Improving diagnosis of acute coronary syndromes in an emergency setting: A machine learning approach

Green, Michael

2008

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Improving diagnosis of acute coronary syndromes in an emergency setting: A machine learning approach

Michael Green

Computational Biology & Biological Physics group
Department of Theoretical Physics
Lund University

Thesis for the degree of Doctor of Philosophy

Thesis Advisor: *Mattias Ohlsson*
Faculty Opponent: *Paulo Lisboa*

To be presented, with the permission of the Faculty of Science of Lund University, for public criticism in lecture hall F of the Department of Theoretical Physics on Wednesday, the 18th of June 2008, at 10.15.

Organization **LUND UNIVERSITY** Department of Theoretical Physics Sölvegatan 14A SE–223 62 LUND Sweden	Document name **DOCTORAL DISSERTATION** Date of issue May 2008 Subject designator Project name
Author Michael Green	Sponsoring organization

Document title
Improving diagnosis of acute coronary syndromes in an emergency setting: A machine learning approach

Abstract: Acute coronary syndrome (ACS) is the biggest people killer in the western world today. Despite well trained physicians and reliable diagnostic tools, diagnosing ACS early in the emergency departments (ED) remains a challenge. In this thesis we used machine learning, via logistic regression models and artificial neural network ensembles, to investigate the possibility of predicting ACS at an early stage using electrocardiogram data. Thorough comparisons were made to several expert physicians, currently working in the ED, to verify the models. In the context of neural networks we developed methods for the case based explanation of their decisions.

Sammanfattning: Akut koronart syndrom (AKS) är den största folkdödaren i väst idag. Trots välutbildade läkare och bra diagnostiska verktyg så är det forfarande svårt att ställa en diagnos tidigt på sjukhusens akutavdelningar. I den här avhandlingen undersöker vi möjligheter att i ett tidigt skede förutsäga AKS med hjälp av maskininlärning. Främst användes logistiska regressionsmodeller och kommitteer av artificiella neurala nätverk (ANN). Jämförelser med expertläkare genomfördes kontinuerligt som en kvalitetskontroll. Vi utvecklade även praktiska patientbaserade förklaringsmodeller för ett ANNs beslutsprocess.

DOKUMENTDATABLAD
enl SIS 61 41 21

Key words
artificial neural network, ensemble, machine learning, acute coronary syndrome, electrocardiogram, decision support system, case based explanation

Classification system and/or index terms

Supplementary bibliographical information	Language English
ISSN and key title	ISBN 978-91-628-7434-6

Recipent's notes	Number of pages 147	Price
	Security class	

Distributor
Michael Green
Department of Theoretical Physics, Sölvegatan 14A, SE–223 62 Lund, Sweden

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources the permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature Michael Green Date 2008-05-21

For Josefin and Grandmother —
the two most amazing women I have ever known

This thesis is based on the following publications[1]:

I. Michael Green, Jonas Björk, Jakob Lundager Forberg, Ulf Ekelund, Lars Edenbrandt and Mattias Ohlsson
Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room
Artificial Intelligence in Medicine **38**, 305–318 (2006)

II. Michael Green, Mattias Ohlsson, Jakob Lundager Forberg, Jonas Björk, Lars Edenbrandt and Ulf Ekelund
Best leads in the standard electrocardiogram for the emergency detection of acute coronary syndrome
Journal of Electrocardiology **40**, 251–256 (2007)

III. Michael Green and Mattias Ohlsson
Comparison of standard resampling methods for performance estimation of artificial neural network ensembles
In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare*, Plymouth, England, E. Ifeachor (ed.), **31** (2007)

IV. Jakob Lundager Forberg, Michael Green, Jonas Björk, Mattias Ohlsson, Lars Edenbrandt, Hans Öhlin and Ulf Ekelund
In search of the best method to predict acute coronary syndrome using only the electrocardiogram from the emergency department
LU TP 07-42 *submitted*

V. Michael Green, Ulf Ekelund, Lars Edenbrandt, Jonas Björk, Jakob Lundager Forberg and Mattias Ohlsson
Exploring new possibilities for case based explanation of artificial neural network ensembles
LU TP 07-43 *submitted*

VI. Michael Green, Ulf Ekelund, Lars Edenbrandt, Jonas Björk, Jakob Lundager Forberg and Mattias Ohlsson
Explaining artificial neural network ensembles: A case study with electrocardiograms from chest pain patients
LU TP 08-06 *submitted*

[1] For collaborations within Computational Biology & Biological Physics, or with other theory groups, authors are listed alphabetically. For collaborations with medical groups, the ordering follows other conventions.

During my time as a PhD student, I have also co-authored the following publications:

- Jonas Björk, Michael Green and Ulf Ekelund
  Comments on 'Practical experiences on the necessity of external validation' by IR König, JD Malley, C Weimar, H-C Diener and A. Ziegler
  Statistics in Medicine (online early) (2008)

- Michael Green, Jonas Björk, Jakob Hansen, Ulf Ekelund, Lars Edenbrandt and Mattias Ohlsson
  Detection of acute coronary syndromes in chest pain patients using neural network ensembles
  In *Proceedings of the Second International Conference on Computational Intelligence in Medicine and Healthcare*, Lisbon, Portugal, J. Fonseca (ed.), pp. 182–187, (2005)

- Gunnar Andersson, Michael Green and Jan Komorowski
  Estimating risks at individual level using binary classifiers
  LU TP 08-08 (2008)

- Michael Green, Pawel Krupinski, Pontus Melke, Patrik Sahlin and Henrik Jönsson
  Costanza: a segmentation and analyzing tool for three-dimensional confocal data
  LU TP 08-09 (2008)

Contents

i

Introduction

The most valuable of all talents is that of never
using two words when one will do.
— Thomas Jefferson

A friend of mine once said that work is important. I think there may be a lot of truth in that statement. On my behalf though I would say that it depends on who you ask. I'm not saying that some peoples work is not important. Rather, I'm claiming that some of them might not consider it to be. For better or worse I'm actually one of the guys who thinks his work is important. This might make me a stuck up snob or a devoted scientist who believes that he one day might actually do something useful for the future. This thesis is the story of a four year pursuit of that goal.

