vectors is written as a weighted sum of the values of certain kernel function.

The mapping of the data from a lower dimensional input space to a higher dimensional feature space is possible when SVM acts as a binary classifier. This is to make the data linearly separable into two classes. To make the binary classifiers applicable to various multicategory problems, one versus-all approach is used, for which c (number of classes) binary classifiers should be built for SVM to distinguish one class from all the rest of the classes. Similarly, when the one-versus-one comparison approach is used, c(c-1)/2 binary classifiers should be built for SVM in such a way that it must be able to distinguish between every two class combination. Thus, the overall complexity of the classifier increases, when the number of classes c increases.

5.2.1. **SVM Algorithm**

The fundamental idea of SVM can be described as follows:

*Step 1:* The inputs are formulated as feature vectors.

*Step 2:* By using the kernel function, these feature vectors are mapped into feature space.

*Step 3:* A division is computed in the feature space to separate the classes of training vectors.
5.2.2. **Linear SVM Classification**

SVM is one of the best linear classification methods. The transformation of the samples space to high-dimension space is possible by the kernel mapping and the best linear classification surface of samples in this new space is obtained. This Non-linear transformation can be achieved by suitable inner product function. The best linear classification surface function of characteristics space can be described by the following equation:

\[
g(x) = \sum_{j=i}^{n} a_j y_i k(x, x_i) + b
\]  

(5.1)

where \((x, x_i)\) are the two types of sample collection divided in the sample space, \(b\) is the classification threshold and \(k(x, x_i)\) is being the nonlinear kernel function that replace characteristics space and meets Mercer conditions (a real-valued function \(K(x, y)\) is said to fulfill Mercer's condition if for all square integrable functions \(g(x)\) one has \(\int \int K(x, y)g(x)g(y)dxdy \geq 0\)). The best linear classification surface function is obtained by striking the best resolve \(a_i\) where \(i = 1, 2 \ldots n\) of the following function \(Q(a)\).

\[
g(x) = \sum_{j=i}^{n} a_j y_i k
\]  

(5.2)

\[
\max \limits_{a} Q(a) = \sum_{i=0}^{n} a_i - 0.5 \sum_{i=0}^{n} \sum_{j=0}^{n} a_i a_j y_i y_j k(x, x_i)
\]  

(5.3)

\[
\sum_{j=i}^{n} a_j y_i = 0
\]  

(5.4)

\(i = 1, 2 \ldots n\) and \(0 \leq a_i\)
The above equation is solving of quadratic function extreme value on condition that inequality, $Q(a)$ is convex function (A convex function is a continuous function whose value at the midpoint of every interval in its domain does not go beyond the arithmetic mean of its values at the ends of the interval). Because its local optimal solution is global optimal solution, the solution is unique. Thus the best classification function of SVM is:

$$f(x) = \text{sign}(g(x)) = \text{sign} \left\{ \sum_{j=1}^{n} a_j y_j k(x, x_j) + b \right\}$$  \hspace{1cm} (5.5)

### 5.2.3. SVM Kernel Functions

A kernel function and its parameter have to be chosen to build a SVM classifier. Three main kernel functions have been used to build SVM classifiers.

- **1. Linear Kernel function**, $K(x, z) = \langle x, z \rangle$, where $\langle x, z \rangle$ is the dot product of $x$ and $z$.

- **2. Polynomial kernel function**, $K(x, z) = (\langle x, z \rangle + 1)^d$, $d$ is the degree of polynomial.

- **3. Radial basis function**, $K(x, z) = \exp \left( \frac{-|x-z|^2}{2\sigma^2} \right)$, $\sigma$ is the width of the function.

### 5.2.4. Binary SVM Classification

Binary classification SVM is used to find an OSH which generates a maximum margin between two categories of data. To construct an OSH, SVM maps data into a higher dimensional feature space. SVM performs this nonlinear mapping by using a kernel function. Then, SVM constructs a linear OSH between two categories of data in the higher feature space. Data vectors which are nearer to the OSH in the higher feature space are called Support Vectors (SVs) and contains
all information required for classification. In brief, the theory of SVM is as follows.

