Using neural networks and just nine patient-reportable factors to screen for AMI

A. M. Bulgiba and M. H. Fisher

The study investigated the effect of different input selections on the performance of artificial neural networks in screening for acute myocardial infarction (AMI) in Malaysian patients complaining of chest pain. We used hospital data to create neural networks with four input selections and used these to diagnose AMI. A 10-fold cross-validation and committee approach was used. All the neural networks using various input selections outperformed a multiple logistic regression model, although the difference was not statistically significant. The neural networks achieved an area under the ROC curve of 0.792 using nine inputs, whereas multiple logistic regression achieved 0.739 using 64 inputs. Sensitivity levels of over 90 per cent were achieved using low output threshold levels. Specificity levels of over 90 per cent were achieved using threshold levels of 0.4-0.5. Thus neural networks can perform as well as multiple logistic regression models even when using far fewer inputs.

Keywords
acute myocardial infarction, decision support, diagnosis, input variable selection, Malaysia, neural networks

Introduction

Acute chest pain in the adult is a frequently encountered symptom in all healthcare settings [1] and ischaemic heart disease is one of the leading causes. It is also one of the dominant causes of morbidity and mortality in the western world [2] and is increasingly becoming a leading cause of mortality and morbidity in the developing world [3]. In recent
years, there has been renewed interest in computer-aided decision support systems in ischaemic heart disease for reasons of cost and to reduce the number of unwarranted admissions for chest pain. There are a number of published articles on the use of artificial intelligence (AI) in diagnosing chest pain. Work by Kukar et al. [4] and in the Heart Disease Program [5] have contributed much to the use of AI techniques in chest pain diagnosis. The majority of this research has concentrated on the use of neural networks in diagnosing acute myocardial infarction (AMI) and almost all have used biochemical markers and ECG findings in concert with the history and physical examination. For example the work by Ellenius and Groth [6, 7], Bax et al. [8, 9], Itchhaporia et al. [10] and Kennedy et al. [11] has contributed a lot to research in this area. There appears to be little work on the use of artificial neural networks on diagnosing AMI based on signs and symptoms alone. Wang et al. [12] is among the small minority of researchers who have tried to build neural networks capable of AMI diagnosis without the need for any ECG and biochemical markers. Such systems would be of great benefit as a first-line decision support tool when ECGs may not be sensitive enough to pick up AMI and where laboratory tests are not available. An example of this kind of situation is when patients are first seen by paramedics.

Artificial neural networks (ANNs) can be described as methods of representing functions using networks of simple arithmetic computing elements and methods which are able to perform learning such representations from examples [13]. An ANN is composed of a set of neurons or nodes $X_i$ each transforming its total or net input according to an activation function (or transfer function). Each node sends its output to other units through connections (or links) each having a certain effectiveness or weight. The net input to any unit is usually modelled as a sum of all the outputs from other units, weighted by the weights of the respective connections.

Neural networks are often used for statistical analysis and data modelling. They are typically used in problems that may be couched in terms of classification or forecasting and for pattern recognition [14]. There are basically three types of neural network architecture: feedforward networks, recurrent networks and self-organizing maps. Feedforward networks consist of a single or several layers of neurons arranged in sequence. The multilayer perceptron (MLP) is a feedforward network with one or more hidden layers and one output layer. The MLP was first invented by Bryson and Ho in 1969 but it was not until the mid 1980s that it came into prominence [13]. It is a technique commonly used by many researchers in medical diagnostic tasks and has been used by a number of researchers in AMI diagnosis. In this study we wanted to assess the discriminatory power, sensitivity and specificity of MLPs in diagnosing AMI without the use of ECG or lab findings. This has the potential of being incorporated in a screening tool that could be used by paramedics or even the public.

We present work done with neural networks using different input selections in screening for AMI. The performance of the neural networks is benchmarked against that achieved by multiple logistic regression.
Methods

Source of data
We obtained the dataset from Selayang Hospital, a tertiary level hospital in Malaysia. Permission to use these data for this study was obtained from the Ministry of Health, Malaysia. All records of adult patients (18 years or older) seen in the Emergency Department for non-traumatic chest pain from 20 August 1999 (when the hospital opened) to 9 August 2002 and clerked using the chest pain clerking form were selected for this study.

