A Comparison of Expert Systems for the Automated Interpretation of the ECG. Regulatory Implications of the Use of Neural Networks.

Branko Celler¹, Philip de Chazal¹ and Nigel Lovell²

¹Biomedical Systems Laboratory, School of Electrical Engineering,
²Graduate School of Biomedical Engineering, University of N.S.W., Sydney, Australia

In our laboratory we have developed a rule based expert system for the automated interpretation of the clinical ECG. Robust algorithms, validated against the CSE ECG databases, have been designed for the reliable determination of QRS complexes, T waves and P waves. Individual wave morphometries are then classified and analysed to produce a matrix of over thirty parameters, which is then passed to a rule based expert system for final interpretation and diagnosis. The performance of this system is now beginning to approach that of the best commercial systems available. Difficulties exist however in the cross validation of rules obtained from a variety of clinical sources and journal publications, even when highly experienced cardiologists are available as consultants. We are thus investigating alternative methods based on Probabilistic Neural Networks.

We measure over 180 parameters from the ECG derived from a variety of methods including power spectral density, discrete wavelet decomposition and more classical scalar or vectorcardiographic lead analysis. Receiver operator characteristic (ROC) plots are used to optimise the selection of suitable independent parameters by providing an index of the discriminating ability between alternate categories of a parameter over the complete spectrum of the parameter values. Once an optimal set of independent discriminating parameters are determined for each disease condition, they are used as the input nodes to a neural network.

Preliminary results from a database of 500 fully diagnosed ECG records indicates that diagnostic accuracies of 75-80% are achievable. Artificial neural networks, however are relatively new technologies which pose particular challenges regarding the verification and validation of core processing architectures and the neural network topology. Some of these issues are now being considered by the FDA and the TGA in Australia and could result in the imposition of significant constraints on the use of Neural Networks in medical diagnostic devices.

1. Introduction

Medical devices are increasingly supported by embedded or post-processing software incorporating some form of artificial intelligence, typically based on either expert systems or neural networks, for the automated analysis and interpretation of data collected. Expert systems attempt to model the expertise of a highly qualified human operator, such as a cardiologist or other medical specialist into a set of algorithms which can be executed in software. Such a system typically consists of a knowledge base of rules, heuristics or data tables and an inference engine which uses a set of criteria to use the knowledge to achieve an output. In cardiology an example of a rule based expert system is that published in Comprehensive Electrocardiology [1], which uses a complex list of rules to interpret and diagnose the 12 lead clinical ecg according to morphometric, arrhythmic and other criteria. A typical example of such a rule is shown below;
IF (1) QRS duration ≥ 0.10s and < 0.12s
AND (2) absence of Q wave in leads I, V5 and V6
AND (3) R peak time > 0.06s in leads V5 and V6
THEN Incomplete left bundle branch block

Development of the knowledge base for such a rule based expert system is a complex and painstaking process, requiring exhaustive review of the literature, identification and validation of a large number of rules by a committee of experts, and statistical comparison between the results obtained by the experts and those obtained by the expert system. This process needs to be repeated iteratively until a desired level of accuracy and reliability is achieved. Examples of programs which apply a deterministic approach are Glasgow, HP, IBM, Louvian, Lyon, Marquette, Means, Nagoya and Porto and they typically achieve an accuracy of 66-70% in the classification of the ECG into one of 7 classes.

An artificial neural network (ANN) is a data processing structure which is loosely modelled on principles characteristic of biological neural networks. There are many definitions for describing ANN’s. According to the DARPA Neural Network Study [2], “a neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes.” Fundamental to the definition of an ANN is the threshold firing behaviour of each “neuron” forming individual processing layers, and the interneuron connection strength or synaptic weight which is adaptively modified through a learning process. ANN’s are particularly useful for the classification and mapping of output states when large well validated learning data sets exist, and hard rigorous quantitative rules are not easily available. The automated interpretation of electrocardiograms, where diagnostic statements such as “probable left ventricular hypertrophy” must be produced with high specificity in response to a set of input parameters or classifiers is well suited for analysis by ANN.

