**Noname manuscript No.**
(will be inserted by the editor)

# Probabilistic classification of acute myocardial infarction from multiple cardiac markers

**Paul C. Wilson, George W. Irwin[1], John V. Lamont[2], Robert F. Harrison[3]**

[1] Intelligent Systems and Control Group
Electrical and Electronic Engineering
Queen's University Belfast
Belfast BT9 5AH
UK

[2] Randox Laboratories Ltd.
55 Diamond Road
Crumlin
Co. Antrim BT29 4QY
UK

[3] Automatic Control and Systems Engineering
The University of Sheffield
Mappin Street
Sheffield S1 3JD
UK

**Abstract** Logistic regression and Gaussian mixture model (GMM) classifiers have been trained to estimate the probability of acute myocardial infarction (AMI) in patients based upon the concentrations of a panel of cardiac markers. The panel consists of two new markers, fatty acid binding protein (FABP) and glycogen phosphorylase BB (GPBB), in addition to the traditional cardiac troponin I (cTnI), creatine kinase MB (CKMB) and myoglobin. The effect of using principal component analysis (PCA) and Fisher discriminant analysis (FDA) to preprocess the marker concentrations was also investigated.

The need for classifiers to give an accurate estimate of the probability of AMI is argued and three categories of performance measure are described, namely *discriminatory ability*, *sharpness*, and *reliability*. Numerical performance measures for each category are given and applied.

The optimum classifier, based solely upon the samples take on admission, was the logistic regression classifier using FDA preprocessing. This gave an accuracy of 0.85 (95% confidence interval: 0.78–0.91) and a normalized Brier score of 0.89. When samples at both admission and a further time, 1–6h later, were included, the performance increased significantly, showing that logistic regression classifiers can indeed use the information from the five cardiac markers to accurately and reliably estimate the probability AMI.

**Key words** acute myocardial infarction – AMI – cardiac markers – diagnostic aid – probabilistic classification

## 1 Introduction

Coronary heart disease is a narrowing of the coronary arteries which can cause a heart attack (acute myocardial infarction or AMI). This is the leading cause of premature death in the developed world, causing one quarter of all male deaths and one sixth of all female deaths in the UK [1].

An acute myocardial infarction is caused by a coronary artery becoming completely obstructed, normally due to a blood clot, and blood supply being lost to an area of the cardiac tissue. Once the cells in the heart loose their blood supply (cardiac ischemia), their membranes become more permeable and certain chemicals, previously retained within the cell, are able to pass through the cell membrane and enter the blood stream. A number of these chemicals, known as *cardiac markers*, have been found to be useful in diagnosing AMI, especially where electrocardiograph (ECG) readings are inconclusive [2]. Most hospitals will check blood samples from patients with a suspected AMI for one or two of these markers, usually creatine kinase MB isoenzyme (CKMB) or cardiac troponin I (cTnI).

If AMI is detected quickly, treatment can be administered either via drugs or surgery to minimise the effects of the infarct and keep the heart functioning effectively. The treatments, however, are both financially costly and may have side effects rendering them risky for patients who have not had an AMI. It is therefore important to rapidly and accurately assess the condition of a patient for the correct treatment to be administered.

Certain cardiac markers, most notably cTnI, are very specific to cardiac tissue damage, but may take as long as 6–12h to reach diagnostic values. Others, for example myoglobin, can be found in the blood stream at abnormal levels within 2h of an infarct occurring, but are much less sensitive to AMI: that is, elevated levels can be caused by many factors other than AMI [2].

Clinical diagnosis is based upon a combination of ECG data, clinical signs and symptoms, and cardiac marker concentrations. Various artificial intelligence pattern recognition techniques have been used assimilate some of this data to assist clinical decision making. Most have focused on using data from clinical symptoms or ECGs. Perhaps the most influential research in this area was by Goldman *et al.* [3, 4] who presented a decision tree classifier based on studies of over six thousand patients. The Goldman Protocol, as the classifier became known, used information from patient histories, clinical symptoms and ECG measurements, but only returned binary decisions.

Other techniques have mainly used either logistic regression [5, 6, 7, 8] or artificial neural networks [9, 10]. The logistic regression based ACI-TIPI algorithm [5], which attempts to diagnose AMI from computerised interpretation of ECG signals combined with the patient's age, gender and information on chest pain, forms the basis of a commercial software application by Philips Medical Systems (Andover, Massachusetts, USA) to assist in the diagnosis of AMI. Most recently, Ellenius *et al.* have developed a protocol, based primarily upon neural network analysis of sequential measurements of combinations of CMKB, myoglobin and cTnI at 30 minute intervals, which shows promising results [11, 12].

In this work, patient samples were analysed on an Evidence biochip analyser (Randox Laboratories Ltd.). This analyser automatically measures the concentrations of five key cardiac markers: the previously mentioned CKMB, myoglobin and cTnI; and the more recently discovered glycogen phosphorylase isoenzyme BB (GPBB) and fatty acid binding protein (FABP).

In contrast to other publications in this field, which typically use only one or two markers, this work has access to the concentrations of five markers simultaneously, two of which (GPBB and FABP) have not been used in any previous computer based classification system. These concentrations are used to construct classifiers which predict the probability a patient has AMI (rather than a binary 'AMI' or 'non-AMI' classification).

By giving an accurate probability of AMI, such a classifier can be used by clinicians as a *decision aid*, rather than as a replacement for human decision making. However traditional classifier performance measures, such as accuracy, are only viable when classifiers produce a binary decision. Assessing the performance of a probabilistic classifier is a more troublesome task, and is seldom discussed in the literature. Techniques to accomplish this are reviewed.

Two different classifier structures are used, logistic regression and Gaussian mixture models, together with several preprocessing techniques, in order to determine the optimal classifier design. Samples were available from patients upon their admission to hospital and, in many cases, at a second time 1–6h later. Incorporating a second sample (from the same patient) into the classifier is shown to improve diagnostic performance, but at the price of delaying the time before the diagnosis is available.

The long-term aims of this research are to enhance the information available to clinical staff concerning AMI patients to assist with diagnosis and treatment. Initially, the concentrations of five markers, instead of the current one or two, will be provided; this paper is concerned with the next step: processing these raw marker readings to estimate the probability of AMI in new patients.

This paper is organised as follows: section 2 explains the study population, recruitment, sampling and inclusion criteria. Section 3 then discusses the need for probabilistic classifiers (i.e. classifiers which give a probability of AMI rather than simply 'yes' or 'no') and highlights appropriate measures for assessing the performance of such classifiers. In section 4 two classifiers are presented: logistic regression and Gaussian mixture models (GMMs), together with the various preprocessing techniques used. The results when these classifiers were evaluated on the dataset are given and discussed in section 5, the conclusions are drawn in section 6 and section 7 gives suggestions for future work.

## 2 Study details

Samples for this study were collected from patients entering the Department of Emergency Medicine, St. James' Hospital, Dublin, Ireland, between 18 September 2000 and 12 October 2002 as part of a large ongoing study.

A research nurse invited patients entering the department with symptoms indicative of AMI (typically chest pain) to enrol in the study. Those who gave their informed consent had a selection of clinical, demographic and ECG information recorded in a database and an initial blood sample taken.

A protocol was established which required subsequent blood samples to be taken every two hours, up to a maximum of ten samples. Due to the nature of a busy emergency department and the acute nature of the patients' illnesses, it was not always possible to follow the timing or number of samples exactly, and a number of patients, especially those with AMI, withdrew their consent part way through the sampling regime. None of the patients, however, requested complete withdrawal from the study, so samples already taken were retained. The sample taken on admission was labelled T0, and subsequent samples T1, . . . , T9. They were then frozen on site and transported to Randox Laboratories Ltd. where they were processed using the Evidence biochip analyser. Follow-up from patient records and home telephone calls sought to establish the final diagnosis given to each patient, together with surgical and drug treatments as well as 7 and 30 day mortality.

Approximately 8% of the patients in the study had an AMI. Preliminary investigations showed that this small number of patients was insufficient to reliably construct a classifier. In order to enhance the proportion of AMI patients, priority was then given to analysing samples from AMI patients over those from non-AMI ones. Although this meant the prevalence of each disease in the data set was no longer representative of that in the study population, it has proved possible to develop and test various classifier models. Bayes' theorem can be used at a later time to adjust the probabilities given by a classifier to correct for the difference in prevalence of the training population and that in the population on which the classifier will be used [13].

