A confident decision support system for interpreting electrocardiograms

Holger Holst[1], Mattias Ohlsson[2], Carsten Peterson[2] and Lars Edenbrandt[1]

Departments of [1]Clinical Physiology and [2]Theoretical Physics, Lund University, Lund, Sweden

Received 11 November 1998; accepted 23 February 1999

Correspondence: Holger Holst, Department of Clinical Physiology, University Hospital, S-221 85 Lund, Sweden

Summary

Computer-aided interpretation of electrocardiograms (ECGs) is widespread but many physicians hesitate to rely on the computer, because the advice is presented without information about the confidence of the advice. The purpose of this work was to develop a method to validate the advice of a computer by estimating the error of an artificial neural network output. A total of 1249 ECGs, recorded with computerized electrocardiographs, on patients who had undergone diagnostic cardiac catheterization were studied. The material consisted of two groups, 414 patients with and 835 without anterior myocardial infarction. The material was randomly divided into three data sets. The first set was used to train an artificial neural network for the diagnosis of anterior infarction. The second data set was used to calculate the error of the network outputs. The last data set was used to test the network performance and to estimate the error of the network outputs. The performance of the neural network, measured as the area under the receiver operating characteristic (ROC) curve, was 0·887 (0·845–0·922). The 25% test ECGs with the lowest error estimates had an area under the ROC curve as high as 0·995 (0·982–1·000), i.e. almost all of these ECGs were correctly classified. Neural networks can therefore be trained to diagnose myocardial infarction and to signal when the advice is given with great confidence or when it should be considered more carefully. This method increases the possibility that artificial neural networks will be accepted as reliable decision support systems in clinical practice.

Keywords: computer-assisted electrocardiography, diagnosis, myocardial infarction, neural networks (computer).

Introduction

Computer-aided interpretation of electrocardiograms (ECGs) was first studied by Pipberger *et al.* (1961). In the early 1980s, the first commercially available programs for the analysis of the 12-lead ECG were presented by Marquette Electronics Inc. and soon thereafter by Siemens Elema AB. Three different interpretation techniques have been used in interpretation programs. In the beginning, statistical methods were most commonly used. During the last two decades, a method using ECG criteria translated into computer programs has become the most used. This method is called a 'deterministic' method. During the 1990s, the third technique, artificial neural networks, has been studied for ECG interpretation. Artificial neural networks achieves their performance by learning from examples. Hedén *et al.* (1996a, 1997) have shown that artificial neural networks usually perform better than deterministic methods and as well as experienced physicians. Artificial neural networks have recently been incorporated in a commercially available ECG interpretation program.

Today, computer-aided interpretation of ECGs is widespread. An estimated 300 million ECGs are recorded each year, and most of these are interpreted by computerized electrocardiographs. These computer interpretations can support physicians in situations where more experienced colleagues are not present. However, many physicians do not rely upon computer interpretations even though the computer has proved to be highly accurate. The best ECG interpretation programs perform almost as well as human experts (Willems *et al.*, 1991). One reason for the hesitation is probably that the computerized electrocardiographs present advice without information about the confidence of the advice. In contrast, a colleague can give advice with different degrees of confidence – 'this is a typical pattern of myocardial infarction' or 'this is an unusual ECG pattern which I believe is due to left ventricular hypertrophy'. This type of information makes it easier to rely on the colleague than the computer.

Interpretation programs of today present probability estimates, for example 'probable left ventricular hypertrophy' and 'possible myocardial infarction'. These probability estimates describe the uncertainty that arises when there is overlap between different diagnostic groups. For example, ECGs with small R waves in the leads V_2–V_4 can be found both in normal subjects and in patients with anterior myocardial infarction. Therefore, a statement such as 'possible anterior myocardial infarction' is justified even though this is a very common ECG pattern.

The problem is that the interpretation programs present dogmatic advice for unusual ECG patterns. This confidence problem, which is addressed in this study, can be illustrated by Fig. 1. The first ECG (Fig. 1a) represents a common pattern among patients with healed anterior myocardial infarction. There are both Q waves and ST–T changes in many of the anterior leads. In the second ECG, the Q waves in leads V_2 and V_3, and the small R wave in lead V_4 are accompanied by ST–T patterns not typical for anterior myocardial infarction (Fig. 1b). The ECG would fulfil most Q-wave criteria for anterior myocardial infarction but the ECG appearance could also be due to left ventricular hypertrophy, for example. Therefore, a less dogmatic interpretation than 'definite anterior myocardial infarction', would be justified.