The most comprehensive comparison of computer expert systems for ECG diagnosis was carried out by the CSE working party [3]. Using a database of 1220 ECG recordings the working party compared the diagnostic performance of statistically based expert systems, rule based expert systems and cardiologists. An significant outcome of the analysis was that the two statistically based programs achieved a significantly higher (median 76.6%) overall classification rate than the seven deterministic programs (median 69.3%) on the same database which had been independently diagnosed with 100% accuracy. The diagnostic performance of the statistically based programs was comparable to cardiologists (median 76.3%). A limitation of the study is that it only looked at diagnosis of anatomic lesions and pathophysiologic states, eg. myocardial infarction or ventricular hypertrophy. It ignored the important aspect of the electrocardiographic diagnosis of electrophysiologic changes such as arrhythmias and atrio- and ventricular conduction blocks.

In this study we compare the diagnostic accuracy of results obtained using an artificial neural network with inputs carefully selected on the basis of their discriminating ability and lack of correlation with those reported in the literature, and discuss the policy and regulatory implications of incorporating artificial intelligence and expert systems in ECG instruments.

2. Methods

2.1 Database.

From our ECG database we selected all cases with only one disease condition. The resulting set of 486 records used through out this study contained 155 normal (NOR) and 331 abnormal (ABN) cases. The abnormal cases comprised of 73 left (LVH), 21 right (RVH) and 25 bi- (BVH) ventricular hypertrophy and 75 anterior (AMI), 106 inferior (IMI) and 31 combined (MIX) myocardial infarction cases. The database contains from 8-10 seconds of digitally sampled (500
Hz sampling rate) data for each case from simultaneously recorded 12 lead and Frank lead ECGs.
Where smoothing was required a 0.5-40 Hz linear phase digital bandpass filter was applied. We use optimised ninth order filters which have maximally flat pass bands and approximately 50dB/octave roll off. Rejection in the stopband is better than 50dB. Data is applied to the filters in the forward and reverse direction to ensure linear phase of the output.

2.2 QRS typification

Of the 276 parameters we measured, 234 are derived from a representative QRS complex from a sinus beat and thus determination of the QRS onset and offset is necessary first step. QRS typification was divided into three steps: 1. QRS detection 2. global QRS onset and offset estimation 3. individual lead QRS onset and offset determination and individual wave determination [4].

QRS detection: Our QRS detection method is a modification of the method published by B. Yu [5]. Beginning with the pseudo orthogonal leads (II, V2 and V6) we applied the smoothing filter and formed the magnitude vector using $\sqrt{II^2 + V2^2 + V6^2}$. A 60 millisecond median filter was applied to the magnitude vector to remove the QRS complexes. The isolated QRS complexes were then recovered by subtracting the median filtered signal from the unfiltered magnitude vector. An adaptive threshold was then applied to this signal to determine the QRS locations. An exclusion interval of 250 milliseconds was applied after each QRS detection to prevent any false detection of T waves.

Global QRS onset and offset estimation Leads II, V2 and V6 were smoothed and a thresholding signal (THR) was formed from the sum of the absolute values of the first derivatives of each the three filtered leads. The maximum point of THR in the vicinity of each QRS detection was found and a search made immediately before and after the maximum point for the signal levels below an adaptive threshold. These signal points became the global QRS onset (QRSon) and offset (QRSoфф).

Isoelectric level: Each lead was searched in the 40 ms prior to QRSon for a section with less than $2\mu V$/sample slope. The slope was measured over a 20 ms interval to eliminate any mains frequency noise. If the isoelectric section was found then the isoelectric level was calculated by averaging the ten samples centred around the mid point of the isoelectric section.

Individual lead QRS onset: We found that the QRS onset was the most easily identified characteristic point of the ECG signal and our most accurate estimation of this point was obtained when using the unfiltered ECG signal. Leads we successfully identify the isoelectric segment were passed directly to the waveform boundary determination algorithm described below.

For leads with no identifiable isoelectric segment, the slope of the signal immediately before the QRSon was determined. The ECG signal was then translated and rotated so that this region of the signal has zero slope and average. This signal is passed to the waveform boundary determination algorithm.

Individual lead QRS offset: For determination of the QRS offset we first applied an additional low pass filter. The slope of ST segment is measured by determining the least squares linear fit of the samples immediately after QRSoфф. If the slope of the ST segment was greater than 0.5 $\mu V$/sample then the ECG signal was rotated so that the ST segment has zero slope.

