TOWARD IMPROVING EXERCISE ECG FOR DETECTING ISCHEMIC HEART DISEASE WITH RECURRENT AND FEEDFORWARD NEURAL NETS

Georg Dorffner
Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Vienna, Austria
georg@ai.univie.ac.at

Ernst Leitgeb, Heinz Koller M.D.
GTZ, A-3101 St. Pölten, Austria

Abstract. This paper reports about a study evaluating the usefulness of neural networks for the early detection of heart disease based on ECG and other measurements during exercise testing [10]. Data from 350 persons who underwent stress tests consisted of patient demographic data and fifteen time frames of measurements during stress and rest. Three different neural networks, two recurrent and one feedforward using background knowledge for preprocessing, were trained and compared to the performance of skilled cardiologists. It could be shown that the best neural networks can compete with experts in classifying tests as CAD (coronary artery disease) or normal. What concerns an index value expressing the likelihood of disease, to be used for monitoring the success of treatments, the neural networks outperformed classical statistical techniques published previously. This study has thus shown large evidence in favor of using neural nets to improve the exercise ECG as a non-invasive technique for detecting heart diseases.

THE APPLICATION

The electrocardiogram (ECG) is the recording of voltage changes transmitted to the body surface by electrical events in the heart muscle, providing direct evidence of cardiac rhythm and conduction and indirect evidence of certain aspects of myocardial anatomy, blood supply and function. Electrocardiography has been used for many years as a key non-invasive method in the diagnosis and early detection of ischemic heart disease (coronary artery disease, or CAD), which is the leading cause of mortality in Western countries [5,6].

To improve the accuracy of the electrocardiogram and obtain more information on the dynamic state of the heart, exercise testing was introduced [5,11]. During stress testing not only the electrogardiogram is continously registered but also other physiological parameters are monitored (blood pressure, physical symptoms and angina pectoris). According to different established protocols, the workload is increased step by step and the changes of parameters during stress and recovery are recorded and analysed. Skilled cardiologists achieve 65–75% specificity (correctly classified normals) and 75–85% sensitivity (correctly classified CAD cases) in detecting CAD based on the resulting data [5,6].

In patients with suspected angina pectoris, exercise testing may confirm the diagnosis of ischemic heart disease and indicate the severity and prognostic importance of coronary artery lesions. In patients with definite ischemic heart disease, the exercise test is used to follow the progression or regression of the disease and the effect of therapy including drugs, invasive cardiology (e.g. angioplasty, atherectomy,..) or coronary artery surgery. Following myocardial infarction, exercise testing is performed to allow risk stratification, patients identified as being at low risk for death or re-infarction can be reassured and those at high risk can be managed appropriately [6].

If contra-indications (e.g. in the presence of acute, severe illness) are strictly observed, stress testing is a safe, cheap and non-invasive method, and is widely used in hospitals, by cardiologists, and general practitioners in primary care and health care centers. The success of the test is widely determined by the skill of the observer (cardiologist, general practitioner,..) and the patients themselves. Several efforts have been made to minimize these effects [7]. The following list shows a short summary of how automatic methods of CAD detection could improve the value of ECG and stress testing as indicator for heart diseases:

- automatic methods could minimize inter- and intraobserver variability on the test
- they could generally improve the detection of diseases like CAD
- they could contribute to improved monitoring of different therapies
- they could select continously new information on a given data set
- they could improve the acurracy of unskilled observers

Previous approaches to such improvements, such as [2,4,7,9], concentrated on classical statistical techniques and yielded results of up to 79 % sensitivity and 76 % specificity. In this paper we report about studying artificial neural networks with respect to their ability for such improvements. In particular, if neural networks prove to be able to (objectively) classify cases comparably to (partially subjective) expert performance, and if they can provide tools for monitoring therapies, they can be viewed as valuable tools for future diagnostic systems in this domain. As the results below show, neural networks indeed prove to be able to do so.

THE DATA

The data used in this study consisted of patient-demographic parameters and fifteen frames of measurements from stress testing. The former included the person's sex, age, weight, and size, an indication whether a prior infarction is known, the workload that was reached by the person, the duration of the phase of the highest workload, and the expected heart rate, as well as workload to be achieved, computed according to [12,13]. The latter consisted of the above-mentioned measurements — namely heart rate, systolic and diastolic blood pressure, physical symptoms, angina pectoris, and features extracted from the ECG such as ST-segment depression and rhythmic anomalities. These measurements were taken during 11 stress phases (from 0 to 250 W, incremented by 25 W at each phase) and 4 subsequent rest phases (immediately after stress, and after 1, 3, as well as 5 minutes).