One of the study aims is to make an early diagnosis of AMI. With this in mind, patients with a T0 sample taken more than 18h after the onset of pain were discarded. A second data set was constructed containing both T0 and T1 samples. Patients for this set were further rejected if the T1 sample was taken more than 6h after T0. This regime gives rise to potentially large variations in the time between an AMI event occurring and their T0 sample being taken, and also gives the range between first and second samples being taken (T1−T0) of 1–6h. These variations are reflective of the clinical situation in busy emergency departments, and classifiers constructed from such data will incorporate an inherent robustness to practical variations in sampling times.

Finally two studies were omitted from the research because their diagnosis was uncertain. The first was recorded as having had an AMI, but a later clinical review concluded the event may have happened whilst in hospital, probably about 48h after the initial pain onset, and almost certainly after the T0 and T1 samples were taken. The diagnosis is therefore unreliable and this patient has been omitted from the study. For the second, clinical opinion was divided between unstable angina and non-Q wave MI for the final diagnosis.

This left one set of 159 patients with both a valid diagnosis and T0 measurements, 50 of whom had AMI (31%). The second set, containing both T0 and T1 measurements had 132 patients, 38 of whom had AMI (29%).

## 3 Measures of performance

Many automatic classification systems produce some score, or measure of certainty in their output, rather than simply a binary 'yes' or 'no' decision. This score frequently lies between 0 and 1, and is often referred to as a probability. Despite this, the score is often hidden from the user. A threshold is set, above which a value of 1 (or 'yes') is returned, and below which a value of 0 (or 'no') is given.

In the field of medical diagnostics, returning the actual probability of a disease has the potential to be more clinically useful than simply providing a dichotomous decision. Clinicians are used to dealing with uncertainties and performing risk assessment, and a probability can provide evidence to be incorporated the clinician's patient management strategy. Providing a dichotomous decision can be perceived as trying to compete with the clinical judgement and removing decision making from the doctor—de Dombal labelled such a perception "the kiss of death" to many computer-aided diagnostic systems [14].

However, assessing the performance of a probabilistic classifier is more challenging; when classifiers give a binary output, the only possible measures of performance depend upon how many patients are diagnosed correctly. If this is measured over an entire sample population, it is called *accuracy*, or *non-error rate (NER)*. If it is only measured over diseased patients, it is termed *sensitivity*, or if only measured over healthy ones, *specificity*. In many cases, a test or classifier gives a continuous output which is then thresholded to give a binary decision. Receiver operator characteristic curves (ROC curves) [15] are a tool to visualise the relationship between sensitivity and specificity as this threshold is varied, and the area under the ROC curve (AUROCC) provides a numerical measure of the classifiers' performance, which has the advantage of being independent of the threshold.

In the probabilistic realm, however, all of these measures (except AUROCC) break down. If a probabilistic classifier claims that a healthy sample has an 80% chance of being diseased, it is not possible to say the classifier is 'wrong' based on this single sample. Habbema, Hilden and Bjerregaard proposed three distinct, yet related, quantities for measuring the performance of probabilistic classifiers: *discriminatory ability, sharpness* and *reliability* ([16, 17, 18, 19, 20]), each of which must be considered when assessing the performance of a probabilistic classifier.

### *3.1 Discriminatory ability*

Discriminatory ability is a measure of how well a classifier can select between diagnoses. A good classifier should clearly assign high probabilities to the diseases the patients actual have.

In addition to the well known AUROCC, the normalised Brier score ($Q_{31}$ in Habbema *et al.*'s notation [16]) was used to measure discriminatory ability in this work. In the case of two mutually exclusive classes (e.g. AMI and non-AMI), this score is defined as:

$$Q_{31} = 1 - \frac{1}{N} \sum_{i=1}^{N} \left(1 - P_{i,actual}\right)^2 , \qquad (1)$$

where $N$ is the number of patients and $P_{i,actual}$ is the probability assigned to the $i$th patient's actual disease. For example, if patient $i$ is given a 35% chance of AMI then $P_{i,\mathrm{AMI}} = 0.35$ and $P_{i,\text{non-AMI}}$, the probability of non-AMI is therefore 0.65. If clinical diagnosis indicates AMI, then $P_{i,actual} = P_{i,\mathrm{AMI}} = 0.35$, whereas if the clinical diagnosis is non-AMI, then $P_{i,actual} = P_{i,\text{non-AMI}} = 0.65$.

For any given patient, the best possible classification occurs when $P_{i,actual} = 1$, thus the Brier score is based upon the mean square deviation from this ideal. This mean-square deviation, however, decreases as the classifier performance improves, contrary to conventional performance measures such as accuracy. The normalised Brier score, therefore, subtracts this deviation from unity to give a score of 1 to a perfect classifier (i.e. one which always gives a probability of 1 to the patient's actual disease), and a score of 0 to a totally imperfect classifier. The normalised Brier score can be generalised to deal with more than two classifications [17].

### *3.2 Sharpness*

The second performance measure, sharpness, is a measure of the confidence which a classifier has in its outputs, rewarding confident predictions (i.e. close to unity or zero), rather than indecisive ones. Sharpness is defined as the expected value of the discriminatory ability:

$$E\left[Q_{31}\right] = 1 - \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{j \in \{\mathrm{AMI}, \\ \text{non-AMI}\}}} P_{i,j} \left(1 - P_{i,j}\right)^2 . \qquad (2)$$

Note that as this measure is an *expected* value, it does not depend upon the actual disease classification. Here a classifier which assigns a probability of 1 randomly to diseases would be perfectly sharp, although the probabilities would be meaningless in practice. This problem leads to the need for the third performance measure.

### *3.3 Reliability*

Reliability ties together the ideas of discriminatory ability and sharpness. If 100 patients are each assigned an AMI probability of 0.8, about 80 would be expected to actually have that disease. If only 50 in fact had the disease, the classifier is *over confident* and should have assigned a lower probability to each patient. On the other hand, if 99 of the patients did have the disease, then the classifier is *diffident*, or too cautious. In either case, the classifier has a poor reliability.

Reliability, therefore, is the difference between how well the classifier claims it can perform (sharpness) and how well it actually performs (discriminatory ability). The reliability, $Q_3$, is given by:

$$Q_3 = Q_{31} - E\left[Q_{31}\right]. \tag{3}$$

Hilden *et al.*[16] have shown that negative reliability indicates an overconfident classifier, whilst reliabilities greater than zero reveal a classifier to be diffident.

Reliability on its own is insufficient as a measure of performance: consider a classifier which simply assigns every patient the same probability of disease, and that probability is the prevalence of the disease in the population under investigation. For example, if it is known that 10% of a population have a certain disease, then assign every patient a probability of 0.1, regardless of all other evidence. Such a classifier would have perfect reliability, yet would not provide any additional clinical information.

For this reason, reliability is used and reported in conjunction with discriminatory ability in this work. (The relationship expressed in equation 3 shows that including sharpness does not provide any additional information.)

### *3.4 Dot Diagrams*

In addition to the numeric measures above, a useful graphical visualisation technique was presented by Hilden *et al.* [16] in the form of the dot diagram. This shows the variation in the AMI and non-AMI probabilities from the classifier in the form of two rows of points. It therefore provides a rapid method for visualising the spread of the probabilistic outputs for a given disease, see figure 3 for example. Ideally, the points should be bunched to the right hand side of the diagram (meaning that high probabilities were assigned to the correct diseases).

### *3.5 Optimal use of data: Leave-one-out cross-validation*

Assessing the performance of a classifier typically involves presenting the classifier with a number of samples and asking it to classify them. Its classifications are then compared with a gold standard (clinical diagnosis in this case) and the performance measures calculated. Ideally the data used to construct the classifier should be different from that used for assessment as it will almost certainly perform better on samples already seen.

The problem then arises as to how to divide the finite amount of study data available between the training set and the test set. One elegant solution is leave-one-out cross-validation [13]. This entails training a classifier on $N-1$ of the available $N$ studies, then testing the classifier on the remaining one. This process is repeated $N$ times, using a different study for testing on each occasion, resulting in the construction of $N$ unique, yet similar, classifiers. Using the results from these $N$ tests, the performance measures described earlier can then be calculated to provide an unbiased estimate of the performance of a classifier trained using all $N$ input samples.

## 4 Classifiers

From the many techniques available for automatic pattern classification, two were selected for testing in this application. The first, logistic regression (LR), is one of the simplest techniques used in multivariate pattern recognition and aims to estimate the class conditional probability distribution. It is well understood and much less prone to over-fitting when trained on small data sets than most of the alternatives available. LR classifiers can be viewed as drawing a series of parallel, $n-1$ dimensional, probability contours (i.e. hyper-planes) through $n$ dimensional space. This is illustrated in figure 1 for the two dimensional case.