Waveform boundary determination algorithm: The CSE working party recommends that the smallest reliably discernible wave in the QRS complex has an amplitude of 20 $\mu V$ with a duration of 6 ms [6]. Our algorithm uses these minimum wave requirements for noise free ECGs. To determine the QRS onset point, we first measure the noise in each lead by determining three times the standard deviation of the samples immediately before QRSon and if the noise exceeds 20 $\mu V$ the minimum wave amplitude is increased to this figure. If the noise exceeds 50 $\mu V$ we do not attempt to find the fiducial point.
The ECG is differentiated and the maximum absolute slope (MAXSLOPE) between the QRSon and QRSoff is determined. Next we search in the vicinity of QRSon and find the left and right isoelectric crossing points of the first valid wave. Finally we search both the differentiated ECG and the ECG from the left isoelectric crossing point towards the right and find the first set of 3 consecutive points which either have an absolute slope which exceeds 5% of MAXSLOPE or exceeds the minimum wave amplitude. The first point of this set becomes the fiducial point.

The same algorithm is used to determine the QRS offset point except that the signal is reversed in time and the values of QRSon and QRSoff were swapped before being processed.

2.3 Parameters

Two hundred and thirty four parameters were derived from the QRS complex and the remaining parameters were derived from the whole VCG recording. No P or T wave parameters were included.

One hundred and sixty six VCG parameters measured from the representative QRS complex were obtained using methods found in the literature. These parameters were derived from scalar, vector-loop and 3-D loop representations of the VCG. The scalar lead parameters included measurements of heights, angles, time durations and areas relating to the complete QRS complex and the individual QRS waves. The vector loop parameters included measurements of heights, angles and areas of the three QRS vector loops. The 3-D parameters were measurements of heights, azimuth and elevation angles, planar areas and planarity factors associated with the 3-D QRS loop. Some parameters were ratios of other parameters.

The above parameters were supplemented with novel parameters recently investigated in our laboratory [7] which utilise the power spectral density (PSD) and discrete wavelet transforms (DWT). Three additional parameter groups were added (i) PSD estimates of a complete ECG record. Power spectral density estimates were developed from a data window extending from the end of the first QRS to the beginning of the last QRS complex detected. (ii) PSD estimates of a single QRS complex. A single QRS complex was selected, and the baseline set to the equipotential voltage. A total of 125 samples centred on the QRS complex was used, and a Hanning window applied before calculating the discrete Fourier transform (DFT). The absolute value of the DFT provides an estimate of PSD from 0-40Hz at 4Hz intervals. (iii) Single beat DWT. Single beat data was subjected to a six level decomposition using the Haar wavelet. Parameters were formed from the root mean square value of the reconstructed approximation and the six reconstructed detail waveforms.

The age and sex of the patient were also included in the parameter set.

2.4 ROC Analysis

The ROC graph is a plot of all the sensitivity/specificity pairs resulting from continuously varying the decision threshold over the entire range of results observed [8]. The ROC plot thus depicts the overlap between the two distributions by plotting the sensitivity against (1-specificity) for the complete range of decision thresholds. The theoretical plot for a test with perfect discrimination, passes through the upper left hand corner, and the plot for a test with no discrimination, is diagonal line passing through the origin and upper right hand corner. A test with observed greater accuracy would generate a plot to the left and above that for the control. A convenient global way of characterising the diagnostic value of a test is the area under the ROC plot. An area of 0.8 for example, means that a randomly selected value from one disease category has a value either larger or smaller than for the normal category, 80% of the time.

2.5 Rank Correlation

Rank correlation provides a measure of the closeness of association of two variables based on the ranks of the observations. The advantage over the Pearson correlation coefficient is it makes no assumptions concerning the distributions of the two variables. The rank correlation
can have a value between minus one and one. A value of one indicates that ranks of two variables vary identically and a rank correlation of zero indicates the two variables vary independently.

2.6 Neural Network Classifiers

A feed-forward fully connected network trained using the back-propagation algorithm was used to develop the seven class discriminator for the parameter set. We used a three layer network with 20 input nodes forming the input layer, ten nodes forming the middle layer and seven nodes in the output layer. Log-sigmoid transfer functions were used in the middle and output layers. The output node with the largest value became the final classification of the network.

The 486 cases were randomly broken into two data sets - a training set of 325 cases and a testing set of 161 cases. The two data sets were randomly adjusted so that they contained the same relative proportions of normals and the six disease conditions.