Data cleaning and preprocessing
Data cleaning and preprocessing were performed before we constructed our neural networks. This involved accuracy checking, treatment of missing values, recategorization and recoding of fields. The proportion of missing values was not high (between 1 and 5%) and it did not seem to unduly affect our experiments. We split the data into two sets: a validation set and a cross-validation set. The validation set consisted of 177 records and was used to determine the number of hidden neurons in our neural nets. It was also used as a validation set for early stopping during our cross-validation experiments. The cross-validation set contained 710 records and was used for training and testing the optimal MLP architecture once this had been determined in the earlier experiments.

Diagnostic criteria
We used the diagnosis on discharge as the definitive diagnosis. This is the diagnosis as confirmed by specialist physicians after taking into consideration ECG readings and other laboratory investigations.

Input variable selection
Only patient-reportable and examination variables or inputs were used in this study. Four possible combinations of inputs were used in these experiments. These four combinations were chosen to see whether having fewer inputs would have the same performance as having more inputs or having all inputs. Fewer inputs would theoretically make it easier to construct a quick screening tool for use by paramedics in the field to screen for AMI. The four types of input combinations used were:

1 All inputs selection (94 inputs). This consists of all history and examination inputs (Table 1). There were originally only 64 inputs but the number increased from 64 to 94 due to some necessary transformation and recoding.

2 Expert selection I (9 inputs). We performed multiple logistic regression (MLR) analysis on the dataset using all 64 inputs. The most significant inputs from the MLR analysis was designated as the expert selection I [15]. The input variables for this selection were sex (male), race (Indian), sudden onset of pain, persistent pain, crushing pain, pain that was relieved by other means, associated sweating, history of diabetes mellitus, and a history of heart disease on treatment. None of the examination inputs made it into this selection, thus making this selection independent of physical examination.
<table>
<thead>
<tr>
<th>Table 1</th>
<th>Input variables</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group</strong></td>
<td><strong>Fields</strong></td>
</tr>
<tr>
<td>Demographic</td>
<td>age, citizen, race, sex, marital status</td>
</tr>
<tr>
<td>Nature of chest pain</td>
<td>location, onset, pattern, quality</td>
</tr>
<tr>
<td>Radiation of pain</td>
<td>jaw, left arm, laterally, neck, locally, other parts</td>
</tr>
<tr>
<td>Relieving factors</td>
<td>leaning forward, sitting up, GTN, rest, other means</td>
</tr>
<tr>
<td>Aggravating factors</td>
<td>posture, meals, coughing, inspiration, exertion</td>
</tr>
<tr>
<td>Associated heart/lung symptoms</td>
<td>cough, dyspnoea, oedema, orthopnoea, palpitations</td>
</tr>
<tr>
<td>Other associated symptoms</td>
<td>collapse, headache, dizziness, fever, numbness, nausea, sweating, vomiting, fainting</td>
</tr>
<tr>
<td>Cardiac risk factors</td>
<td>age &gt; 40, diabetes mellitus, family history, hypertension, physical inactivity, obesity, smoking, known case defaulted treatment, known case on treatment, high cholesterol levels</td>
</tr>
<tr>
<td>General examination</td>
<td>pulses, pulse rate, respiratory rate, systolic BP, diastolic BP</td>
</tr>
<tr>
<td>Heart/lung examination factors</td>
<td>air entry, breath sounds, chest expansion, chest wall, crepitations, heart sounds, JVP, percussion, pleural rub, praecordium, rhonchi</td>
</tr>
<tr>
<td>Other examination factors</td>
<td>abdomen, central nervous system (CNS), eye, face</td>
</tr>
</tbody>
</table>

3. **Expert selection II (23 inputs).** These consist of inputs found by univariate analysis to be significantly related to AMI [15]. The input variables were age, sex (male), race (Indian), retrosternal pain, right-sided pain, sudden onset of pain, persistent pain, crushing pain, pain that was relieved by other means, associated palpitations, associated nausea, associated sweating, associated vomiting, pain relieved by rest, pain aggravated by posture, cough, inspiration and exertion, history of diabetes mellitus and a history of heart disease on treatment, being a smoker, and abnormal chest wall and face examination. All the inputs in this case were patient-reportable or history factors except for two inputs (abnormal chest wall and face examination).