I belong (at least for a few more days) to the department of Theoretical Physics. That being said, I don't actually do any physics of that kind in my work. Smashing particles together and observing what happens has never been my idea of fun. My skills lie primarily in the field of machine learning and artificial intelligence.

Over the next few pages I will throw you into the world of machine learning and heart related emergency medicine, and at least try to explain why these two, seemingly very different subjects, are worth combining. Naturally I will try to explain the medical problem we're faced with in a coherent manner, but I would like to stress that the explanations are from a modeling perspective, and as such they may lack an excrusiating amount of details.

i.1 The heart

Few things are harder to put up with
than the annoyance of a good example.
— Mark Twain

The heart is an amazing thing. It is vital to most processes in the body since it provides our other organs with blood carrying oxygen and nutrients by pumping it through the body via channels called vessels . These channels form a network that allows the blood to flow through the entire body. This is of course not entirely true. Obviously our nails do not have blood vessels, but you get the point, I'm sure. Anyway, the heart keeps us alive.

As you can see in Figure *i*.1 the heart basically consists of two compartments, where one of them is responsible for delivering oxygenated blood to the body, and the other one for pushing blood into the lungs for reloading. In principle the left compartment receives blood filled with oxygen from the lungs which it then pumps out through the aorta where the blood travels further to the upper and the lower part of the body in more and more finely grained vessels. Other equally small vessels take the de-oxygenated blood and transports it back to the heart where it enters the right compartment via the inferior and superior *venae cavae*. Once there, the blood is pumped out to the lungs via the pulmonary artery , where it is refilled with oxygen and then travels the pulmonary veins back to the left compartment, where the process began.

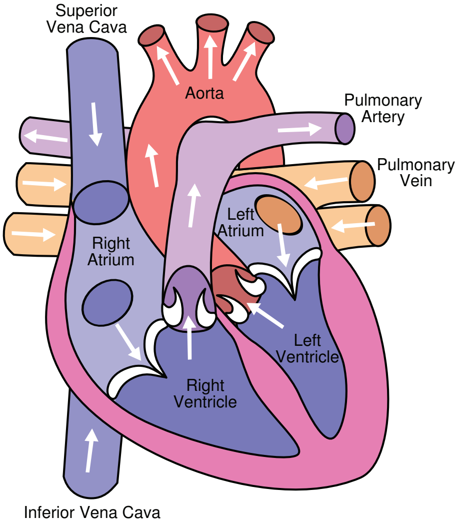

Figure *i*.1: Schematic overview of the compartments of the heart and the corresponding vessels leading into them.
Image by Eric Pierce.

The most active part of the heart is known as the myocardium which consists of a lot of muscle cells that allow it to contract. This contraction is what pumps the blood from the heart to the rest of the body. How does this really work though? How can a bunch of cells make sure that the heart pumps? Not only do the muscle cells have to contract, they have to do it in synchronized manner. In order to explain this I have to tell you about how a single muscle cell in the myocardium can contract. Are you ready? OK, a cell is basically a membrane separating some internal goo from the outside. The ion channels in this membrane are sensitive to differences in electric potential, *i.e.*, they open and close depending on the magnitude of this difference. This is useful since it allows the cell to maintain concentration differences between itself and the external environment. Indeed a vital property for a number of processes in cells in general. In order for the cell to contract it needs ions of potassium, calcium and sodium in concentrations varying over time. These ions are abundant around myocardial cells. When a stimulus reaches the cell it manipulates the ion channels permitting these ions to flow into or out from the cell due to the concentration differences. This in turn starts a process that makes the cell contract.

So now we have the basic picture of what happens during the contraction of a cell in the myocardium. However, the cells contract in a very distinct manner in the heart. This is accomplished by the fact that the contraction of a muscle cell stimulates its neighboring cells to contract as well. Effectively creating a cascade of contracting cells. It is this collaborative effort that pumps the blood.

I have neglected a lot of details here on purpose since describing them is way beyond the scope of this thesis, as well as my knowledge. In the next section I will go through what happens when the blood flow in the heart is restricted, and why this restriction occurs in the first place.

i.1.1 Acute coronary syndromes

A collection of heart diseases, or symptoms thereof, known as acute coronary syndromes (ACS) is the largest people killer in the western world today [1].

The basis for this disease is due to cholesterol getting stuck in the wall of an artery. Many believe that this only applies to the bad cholesterol [2, 3]. The presence of bad cholesterol suggests that there must exist some good cholesterol as well, and this is indeed the case. It mainly comes into play via a process called reversed cholesterol transport [4] where it removes cholesterol from the arterial wall back to the liver and then further excreted through the bile. Anyway, if enough cholesterol is gathered inside the wall an inflammation can occur, initiating the immune system to call in the cavalry, which in this case are the macrophages. These guys enter the arterial wall and start eating the cholesterol. Unfortunately they cannot process it, and keep eating until they eventually rupture releasing the cholesterol back into the arterial wall which again triggers more macrophages to come to the rescue. This creates a vicious cycle where more and more well fed macrophages collect and eventually form plaque that increases the pressure on the arterial wall and effectively reduces the amount of accessible volume of the vessel. This disease is called atherosclerosis [3] and we all have it to some extent. In the presence of severe atherosclerosis the amount of blood that can flow through the artery is limited, and renders that particular section of the tissue less efficient, a condition known as ischemia. If the pressure on the arterial wall is too high it may break, releasing the soup of macrophages and other stuff into the artery. This process causes a coagulation of the blood which creates a blood clot preventing any blood from flowing through. Thus no oxygen can be transported past this point in the artery. This is precisely what happens in the heart during a myocardial infarction, commonly known as heart attack, where parts of the heart muscles die due to shortage of oxygen. An illustration of what happens in the heart during the infarction is displayed in Figure *i*.2.