The training set was further divided into an estimation set and a verification set. Cross validation was used to avoid over training. Training was stopped on the estimation set when the sum square error on the verification set was minimised. Five hundred networks were trained using randomly chosen estimation and verification sets for the parameter sets. The diagnostic performance of each network was found on the test set and analysis was made on averaged results.

2.7. Combining Multiple Neural Networks

Recent research [9] has shown that combining multiple classifiers which were derived from the same training data but with the addition of a randomising element improved the classification performance of unstable classifiers such as neural networks. We used the “bagging” technique which perturbs the training set of size N by placing equal probabilities on each record in the training set. An estimation set is formed by sampling with replacement N times from the original training set. Any given record has a probability of approximately 0.632 of being in the estimation set. Thus some records do not appear in the estimation set while others appear more than once. The records which did not appear in the estimation set formed the verification set. The proportion of the two classes in the estimation and verification sets were equalised by randomly duplicating cases in the minority classes. The effect of this was to approximately equalise the sensitivity and specificity results on the testing set.

The final decision of the combined networks was found using a “voting” method whereby the majority decision of the individual networks is chosen. Our multiple network configurations combined ten individual networks.

3. Results

Using the ROC and rank correlation analysis we choose the 3 most discriminating and uncorrelated parameters for the seven bigroups. The ROC areas for these 21 parameters are shown in table 1. Two of the bigroups IMI/nonIMI and MIX/nonMIX had a common parameter, so there were 20 unique parameters. These 20 parameters formed the inputs to the neural network and the classification matrix from the test set is shown in table 2.

<table>
<thead>
<tr>
<th>NOR</th>
<th>LVH</th>
<th>RVH</th>
<th>BVH</th>
<th>AMI</th>
<th>IMI</th>
<th>MIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.82</td>
<td>0.80</td>
<td>0.81</td>
<td>0.78</td>
<td>0.82</td>
<td>0.79</td>
<td>0.84</td>
</tr>
<tr>
<td>0.71</td>
<td>0.77</td>
<td>0.81</td>
<td>0.76</td>
<td>0.76</td>
<td>0.76</td>
<td>0.83</td>
</tr>
<tr>
<td>0.70</td>
<td>0.74</td>
<td>0.79</td>
<td>0.75</td>
<td>0.76</td>
<td>0.76</td>
<td>0.80</td>
</tr>
</tbody>
</table>
Table 1: The area under the ROC curve for the 3 most discriminating and uncorrelated parameters for the seven bi-groups: normal/abnormal; LVH/non-LVH; RVH/non-RVH; BVH/non-BVH; AMI/non-AMI; IMI/non-IMI and MIX/non-MIX.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Size</th>
<th>NOR</th>
<th>LVH</th>
<th>RVH</th>
<th>BVH</th>
<th>AMI</th>
<th>IMI</th>
<th>MIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOR</td>
<td>52</td>
<td>80</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>LVH</td>
<td>24</td>
<td>5</td>
<td>67</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>RVH</td>
<td>7</td>
<td>34</td>
<td>3</td>
<td>49</td>
<td>0</td>
<td>12</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>BVH</td>
<td>8</td>
<td>24</td>
<td>24</td>
<td>21</td>
<td>29</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AMI</td>
<td>25</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>73</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>IMI</td>
<td>35</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>84</td>
<td>2</td>
</tr>
<tr>
<td>MIX</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>47</td>
<td>2</td>
<td>52</td>
</tr>
<tr>
<td>Total</td>
<td>161</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total accuracy 72%  Partial accuracy 82%
Nor/Abn accuracy 88%

Table 2: Percentage correct for seven way classification of the Frank lead ECG on the testing set. See text for description of partial accuracy and normal/abnormal accuracy. Cells underlined are used in the calculation of partial accuracy.

The total accuracy was 72%. Partial accuracy was calculated by counting the diagnosis of a LVH or RVH as BVH; BVH as LVH or RVH; AMI or IMI as MIX nd MIX as AMI or IMI as being partially correct. Normal/abnormal accuracy was calculated by considering as correct a diagnosis of an abnormal recording into any of the six disease conditions.