4. **PCA selection (11 inputs).** This was selected by principal component analysis (PCA). Only inputs that contributed to at least 2 per cent of the variation in inputs were selected.

**Determination of hidden layer**

The multilayer perceptron (MLP) was used for all neural net experiments. We used the MLP as we believe that it is a relatively simple technique to use and it would be easier to justify to medical professionals how MLPs worked rather than an unsupervised method, especially when some form of life-saving decision needs to be made. We used one input layer, one hidden layer and one output layer (this is considered as a two-layer MLP as the input layer is not counted). All MLP algorithms were tested with 2, 4, 8, 16, 32 and 64 neurons in the hidden layer. We determined the optimal number of hidden neurons by
Table 2  Optimal number of hidden neurons

<table>
<thead>
<tr>
<th>Selection</th>
<th>GDX</th>
<th>CGF</th>
</tr>
</thead>
<tbody>
<tr>
<td>All inputs</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Expert I</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Expert II</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>PCA</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

creating 20 MLPs for each architecture and input selection. An ROC (receiver operating characteristic) curve analysis [16] was performed and the architecture with the highest AUC (area under the curve) was used for further (cross-validation) experiments. We used the validation dataset for this part of the experiments. The number of optimal neurons found for each input selection is shown in Table 2.

**Cross-validation experiments**

After determining the optimal number of hidden neurons, we created an MLR (multiple logistic regression) model and a number of MLPs using a cross-validation and committee approach. The cross-validation dataset was split into 10 segments ($D_1$–$D_{10}$). Training of a network was carried out using data from nine segments and the performance was tested on the one remaining segment. This process was repeated for each of the 10 possible choices for the segment that was omitted from the training process. The partitioning of the data is illustrated in Figure 1. Such a procedure allowed us to use 90 per cent of the data for training the networks, while also making use of all data points in evaluating the cross-validation error. In order to ensure that variations in the performance of the neural networks were cancelled out, a committee of 20 networks was created for every possible combination of training/test set (run). The outputs of all the models in each run were

![Figure 1](division_of_the_cross-validation_set_into_10_segments)

**Figure 1** Division of the cross-validation set into 10 segments
averaged out and all 10 averages were combined into one, thus making an output for the entire cross-validation set. The final result of using cross-validation is that all 710 records were used for training and testing the neural networks and a total of 200 neural networks were created for each architecture. We trained all MLPs over a maximum of 1000 epochs for each combination of input and target using early stopping to prevent overfitting. Early stopping is a technique where the error of a separate validation set is monitored at the same time as the error of the training set. Typically the error of the training set will continue to decrease without stopping during training of a neural network. This leads to overfitting of the neural network, which leads to poor generalization performance, and is the biggest problem with neural networks [14]. Early stopping stops this happening by halting training when the error on the validation set starts to rise in comparison to the error on the training set. None of the MLPs ever reached 1000 epochs as early stopping prevented them from doing so. This prevented overfitting of the MLPs.

**Measure of error**

We used the cross-entropy error instead of the mean-squared error as the performance function in these MLP experiments as the mean-squared error is not appropriate for classification problems [14, 17]. We used Matlab version 6.5 with the Neural Network and Statistics Toolboxes for all experiments with the MLR model and MLP.

**MLP algorithms**

We compared two optimization algorithms:

1. **GDX**: gradient descent [18] with adaptive learning rate and momentum.
2. **CGF**: conjugate gradient descent with Fletcher–Reeves update [19].

These two algorithms were chosen after experimenting with 14 different algorithms and were decided on as they were considered to be good compromises between speed and accuracy. Some of the 14 optimization algorithms were computationally expensive with no advantage in accuracy and so were dropped from subsequent experiments.