The second method employs a different strategy: rather than trying to create a boundary in the data space, with points near the boundary assigned probabilities around 0.5 and those further away given probabilities closer to 0 or 1, Gaussian mixture models (GMMs) estimate the probability density functions (pdfs) for both the AMI and non-AMI datasets and combine the results using Bayes' theorem. These pdfs contain information about the probability density anywhere in the input space, based on the concentration of training points in that locality. This means that the GMM can model arbitrarily complex functions (pdfs), but at the cost of being highly parameterised and much more likely to overfit the data.

### *4.1 Data preparation*

From the measurements on the 291 unique blood samples used in this work, some marker concentrations in some samples were found to be too high to be accurately measured. Eleven CKMB, 30 FABP, 15 myoglobin and 5 cTnI measurements fell into this category. To facilitate classifier construction, these out-of-range readings were replaced with values drawn randomly from a uniform distribution ranging from the maximum value accurately recorded to 1.1 times this maximum value.

Given the nature of the data and disease, this is not as significant an issue as it may first appear: the markers are released primarily as a result of damage to the myocardium (heart muscle); very high levels of markers therefore are almost certainly indicative of AMI. In fact, only two non-AMI patients had any marker readings above this threshold: 213 and 267. Patient 213 had a myoglobin reading out of range, and suffered from a cardiomyopathy, which would cause damage to the cardiac tissue. Patient 267 suffered from unstable angina, a condition which also causes damage to the myocardium.

The marker values were then normalised to give each zero-mean and unit variance, with markers from T0 treated separately to those from T1 during normalisation.

### *4.2 Pre-processing and logistic regression*

There are numerous techniques for pre-processing data in order to either remove noise or reduce the dimensionality of datasets (or both). Two procedures were employed on the normalised marker data: firstly principal component analysis (PCA), and secondly Fisher discriminant analysis (FDA) [13].

Principal component analysis is commonly used for dimensionality reduction and involves selecting principal components (PCs) which are linear combinations of the input variables. The first principal component, PC1, is selected so as to describe the maximum variation in the data. The second is selected to contain the maximum variation in the data, subject to the constraint that it is perpendicular to PC1, and so on. Often the majority of the variation in the data is captured by the first few PCs.

Fisher discriminant analysis (also known as canonical analysis) can be viewed as a linear transformation akin to PCA, except that rather than trying to maximise the variation captured by PC1, it uses class labels to maximise the separation between classes (AMI and non-AMI in this case). A detailed description is given in [21].

Logistic regression classifiers were designed and tested using all possible combinations of the normalised markers, different numbers of PCs and finally FDA.

### *4.3 Gaussian mixture models*

Bayes' theorem provides a method for determining the probability, $P(\mathrm{AMI}|X)$, that a patient with a vector of five marker concentrations, $X$, has AMI, given only information about the probability density function (pdf) of marker concentrations in the population of AMI patients $P(X|\mathrm{AMI})$, the pdf of a similar non-AMI population $P(X|\text{non-AMI})$ and the prevalence of AMI in the total population under study, $P(\mathrm{AMI})$. Mathematically:

$$P(\mathrm{AMI}|X) = \frac{P(X|\mathrm{AMI})}{P(X|\mathrm{AMI}) + P(X|\text{non-AMI}) \cdot (1 - P(\mathrm{AMI})/P(\mathrm{AMI}))}. \quad (4)$$

The prevalence of the disease can easily be calculated by counting the number of AMI patients in the sample. Modelling the pdfs of the AMI and non-AMI populations is somewhat more difficult; in this case Gaussian mixture models (GMMs) [13] were used.

A GMM of a pdf is created using a combination of Gaussian kernels, similar to a radial basis function neural network. A GMM was created for each of the two classifications. One is an estimate of the probability density function (pdf) for non-AMI patients, the other an estimate of the pdf for AMI patients. In addition to the number and combination of inputs needed, the number of centres, or Gaussians, used to construct the model had to be decided. Once these meta-parameters were selected, the centres and variances of each Gaussian were determined by the expected maximisation training algorithm [22].

A sample GMM pdf with three centres is shown in figure 2. When a new patient is seen with a vector of markers concentrations, $X$, the probabilities $P(X|\text{AMI})$ and $P(X|\text{non-AMI})$ can be determined from the GMMs, and therefore $P(\text{AMI}|X)$ follows from equation 4, $P(\text{AMI})$ having been previously calculated.

## 5 Results and discussion

Results from a number of different classifiers are described in this section, together with a discussion of the merits and drawbacks of each, mainly concerning the sources of bias in the performance estimates. These results are summarised in tables 1 and 2. All computation was carried out using Matlab 6 [23] and, where appropriate, the Netlab toolbox [22].

### *5.1 Logistic regression at admission*

Although concentrations of five markers were available at admission (T0), reducing the dimensionality of a classifier can usually improve its performance on unseen data, as it is less likely to over-fit the training set. Initially, this dimensionality reduction was accomplished by omitting markers from the data set. For five markers, there are $2^5 - 1 = 31$ unique combinations, if the trivial case of no inputs to the classifier is omitted.

From these 31 possible logistic regression classifiers trained using combinations of markers taken on admission, the optimum performance (largest Brier score and NER) occurred with the classifier trained using two of the five available markers: CKMB and GPBB. In this case, the classifier had a NER of 0.868 (95% confidence interval: 0.805–0.916), Brier score of 0.904 (with an associated reliability of -0.013) and AUROCC of 0.913 (standard error: 0.029). The classifiers were ranked according to Brier score to investigate the relationship between the markers used in construction of each classifier and its performance. The first notable pattern was that the 16 classifiers which included CKMB ranked higher than the 15 which did not. Also, FABP proved least useful, occurring only twice in the top ten, while GPBB, myoglobin and cTnI appeared 5, 4 and 4 times respectively.

Figure 3 shows the dot diagram for the optimum classifier, where the good performance obtained in the majority of patients can easily be seen. Three non-AMI patients, 320, 379 and 239 with P(AMI) below 0.4 were badly diagnosed. Patient 320 suffered from unstable angina followed by pulmonary oedema, and proceeded to suffer an MI one week later. Patient 379 was diagnosed with stable angina, cardiac arrthymia and heart failure. Patient 239 was assigned about a 70% chance of AMI, despite having a 'non-cardiac' final diagnosis. The reason for this anomolous classification is unclear since, although the patient had an MI 3 years previously, this is unlikely to have affected the results.

A larger number of AMI cases were misclassified, with some 15 out of 50 given less than a 50% chance of AMI. This is primarily due to the first blood sample having been taken very soon after the AMI, before the markers had risen to diagnostic concentrations.

### *5.2 Logistic regression using two samples*

This section reports the performance of classifiers constructed using the concentrations of markers on admission (T0) and at a second time (T1), 1–

6 hours later. The marker concentrations from T0 and T1 were presented as separate inputs to the logistic regression classifiers with any combination of up to 10 inputs. This resulted in $2^{10}-1 = 1023$ candidate classifiers. From all these, the best performance (as measured by the Brier score) was produced when using FABP and myoglobin concentrations at T0 and CKMB, FABP and cardiac troponin I at T1. This classifier resulted in a NER of 0.955 (CI: 0.904–0.983), Brier score of 0.960, reliability of -0.018 and AUROCC of 0.959 (SE: 0.023).

The dot diagram for this classifier (figure 4) shows virtually all patients well classified (i.e. a high probability was assigned by the classifier to the patient's actually diagnosis). Only three AMI patients were badly diagnosed ($P(\text{AMI}) < 0.4$), and similarly, only two non-AMI patients were badly classified.

The AMI patients in question were 380, 439 and 963. From these, the first had a T0 reading taken 5h 14min post pain onset, and a T1 sample taken 6h 29min post pain onset. Both these samples were taken at an early stage and in very close succession. These two factors may work together to make it difficult for the classifier to reliably diagnoise AMI. Patient 439 again had markers recorded at a very early stage, T0 at 4h 30mins after pain onset and T1 at 6h 30. This patient's markers were also unlikely to have risen enough in this short space of time to be reliably detected by the classifier. Finally, patient 963 again had T0 taken very early (4h 09mins post pain onset) and had a relatively minor (non Q waves, no ST elevation) infarction.

The non-AMI patients misclassified were patients 267 and 239. Patient 267 had had 4 previous MIs and was diagnosed with unstable angina, a condition which causes some damage to cardiac tissue. This patient had elevated levels of all markers, especially myoglobin and cardiac troponin I, and also showed the rise and fall profile characteristic of AMI patients [24]. Patient 239 has been discussed previously (section 5.1).