A reliable computerized electrocardiograph should present interpretations with probability estimates but also signal when the advice is given with great confidence or when it should be considered more carefully. It is well known that artificial neural networks can present probability estimates given that a number of conditions are fulfilled (Richard & Lippman, 1991). The most important requirement is that the training cases are good representatives of the actual use or test conditions. The problem is that such a test ECG may be different from the ECGs used in the training phase. In other words, even though network outputs in principle have probabilistic interpretations, in reality this may not be the case due to mismatch between training and test sets. For this reason, one often needs to augment the procedure with some estimate of how close to the training set a test data point is located.

The purpose of this study was to develop a method to validate the advice of a computer by estimating the error of an artificial neural network output. The method, which is of a general nature, was applied in an electrocardiographic classification task.

The type of confidence assessment presented in this study is different from the probability estimates presented by the interpretation programs of today. In order to illustrate this difference, the probability estimates of a neural network and a widely used set of rule-based criteria were also studied.

Methods

Study population

A total of 1249 ECGs from the Bowman Gray School of Medicine Data Base (Pahlm *et al.*, 1991) recorded on patients who had undergone diagnostic cardiac catheterization were studied. The ECGs were used to train and test an artificial neural network for the diagnosis of healed anterior myocardial infarction. Anterior myocardial infarction was defined by the presence of 75% diameter stenosis of the left main coronary artery, the left anterior descending artery or its major diagonals, and akinesia or dyskinesia of the anterior–superior wall on the right anterior oblique ventriculogram. Patients with normal coronary arteries, normal contrast left ventriculogram, no evidence of valve dysfunction or congenital heart disease, ejection