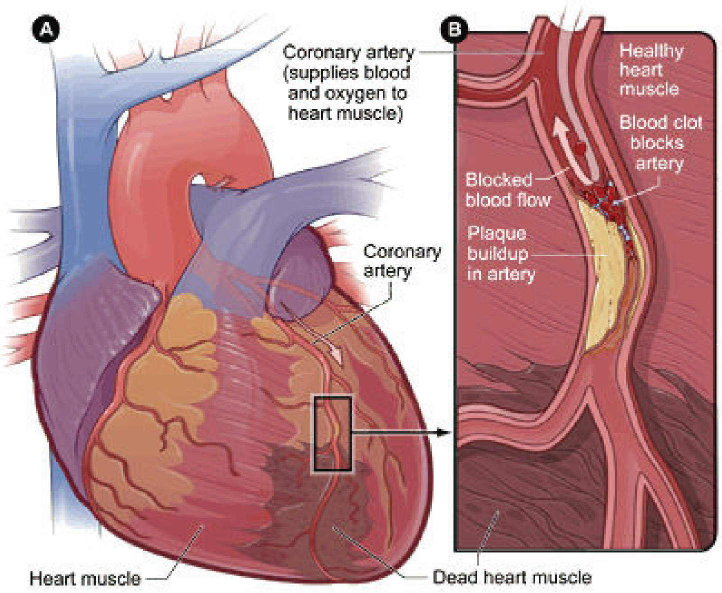

Figure *i*.2: Illustration of a myocardial infarction occurring due to an occlusion in a coronary artery.
Image by NHLBI Disease and Conditions Index topic, Coronary Artery Disease.

If the heart muscle is not damaged the condition is known as unstable angina which physically manifests itself in the same way. Now we are ready to state more clearly what acute coronary syndrome is. It is defined as either myocardial infarction and/or unstable angina.

i.1.2 The 12-lead electrocardiogram

In the early 20th century a physiologist named Willem Einthoven used a string galvanometer to measure the electric activity of the heart. He also assigned letters to the different features of this measurement. The name of the method was coined electrocardiogram (ECG). As of today it is still the most widespread way of examining the functional status of the human heart. It measures the electric activity of the heart, over time, by attaching 10 electrodes on the body; one for each arm and leg (extremity), and six on the chest (precordial). These electrodes are used to form 12 *leads*. Each of these leads can extract a time series, describing how the electric potential of the heart varies over time. These time series are also known as complexes. The leads are given distinctive names depending on which electrodes were used to record them. See Figure *i*.3 for an illustration.

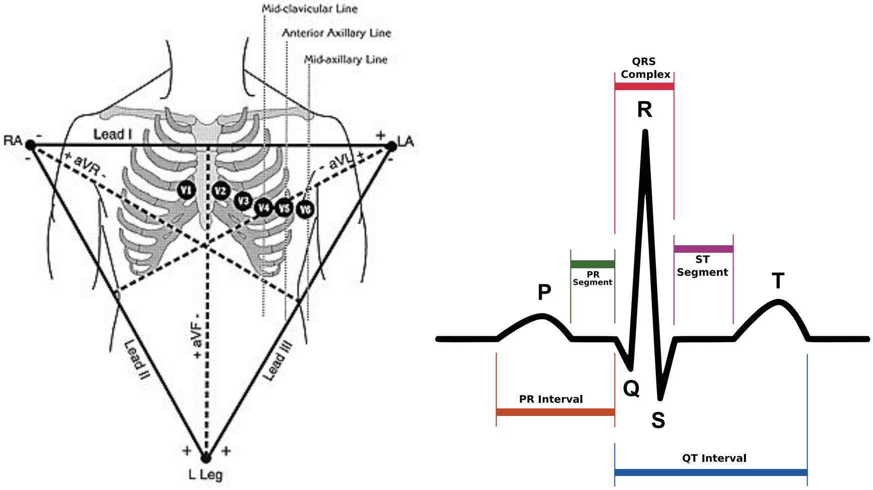

Figure *i*.3: The left picture illustrates the positioning of the electrodes used for recording the 12 ECG leads. The picture on the right shows a schematic overview of the different segments a typical lead in the electrocardiogram is divided into.
Images by F. G. Yanowitz and A. Atkielski.

Each lead is dived into segments, waves and intervals in order to facilitate the interpretation of it. The anatomy of a given ECG lead is presented to the left in Figure *i*.3 together with the typical shape of the time series. Einthoven managed to associate a number of heart diseases to specific deviations from this shape [5]. Even though more deviations and their corresponding diseases has been discovered in electrocardiography, the basic idea is still the same when trying to detect a coronary illness in the heart in the modern clinics today.

The standard ECG consists of 12 leads, six extremity and six precordial ones. The extremity leads are named I, II, III, aVL, aVR and aVF, meanwhile the precordial ones are called V_1-V_6. These leads and the corresponding electrodes used to record them are shown in Figure *i*.3. All of these leads view the heart from different angles which helps us to get a better picture of what is going on. The ECG may very well appear completely normal in one of the leads, while another may show pathological changes.