Although these results initially appear excellent, they must be tempered by the knowledge that picking the best classifier from a selection of over one thousand, based on its performance over test points from leave-one-out cross validation, then using the *same* test set of patients to measure the performance will lead to bias in the performance measure. Ideally, the available data should be split, one part being used to train and test the models and the results used to select the best classifier. This classifier should then be tested on the remaining part of the data set to provide an unbiased measure of performance. If this final step is omitted bias is inevitably introduced in the performance measures. The more models available to choose from (1023 in this case), the larger the effect of this bias.

### *5.3 Principal component analysis*

The optimum classifiers in each of the previous two sections were produced from a subset of the five available markers, rather than from the complete set. This was deliberately done because of the well-documented effect of overfitting with high-dimensional data (e.g. [13]). Omitting markers is the simplest way to reduce the dimensionality of the input space but it is unlikely to provide the best approach since potentially important information is discarded. Principal component analysis (PCA) often proves a better technique for reducing the dimensionality of a classifier. It was applied to the marker concentrations for each patient and the logistic regression analysis was then repeated, firstly using the T0 sample, then using both the T0 and T1 samples.

Table 3 and figure 5 show the results of this analysis, where the optimum performance (highest Brier score and NER plus a reasonable reliability) resulted from using two PCs. The Brier scores for the best marker combinations were very similar (table 1), with PCA being just slightly poorer. PCA, however, may have a slight edge in terms of producing less bias in the performance measures of the optimum classifier. Here the best classifier from 5 possibilities is being selected, rather than the best one from 31 when markers are not pre-processed.

The four non-AMI patients badly classified in figure 5 are from patients 320, 322, 267 and 308. From these, patients 320 and 267 have already been discussed (section 5.2), patient 308 had elevated marker levels and was diagnosed with unstable angina, which results in damage to cardiac tissue. Patient 322 had elevated marker levels, especially at T0, and was diagnosed with digoxin toxicity and has a history of angina, which may have caused minor cardiac damage.

When markers from both T0 and T1 were used (table 4, figure 6), the results were less clear. Generally speaking, the reliability decreased as more principal components were included. This is because the number of model parameters to be determined in training increased with the number of inputs and the model was then likely to over-fit the data. The poorer figures for reliability signify that the classifier is more likely to push probabilities to the extreme values of 0 or 1 regardless of whether that is the correct diagnosis or not. That is, the classifier's *sharpness* increases without a corresponding increase in its *discriminatory ability.*

Taking this into account, the optimum classifier can be found by considering the trade-off between discriminatory ability and reliability, which probably leaves the best classifier as being the one which uses either 3 or 5 PCs.

The performances of the best classifiers using PCA with inputs from two consecutive blood samples, however, are poorer than the best classifier found by crudely omitting markers (table 2). There are two possible contributing factors: firstly the assumption that directions (principal components) which capture the maximum variation of the data are the best directions in which to project the data to provide the optimum separation of two sets of labelled data may not be valid. Fisher discriminant analysis (section 5.4) attempts to circumvent this shortcoming but it is not without its own difficulties.

The second factor is the bias in the performance measurement. Selecting the best classifier, from the set of 10 constructed using PCA preprocessing, produces a smaller bias than selecting the best from a set of 1023. This is the most likely cause of the small difference between the results obtained using, for example the first 4 PCs, and that obtained using the optimum combination of markers listed above.

### *5.4 Fisher discriminant analysis*

The second preprocessing method investigated was Fisher discriminant analysis (FDA), also known as canonical analysis [21]. A summary of the results is presented in table 7.

Here the performance of both FDA classifiers was slightly worse than that obtained using either the optimum combination of markers (sections 5.1 and 5.2), or the optimum number of principal components (section 5.3).

There are three possible reasons for the difference in performance measures. Firstly, the differences are quite small, and certainly well within the confidence intervals for the performance measures, so that they may be due to random effects and would disappear if a larger sample was available.

Secondly, they may be caused by bias in the selection of the other classifiers, as discussed in section 5.3, where an independent test set is required to assess the performance of the optimum classifier in each case. As FDA produces only a single variable, no further selection is required, hence its performance measure is unbiased.

Finally, FDA rests on the assumption that the intra-class characteristics of each data set (AMI and non-AMI) is comparable [25]. This is not the case with the current dataset, as non-AMI patients tend to have almost uniformly low concentrations across all markers, whereas AMI patients have a much larger variation in their marker concentrations.

### *5.5 Gaussian mixture models*

The final classification approach used was Gaussian mixture models (GMM). GMM techniques have many more hyper-parameters to be selected than the previous methods discussed, leaving them much more prone to bias in the selection of the optimum classifier.

When considering T0 samples, GMMs were constructed with between 1 and 5 centres, using spherical covariance matrices only. Using all 31 combinations of markers, and the 25 combinations of AMI/non-AMI centres led to 775 unique classifiers. From these, the best classifier (judged according to Brier score), used only CKMB and GPBB and had four centres in both its AMI and non-AMI models. It had a NER of 0.843 (95% CI: 0.777–0.896), Brier score of 0.898, reliability -0.030 and AUROCC of 0.918 (SE: 0.028). Its dot diagram is shown in figure 9.

These results have a lower NER and Brier score, and poorer reliability than the results from either the simple logistic regression, or the logistic regression with PCA pre-processing. They perform slightly better against the FDA classifier, having a marginally better Brier score, yet a poorer reliability and slightly worse NER.

The experiment was repeated using PCA preprocessing and increasing the number of PCs entered into the classifier whilst varying the number of centres. The results are listed in table 8, with the dot diagram for the case of 5 principal components shown in figure 10. Again, these results have slightly worse diagnostic performance measures and poorer reliability scores than the logistic regression classifiers.

Given the results are not as good as other, simpler, methods and that the bias in the results has the potential to also be much larger, it was decided not to pursue the GMM analysis with two sets of marker measurements.

There is another reason for not investigating GMMs further: the nature of the Gaussian kernels from which the model is built. Cardiac markers operate in a moderately predictable manner: as the amount of myocardial damage increases, the amount of marker released and hence its concentration in the blood increase. This prior knowledge indicates that if the concentration of any marker increases, the probability of AMI returned by the classifier should increase too. The shape and nature of a Gaussian kernel mean that data points far from its centre are assigned a probability density of virtually zero. Thus GMMs only cover part of the input space, and are capable of dividing it up in a smooth, though somewhat arbitrary fashion if required, rendering them well-suited to multiple classification problems. Logistic regression, however, divides the entire input space in a manner consistient with this prior knowledge, see figure 1.

## 6 Conclusions

Good performance in estimating the probability of AMI from cardiac marker concentrations from patients admitted to hospital with chest pain has been demonstrated. The benefit of measuring the concentrations of several markers has also been shown, as classifiers with inputs from multiple markers performed better than those with data from only a single marker.

Including marker concentrations from two sequential times also leads to a marked improvement in the discriminatory ability of classifiers. This is probably due to two reasons: firstly, as there is a time lag between an AMI event and markers entering the bloodstream, the second measurement will almost certainly have higher marker concentrations than the first, which may be enough to improve the performance. The second reason is that knowing how marker concentrations change with time can provide more information than either marker measurement individually [26, 6].

The new FABP marker has shown itself as a useful addition for diagnosing AMI based on two marker readings. From the 1023 possible marker combinations using samples from both T0 and T1, the top 147 (ordered by Brier Score) included at least one FABP measurement in their combination. From those combinations that did not include FABP, the highest Brier Score was 0.932, compared to 0.955 from the best combination which included FABP. This large increase in performance provides a basis for suggesting the inclusion of FABP in regular hospital testing, although the reasons why it failed to perform as well when using only T0 samples are unclear.

Estimating the probability of AMI, rather than simply opting for a binary (‘yes’ or ‘no’) classifier, has the potential for higher clinical acceptability and utility as it aims to assist physician diagnosis rather than compete with it. Also, estimates of the probability of AMI produced by a classifier can be adjusted, via Bayes’ theorem, to accommodate variations in the prevalence of AMI between different populations. This could assist in making the classifier portable between different locations.

Evaluating the performance of a probabilistic classifier, however, is significantly more difficult than for binary one, as demonstrated by the identification of multiple performance measures, namely discriminatory ability, sharpness and reliability, in section 3.

This need for multiple performance measures is best observed by the fact that there is no significant difference between the measures of discriminatory ability (i.e. non-error rates (or accuracies), Brier scores or AUROCCs) of the optimum classifier in each section. Alternative criteria must be used to differentiate between these classifiers. The reliability measure, together with the complexity of the classifier and appropriate sources of bias must also be considered. This points to logistic regression with FDA preprocessing as being a good candidate solution. This lack of difference in certain performance measures can be viewed as reassuring—all classifiers perform similarly, and selecting a good candidate does not involve hitting upon a golden preprocessing algorithm.