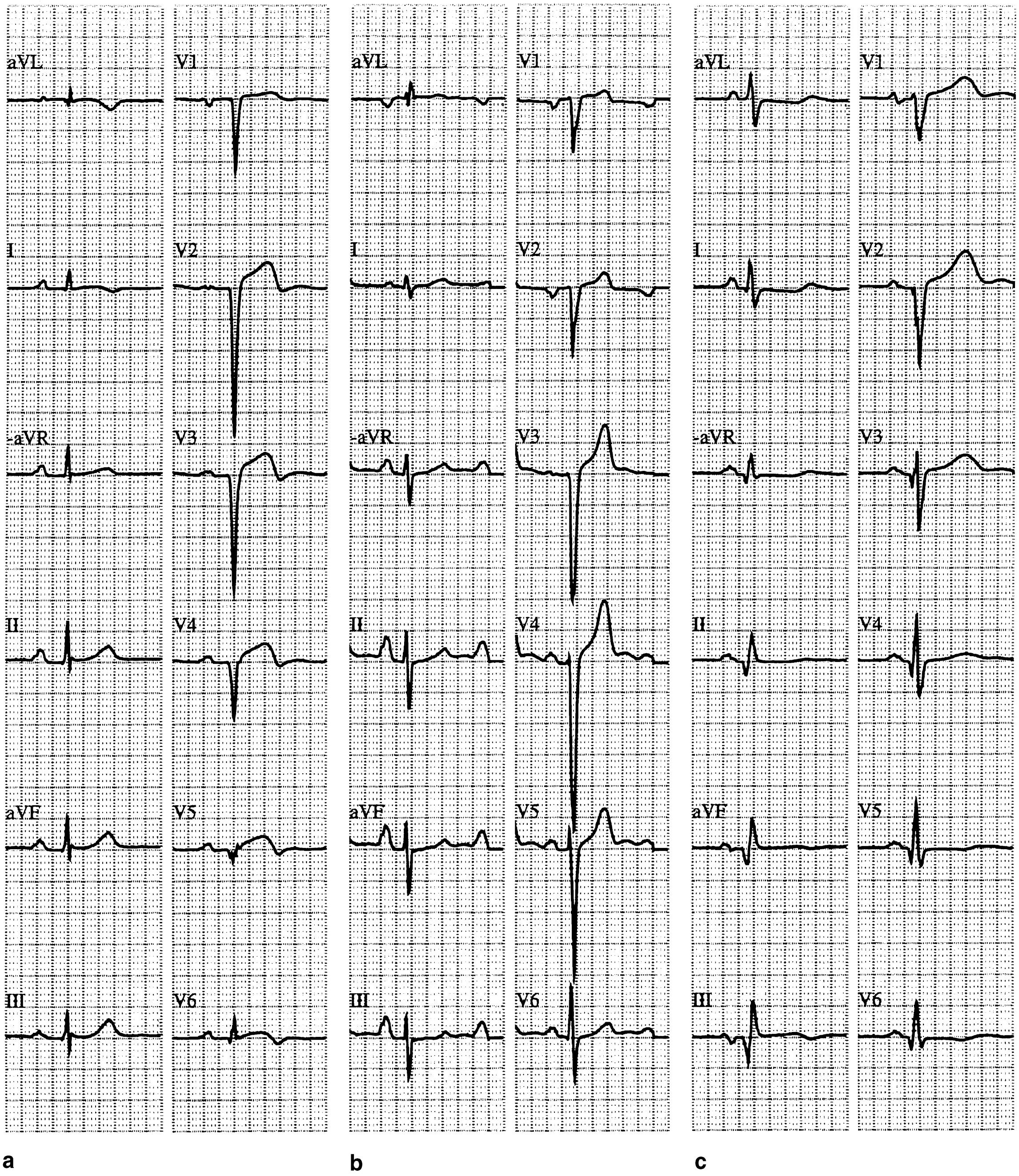

Figure 1 (a) One ECG from the infarct group and (b, c) two ECGs from the control group. The network outputs for the ECGs were 0·995, 0·979 and 0·981, respectively. The error estimate measures were low for ECGs (a) and (c), and high for ECG (b).

fraction 50%, and an overall study evaluation of 'normal' were classified as 'cath normal'. Inferior myocardial infarction was defined by the presence of 75% diameter stenosis of the right coronary artery, and akinesia or dyskinesia of the inferior wall on the right anterior oblique ventriculogram.

The differentiation of interest in the clinical situation is often whether a diagnosis, for example anterior myocardial infarction, is present or not rather than differentiation between the diagnosis and normal. Therefore, ECGs recorded on patients with anterior myocardial infarction, singly or in combination with inferior myocardial infarction, were denoted the anterior infarct group, while cath normals and patients with single inferior myocardial infarction comprised the control group. In both the anterior infarct group and the control group, ECG patterns were found which indicated other types of heart disease, for example left ventricular hypertrophy. However, only the diagnosis of anterior myocardial infarct was considered in the present study.

The material was randomly divided into three groups, one each for the training, validation and test sessions described below. The numbers of ECGs in the different groups are presented in Table 1.

The 12-lead ECGs were recorded using computerized electrocardiographs. The recording technique of the electrocardiographs was in accordance with AHA specifications. The frequency range was 0·05–100 Hz and noise reduction was performed by time coherent averaging. Averaged complexes were transferred to a computer and stored for further analysis. Measurements of amplitudes and durations of the electrocardiographic complexes were performed using an analysis program developed at our department. The definitions of measurements follow the recommendations of the CSE Working Party (1985).

Conventional ECG criteria for the diagnosis of anterior myocardial infarction use the leads V_2–V_4. Therefore, the following automated measurements from leads V_2–V_4 were used as inputs to the artificial neural network: Q, R, and S amplitudes, Q and R durations, and three amplitudes within the ST–T segment. The interval between the ST junction and the end of the T wave was divided into six segments of equal duration and the amplitudes at the end of segments 1, 3 and 5 were used as network inputs.

Table 1 Study population.

	Training	Validation	Test	Total
Anterior infarct group	145	129	140	414
Control group	287	271	277	835
Total number of patients	432	400	417	1249

Training

The 432 ECGs of the training set were used in two separate procedures. One was to train an artificial neural network and the other was to create a partitioning of the multi-dimensional data space in which any single ECG is represented by a single data point.