That being said, there are other systems available using fewer leads, and indeed recent studies have shown that the 12 leads contain a lot of redundant information [6–8]. Starting from the 12 leads one can effectively reduce it down to 8 by using Einthoven's relations, which states that any two extremity leads can be used to derive the remaining four. So any reduced lead set derived from these relations would at least contain leads V_1-V_6, and two extremity leads.

Remember that I talked about the different segments you divide the time series from a given lead into? Well now it is about time we learn exactly what these segments mean and how they are in fact generated. Each of the segments represent a specific event of a heart beat, *i.e.*, the right picture in Figure *i*.3 illustrates the electric activity in the heart during one heart beat as viewed from a specific angle. The muscle cells in the myocardium are surrounded by a conductive medium which means that when one muscle cell is hit by the activation potential this activity is transferred to its immediate neighbors. This creates a synchronous flow of depolarizations in the heart, resulting in a pumping phenomenon. This flow can be described in terms of an electrical vector. If you don't know what a vector is, don't worry. It is basically just an arrow pointing in a direction of interest. Usually the length of that arrow is important as well. At least in mathematics and physics. In this description we can neglect it though. Nice huh? Anyway, this vector describes the direction in which the depolarization wave is traveling at the moment (see Figure *i*.4). The different segments of the recording from an ECG lead can be explained in terms of this vector.

The P wave (see Figure *i*.3) is generated during the depolarization of the right and left atrium, when the vector is pointing approximately 30° downwards. Just after this depolarization there is a delay before the right and left ventricle start their process. It starts with the depolarization of the left part of the wall separating the two ventricles from each other. This is the start of the QRS segment where the electrical vector is pointing horizontally from the left ventricle to the right, and gives rise to the Q wave. As the depolarization continues to spread to the rest of the walls surrounding the ventricles, it results in a positive reading in the ECG which gives us the R wave, and later the S wave. The S wave which is a manifestation of the depolarization of the basal parts of the left ventricle wall, is not always present in the ECG though. During this part the electrical vector is directed approximately 45° upwards. The last part of the ECG is the T wave, which describes the repolarization of the ventricles. Fine, but something is missing from this picture. What about the repolarization of the right and left atrium? Well their repolarization occur at the same time as the depolarization of the ventricles which means that it will not show on the ECG.

That is basically it. Now you have a fairly detailed description of what the different parts of an ECG are generated from. I've described the process from one lead only and indeed the ECG will look slightly different when recorded from other leads because they view the heart from different angles. Still, the overall procedure is the same.

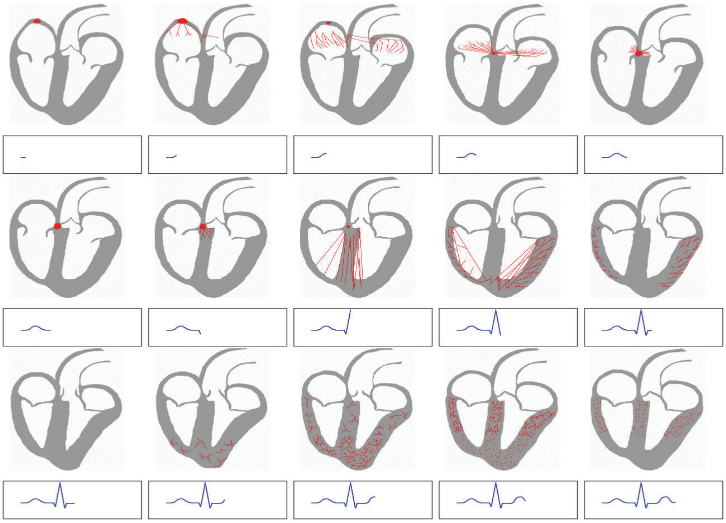

Figure *i*.4: Illustration of how the depolarization wave travels through the heart and the corresponding potential differences this causes as seen from an individual lead in an electrocardiogram.

i.2 Machine learning and the art of prediction

I never think of the future. It comes soon enough.
— Albert Einstein

Now that you know just how amazing the heart is and all the horrible malfunctions it can suffer from, it is time to start thinking about what we can do to prevent them from happening. The best way to recover from a heart disease is to discover it in the early stages. In the later stages, when parts of the heart has actually started dying, the problems, whatever they may be, are always more difficult to fix. Fortunately the ECG lets us peek into the activity of the heart so that we can see what is going on. Fine, so we know that we can monitor the heart and that deficiencies should appear in one or several of the ECG leads. I use the word "should" here since it is well known that about 50% of the patients with acute myocardial infarction (part of acute coronary syndrome) show no apparent ECG changes [9–12]. Even so the physicians still use the ECG when assessing whether a patient has acute coronary syndrome or not, but often combine this information with biomarkers from blood samples, patient history and so on [13]. When assessing the ECG, there are a number of criteria that can help identify different heart diseases, and every physician learns this during their education. Expert physicians tend to develop their own set of criteria, or gut feeling really, built from years of experience interpreting these kind of data. What I would like to do is to tap into their knowledge by using machine learning.

i.2.1 Machine learning

In science these days everyone is more or less exposed to the field of machine learning, though they might not realize it of course. Machine learning is basically anything that involves allowing a computer to learn from data by discovering patterns. Though I understand that the notion of a pattern might seem elusive to some readers, it is basically very simple. Let us go through an example. Say that we are given the weight and height of 1000 newly born babies. By plotting the weight against the height we might discover that in general really tall babies seem to weigh more than really short ones. This is a pattern. This whole thesis deals with the search for these kind of patterns in chest pain patients visiting an emergency department.