Classifiers based on GMMs were over-parameterised which led their poor reliability scores, reflecting a tendency to overfit the data, meaning that their probabilities could not be trusted.

Compared to many pattern classification problems (for example, visual recognition), diagnosing AMI is relatively straightforward—as the concentration of a marker increases, the probability of AMI should increase too. Simple classification techniques therefore, such as linear regression, are able to capture these differences. This is demonstrated in the results presented, and in unpublished preliminary work, in which more complex artificial neural networks showed a strong tendency to overfit data and gave poor reliability.

This work has shown good initial progress in estimating the probability of a patient having suffered an AMI, based upon the concentration of cardiac markers in their blood. It has been limited by the number of patients available to it, but classifier performance is expected to improve as more data becomes available from ongoing clinical trials.

## 7 Future work

The data used in this study is part of a large ongoing hospital study which, as of March 2004, has collected 4892 blood samples from some 1500 eligible patients. Many of these samples are awaiting analysis. As more data become available, investigations will be extended into three areas: the first is detecting AMI as early as possible to enable rapid appropriate treatment of patients. The second is sub-classification of patients along the continuum coronary syndromes from non cardiac chest pain, through stable angina and unstable angina to non-Q wave infarctions, where typically a thinner layer of the heart is damaged, and finally Q wave infarctions which involve the death of a large area of cardiac tissue. Currently a sharp division is drawn between unstable angina (designated non-AMI) and non-Q wave MI (designated AMI). The final area for additional research is for risk stratification or prognostic information. Eventually it is hoped that a classifier (or classifiers) may be integrated with the Evidence hardware to produce a device capable of automatically analysing blood samples and giving relevant information on a patient's coronary health care to assist clinicians.

## Originality and Contribution

Logistic Regression and Gaussian Mixture Model classifiers have been trained to estimate the probability of acute myocardial infarction (AMI) based upon the concentrations of cardiac markers obtained from hospital patient blood samples. In contrast to previous studies, this work had access to measured data on five markers simultaneously (CKMB, myoglobin, cTnI, GPBB and FABP). The last two are more recently discovered and have not been used in any previous classification system for coronary heart disease.

Logistic Regression, with data pre-processing by Fisher Discriminant Analysis, provided the best overall classifier based only on blood samples taken on admission. This gave an accuracy of 0.85 (with a 95% confidence interval of 0.78-0.91) and a normalized Brier score of 0.89. When a second sample taken 1 to 6 hours later was included in the classifier, the performance increased significantly.

The benefit of measuring the concentrations of several markers is shown, as classifiers with inputs from multiple markers outperformed those for a single marker. In particular, the new FABP marker proved to be a useful addition to the classifier.

## 8 Acknowledgements

The first author would like to the Department for Employment and Learning, Northern Ireland, for providing a PhD studentship, and Randox Laboratories for financial support and provision of data.

## References

1. Peterson, S., Rayner, M., Coronary Heart Disease Statistics: 2002 edition, British Heart Foundation, 2002.
2. Wu, A. (Ed.), Cardiac Markers, Pathology and Labratory Medicine, Humana Press, Totowa, New Jersey, 1998.
3. Goldman, L., Cook, E., Brand, D., Lee, T., Rouan, G., Weisberg, M., Acampora, D., Stasiulewicz, C., Walshon, J., Terranova, G., Gottlieb, L., Kobernick, L., Goldstein-Wayne, B., Copen, D., Daley, K., Brandt, A., Jones, D., Mellors, J., Jakubowski, R., A computer protocol to predict myocardial infarction in emergency department patients with chest pain, The New England Journal of Medicine 318 (13) (1998) 797–803.
4. Goldman, L., Weinberg, M., Weisberg, M., Olshen, R., Cook, E., Sargent, R., Lams, G., Dennis, C., Wilson, C., Deckelbaum, L., Fineberg, H., Stiratelli, R., A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain, The New England Journal of Medicine 307 (10) (1982) 588–596.
5. Selker, H., Beshansky, J., Griffith, J., Aufderheide, T., Ballin, D., Bernard, S., Crespo, S., Feldman, J., Fish, S., Gibler, W., Kiez, D., McNutt, R., Moulton, A., Ornato, J., Podrid, P., Pope, J., Salem, D., Sayre, M., Woolard, R., Use of the Acute Cardiac Ischemia Time-Insensitive Predictive Instrument (ACI-TIPI) to assist with triage of patients with chest pain or other symptoms suggestive of acute cardiac

ischemia: A multicenter, controlled clinical trial, Annals of Internal Medicine 129 (1) (1998) 845–855.

6. Fesmire, F., Hughes, A., Fody, E., Jackson, A., Fesmire, C., Gilbert, M., Stout, P., Wojcik, J., Wharton, D., Creel, J., The Eulanger chest pain evaluation protocol: A one-year experience with serial 12-lead ECG monitoring, two-hour delta serum marker measurements, and selective nuclear stress testing to identify and exclude acute coronary syndromes, Annals of Emergency Medicine 40 (6) (2002) 584–594.
7. Pozen, M., D'Agostino, R., Selker, H., Sytkowski, P., Hood, W., A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease, The New England Journal of Medicine 310 (20) (1984) 1273–1278.
8. Tierney, W., Roth, B., Psaty, B., McHenry, R., Fitzgerald, J., Stump, D., Anderson, K., Ryder, K., McDonald, C., Smith, D., Predictors or myocardial infarction in emergency room patients, Critical Care Medicine 13 (7) (1985) 526–531.
9. Harrison, R., Marshall, S., Kennedy, R., Minimum-risk decisions in the management of suspected heart attack: an application of the Boltzmann perceptron network, in: Rogers, S. (Ed.), Proceedings of SPIE. Applications of Artificial Neural Networks III, Vol. 1709, 1992, pp. 1701–1082.
10. Fraser, H., Pugh, R., Kennedy, R., Ross, P., Harrison, R., A comparison of back propagation and radial basis functions in the diagnosis of myocardial infarction, in: Proceedings of the International Conference on Neural Networks and Expert Systems in Medicine and Health Care, 1994, pp. 76–84.
11. Ellenuis, J., Groth, T., Lindahl, B., Wallentin, L., Early assessment of patients with suspected acute myocardial infarction by biochemical monitoring and neural network analysis, Clinical Chemistry 43 (10) (1997) 1919–1925.
12. Groth, T., Ellenius, J., Provision of decision support for acute myocardial infarction, United States Patent: 6 443 889 (3 September 2002).

13. Bishop, C., Neural Networks for Pattern Recognition, Oxford: Clarendon Press, 1995.
14. de Dombal, F., Medical Informatics: The Essentials, Butterworth-Heinemann, Linacre House, Jordan Hill, Oxford OX2 8DP, UK, 1996.
15. Hanley, J., McNeil, B., The meaning and use of the area under a receiver operator characteristic (ROC) curve, Radiology 143 (1982) 29–36.
16. Hilden, J., Habbema, J., Bjerregaard, B., The measurement of performance in probabilistic diagnosis: II. Trustworthiness of exact values of the diagnostic probabilities, Methods of Information in Medicine 17 (4) (1978) 227–237.
17. Hilden, J., Habbema, J., Bjerregaard, B., The measurement of performance in probabilistic diagnosis: III. Methods based on continuous functions of diagnostic probabilities, Methods of Information in Medicine 17 (4) (1978) 238–246.
18. Habbema, J., Hilden, J., Bjerregaard, B., The measurement of performance in probabilistic diagnosis: I. The problem, descriptive tools , and measures based on classification matricies, Methods of Information in Medicine 17 (4) (1978) 217–226.
19. Habbema, J., Hilden, J., The measurement of performance in probabilistic diagnosis: IV. Utility considerations in therapeutics and prognostics, Methods of Information in Medicine 20 (2) (1981) 80–96.
20. Habbema, J., Hilden, J., Bjerregaard, B., The measurement of performance in probabilistic diagnosis: V. General recommendations, Methods of Information in Medicine 20 (2) (1981) 97–100.
21. Huang, P., Harris, C., Nixon, M., Comparing different template features for recognising people by their gait, in: Proceedings of the Ninth British Machine Vision Conference, Vol. 2, 1998, pp. 639–648.
22. Nabney, I., Netlab: Algorithms for Pattern Recognition, Advances in Pattern Recognition, Springer, 2002.
23. Inc., T. M., Matlab version 6.1 release 12.1, Software (2001).
24. Armstrong, G., Petry, C., Wagner, G., Wu, A., Reflex algorithm for early and cost effective diagnosis of myocardial infarctions suitable for automated diagnostic platforms, European Patent Application number:

99110216.1 (May 1999).