Data from 350 persons was available, including 107 normals and 243 with coronary artery disease, ranging from single to three vessel diseases. Among the 107 normals, data from 31 athletes were included. As compared to the other normals, these constitute "ideal normals," since all other persons undergoing stress testing were at least suspected of CAD and thus had a non-negligible prior probability for the disease. This is a well-known problem in using techniques like neural networks that rely on available data material. In many cases, normals are too similar to the pathologicals to permit clean separation. Non-evasive stress testing, on the other hand, can without risk be applied to persons with a negligible prior probability for the disease. The following table depicts the distribution of all cases, including a distinction according to the persons' sex:

	total	females	males
athletes	31	2	29
other normals	76	33	54
1 vessel CAD	60	16	44
2 vessel CAD	80	13	67
3 vessel CAD	103	14	89

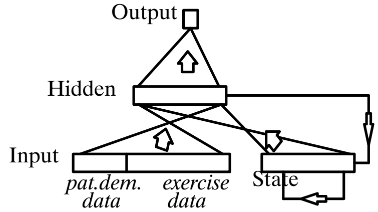

Figure 1

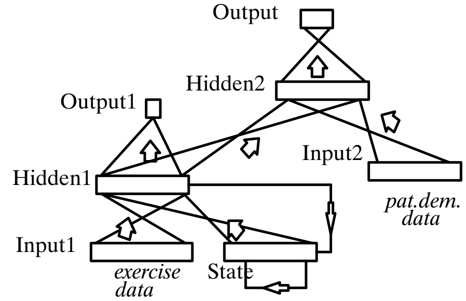

Figure 2

THE NEURAL NETWORKS USED

The task of this study was to evaluate the ability of neural networks to indicate coronary artery disease based on the data described above. Three types of neural network were used. The first and the second network were recurrent (roughly Elman-type) networks to account for the fact that the fifteen frames of stress test measurements form a time series with temporal evolution of all parameters. The third was a multilayer perceptron applied to preprocessed data (using knowledge about the domain).

Neural network 1: The first attempt of applying a neural network to the data was a somewhat “blind” training using only little background knowledge about the domain. An input layer of 19 units was used encoding patient demographic and stress test data for each time frame. This layer fed a recurrent network, somewhat similar to [3], as depicted in figure 1. At each update step through the network, the hidden layer activations were fed back to a state layer of the same size with weighted but fixed one-on-one connections. Each unit in this state layer was connected onto itself with a fixed weight. State and input layer together formed the input for the hidden layer, which in turn spread activation to an output layer of one unit. Aside from the feedback via the recurrent connections the network was considered as a multilayer perceptron and thus trained by backpropagation at each of the fifteen time frames. The target for the output unit was chosen 1 for pathological cases and 0 for normal ones. To account for the temporal evolution during the fifteen time frames, the target for pathological cases was continuously raised from 0 to 1 between the start of the sequence and the last stress phase reached by the patient (i.e. the highest workload successfully passed). After that, for the remaining time frames, it was clamped at 1. A similar method for classifying sequences has been suggested elsewhere (e.g. [1]).

The hidden layer size was varied between 8 and 20. The weights on the one-on-one connections between hidden and state layers were kept fixed at 1, the weights of the state layer units onto themselves were all set at 0.5 (thus preserving 50 % of the previous history at each time frame in the sequence — compare [16]). Roughly half of the 350 cases (173) were chosen as a training set that was kept fixed for all training runs reported in this paper. It was chosen such that the distribution of athletes vs. other normals vs. pathologicals was roughly the same for training and test set, and such that no significant difference in the distribution of the patient-demographic parameters occured between training and test sets. Other than that, the selection was random. Each training run consisted of between 60,000 and 100,000 presentations of single cases (each consisting of the full fifteen time frames), picked randomly from the training set, with a learning rate of 0.01, and of between 60,000 and 100,000 further presentations with a learning rate of 0.001. This simple schedule of lowering the learning rate had proved sufficient for reaching convergence in several preliminary training runs, and was also fixed for all runs reported here. In addition, a momentum term (according to [15]) with scaling factor 0.9 was used.

One problem with this blind application of a recurrent network might be the over-representation of patient demographic data, which did not change during the temporal sequence. Thus, in several variations of this network scheme, the input units corresponding to this part of the data were activated only either at the beginning time frame, the final two time frames, or at both such ends of the sequence, while being clamped at 0 for the other time frames.

Neural network 2: To solve that possible problem of over-representation of patient demographic data in a more elaborate way, a second network architecture was devised and tested. It consists of two modules explicitly separating the data changing over time from the time-independent data, depicted in figure 2. The first module is another recurrent network as described in the previous section, but which was only fed with the time-changing stress test data. The second module is a multilayer perceptron with two input layers — a layer encoding the patient demographic data similar to above, and the hidden layer of the recurrent network after complete update cycles through the sequence. Training consisted of two phases — first of training the recurrent network as above, and secondly, of training the multilayer perceptron by backpropagation.