An artificial neural network with a multi-layer perceptron architecture (Rumelhart & McClelland, 1986) was used. A more general description of neural networks can be found elsewhere (Cross *et al.*, 1995). The neural network consisted of one input layer, one hidden layer, and one output layer. There were 24 units in the input layer, one for each of the input variables, i.e. eight measurements from each of three leads. The hidden layer contained five units. The output layer contained a single unit that encoded whether the ECG was classified as anterior myocardial infarction or not. During a training process, the connection weights between the units were adjusted using the Langevin extension of the back-propagation updating algorithm (Rögnvaldsson, 1994). The learning rate was decreased geometrically every epoch from a start value of 0·5 to an end value of 0·1. The momentum was set to 0·7. The network training was stopped when the error in the training set reached a pre-defined error threshold in order to avoid 'overtraining'. This error threshold was decided using a threefold cross validation procedure. The network weights were frozen after the training process. All calculations were performed using the JETNET 3·0 package (Peterson *et al.*, 1994).

The partitioning of the multi-dimensional data space was performed using a modified *k*-means clustering technique (MacQueen, 1967). The training ECGs were divided into groups, or clusters, such that ECGs with similar appearance were assigned to the same cluster. In this application, a 24-dimensional data space was studied, where each of the 24 ECG measurements is one dimension. A cluster is characterized by a reference point, the cluster centre, and each data point (training ECG) is assigned to the closest cluster centre. The clustering procedure was performed in the following way. First, a number of cluster centres were assigned random positions in the 24-dimensional data space. Thereafter, the positions of the cluster centres were adjusted in an iterative process in order to minimize the squared distance between data points and cluster centres. It should be noted that the number of ECGs assigned to the different clusters changes during the iterative process. The result of the clustering procedure, the partitioning of the data space, is embedded in the final positions of the cluster centres.

Validation

The 400 ECGs of the validation set were processed through the network and a validation error, i.e. the difference between the network output and the desired output, was calculated for each case. Furthermore, each case in the validation set was assigned to the closest cluster centre. The individual validation errors for all cases assigned to the same cluster were used to calculate a mean validation error for that particular cluster. Hence, the result of this session was a validation error for each of the clusters, i.e. an estimation of how reliable the network outputs were for a group of similar ECGs. A more detailed description of the validation session is presented in the Appendix.

The estimation of a validation error for a cluster must be based on a fair number of individual cases. If less than 20 validation cases were assigned to one of the clusters, the clustering procedure in the training session was repeated with a smaller number of clusters. When seven clusters were used in the electrocardiographic application, the smallest number of validation cases assigned to a single cluster was 21.

Test

In the test session, 417 ECGs were processed through the network and the resulting output values were used to assess the performance of the network. The output values of the network were in the range from 0 to 1. The desired output was 0 for a control ECG and 1 for an anterior myocardial infarction. A threshold in this interval was used, above which all values were regarded as consistent with anterior myocardial infarction. By varying this threshold between 0 and 1, a receiver operating characteristic (ROC) curve was obtained. Areas under the ROC curves were calculated as measures of performance. The 95% confidence limits of the areas were estimated by a bootstrap technique.

An error estimate was also computed for each case in the test set. The distance from a test case to each cluster centre and the validation error of the same cluster were taken into account in these calculations. The validation error of a cluster centre close to the test case influenced the error estimate more than that of a cluster centre at a longer distance. A low error estimate indicates that the network is presenting advice with great confidence whereas a high error estimate indicates that the network output should be considered more carefully.

Conventional criteria

A set of rule-based criteria for anterior myocardial infarction was adopted from the Glasgow Royal Infirmary (GRI) program. These criteria are described in the textbook 'Comprehensive Electrocardiology' (Macfarlane & Lawrie, 1989). This widely used program has been evaluated in a large European study (Willems *et al.*, 1991). The program consists of three different sets of criteria for the diagnosis of anterior myocardial infarction. By applying these criteria, each ECG was classified into one of the following four groups or probability estimates:

- 'anterior myocardial infarction'
- 'probable anterior myocardial infarction'
- 'possible anterior myocardial infarction'
- 'no anterior myocardial infarction'

The criteria were based on automated measurements from the leads V_2–V_4, and were applied to the 417 ECGs of the test set.