25. Magee, D., Boyle, R., Improving class seperation in principal component analysis using delta analysis, research Report: School of Comnputer Studies, University of Leeds (December 1998).
26. Fesmire, F., Percy, R., Bardoner, J., Wharton, D., Calhoun, F., Serial creatinine kinase (CK) MB testing during the emergency department evaluation of chest pain: Utility of a 2-hour delta CK-MB of +1.6ng/ml., American Heart Journal 136 (2) (1998) 237–244.

## List of Figures

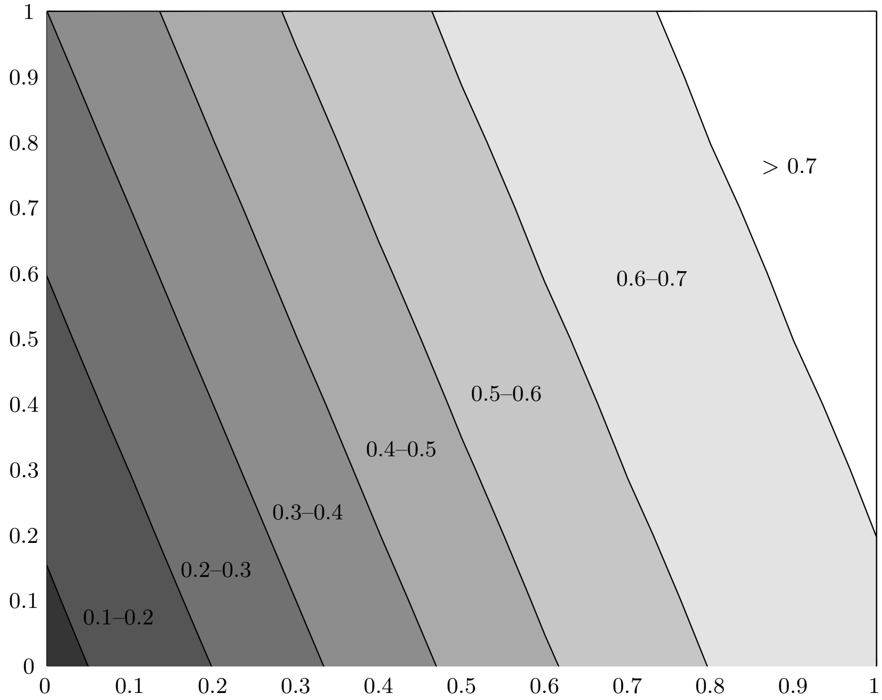

**Fig. 1** Example of probability contours for logistic regression classifier.

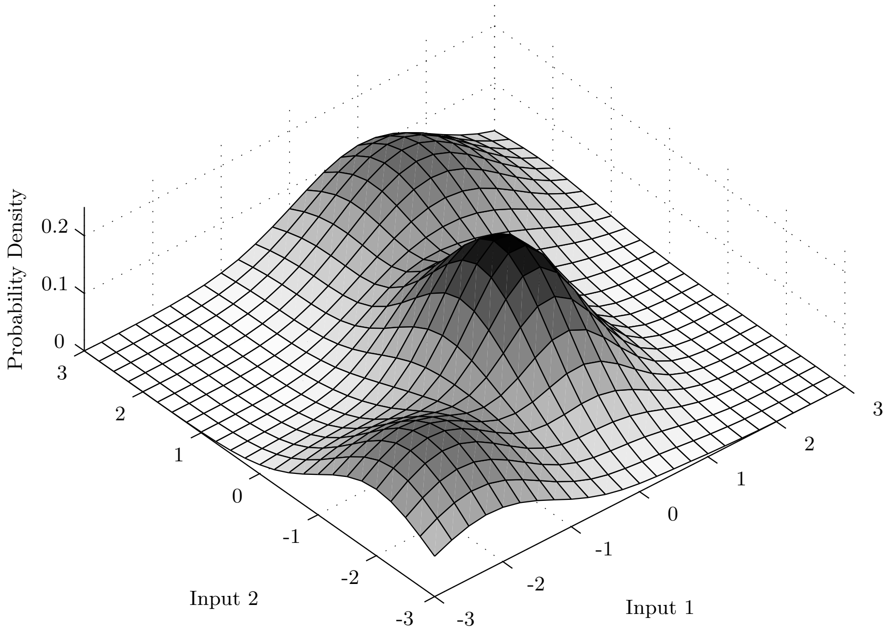

**Fig. 2** Example of probability density function for GMM with three centres.

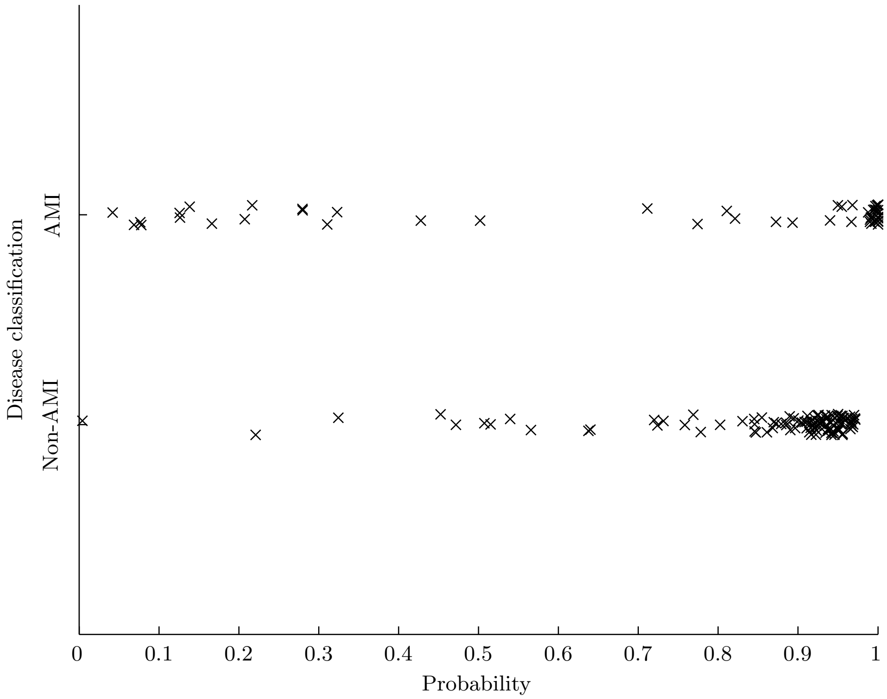

**Fig. 3** Dot diagram for logistic regression classifier using CKMB and GPBB at admission

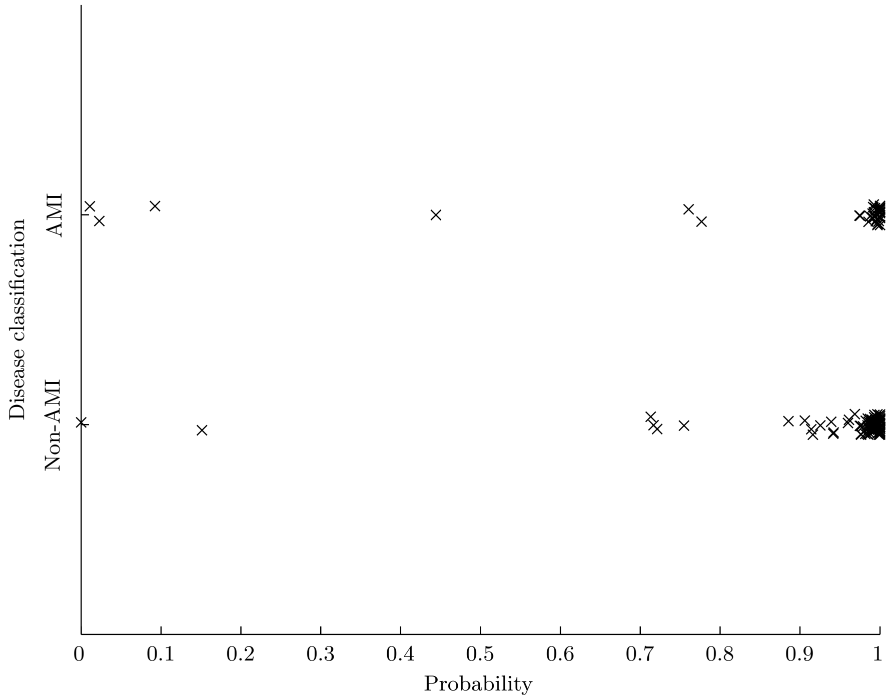

**Fig. 4** Dot diagram for logistic regression classifier using FABP, myoglobin on admission and CKMB, FABP and troponin I at $T1$