The problem of learning for a computer is very similar to those of humans. Either we have a teacher that will provide us with examples and explain them to us. Or we are faced with a problem without any teacher present. These two cases are called supervised and unsupervised learning respectively. The unsupervised approach is mainly useful when we have little or no intrinsic knowledge about the data we are trying to learn something from. So in this thesis I will only deal with supervised algorithms. Further, I will also limit the scope of this introduction to binary classification problems.

I want to start by apologizing for the technical nature of this part of the thesis. In my defense I have to say that talking about machine learning without equations is a bit like brushing your teeth without tooth paste. You can do it, but the fresh satisfying sensation afterwards will not be there.

i.2.2 Generalized linear models

In this section I will describe a neat family of classifiers called generalized linear models (GLM) [14]. These models all take the following form

$$y(\boldsymbol{x}, \boldsymbol{\omega}) = f\left(\sum_{i=1}^{M} \omega_i \varphi_i(\boldsymbol{x}) + \omega_0\right) = f\left(\sum_{i=0}^{M} \omega_i \varphi_i(\boldsymbol{x})\right) \qquad (i.1)$$

where ω_i are the parameters we would like to fit, φ_i are the basis functions, and $M+1$ is the number of parameters in our model. In the last step of Equation *i*.1 I have included the bias ω_0 in the sum and defined $\varphi_0(\boldsymbol{x}) = 1$ in order to make the notation more compact. This allows me to write it down in terms of a dot product

$$y(\boldsymbol{x}, \boldsymbol{\omega}) = f\left(\boldsymbol{\omega} \cdot \boldsymbol{\varphi}(\boldsymbol{x})\right).$$

The basis functions can be chosen arbitrarily and may of course be used to transform our data to anything we want. This also includes non linear transformations which actually makes you wonder why we call it linear models. Well the reason is that the models are still linear with respect to the ω parameters. As long as this condition is true, we are still in the class of linear models. This is also independent of the choice of f. Speaking of which, what would be a sensible choice for that function? In the case of classification the preferred choice is usually the logistic sigmoid

$$f(x) = \frac{1}{1 + e^{-ax}} \tag{i.2}$$

where a is a parameter that is often set to unity. This effectively squashes the linear combination of the variables into an interval between 0 and 1 which is neat since we can interpret it as a probability. In this thesis all GLMs have been developed using the basis functions $\varphi_i(\boldsymbol{x}) = x_i$, which transforms equation *i*.1 to

$$y(\boldsymbol{x}, \boldsymbol{\omega}) = f(\boldsymbol{\omega} \cdot \boldsymbol{x}) \, . \tag{i.3}$$

These kinds of linear learning algorithms are known as logistic regression models, and are common in statistical prediction literature. Due to their simplicity they are relatively easy to analyze. For instance, you can assign error bars to any given prediction giving you a measure of how confident the method is for a specific instance. As we move into a more powerful and intrinsically more complicated method called artificial neural networks, this neat property tougher to achieve [15].

i.2.3 Artificial neural networks

One of the possibly most famous families of machine learning algorithms in the world is known as Artificial Neural Networks (ANN) [16]. Mostly they are known due to their unfortunate association to the human brain. They have actually little to do with the way the human brain works, and even though they are mentioned in the Terminator movies they are also seldom used to create killing machines. That being said, they are quite powerful learning machines. Especially the multi-layered perceptron (MLP) has been given a lot of attention, which, given enough data, can learn almost any problem really well [17,18]. Indeed, in this thesis, I will make no distinction whatsoever between ANN and the MLP and will consequently use both terms interchangeably. The MLP is a generalization of the perceptron [19, 20], a computational model inspired by the neural cells in the brain, which was introduced in the late 1950's by Frank Rosenblatt. This brilliant man also provided a learning strategy for it that guaranteed convergence on linearly separable problems [21].

A graphical illustration of the MLP is shown in Figure *i*.5. Here each blue unit, in the hidden layer, carries the same computational power as the perceptron, and has the functional form of equation *i*.3. So each of them can be viewed as a logistic regression model that we feed the inputs, in the green layer, to. What the MLP does is that it connects all of the outputs from these hidden units into a final output in the purple layer in Figure *i*.5. This three-layer architecture is very typical for MLPs in general but in principle, we could have several hidden layers allowing for even more complex

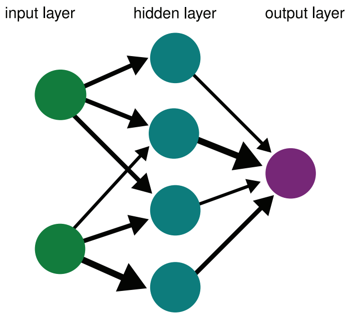

Figure *i*.5: Illustration of a feed-forward multi-layer perceptron. The thickness of the arrows indicate the magnitude of the weights leading into a given unit.

models. In practice however, one hidden layer provides us with more than enough complexity since we can add as many hidden units as we want within a given layer. In fact, when training the MLP it is the complexity, rather than the lack of it, that messes things up for us.

There is, in my opinion, an even nicer way of viewing neural networks. Take another look at the linear model in equation *i*.1, and recall that we could choose the basis functions φ_i any way we wanted to. Well there is nothing stopping us from defining them as a GLM themselves that depend on another set of adaptable parameters $\tilde{\boldsymbol{\omega}}$. This way the linearity of the trainable parameters is lost and we have left the class of linear models and moved into neural networks. Equation *i*.1 would then have turned into

$$y(\boldsymbol{x}, \boldsymbol{\omega}) = f\left(\sum_{i=0}^{M} \omega_i \cdot \varphi_i\left(\sum_{j=0}^{J} \tilde{\omega}_{ij} \cdot x_j\right)\right) \qquad (i.4)$$

where M is the number of hidden nodes in the network, J is the number of variables, and the $\tilde{\omega}_{ij}$'s is the new set of parameters that are absorbed into $\boldsymbol{\omega}$.