Results

The network outputs and error estimates of the 417 test ECGs are presented in Fig. 2. High error estimates were found for output values in the range 0–1. Only ECGs with network outputs very close to 0 or 1 had low error estimates. The performance of the neural network in the test set is presented as ROC curves in Fig. 3. The area under the curve was 0·887 (0·845–0·922). The test ECGs with the lowest error estimates were also studied separately. An ROC curve was calculated from the 25% test ECGs (n = 104) with lowest error estimates (Fig. 2). The area under this ROC curve was as high as 0·995 (0·982–1·000), i.e. almost all of these ECGs were correctly classified. The area under the corresponding ROC curve for the remaining 75% (n = 313)

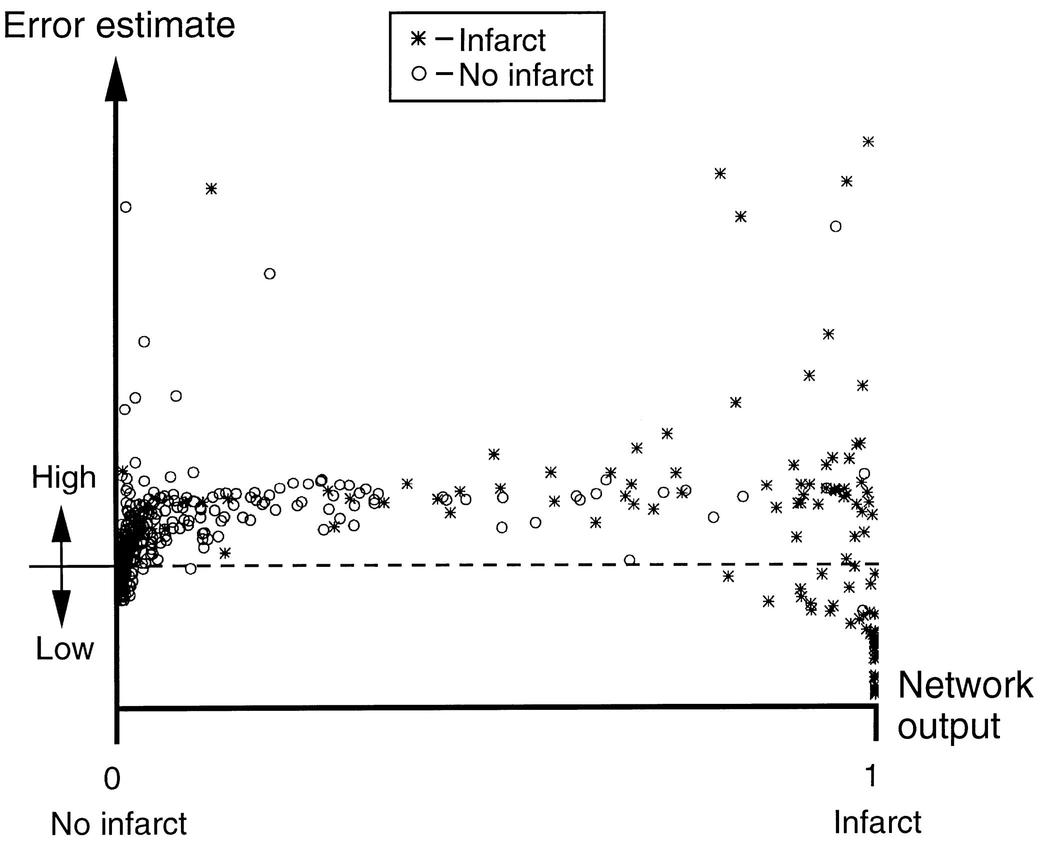

Figure 2 Network outputs and error estimates from eqn 4 (see Appendix) of the 417 test ECGs. The broken line shows where the limit between the 25% test ECGs with the lowest error estimate and the 75% test ECGs with the highest error estimate is located. A low error estimate was found in combination with network outputs close to 0 or 1.

ECGs was 0·796 (0·728–0·858). Thus the error estimates can be used to identify a subgroup of correctly classified test ECGs.

The 25% test ECGs with network outputs closest to 0 or 1 (0·00–0·01 or 0·99–1·00), i.e. those with very low or very high probability for infarction, produced an ROC curve with an area of 0·928 (0·853–0·981). This area was larger than that for the total test set but smaller than that for the ECGs with lowest error estimates, indicating that the confidence for this subgroup is only slightly higher than that for all test ECGs.

The conventional criteria were applied to the 417 test ECGs with the following results. A total of 57 ECGs of the anterior infarct group were classified as anterior myocardial infarction, while 24 ECGs were classified as either possible or probable anterior myocardial infarction and the remaining 59 ECGs were false-negative. In the control group, four ECGs were falsely classified as anterior myocardial infarction, while five ECGs were falsely classified as either possible or probable anterior myocardial infarction. A true negative classification was found in the remaining 268 control ECGs. The total accuracy was calculated as 83·7% (57 + 24 + 268)/417 in the total test set. When only test ECGs with highest or lowest probability estimates for anterior myocardial infarction, i.e. the ECGs classified as possible or probable infarction, were excluded, the total accuracy was 83·8% (57 + 268)/388. Thus, it was not possible to identify a subgroup of correctly classified ECGs by using the conventional criteria classifications.