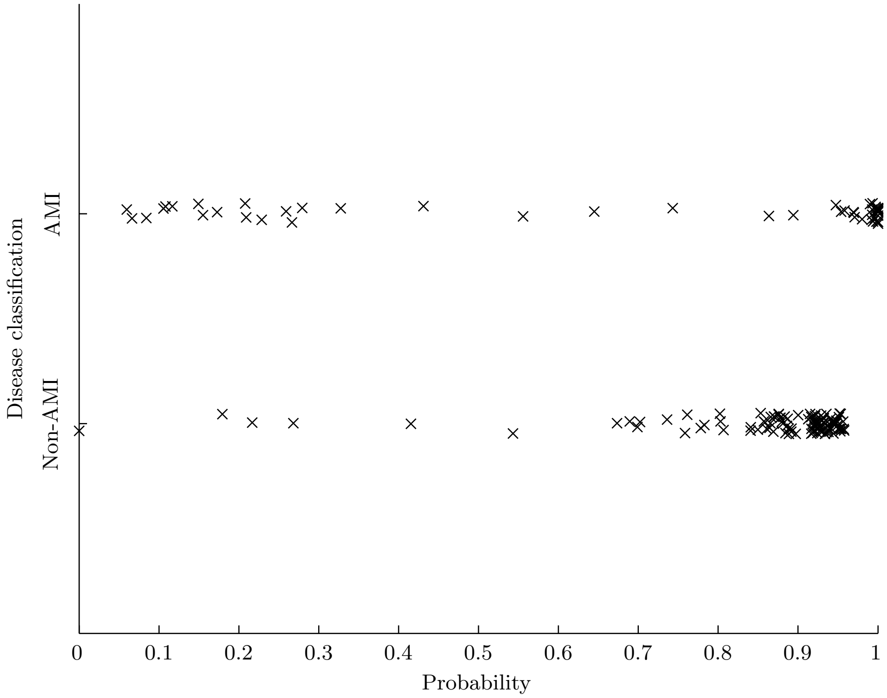

**Fig. 5** Dot diagram for logistic regression classifier using first two principal components at admission

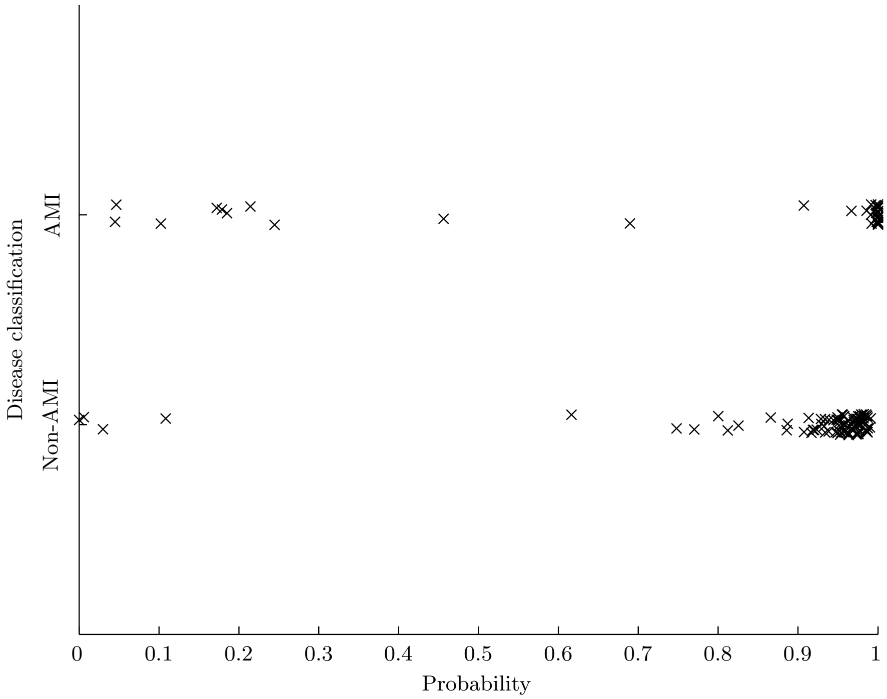

**Fig. 6** Dot diagram for logistic regression classifier using first five principal components at T0 and T1

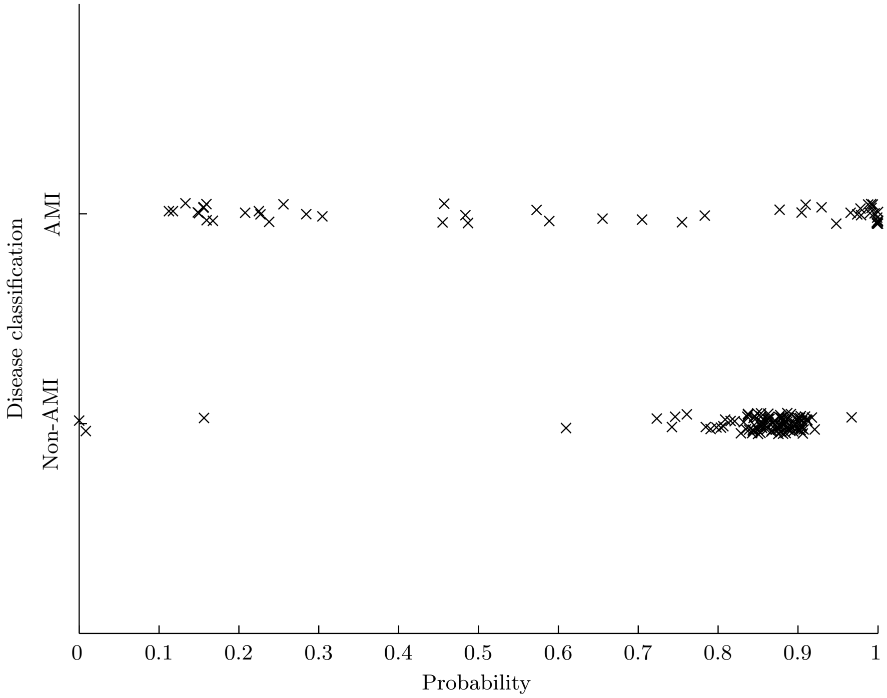

**Fig. 7** Dot diagram for logistic regression classifier using FDA pre-processing at T0

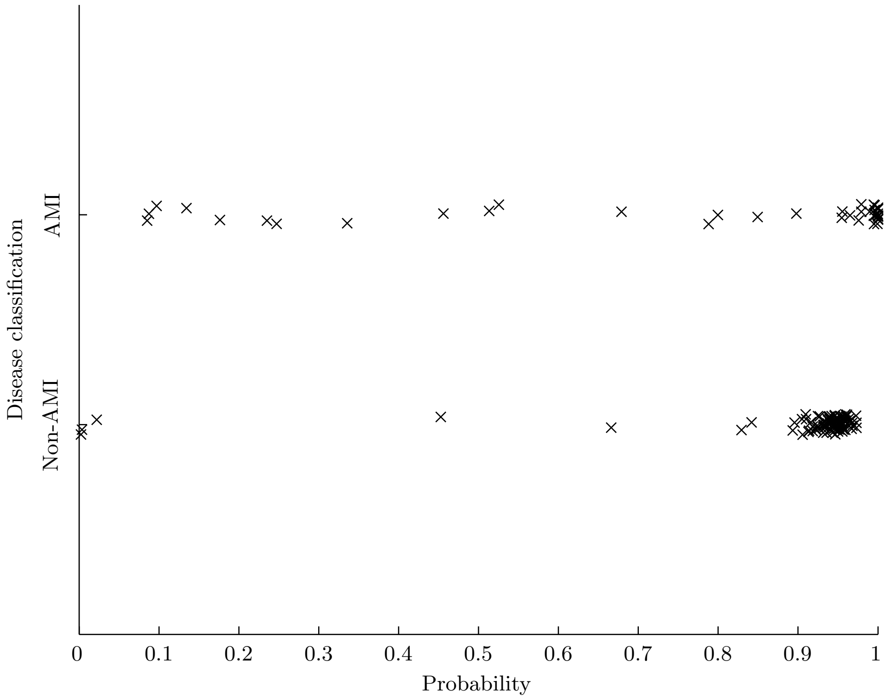

**Fig. 8** Dot diagram for logistic regression classifier using FDA pre-processing at T0 and T1

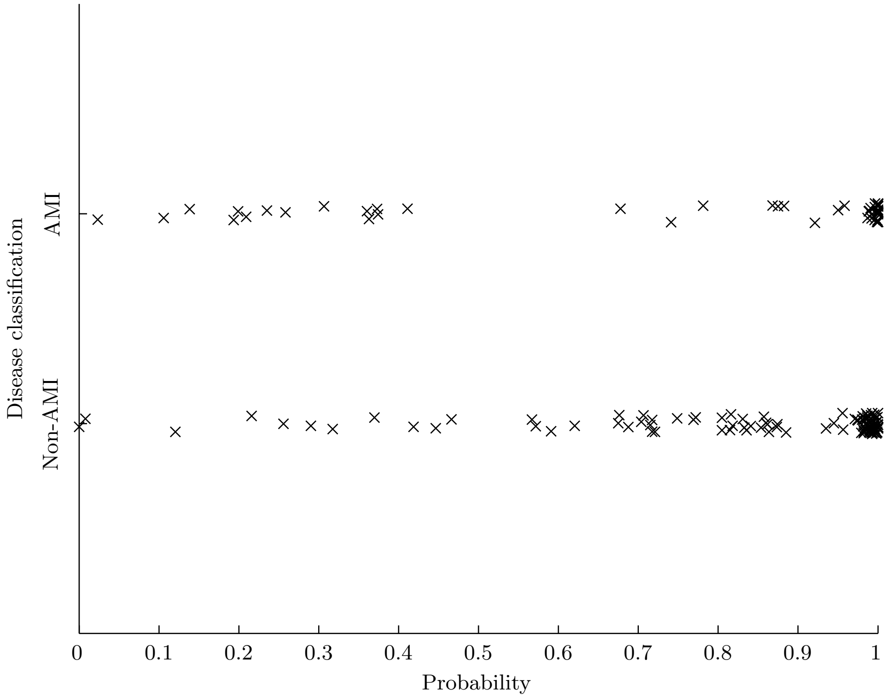

**Fig. 9** Dot diagram for GMM classifier using CKMB and GPBB at admission

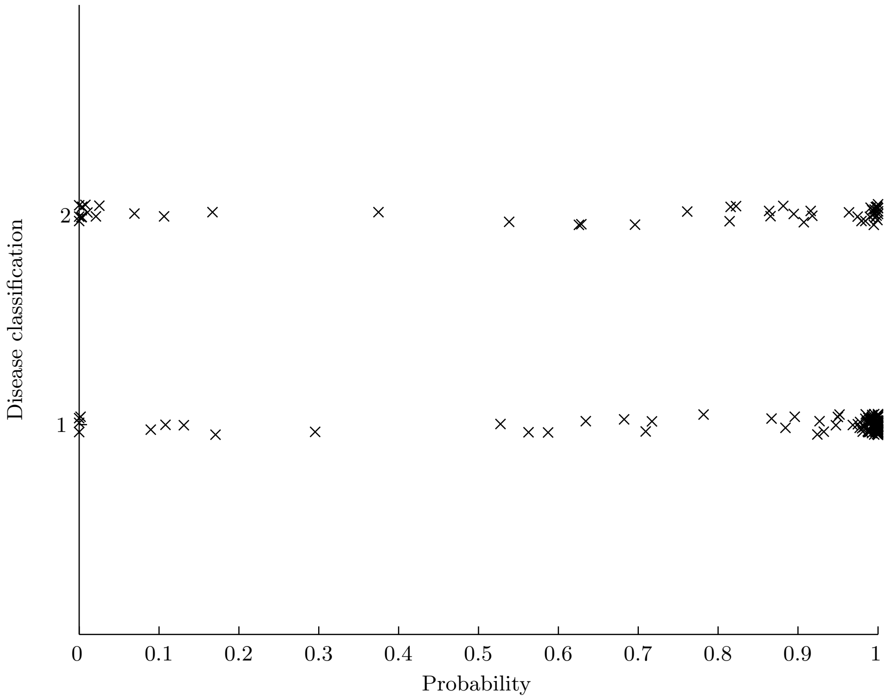

**Fig. 10** Dot diagram for GMM classifier using five principal components at admission

## List of Tables

| Classifier | NER | AUROCC | Brier score | Reliability |
|---|---|---|---|---|
| Logistic Regression (LR) | 0.868 | 0.913 | 0.904 | -0.013 |
| LR with PCA | 0.862 | 0.906 | 0.895 | -0.018 |
| LR with FDA | 0.849 | 0.908 | 0.887 | -0.001 |
| GMM | 0.843 | 0.918 | 0.896 | -0.030 |
| GMM with PCA | 0.855 | 0.840 | 0.861 | -0.099 |

**Table 1** Comparison of classification techniques based upon blood samples taken at admission (T0)

| Classifier | NER | AUROCC | Brier score | Reliability |
|---|---|---|---|---|
| Logistic Regression (LR) | 0.955 | 0.959 | 0.960 | -0.018 |
| LR with 5 PCs | 0.902 | 0.956 | 0.920 | -0.035 |
| LR with FDA | 0.902 | 0.923 | 0.923 | -0.014 |

**Table 2** Comparison of classification techniques based upon blood samples taken at T0 and T1

| Number of PCs | NER | Brier Score | Reliability |
|---|---|---|---|
| 1 | 0.843 | 0.883 | -0.009 |
| 2 | 0.836 | 0.877 | -0.017 |
| 3 | 0.862 | 0.895 | -0.018 |
| 4 | 0.855 | 0.895 | -0.019 |
| 5 | 0.855 | 0.896 | -0.023 |

**Table 3** Performance of logistic regression classifier with increasing numbers of principal components as inputs using samples on admission

| Number of PCs | NER | Brier Score | Reliability |