In addition, three output units instead of one were used encoding the more detailed cases of normal (all units 0), one, two or three vessel disease (first, first and second, or all three units active at 1). For evaluation, still only the distinction between normal and CAD was considered. The expected effect of the two additional units was improved discriminability through the extra information in the target (this was reported previously as improving network performance, e.g. [14]). Both hidden layer sizes were varied between 10 and 20.

Since many patients could not finish the stress test up to the highest workload (which itself is a certain indicator for CAD), many time frames consisted of zero measurements. Thus, in a further extension, the sequential update of the recurrent network was adjusted such as to skip those null frames, making the length of each sequence variable.

Neural network 3: The third attempt at a neural network solution involved an additional amount of background knowledge, which was mainly used to preprocess the data. The major difference to above was that no longer a recurrent network, but instead a multilayer perceptron with three input layers was used. The information in the time sequence was explicitly encoded by making use of previous methods of arriving at an indicator for CAD from the same kind of data [8]. There, each time frame was evaluated separately, and the contributions (basically a weighting of several factors considered as possible single indicators for CAD) of all time frames were summed. For the computation of a final index, which can be shown to highly correlate with CAD (see also below) only those sums were used. In addition, an explicit distinction between stress and rest phases was made.

According to these expert decisions, the third neural network was fed with the sums of the following indicators (taken from [5,6]; as in [8]):

- a deviation of the change in heart rate from a given tolerance interval
- a decrease in systolic blood pressure
- the presence one of several critical phsysical symptoms
- the presence of angina pectoris
- ST-segment depression
- the presence one of several critical rhythmic anomalities

In distinction to [8], the first two were included as scaled values, instead of binary decisions about their presence. Furthermore, the following informations were also included [5,6]:

- a decrease in diastolic blood pressure
- pathological systolic blood pressure (larger than 140)
- pathological diastolic blood pressure (larger than 90)

again as scaled values. This was done separately for the stress and rest phases, leading to the activations of two of the three input layers. The third input layer encoded the patient demographic data as above. While many of these indicators were also used for network 2, here they were specifically tuned according to literature and, above all, explicitly summed up (rather than accumulated in the recurrent network).

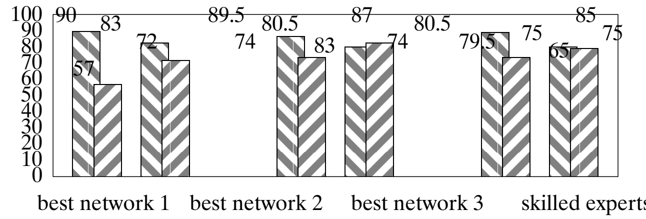

Figure 3

Each training step consisted of one presentation of input patterns and one learning cycle with backpropagation. The hidden layer size was varied between 10 and 20. Again three output units were used.

THE RESULTS

In this study the neural networks were evaluated against two criteria:

(1) their ability to correctly classify cases into CAD and normal.

(2) their ability to produce an index expressing the likelihood of disease, which can be used to monitor the success of treatments (a decreasing index after treatment would indicate less likelihood of CAD and thus success of treatment).

Figure 3 shows an overview of the results concerning criterion (1). It depicts

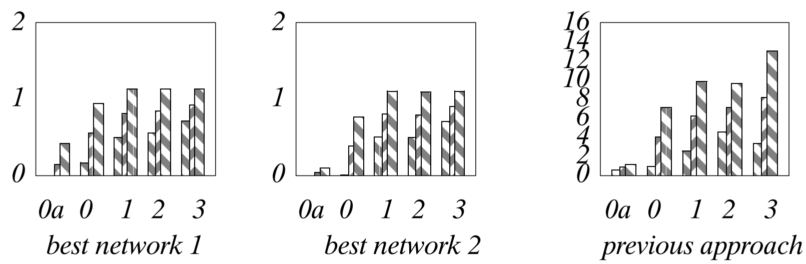

Figure 4

the best performances of the three networks, drawn as sensitivity (correct positives — white bars) and specificity (correct normals — black bars) in percentages. Since through varying the decision threshold at the output unit these two values can be changed, for all results two pairs of values are depicted — one with relatively high sensitivity (always the result at the default threshold 0.5), the other with relatively high specificity (such that sensitivity stays above 80 %). For comparison, the range of the best performances of skilled cardiologists in interpreting the same data is shown (the two pairs of values corresponding to worst and best performance, i.e. the 75/65 % and 85/75 % mentioned above).