Fig. 1(a) illustrates a correctly classified infarct ECG (high network output) with the lowest error estimates of all test ECGs. The control ECG shown in Fig. 1(b) was falsely classified as infarct by the network (high network output). This ECG also

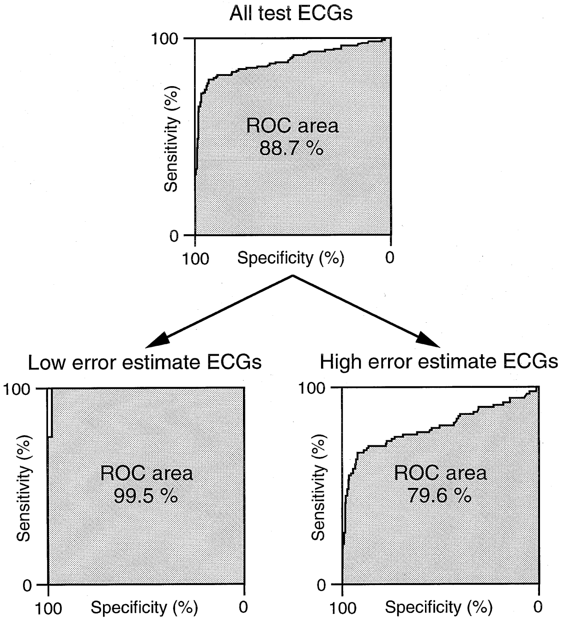

Figure 3 A nearly perfect ROC curve presenting the network performance in the group of test ECGs with low error estimates; the corresponding curves for the total test set and the group of test ECGs with high error estimates are also presented.

fulfilled the conventional Q-wave criteria. The error estimate for this example was high, indicating that the advice from the neural network should be considered more carefully. These two ECGs illustrate the advantage of the confidence method. One of the 104 test ECGs with low error estimate was falsely classified by the network (Fig. 1c). This ECG was interpreted as an anterior infarction by the network (high network output) but according to the gold standard the ECG was recorded on a patient with isolated inferior myocardial infarction.

Discussion

Main findings

The output values of an artificial neural network could be interpreted as Bayesian probabilities and translated into verbal statements such as 'definite', 'probable' and 'possible' (Hedén *et al.*, 1996b). However, output values close to 0 or 1, i.e. with a very low or very high probability for a certain diagnosis, are not always correct, the reason being that the training set is not entirely representative of the test cases. Such a condition needs to be fulfilled in order to have a probabilistic interpretation of the output signal (Richard & Lippman, 1991). These mistakes are not common but it makes it difficult to rely on the computer-based advice. The results of the present study show that computer advice can be validated by estimating the error of an artificial neural network output. ECGs with a low error estimate were almost always correctly classified.

The results of the conventional criteria show that the statements 'anterior myocardial infarction' and 'no anterior myocardial infarction' were not correct more often than the statements 'possible' or 'probable' anterior myocardial infarction. Thus, presently used interpretation programs based on rule-based criteria do not signal when an interpretation is given with great or low confidence, nor do neural networks without the error estimates.

Neural networks

ECG interpretation programs are widespread, and most of these programs use a deterministic approach, i.e. human experts construct rules or criteria. A small number of programs are based on statistical methods. Yang *et al.* (1994) have implemented artificial neural networks for the diagnosis of myocardial infarction in the GRI program. These networks improved the performance and therefore networks are now part of the GRI program. A common feature of computer-based decision support systems is the trade-off between accuracy and transparency. A small number of rules can be easy to follow but the accuracy is often not satisfactory. The performance of the rule-based ECG interpretation programs has been improved by the construction of highly complex criteria. The complexity often makes it very difficult or impossible for the ordinary user to find out the exact reasons for a diagnostic statement. Neural networks are even more of black boxes compared to the deterministic approach, but the networks have out-performed criteria in a number of applications in the medical field (Hedén *et al.*, 1994, 1997). Neural networks are well suited for pattern recognition tasks which are common in the interpretation of ECGs. Pattern recognition tasks take place intuitively in humans and therefore a very accurate but not very informative network is acceptable in many situations. The method used to validate the network outputs presented in this study will not improve the transparency of the networks but may make it easier to rely on the network. In this study, a bi-group classification task, anterior myocardial infarction or not, was used for the development and testing of the method. However, the method is also applicable to multi-classification problems, which are common in clinical practice.