To build a SVM classifier, a kernel function and its parameters need to be chosen. In this work, the following kernel function has been applied to build SVM classifiers:

Radial basis function $K(x, z) = \exp\left\{-\frac{||x-z||^2}{2\sigma^2}\right\}$, $\sigma$ is the width of the function.

5.2.5. Advantages of SVM

The significance of the classification approaches are based on the data which are being examined and thus have a relative relevance. SVM can be a valuable tool for insolvency analysis, in the case of non-regularity in the data, for example when the data are not regularly distributed or have an unknown distribution. The advantages of the SVM technique can be summarized as follows:

Because of the usage of kernel, SVMs increase the flexibility in the option of the form of the threshold separating affected from unaffected signals, which need not be linear and even need not have the same functional form for all data, as its function is non-parametric and works locally [99].

1. As the kernel consists of a non-linear transformation, no statement about the functional form of the transformation, which builds data linearly separable, is very essential. The transformation occurs absolutely on a dynamic theoretical basis and human knowledge judgment before is not required.

2. SVMs offer a good out-of-sample generalization, if the regularization parameter is chosen appropriately. Thus, if a suitable generalization grade is chosen, then SVMs can be very potential, even if the training sample has some bias.
3. SVMs deliver a unique solution, as the optimality problem is convex (Convex problem is quadratic, in which SVMs maximize the margin width while minimizing errors). This is an advantage compared to Neural Networks, which have multiple solutions connected with local minima and for this reason; it may not be robust over various samples.

5.3. APPLICATION OF SUPPORT VECTOR MACHINE FOR ECG CLASSIFICATION APPROACH

5.3.1. Support Vector Machines

SVM is usually used for classification tasks introduced by Vapnik [103]. The morphological and temporal features extracted from the previous phase are given as the input vector to the SVM classifier. For binary classification SVM is used to find an Optimal Separating Hyper Plane (OSH) which generates a maximum margin between two categories of data. To construct an OSH, SVM maps data into a higher dimensional feature space. SVM performs this nonlinear mapping by using a kernel function. Then, SVM constructs a linear OSH between two categories of data in the higher feature space. Data vectors which are nearest to the OSH in the higher feature space are called Support Vectors (SVs) and contain all information required for classification. In brief, the theory of SVM is as follows [103].

Consider training set \( D = \{(x_j, y_j)\}_{i=1}^{L} \) with each input \( x_i \in R^n \) and an associated output \( y_i \in \{-1, +1\} \). Each input \( x \) is firstly mapped into a higher dimension feature space \( F \), by \( z = \phi (x) \) via a nonlinear mapping \( \phi: R^n \rightarrow F \). When data are linearly non-separable in \( F \), there exists a vector \( w \in F \) and a scalar \( b \) which define the separating hyper plane as:

\[
y_i(w \cdot z_i + b) \geq 1 - \xi_i, \forall i
\]  

(5.6)
where $\xi_i$ is slack variable. The hyper plane that optimally separates the data in $F$ is one that

$$\text{minimise } \frac{1}{2}w^Tw + C. \quad (5.7)$$

$$\text{subject to } y_i(w^Tz_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \forall i$$

where $C$ is called regularization parameter that determines the tradeoff between maximum margin and minimum classification error. By constructing a Lagrangian, the optimal hyper plane according to (5.7) may be shown as the solution of

$$\text{maximize } W(\alpha) = \sum_{i=1}^{L} \alpha_i - \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (5.8)$$

$$\text{subject to } \sum_{i=1}^{L} y_i \alpha_i = 0, 0 \leq \alpha_i \leq C, \forall i$$

where $\alpha_1, \ldots, \alpha_L$ are the nonnegative Lagrangian multipliers. The data points $x_i$ that correspond to $\alpha_i > 0$ are SVs. The weight vector $w$ is then given by

$$w = \sum_{i \in SVs} \alpha_i y_i z_i \quad (5.9)$$

For any test vector $\in R^n$, the classification output is then given by

$$y = \text{sign}(w.z + b) = \text{sign} \left( \sum_{i \in SVs} \alpha_i y_i K(x_i, x) + b \right) \quad (5.10)$$