4. Conclusion

We have demonstrated that be careful selection of highly discriminating and uncorrelated parameters as inputs to a three layer artificial neural net with only 20 inputs, diagnostic accuracies (72% total and 82% partial accuracy) approaching those achieved by cardiologists (median 76.3%) can be achieved with relatively modest computational power. Better results could be achieved with the availability of a much larger training set of fully (100% accurately) diagnosed ECG recordings. Such data sets however are not readily available, and as the data collection and validation process to ensure 100% accuracy is extremely costly, an expensive multinational effort would be required. Many hundreds of thousands of ECG recordings however are made every year in Australia and most are interpreted by at least one specialist cardiologist, presumably with an accuracy approaching 76%. As artificial neural networks are especially tolerant of some imprecision, very high diagnostic accuracies could be obtained if the training set was large enough (>10,000 records) and diagnostic accuracy was typically of the order of 60-80%.

In this study we selected parameters as inputs to the neural net from a wide variety of sources. These included, simple scalar parameters, those from vector cardiographic analysis and yet others from power spectral and wavelet transform analysis. There is however a substantial cost both in developing the algorithms to reliably measure these parameters and in computing their value with high reliability for poor and atypical ecg recordings. A significant challenge is therefore the identification of parameters which can be easily measured from the eight independent leads of the 12 lead clinical ecg, without requiring the detailed measurement of parameters describing the beat by beat morphometry of individual beats. Our preliminary studies suggest that parameters from power spectral density measures or discrete wavelet transforms could provide such classifiers if the rule for low correlation was relaxed and compensated for by selecting a neural net with a much larger number of inputs, typically in the order of 80-160. The need to identify and exclude atrial and ventricular ectopic beats, and a
range of other A-V conduction defects, however will always require that some beat-by-beat arrhythmia processing be carried out before global processing for parameters to use as classifiers to a neural network can take place.

The use of Artificial Intelligence, Expert Systems and Neural Networks pose special challenges regarding the verification and validation of the core processing architecture for regulatory and software quality control requirements. This core processing architecture may be the knowledge base and inference engine for expert systems and the neural net engine for the neural network system. These issues are identified and discussed in the U.S. Food and Drug Administration (FDA) Center for Devices and Radiological Health (CDRH) software guidance document [10], “ODE Guidance for the Content of Premarket Submission for medical Devices Containing Software” and were further the focus of the FDA Software Policy Workshop held at the National Institute of Health in Bethesda in September 3-4, 1996. Although neither the FDA or the Therapeutic Goods Administration (TGA) in Australia have as yet established policies for the registration and regulation of computer aided diagnostic and decision support systems, some general guidelines are available. For rule based expert systems these include, the need to;

- Verify the accuracy of the knowledge base and the relationships between data objects and object classes
- Analyse the heuristics and rules underlying the inference engine to ensure that there are no common sense contradictions and / or logical paradoxes
- Provide the reasoning path followed by the software to reach a particular conclusion so that the user may question and possibly suggest alternative and possibly more appropriate reasoning paths.

Artificial neural networks are by their structure and implementation difficult to validate using traditional software engineering methodologies. The strength of the ANN, the ability to learn by example and to self adjust on exposure to a new training set, makes analysis of the reliability of the neural network engine difficult. As the performance of the ANN is determined by its selective exposure to alternative training sets and is only approximately related to statistical techniques such as projection pursuit regression, kernel discriminant analysis, k-means cluster analysis or principal component analysis, behaviour can be non-deterministic and unpredictable.

The potential benefits from use of neural networks are however, resulting in their increasing incorporation in medical devices. The FDA in a draft document released for discussion [10] suggests that the following requirements may become mandatory before devices using embedded neural networks are licenced to enter the market;

- Design, assumptions, learning method and training set data need to be evaluated for appropriateness and correctness
- The designers are required to explain the choices made for the ANN model, topology and training sets with reference to the data set class it is meant to operate on
- Explanation of how over fitting or over training is avoided and the final ANN design selected
- Testing of performance of the ANN on additional data sets with varying distributions of features to be extracted
- Testing to ensure that the ANN has not been trained to detect a particular peculiarity of the training set rather than the intended features
- The raw data processed by the ANN should be available to the user for comparison with the output provided
A significant advantage of rule based over a statistically based expert systems is thus the ability to provide reasoning in a form comprehensible to a cardiologist to support a particular diagnosis. A statistically based program such as a neural network decision system can at best report the intermediate coefficients it used to develop the final diagnosis. The coefficients would in general not supply any useful information to a cardiologist.

5. References

2. DARPA Neural Network Study