Limitations of the study

The performance of a neural network depends on the size and composition of the material used for training. The method presented in this study divides the material into a training, a validation and a test set. The training set must contain sufficient cases that the *n*-dimensional data space can be divided into a number of clusters, and in each cluster a reasonable number of cases must be present in order to obtain a good estimate of the validation error. Therefore, this method could only be successfully applied to diagnostic problems where large databases are available. Automated ECG interpretation is an area where neural networks have been trained with thousands of examples (Hedén *et al.*, 1997). Even though the training and validation sessions are complicated and time-consuming, it should be stressed that a neural network and the method presented in this study will be easy to implement in computerized electrocardiographs worldwide once the method has been developed.

In this initial work, we have illustrated the feasibility of a new method, but we have not established criteria for how different error estimates could be translated into statements. We defined 'low' error estimate as those 25% ECGs with the lowest error estimate, which is not a true *a priori* definition but a reasonable part of the test group. Further studies are needed to establish how the error estimate could be used clinically.

Conclusions

Artificial neural networks have proved to perform well in pattern recognition tasks, for example in ECG interpretation. The networks can be regarded as a black box method, but the results of this study show that the network outputs can be validated by estimating the error. ECGs with a low error estimate were almost always correctly classified. This method increases the possibility that artificial neural networks will be accepted as reliable decision support systems in clinical practice.

Acknowledgements

This study was supported by grants from the Swedish Medical Research Council (K98-14X-09893-07A), the Swedish National Board for Industrial and Technical Development, the Swedish Natural Science Research Council, and Siemens-Elema AB, Solna, Sweden.

References

CSE Working Party (1985) Recommendations for measurement standards in quantitative electrocardiology. *Eur Heart J*, **6**, 815–825.

Cross S. S., Harrison R. F. & Kennedy R. L. (1995) Introduction to neural networks. *Lancet*, **346**, 1075–1079.

Hedén B., Edenbrandt L., Haisty W. K. Jr & Pahlm O. (1994) Artificial neural networks for the electrocardiographic diagnosis of healed myocardial infarction. *Am J Cardiol*, **74**, 5–8.

Hedén B., Ohlsson M., Holst H., Mjöman M., Rittner R., Pahlm O., Peterson C. & Edenbrandt L. (1996a) Detection of frequently overlooked electrocardiographic lead reversals using artificial neural networks. *Am J Cardiol*, **78**, 600–604.

Hedén B., Ohlsson M., Rittner R., Pahlm O., Haisty W. K., Petterson C. & Edenbrandt L. (1996b) Agreement between artificial neural networks and human expert for the electrocardiographic diagnosis of healed myocardial infarction. *J Am Coll Cardiol*, **28**, 1012–1016.

Hedén B., Öhlin H., Rittner R. & Edenbrandt L. (1997) Acute myocardial infarction detected in the 12-lead ECG by artificial neural networks. *Circulation*, **96**, 1798–1802.

MacFarlane P. W. & Lawrie T. D. V. (1989) Diagnostic criteria. In: *Comprehensive Electrocardiology*, Volume 3, pp. 1527–1551. Pergamon Press Inc, Oxford.

MacQueen J. (1967) Some methods for classification and analysis of multivariate observations. In: *Proceedings of the 5th Berkeley Symposium on Math Stat And Prob* (eds Lecam L. M. & Neyman J.), University of California Press, Berkeley.

Pahlm O., Haisty W. K., Wagner N. B., Pope J. E. & Wagner G. S. (1991) Specificity and sensitivity of QRS criteria for diagnosis of single and multiple myocardial infarcts. *Am J Cardiol*, **68**, 1300–1304.

PETERSON C., RÖGNVALDSSON T. & LÖNNBLAD L. (1994) JETNET 3.0: a versatile artificial neural network package. *Comp Phys Commun*, **81**, 185–220.