Training the neural network

The ANN is only useful once trained on examples. In other words, once we have fitted the parameters of the ANN to data. The first thing we need to concern ourselves with is choosing an error function that will provide us with information of how well the network fits the data. Since I am dealing only with classification in this thesis the most appropriate choice is the cross entropy error function [22]. Why? Consider the problem at hand. We are given a data set consisting of a set of variables and their corresponding class labels. These labels can be interpreted in a win or lose situation. Imagine the following scenario, you are on your way to limbo [23] when a strange creature gives you a chance of returning home without harm. Naturally you take this chance. The creature explains that he will toss a coin. Heads you win, tails you lose. Since there are only two possible outcomes of this game the only thing you have to fear is the fairness of the coin. Being nifty with math you decide to turn to the Bernoulli distribution for help. This distribution tells you the probabilities of the two different outcomes and looks like this

$$p(t) = y^t (1 - y)^{1-t} \tag{i.5}$$

where $t = 1$ or $t = 0$ if the coin comes up heads or tails respectively, and y is the probability of heads. But wait! This does not really help you since you have no more information about y. It could be anything. Even though it might seem like a good idea to say that the probability is 0.5 for both heads and tails, it could very well be that the coin is constructed such that it ends up heads two times out of three. So how would we know what to guess? Well to tell you the truth you need data. So it looks like there might be a one way ticket left to limbo for you. But if you had data, basically a lot of t_n's, you could use them to infer the most probable value of y given that data. We can use this idea with our variables and labels as well for the training of our ANN. But we would have to take it one step further. We still use the Bernoulli distribution but we need to use it for each data point we have access to and push it together into a likelihood function

$$\mathcal{L} = \prod_{n=1}^{N} p(t_n). \tag{i.6}$$

This requires that our data are independent and identically distributed. Maximizing this function would give us a fair chance of tuning the parameters of the ANN. In this setting $y_n = y(\boldsymbol{x}_n, \boldsymbol{\omega})$ is still the probability of heads, *i.e.*, a win, but modeled by our ANN as a function of a data point $\boldsymbol{x}_n$ and some parameters $\boldsymbol{\omega}$. This function looks a bit nasty though so we instead minimize the negative log likelihood, which is way easier to deal with mathematically. Anyway, by negating the logarithm of equation *i*.6 and expanding we get

$$\begin{aligned} E &= -\ln \mathcal{L} = -\ln \prod_{n=1}^{N} p(t_n) = -\sum_{n=1}^{N} \ln\left(y_n^{t_n} (1 - y_n)^{1-t_n} \right) \\ &= -\sum_{n=1}^{N} \left[t_n \ln y_n + (1 - t_n) \ln(1 - y_n) \right] \end{aligned} \tag{i.7}$$

which is known as the cross entropy [24] error function for two classes. Thus, maximizing the likelihood of our target variables is equivalent to minimizing the cross entropy error function. This is interesting since it provides us with two different ways of viewing the same thing. Either we try to maximize the probability that the data we have were generated from our model. Or we minimize the error our model makes when predicting the targets. I hope that this small excursion has convinced you that cross entropy is the way to go when dealing with classification in ANN.

So now there is an error function for us to use and the next step is to minimize this error. Though there are numerous such algorithms out there, I will stick to the ones I know. The most naive thing to do is to just use steepest descent [25], which basically lets the network slide down to the nearest valley in the energy (error) landscape (see Figure *i*.6). This approach can be spiced up by taking into account how far you want to slide, how far you slid last time etc. Still we are only taking first order information into account here, since basically only information from the gradient is used. If you want to get more fancy you could use a second order approach called Quasi-Newton [25, 26] which approximates the energy landscape as a big bowl. This approximation can be horrible in some cases which means that far from all minimization problems are suitable for this method. In my experience it rarely works well.

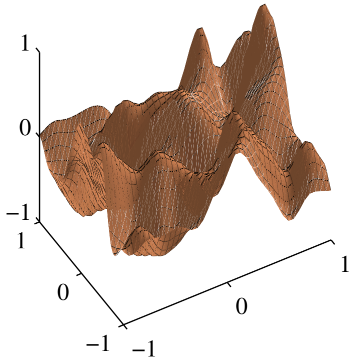

Figure *i*.6: Illustration of a possible energy (error) landscape where the goal is to find the deepest valley. To understand what a difficult task a minimization algorithm is faced with you should imagine putting a guy with limited eye sight in this landscape and tell him to find the deepest valley.
Image by Chris Bishop.

Controlling the complexity of the network

Neural networks is a powerful classification tool and can easily adapt to almost any pattern it encounters in the data. However, since data is scarce and noisy we seldom want the network to learn blindly from it. In principle the network could memorize the data and have excellent performance. The problem is when the model is exposed to new data that was not present during the training. The performance of the network on this new data set could be awful since it didn't really learn anything, it just memorized. This is why the complexity of a neural network has to be controlled, since it will otherwise often overfit the data. What happens during an overfit is that the parameters get large positive or negative values which allows the neural network to adapt perfectly to each data point. Because we never provided any restriction in the sizes the parameters may typically take, they will grow as much as they can. If you remember the likelihood function from our discussion about training the neural network you will see that we indeed did not enforce any restrictions on the parameters. The likelihood function in equation *i*.6 can be more formally written as

$$\mathcal{L}(\boldsymbol{\omega}) = p(\boldsymbol{t}|\boldsymbol{\omega}) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n} \qquad (i.8)$$