To build an SVM classifier, a kernel function and its parameters need to be chosen. So far, no analytical or empirical studies have established the superiority of one kernel over another conclusively. The kernel must satisfy the condition
stated in Mercer’s theorem [120] so as to correspond to some type of inner product in the transformed (higher) dimensional feature space $\Phi(X)$ [104]. A typical example kernels used is represented by the following Gaussian function:

$$K(x_i, x) = \exp (-\gamma ||x_i - x||^2)$$ (5.11)

Where $\gamma$ is a parameter which is inversely proportional to the width of the Gaussian kernel.

As described before, SVMs are intrinsically binary classifiers. But, the classification of ECG signals often involves the simultaneous discrimination of numerous information classes. In order to face this issue, a number of multiclass classification strategies can be adopted [105], [106]. The most popular ones are the One-Against-All (OAA) and the One-Against-One (OAO) strategies. The former involves a reduced number of binary decompositions (and thus, of SVMs), which are, however, more complex. The latter requires a shorter training time, but may incur conflicts between classes due to the nature of the score function used for decision. Both strategies generally lead to similar results in terms of classification accuracy. In this work, the OAA strategy is considered. Briefly, this strategy is based on the following procedure. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_T\}$ be the set of $T$ possible labels (information classes) associated with the ECG beats that desired to classify. First, an ensemble of $T$ (parallel) SVM classifiers is trained. Each classifier aims at solving a binary classification problem defined by the discrimination between one information class $\omega_i (i = 1, 2, \ldots, T)$ against all others (i.e., $\Omega - \{\omega_i\}$). Then, in the classification phase, the new rule is used to decide which label to assign to each beat which is “winner-takes-all” rule. This represents that the winning class is the one that corresponds to the SVM classifier of the ensemble that shows the highest output (discriminant function value).

The first experiment focused in estimating the performance of the classifiers in classifying ECG signals directly in the whole original hyper
dimensional feature space (i.e., by means of all the available features). In order to feed the classification process, the two following kinds of features are adopted: 1) ECG morphology features and 2) three ECG temporal features, i.e., the QRS complex duration, the RR interval (the time span between two consecutive R points representing the distance between the QRS peaks of the present and previous beats) and the RR interval averaged over the ten last beats [4]. In order to extract these features, first the QRS detection is performed and ECG wave boundary recognition tasks by means of the well-known eegpuwave software available on [109].

Then, after extracting the three temporal features of interest (QRS), normalized to the same periodic length the duration of the segmented ECG cycles [110]. To this purpose, the mean beat period was chosen as the normalized periodic length, which was represented by 300 uniformly distributed samples. Consequently, the total number of morphology and temporal features equals 303 for each beat.

The second experiment discovers the behavior of the proposed classifiers when incorporated within a standard classification technique based on DWT-AR feature reduction. In order to obtain reliable assessments of the classification accuracy of the investigated classifiers, in all the following experiments, three different trials are performed, each with a new set of randomly selected training beats, while the test set was kept unchanged. The results of these three trials obtained on the test set were thus averaged. The ECG beats are extracted from the patient records which are available in MIT-BIH arrhythmia database.

The following command is used to carry out the ECG signal classification using SVM. The extracted morphological and temporal features are given as input to the SVM classifier.
Function [beats_detected] = svmmulticlassoneagainstall(testdata, testtarget, kerneloption);

% testdata is the extracted morphological and temporal features
% testtarget denotes the disease label
% kerneloption to select the kernel type
% beats_detected is the detected beats

5.4. EXPERIMENTAL RESULTS

This section presents the experimental evaluation and the results of the proposed approach with the existing approach.