**Statistical analysis**

ROC curves with AUC [16] were computed for the MLR model and MLPs. Sensitivity and specificity with 95 per cent confidence intervals were calculated using standard formulas [20]. Continuity correction was used for all calculations of sensitivity and specificity to avoid confidence intervals exceeding 100 per cent [21]. We used SPSS version 11 to analyse the results.

**Results**

The two highest AUCs (area under the curve) of the ROC curves are achieved by CGF (0.792; 95% CI 0.754, 0.829) and GDX (0.773; 95% CI 0.734, 0.812) with just nine inputs (expert selection 1) utilizing just two hidden neurons (Table 3). CGF is better than GDX using just nine inputs but the difference in AUC is not statistically significant at
Table 3  Area under the curve (AUC) of MLPs

<table>
<thead>
<tr>
<th>Input selection</th>
<th>GDX (95% CI)</th>
<th>CGF (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All inputs</td>
<td>0.761 (0.722, 0.800)</td>
<td>0.741 (0.700, 0.781)</td>
</tr>
<tr>
<td>Expert I</td>
<td>0.773 (0.734, 0.812)</td>
<td>0.792 (0.754, 0.829)</td>
</tr>
<tr>
<td>Expert II</td>
<td>0.762 (0.724, 0.800)</td>
<td>0.767 (0.730, 0.805)</td>
</tr>
<tr>
<td>PCA</td>
<td>0.757 (0.717, 0.797)</td>
<td>0.749 (0.709, 0.790)</td>
</tr>
</tbody>
</table>

$p = 0.05$. The expert selection II (23 inputs) also seems to work well enough for both MLP algorithms and has the next two highest AUCs, but the AUCs are very closely matched and there is little difference between them. Using all inputs or using PCA to reduce the number of inputs does not seem to work that well and the AUCs are not remarkable.

For GDX, the best curve seems to be expert selection I (nine inputs) although there is in reality little difference between this and the other selections in GDX (Figure 2). For CGF, the best curve seems to be expert I too, with this curve clearly outperforming the other selections at most points (Figure 3).

If we look at using the expert selection I (nine inputs), it does seem that CGF has the slight edge here, with its curve outperforming GDX at most points and clearly outperforming MLR at all points (Figure 4).

Figure 2  ROC curve for GDX
Displayed in Table 4 are the threshold levels at which the sensitivity levels are greater than 0.8. Above this threshold, the sensitivity drops to less than 0.8. There are differences between the various combinations but it is clear that for some algorithms, the sensitivity drop is more dramatic than others. It is clear that the expert selections I and II are as sensitive as using all 94 inputs or using PCA to reduce the number of inputs.

When we look at the specificity of the various combinations, it is clear that unlike sensitivity, specificity is not a problem for all of them (Table 5). Table 5 displays the lowest threshold at which specificity levels of 0.9 or higher are achieved. All combinations are able to achieve specificity higher than 0.9 and the specificity at these thresholds range from 0.9010 to 0.9690. What is more striking is that these specificity levels are achieved at thresholds of around 0.4 and 0.5, where the accuracy is highest.

Limitations

There are a few limitations to these experiments. These MLPs have not been tested on a real-time basis but there is no reason to think that they might not work in this kind of situation. This study also did not evaluate all potential patients with ischaemia because it was confined to patients with chest pain in the Emergency Department and thus did not deal with those suffering from silent ischaemia. Finally, this study was carried out at a single institution and we may need to corroborate with patients from different locations.
Figure 4 ROC curve for nine inputs with MLR as benchmark