where $\boldsymbol{t}$ is our target data, and $\boldsymbol{\omega}$ are the parameters we want to fit. As before we denote the ANN output and target, for the nth data point, as y_n and t_n respectively. We can extend this formalism by casting it into a Bayesian framework. In this setting we model

$$p(\boldsymbol{\omega}|\boldsymbol{t}) = \frac{p(\boldsymbol{t}|\boldsymbol{\omega})p(\boldsymbol{\omega})}{\int p(\boldsymbol{\omega},\boldsymbol{t})d\boldsymbol{\omega}} \propto p(\boldsymbol{t}|\boldsymbol{\omega})p(\boldsymbol{\omega}) \qquad (i.9)$$

instead where $p(\boldsymbol{\omega})$ constitute a prior probability distribution of our parameters. We can neglect the integral in the denominator since it is only a normalizing constant. The prior reflects our belief about the values we think are reasonable for the parameters to take, before we have seen any data. If we don't know a lot about the problem we might expect this distribution to be *Normal* with a mean and a variance of 0 and $\frac{1}{\lambda}$ respectively. Because the parameters in the neural network can adapt both positive and negative values it's reasonable that they on average should be zero. In addition to this the variance $\frac{1}{\lambda}$ controls how much they are allowed to deviate from the mean. As before, instead of maximizing $p(\boldsymbol{\omega}|\boldsymbol{t})$ we minimize its negative logarithm and end up with the following expression

$$-\ln p(\boldsymbol{\omega}|\boldsymbol{t}) = \underbrace{-\sum_{n=1}^{N}\left(\ln y_n^{t_n} + \ln(1-y_n)^{1-t_n}\right)}_{E} + \underbrace{\frac{\lambda}{2}\boldsymbol{\omega}^T\boldsymbol{\omega}}_{E_\omega} + const \qquad (i.10)$$

which we recognize as equation *i*.7 with an added term, for the parameters $\boldsymbol{\omega}$, in the end. The $const$ term we can neglect since it does not depend on our parameters and will consequently have nothing to do with the training process. So by putting a prior on our parameter distribution, we defined a new error function $E_{reg} = E + E_\omega$ where E and E_ω is defined by equation *i*.7 and *i*.11 respectively.

$$E_\omega = \frac{\lambda}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} = \frac{\lambda}{2}\sum_{i=1}^{I}\omega_i^2. \qquad (i.11)$$

Let's take a moment to think about the implications here. We have an error function that we want to minimize that consists of the cross entropy error and an error term penalizing large parameters. How much we penalize is controlled by λ which is also the inverse variance of the prior distribution of those same parameters. So by cranking up the value of λ we penalize more and consequently decrease the amount our parameters are allowed to deviate from 0. In the neural networks literature this particular E_ω is known as *weight decay* [14, 27]. In this thesis, however, I have used an modified version called *weight elimination* [27]

$$E_\omega = \frac{\lambda}{2} \sum_{i=1}^{I} \frac{\omega_i^2}{\omega_0^2 + \omega_i^2}$$

which can be interpreted as a prior consisting of a mixture of a Normal and a uniform distribution [28]. The price you have to pay for introducing this penalty term is another parameter to tune. In practice though you set $\omega_0 = 1$, and only worry about tuning λ.

The modification of the cross entropy error function to include a term for controlling the size of the parameters gave us an effective way to govern the complexity of a neural network. In fact, we can be generous with the number of nodes in the hidden layer and then penalize as much as is needed in order to make the network generalize well.

i.2.4 On combining classifiers

A fairly recent trend in machine learning is to combine the predictions of many different models in order to improve performance. These new models are known as ensembles or committee machines. Another commonly used name is Hybrid machines, which often refers to the combination of models from different families. Though the combination of classifiers might not be immediately obvious there are many reasons why it is a good idea. For instance, in the context of neural networks, it provides a smoothening effect that may help increase the reproducibility of the results. The most interesting thing with ensembles though is that they improve the performance. In fact it can be shown that the average error an ensemble makes is always less than the average error made by its members [29]. So on average we always increase the performance by combining models. This can be shown by decomposing [30] the error function into bias and variance. In order to gain something from combining the individual models they have to differ in their predictions. In fact the more dissimilar they are the better the effect of combining them. A number of approaches for creating diverse ensemble members from a data set has been proposed [31], and many of them are based on resampling (see section *i*.3.2). Examples are bagging [32] and cross validation ensembles [33] where the idea is that you pick several, not too similar, subsets of the data and use them to generate the models. The way you perform this resampling actually matters to the performance of the ensemble [33].

Assuming that you are able to generate a lot of predictive models, how will you go about combining them? Well, the most simple thing you can do is to just let the ensemble members contribute equally to the final prediction by using the average of their individual outputs.

$$y_{ens} = \frac{1}{M} \sum_{m=1}^{M} y_m \qquad (i.12)$$

This is democracy at work, since you do not let a given model's individual performance affect its contribution. This could of course be a really stupid strategy since some of the models may have a track record of making bad predictions. Should you trust it anyway? I think not. In cases where you might suspect one of your models to be less than fit, there is a generalization of the combination strategy in equation *i*.12 that will allow you to lower the impact of that particular model's decision. It works by assigning a weight α_m to each ensemble member's prediction

$$y_{ens} = \sum_{m=1}^{M} \alpha_m y_m \qquad (i.13)$$

where you require the weights to sum to unity. This scheme is flexible enough to let you select only the well performing models and combine their predictions. In my experience though you rarely get models that perform badly enough to warrant exclusion. Some authors claim otherwise [34].

i.3 Assuring the quality of a machine

> *The trouble with the world is not that people know too little,*
> *but that they know so many things that ain't so.*
> — Mark Twain

i.3.1 Quantifying a models ability to generalize

In a perfect world we would always have all the data we need to represent a given classification problem. However, the world is far from perfect in this respect. This means that data is usually scarce and we have to be very careful not to overfit our models. Overfitting essentially occurs when a model begins to adapt too well to the specific training set. This will lead to an increase of the error when measured on an independent data set not previously seen by the model. This independent data set is often called validation set or test set. In principle they are are used for different stages during the construction of a prediction model, but in this thesis there is no distinction. An example of overfitting is presented in Figure *i*.7 where the error the model makes, during training and testing, is plotted against its complexity.