5.4.1. Experimental Setup

In the experiments, the nonlinear SVM is considered based on the popular Gaussian kernel (referred to as SVM-RBF or simply SVM). The related parameters $\gamma$ and $C$ for this kernel were varied in the arbitrarily fixed ranges $[10^{-3}, 200]$ and $[10^{-3}, 2]$ so as to cover high and small regularization of the classification model and fat as well as thin kernels, respectively. In addition, for comparison purpose, the SVM classifier with feature selection preprocessing approach is compared with SVM-linear.

In this experiment, the SVM classifier is trained based on the Gaussian kernel, which proved in the previous experiments to be the most appropriate kernel for ECG signal classification, in feature subspaces of various dimensionalities. For sample, the detected beats of record 100 is shown in figure 5.3 and the number of beats detected for twenty patient records are shown in figure 5.4. Figure 5.5 shows the beat classification.
Figure 5.3: ECG Wave of Record 100 with Detected Beats

Figure 5.4: Number of Beats Detected for Twenty Patient Records
The total number of training beats was fixed to 500, as reported in Table 5.1. The desired number of features varied from 30 to 70 with a step of 10, namely, from small to high-dimensional feature subspaces. Feature selection was achieved by the DWT and AR modeling.

Table 5.1

Numbers of Training and Test Beats used in the Experiments

<table>
<thead>
<tr>
<th>Class</th>
<th>N</th>
<th>A</th>
<th>V</th>
<th>RB</th>
<th>/</th>
<th>LB</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training beats</td>
<td>150</td>
<td>100</td>
<td>100</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>500</td>
</tr>
<tr>
<td>Test beats</td>
<td>24000</td>
<td>245</td>
<td>3789</td>
<td>3893</td>
<td>6689</td>
<td>1800</td>
<td>40416</td>
</tr>
</tbody>
</table>

5.4.2. Performance Evaluation

The performance of the SVM approach is evaluated based on the accuracy of classification and average number of detected features. The linear SVM is compared with the SVM classifier with feature selection approach with DWT-AR modeling feature selection.
A. Classification Accuracy

It is observed from the table 5.2 that the SVM classification with DWT-AR modeling approach provides better accuracy when compared with the linear SVM classification.

The confusion matrix shows how the predictions are made by the classification system. The rows stand for the known class of the data, i.e. the labels in the data. The columns stand for the predictions made by the model. The value of each of element in the matrix is the number of predictions made with the class corresponding to the column for examples with the correct value as represented by the row. Thus, the diagonal elements show the number of correct classifications made for each class, and the off-diagonal elements show the errors made.

The performance of the classifier is evaluated using this matrix. This allows more detailed analysis than mere proportion of correct guesses. To better understand the confusion matrix, consider the three classes A, B and C and the confusion matrix is framed as below.

\[
\begin{array}{ccc}
A & B & C \\
A & t_p_A & e_{AB} & e_{AC} \\
B & e_{BA} & t_p_B & e_{BC} \\
C & e_{CA} & e_{CB} & t_p_C \\
\end{array}
\]

In the above confusion matrix,

- True Positive (TP) = \( t_p_A \)
- True Negative (TN) = \( t_p_B, e_{BC}, e_{CB}, t_p_C \)
- False Positive (FP) = \( e_{BA}, e_{CA} \)
- False Negative (FN) = \( e_{AB}, e_{AC} \)

\[
\text{Accuracy} \, (\%) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \quad (5.12)
\]
where
‘TP’ represents True Positive- Healthy people correctly diagnosed as healthy,
‘TN’ represents True Negative- Sick people correctly diagnosed as sick,
‘FP’ represents False Positive- Healthy people wrongly diagnosed as sick, and
‘FN’ represents False Negative- Sick people wrongly diagnosed as healthy.

For instance, consider the patient ID 202 from MIT-BIH database. This signal file contains 2061 normal beats and 36 AF beats. The beat classification is done by the Linear SVM and SVM with preprocessing and feature selection. The accuracy is calculated using the obtained results of confusion matrix.