|---|---|---|---|
| 1 | 0.879 | 0.913 | -0.013 |
| 2 | 0.879 | 0.913 | -0.018 |
| 3 | 0.879 | 0.910 | -0.023 |
| 4 | 0.879 | 0.904 | -0.034 |
| 5 | 0.902 | 0.920 | -0.035 |
| 6 | 0.902 | 0.920 | -0.043 |
| 7 | 0.894 | 0.912 | -0.054 |
| 8 | 0.917 | 0.928 | -0.044 |
| 9 | 0.924 | 0.935 | -0.046 |
| 10 | 0.909 | 0.927 | -0.056 |

**Table 4** Performance of logistic regression classifier with increasing numbers of principal components as inputs and two time samples

| PC | Fractional Variance | Cumulative |
|---|---|---|
| 1 | 0.58 | 0.58 |
| 2 | 0.21 | 0.79 |
| 3 | 0.15 | 0.94 |
| 4 | 0.03 | 0.98 |
| 5 | 0.02 | 1.00 |

**Table 5** Variation captured by each principal component using samples on admission.

| PC | Fractional Variance | Cumulative |
|---|---|---|
| 1 | 0.58 | 0.58 |
| 2 | 0.14 | 0.72 |
| 3 | 0.10 | 0.82 |
| 4 | 0.08 | 0.90 |
| 5 | 0.03 | 0.93 |
| 6 | 0.02 | 0.96 |
| 7 | 0.02 | 0.97 |
| 8 | 0.01 | 0.99 |
| 9 | 0.007 | 0.99 |
| 10 | 0.006 | 1.00 |

**Table 6** Variation captured by each principal component using samples at T0 and T1.

| Number of samples | NER | AUROCC | Brier Score | Reliability |
|---|---|---|---|---|
| 1 | 0.8491 (0.784–0.901) | 0.908 (0.029) | 0.8874 | -0.0011 |
| 2 | 0.9015 (0.837–0.947) | 0.923 (0.022) | 0.9228 | -0.0142 |

**Table 7** Performance of classifiers trained using FDA preprocessing. The ‘number of samples’ column refers to the number of blood samples used from a given patient, so that ‘1’ indicated only blood taken on admission was used, and ‘2’ indicates blood taken on admission and at a later time.

| PCs | AMI centres | Non-AMI centres | NER | Brier score | Reliability |
|---|---|---|---|---|---|
| 1 | 1 | 2 | 0.811 | 0.842 | -0.083 |
| 2 | 4 | 5 | 0.811 | 0.853 | -0.065 |
| 3 | 1 | 3 | 0.843 | 0.856 | -0.110 |
| 4 | 1 | 3 | 0.811 | 0.852 | -0.111 |
| 5 | 4 | 5 | 0.855 | 0.861 | -0.099 |

**Table 8** Performance of GMM classifier with PCA pre-processing