Table 4 Sensitivity to AMI

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity (95% CI)</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDX23-4</td>
<td>0.9800 (0.9690, 0.9896)</td>
<td>0.2</td>
</tr>
<tr>
<td>GDX94-2</td>
<td>0.9640 (0.9495, 0.9770)</td>
<td>0.2</td>
</tr>
<tr>
<td>GDX9-2</td>
<td>0.9490 (0.9321, 0.9645)</td>
<td>0.2</td>
</tr>
<tr>
<td>CGF94-2</td>
<td>0.9240 (0.9038, 0.9428)</td>
<td>0.2</td>
</tr>
<tr>
<td>CGF11-4</td>
<td>0.8880 (0.8641, 0.9105)</td>
<td>0.2</td>
</tr>
<tr>
<td>CGF23-8</td>
<td>0.8880 (0.8641, 0.9105)</td>
<td>0.2</td>
</tr>
<tr>
<td>GDX11-4</td>
<td>0.8680 (0.8424, 0.8922)</td>
<td>0.2</td>
</tr>
<tr>
<td>CGF9-2</td>
<td>0.8430 (0.8155, 0.8691)</td>
<td>0.2</td>
</tr>
<tr>
<td>MLR</td>
<td>0.8320 (0.8038, 0.8588)</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Table 5  Specificity to AMI

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity (95% CI)</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDX23-4</td>
<td>0.9690 (0.9122, 1)</td>
<td>0.4</td>
</tr>
<tr>
<td>GDX94-2</td>
<td>0.9630 (0.9064, 1)</td>
<td>0.4</td>
</tr>
<tr>
<td>GDX9-2</td>
<td>0.9610 (0.9044, 1)</td>
<td>0.4</td>
</tr>
<tr>
<td>GDX11-4</td>
<td>0.9430 (0.8918, 0.9928)</td>
<td>0.5</td>
</tr>
<tr>
<td>CGF9-2</td>
<td>0.9300 (0.8791, 0.9795)</td>
<td>0.5</td>
</tr>
<tr>
<td>MLR</td>
<td>0.9200 (0.8807, 0.9579)</td>
<td>0.7</td>
</tr>
<tr>
<td>CGF94-2</td>
<td>0.9160 (0.8608, 0.9698)</td>
<td>0.4</td>
</tr>
<tr>
<td>CGF23-8</td>
<td>0.9140 (0.8588, 0.9678)</td>
<td>0.4</td>
</tr>
<tr>
<td>CGF11-4</td>
<td>0.9010 (0.8462, 0.9544)</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Discussion

There have been a number of published articles involving the use of MLPs in predicting AMI or IHD or just simply coronary artery stenosis. Some like work done by Lapuerta et al. [22] concentrated on inputs from lab data, while others like Xue et al. [23] used ECG data to predict coronary events. Others like Itchhaporia et al. [10, 24], Baxt et al. [8, 9, 25, 26] and Wang et al. [12] have used clinical history and examination findings to detect AMI. The results are variable and not all are directly comparable due to differences in the methodology, but at least for those where the methods and inputs are similar, the results from this research have compared quite favourably.

It would thus appear that a carefully handcrafted input selection such as expert selection I performs best with CGF and GDX. This is not really surprising as these input variables are proven by multiple logistic regression to have a clear relationship between the predictor and output variables. Adding in some noise like using the expert selection II (which is basically expert selection I plus statistically significant univariate inputs) proves to be just as good as using all inputs but not as good as using the expert selection I. The poor performance of the PCA selection (using a cutoff point of 2% variability) and the all inputs selection compared with the expert selections I and II would seem to indicate that these are not superior to handcrafted inputs. Why should this be so? One possible reason is that medical data as in this case have a tendency to have missing values and, although there are more inputs in the all inputs selection, most of the inputs are not relevant to the diagnosis and their presence or absence adds little to the relationship between predictor (input) variables and diagnosis.

One huge advantage of using expert knowledge is the ability to control the number of predictor variables in the system and thus control the amount of time necessary to train the system. Another advantage is that as more knowledge is gained about the disease, the number as well as types of input variables can be altered to reflect this new knowledge and the system can be improved with this new knowledge.

One unexpected benefit from the finding that expert selection I is the best input selection is that it does not include any examination findings and is confined to just patient-reportable factors. In other words, this would be ideal for a system that does not have any need for examination by a physician. An example of such a system would be a
web-based system for use by the public or for a paramedic in the field where early ECG findings are non-specific and where lab investigations cannot be performed.