![Image of confusion matrix]

**Figure 5.6: Confusion Matrix for SVM Classifier**

According to the confusion matrix output which is shown in figure 5.6, the diagonal values show the correctly classified beats. Here, 1679 normal beats and 28 AF beats are correctly classified. 419 normal beats are misclassified as 125 AF beats, 205 V beats and 89 as LB beats. Sum of these misclassified values constructs the FP. 8 AF beats are misclassified as normal which constructs the FN value. Sum of remaining 28 beats construct the TN value.
**Accuracy of Linear SVM**

In our case, TP=1679, TN=28, FN=8, FP=419

\[
\text{Accuracy (\%)} = \frac{1679 + 28}{1679 + 28 + 8 + 419} \times 100 = 80.00\%
\]

**Accuracy of SVM with Preprocessing and Feature Selection**

TP=1872, TN=30, FN=6, FP=226

\[
\text{Accuracy (\%)} = \frac{1872 + 30}{1872 + 30 + 6 + 226} \times 100 = 89.12\%
\]

**Table 5.2**

**Classification Accuracy Comparison of Linear SVM and SVM with Feature Selection**

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Linear SVM</th>
<th>SVM with Preprocessing and Feature Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>202</td>
<td>80.00</td>
<td>89.12</td>
</tr>
<tr>
<td>203</td>
<td>81.03</td>
<td>88.41</td>
</tr>
<tr>
<td>208</td>
<td>80.47</td>
<td>87.65</td>
</tr>
<tr>
<td>212</td>
<td>81.89</td>
<td>89.53</td>
</tr>
</tbody>
</table>

Figure 5.7 shows the SVM classification results. The accuracy of the SVM classifier is observed to be higher for all the patients when compared with the SVM linear classifier. The results are taken up for four patients whose IDs are 202, 203, 208 and 212. For instance, for the patient ID 202, the accuracy of the SVM classifier is 89.12% where as the accuracy of the SVM-linear classifier is 80.00%.
Figure 5.7: Classification Accuracy Comparison of Linear SVM and SVM with Feature Selection for Patient ID: 202

B. Number of Features Detected

The number of features selected for different diseases using DWT and AR modeling are listed in table 5.3. The result of the DWT with AR modeling approach is compared with PSO-SVM approach [115].

<table>
<thead>
<tr>
<th>Feature Selection Techniques</th>
<th>Class</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>A</td>
<td>V</td>
<td>RB</td>
<td>/</td>
<td>LB</td>
<td>AVERAGE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSO-SVM</td>
<td></td>
<td>63</td>
<td>43</td>
<td>35</td>
<td>43</td>
<td>40</td>
<td>42</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM-DWT and AR Modeling</td>
<td></td>
<td>70</td>
<td>46</td>
<td>50</td>
<td>43</td>
<td>48</td>
<td>44</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.3 shows the number of features detected automatically to
discriminate each class from the others. The average number of features required
by the SVM classifier is 60. The average number of features shown in the above
table is calculated from 5 runs, while the minimum and maximum numbers of
features were obtained for the Ventricular premature (V) and Normal (N) classes
with 50 and 70 features, respectively. Auto-Regressive Model parameters and the
variance of discrete wavelet transform are used as feature selection in this
approach. The average number of detected features obtained by PSO-SVM is 49
which is very less when compared to SVM with DWT and AR modeling.

C. Overall Accuracy (OA) Vs. Number of Selected Features

Overall Accuracy (OA) of a classifier is completely based on the number of
features selected. The desired number of features varied from 30 to 70 with a step
of 10 namely, from small to high dimensional feature subspaces while it keeps OA
at the high level. Figure 5.8 shows the effect of OA against the number of features
selected for the test beats with the PSO–SVM and SVM-DWT and AR Modeling.

![Overall Accuracy (OA) versus number of selected features](image)

**Figure 5.8: Overall Accuracy (OA) versus number of selected features
achieved on the test beats with the PSO–SVM and SVM-DWT and AR Modeling**
The OA of SVM-DWT and AR Modeling is very high when comparing with the PSO–SVM. SVM-DWT and AR Modeling shows low sensitivity to the curse of dimensionality as compared to the PSO–SVM.