Concerning criterion (2), the original output value (which is simply compared to a threshold for the former criterion) appears to be usable as an index expressing the likelihood of disease. To demonstrate this, in figure 4 the mean (black bars) and standard deviations (white bars) of the output value for the five classes *athletes (0a)*, *other normals(0)*, *one*, *two*, and *three vessel disease* are shown. In the case of three output units the activation values of all units was averaged. This depiction shows a significant correlation between the index produced by the network and the extent of the disease. For comparison, the same five ranges (although on a different scale) are shown for a previously published statistical method for computing such an index [8].

DISCUSSION

The results show that neural networks can reach the upper ranges of expert performance, in some cases they can even perform slightly better. The second recurrent network using less background knowledge than the feedforward network but with the ability to exploit the time series based on the training data achieved best performance, although closely followed by the feedforward network. Neural networks 1 and 2 could also outperform previous non-neural approaches [2,9].

With respect to an index for monitoring the success of treatment, neural networks appear superior to traditional statistical methods. Standard deviations are smaller and the separation between normals and pathologicals involves fewer overlap.

CONCLUSION

In this paper we have demonstrated the usefulness of neural networks in early detection of heart disease based on measurements during exercise testing. Recurrent networks which can exploit temporal dependencies appear as the best solution at the moment. Future research will investigate the combination of the recurrent approach with the type of background knowledge used in the feedforward case (e.g. through initialization), and the use of neural networks in hybrid neural/rule-based diagnostic systems. The results so far show great promise for significant contributions to making non-invasive ECG measurements during stress testing a prominent method for detecting one of today's most fatal diseases.

ACKNOWLEDGMENTS

The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry of Science and Research. We thank the Economic Chamber of Lower Austria for supporting this research, designed to lead to improved software systems.

REFERENCES

[1] Anderson S., Merrill J., Port R.: Dynamic Speech Categorization with Recurrent Networks, Indiana University, Computer Science Dept., Techn. Report No. 258, 1988.

[2] Deedwania P.C., Joshi B., Cabajal E.V.: Analysis of Temporal Electrical Heterogeneity by a New 22 Lead ECG System Accurately Identifies Patients with Coronary Artery Disease, suppl. to *Circulation* 84, no. 4, 1991.

[3] Elman J.L.: Finding Structure in Time, *Cognitive Science* 14(2)179-212, 1990.

[4] Greenberg P.S., Cangiano H., Leamy L., Ellestad M.H.: Use of the multivariate approach to enhance the diagnostic accuracy of the treadmill test, *J. Electrocardiology* 63, pp.987-1000.

[5] Hurst W.: *The Heart*, 7th ed., New York: McGraw Hill, 1990.

[6] Julian D.G., Camm A.J., Fox K.M., Hall R.J.C., Poole-Wilson P.A.: *Dieseases of the Heart,* London: Bailliere Tindall, 1989.

[7] Kligfield P., Ameisen O., Okin P.: Heart rate adjustment of ST-segment depression for improved detection of coronary artery diesease, *Circulation* 79, pp. 245-255, 1989.

[8] Koller H., Leitgeb E.: Ist eine Verbesserung der Aussagekraft und Reproduzierbarkeit des Belastungs-EKG erzielbar?, *Wiener Medizinische Wochenschrift* 143(5), pp.110-117, 1993.

[9] Mark D.B., et al.: Prognostic Value of a Treadmill Exercise Score in Outpatients with Suspected Coronary Artery Disease, *New England Journal of Medicine* 325 (12), pp.849-853, 1991.

[10] McNeer J.F., Margolis J.R., Lee K.L., et al.: The role of exercise test in the evaluation of patients for ischemic heart disease, *Circulation* 57, pp.64-70, 1978.

[11] Mellerowicz H., Maidorn K., Matzdorff F., Nowacki P., Rittel H.F., Schmutzler H., Schön F.A., Stoboy H., Waterloh E., Zapfe H.: *Ergometrie*, München/Wien/Baltimore: Urban & Schwarzenberg, 1979.

[12] Niederberger M.: Grundlagen der Ergometrie, *Österreichische Ärztezeitung* 4., pp. 21-36, 1978.

[13] Niederberger M.: Belastungsuntersuchungen in der Kardiologie, *Herz* 7, pp.1-19, 1982.

[14] Prem E., Mackinger M., Dorffner G., Porenta G., Sochor H.: Concept Support as a Method for Programming Neural Networks with Symbolic Knowledge, in Ohlbach H.J.(ed.), *GWAI-92: Advances in Artificial Intelligence*, Berlin: Springer, Lecture Notes in AI, Vol.671, 1993.

[15] Rumelhart D.E., Hinton G.E., Williams R.J.: Learning Internal Representations by Error Propagation, in Rumelhart D.E., McClelland J.L.(eds.), *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*, Vol 1: Foundations, Cambridge, MA: MIT Press, 1986.

[16] Ulbricht: Multi-reccurent networks for traffic prediction, to appear in *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Seattle, 1994.