The use of statistical models particularly MLR as benchmarks is not new and it seems to be the baseline method of choice for comparison in a number of publications on work involving neural networks for diagnosing chest pain. Great pains have been taken to demonstrate that neural networks are superior to MLR in all these publications. However, without exception the MLR has been optimized on the entire dataset and then used to predict the cases on which it has been optimized. This is not a fair comparison with the way the MLP has been used. In this research, comparisons with MLR models are on a like-for-like basis, meaning MLR models were trained and tested using the same cross-validation method used for MLPs.

It has been shown by multivariate statistical analysis using the same data that examination factors are not really important in the diagnosis of AMI [15] and angina [27]. One can just make do with the nine most significant factors to predict AMI with accuracy comparable to using all 94 inputs. Hollander et al. [28] studied the effect of having neural network feedback to physicians on decisions to admit or discharge patients with chest pain and found that it did not influence physicians' decisions because lab markers were necessary for the neural networks to work and the lab investigations take too long to complete. One therefore would tend to conclude that an MLP-based decision support system that did not rely on time-consuming lab markers would be more influential on physicians than the one Hollander et al. [28] evaluated.

Sensitivity and specificity levels achieved in this research compare quite favourably with those studies reviewed by Lau et al. [29] and that obtained by Baxt et al. [8, 9] and Kennedy et al. [11] despite the lack of ECG and lab biomarkers. Although it is possible to manipulate these sensitivity levels, like Kukar et al. [4] and Mobley et al. [30] have done, there is a tradeoff. High sensitivity levels can be achieved at low output thresholds, but this comes at a cost and the cost is low specificity. For example Mobley et al. [30] decided on a cutoff point of 0.25 in order to achieve a sensitivity of 1, but the specificity was dismally low at 0.26. It is not easy deciding on the optimum threshold level as this is a local assessment from the ROC graph [16].

The results of experiments here have shown that a low output threshold has the effect of increasing sensitivity and that this is at the expense of specificity. In order to have a sensitivity of 0.8, the output threshold in the algorithms tested here must not on the whole be higher than 0.2. However, in order to have a specificity of 0.9, the output threshold is around 0.4 or 0.5. This seems to indicate that while sensitivity is a problem as it is achieved at pretty low thresholds, specificity is not an issue. The conclusion then is that the selection of input variables as used in this research is good for ruling out AMI rather than for ruling in AMI.

A US study in 1998 reveals that the first-year direct medical cost for fatal AMI was estimated at $17,532 while the first-year direct medical cost for non-fatal AMI was $15,540 [31]. The problem of cost/benefit ratio is particularly relevant for low-risk patients and the ROMIO study found that the initial cost of admitting a patient to rule out AMI or unstable angina was significantly higher at $1349 compared to just $893 if an emergency-department-based strategy for ruling out AMI or unstable angina was used [32]. In other words, a system for early ruling out of AMI would have enormous benefits on direct medical costs. No neural-network-based decision support has been evaluated with regards to cost/benefit of early ruling out of AMI but the potential for such a system certainly exists.
Conclusion

The multilayer perceptron using only nine inputs without any ECG or lab markers can perform as well as multiple logistic regression using 64 inputs in screening for AMI. A decision support system for AMI screening based on a neural network that does not rely on time-consuming lab markers may be worth evaluating.

Acknowledgements

We would like to thank the Ministry of Health, Malaysia and the Selayang Hospital for permission to extract data from their database for this project and for allowing us to publish this article.

References

Bulgiba and Fisher  Neural networks in AMI screening


21 Deeks J J, Altman D G. Sensitivity and specificity and their confidence intervals cannot exceed 100%. *British Medical Journal* 1999; **318** (7177); 193b.


**Correspondence to:** Dr A. M. Bulgiba

---

**A. M. Bulgiba MPH MappStats**  
Department of Social and Preventive Medicine,  
Faculty of Medicine, University of Malaya,  
50603 Kuala Lumpur, Malaysia  
E-mail: awang@um.edu.my

**M. H. Fisher PhD**  
School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK  
E-mail: Mark.Fisher@uea.ac.uk