PIPBERGER H. V., ARMS R. J. & STALLMAN F. W. (1961) Automatic screening of normal and abnormal electrocardiograms by means of a digital electronic computer. *Proc Soc Exp Biol Med*, **106**, 130–132.

RICHARD M. & LIPPMANN R. (1991) Neural networks estimate of Bayesian *a posteriori* probabilities. *Neural Comput*, **3**, 461–483.

RÖGNVALDSSON T. (1994) On Langevin updating in multilayer perceptrons. *Neural Comput*, **6**, 916–926.

RUMELHART D. E. & MCCLELLAND J. L. (1986) *Parallel Distributed Processing*, *Volumes 1 and 2* MIT Press, Cambridge, Massachusetts.

WILLEMS J. L., ABREU-LIMA C., ARNAUD P., *et al.* (1991) The diagnostic performance of computer programs for the interpretation of electrocardiograms. *N Engl J Med*, **325**, 1767–1773.

YANG T. F., DEVINE B. & MACFARLANE P. W. (1994) Use of artificial neural networks within deterministic logic for the computer ECG diagnosis of inferior myocardial infarction. *J Electrocardiol*, **27** (Suppl.), 188–193.

Appendix

Estimating confidence limits from validation set errors and clustering

Below we present the strategy we have used for estimating confidence limits, based on defining closeness in terms of a clustering algorithm and normalizing the confidence limits by means of an error estimate for each cluster. The idea is to characterize the training data in terms of a set of K clusters and then use a validation set that has not been 'infected' by the training process, to correlate observed errors with the distance between the validation set data points and the clusters. Once this relation is established, the confidence levels of real test set data points can be estimated. The procedure is as follows.

- Divide the labelled data into a training set D_{Tr}, a validation set D_{V} and a test set D_{Te}. Each multidimensional data point (in this study measurements from an ECG) is denoted x_{i} below.
- With a pre-defined number of clusters K, assign each data point x_{i} of the training set, to a cluster centre y_a ($a = 1,\ldots,K$) using a modified K-means clustering procedure (MacQueen, 1967). In this method, cluster assignments are computed by minimizing the squared distance between data points x_{i} and cluster centres y_a, where the latter are the parameters to be determined.
- Train a network using the set D_{Tr} and then freeze the trained weights.
- Process the validation set D_{V} through the network, and, for each datapoint, record the validation error E_{k}. E_{k} is simply the absolute value of the difference between the network output and the corresponding target value for validation point x_{k}. Since the cluster assignments for each of the validation set points are known from above, one can compute the errors E_a corresponding to the different clusters y_a according to

$$E_a = \frac{\sum_{\mathrm{k}\epsilon a} E_{\mathrm{k}}}{n_a} \qquad (1)$$

where the sums runs over all data points in cluster a. n_a is the number of data points in cluster a. This is the key relation, where cluster assignment gives an error estimate. E_a is simply related to the confidence limit CL_a for cluster a through the definition

$$CL_a = \frac{t_{95} E_a}{\sqrt{n_a}} \qquad (2)$$

in the case of 95% confidence level, where $t_{95} = 1{\cdot}96$

- The next step is to define a probability $P_{\mathrm{i}a}$ that datapoint x_{i} belongs to cluster centre y_a. $P_{\mathrm{i}a}$ obviously must fulfil the normalization condition $\Sigma\ _a\mathrm{P}_{\mathrm{i}a} = 1$. $P_{\mathrm{i}a}$ is given by

$$P_{\mathrm{i}a} = \frac{\mathrm{e}^{-(x_{\mathrm{i}}-y_a)^2/T}}{\sum_b \mathrm{e}^{-(x_{\mathrm{i}}-y_b)^2/T}} \qquad (3)$$

The parameter T governs the degree of fuzziness for the probability. For a large T, all clusters are equally probable for a given data point ($P_{\mathrm{i}a} = 1/K$). On the other hand, in the limit $T \to 0$, an either/or situation is obtained ($P_{\mathrm{i}a'} = 1$ and $P_{\mathrm{i}a} = 0, a \neq a'$). In this study, $T = 1$ is used.

- Finally, one can compute CL_l for a given data point l in the test set as follows:

$$CL_l = \sum_a P_{la} CL_a \qquad (4)$$

where P_{la} has been computed using the cluster centres y_a and the formula given above. CL_l is termed the error estimate measure in the text.