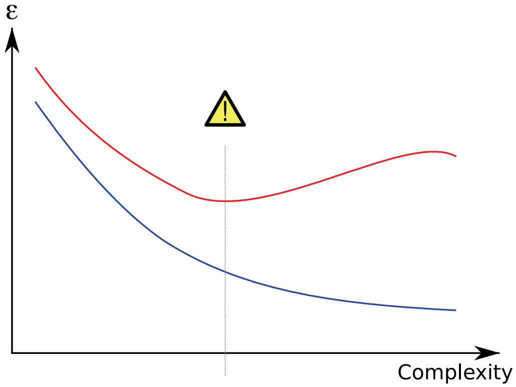

Figure *i*.7: Plot of the training and validation error against the complexity of the model. The blue and red line represents the error from the training and validation data respectively. The dashed line marks the point where overfitting starts to occur.

This means that when developing models and later adapting them to data, we need to have an independent dataset to validate the performance on. Otherwise we would never know if our model just memorized the data or if it actually learned something. Fine, so all we have to do when given a classification problem and a dataset is to divide it into a training set and a validation set right? Well not quite. It turns out that you will most likely get different results depending on the way you divide the data. Thus, you would have to do this splitting several times and use the average of all the validation results as the performance measurement. Exactly how we choose to generate these repeated training and validation sets will be discussed in section *i*.3.2.

i.3.2 Resampling and the illusion of having enough data

The whole idea of resampling, at least in the context of this thesis, is to make the most of the data you have by drawing samples from it in more or less clever ways. The purpose of resampling is often that there is some property of the model or the data that you want to estimate, *e.g.*, classification accuracy and its corresponding variance. Another reason for doing resampling is for producing many diverse models to put together into an ensemble classifier. Section *i*.2.4 explained why this is useful. This section, however, will describe some of the more common methods of resampling.

Hold out

So you have all this data and you know that you have to build your model on a subset of the data and test it on the rest. The simplest thing you could do is cut the data in half, using the first half for training and the other one for testing. But it might be that using 50% of the data for testing is a bit much since you won't have enough data to build your model from. In this case you may want to reserve only $\frac{1}{3}$ of the data for testing. In the end this reasoning ends up with you choosing a fraction of the data to put away for testing. This way of splitting data is called Hold out, since you hold out a fraction of the data.

K-fold cross validation

Probably the currently most used method of resampling in order to quantify generalization performance of a classifier is K-fold cross validation [35]. This method basically splits the dataset into K equally sized parts and uses each of these parts as a validation set while training on the rest of the K-1 parts. Obviously this will give K measurements of the classifiers generalization ability, and usually one reports the average of these K values as the final performance measure. An example of this resampling strategy can be found in Figure *i*.8 where I have used $K = 3$.

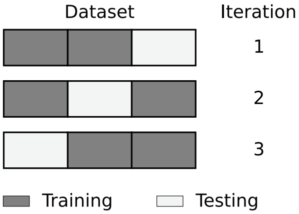

Figure *i*.8: Illustration of a K-fold cross validation with $K = 3$. The dark gray boxes correspond to the training data and the light gray to the data that will be used for testing.

Bootstrap

In an old German legend Baron Münchhausen managed to get himself out of a swamp by pulling himself up by his bootstraps. This is where the resampling method bootstrap [36–38] got its name from. The method samples from the original data with replacement meaning that some data points will be present more than once in the new resampled dataset. However, there will also be some data points left out. These points are used as the validation set. On average, given that the size of the original dataset is N and we sample N points from it, $0.368N$ points will be left for the validation set. Typically this procedure is repeated many times times in order to get as many data sets as we want.

i.3.3 The receiver operating characteristics curve

The area under the receiver operating characteristics (ROC) [39,40] curve is often used in machine learning to evaluate classification performance of classifiers generating class probabilities. For instance, when building predictive models for ACS we typically let the output of our machine learning algorithm model the probability (risk really) of a given patient having the disease. In order to transform this probability into a decision we need to assign a probability cut-off between 0 and 1 where outputs above that cut is classified as ACS. The naive evaluation of a model's performance is *accuracy*, *i.e.*, simply calculating the number of correctly classified patients divided by the total number of patients being predicted. This measure, however, depends on the prevalence of ACS, which is basically a fancy word for describing how common the disease we're modeling is in the data. The dependency on the prevalence can be problematic during resampling since it can, and often will, vary. In the worst case we will end up with a sample from the data containing only instances from one class. A stupid classifier always predicting that particular class will thus get an accuracy of 100%. This is obviously not very useful. One solution to this problem is to make accuracy class-dependent, meaning that within each class we check how many true predictions the classifier has produced. These measures are often called sensitivity (for the class being modeled) and specificity (for the other class), but they have numerous other names [40], *e.g.*, precision, recall, etc.

One important limitation still remains though. We still enforce a cut-off for the classifier's output. This cut-off affects sensitivity, specificity and accuracy alike. Luckily we can get around this problem too, by using the ROC curve. This curve is built by systematically increasing the cut-off, from 0 to 1, and calculating sensitivity and specificity for each cut. Although the ROC curve is interesting in itself it's actually the area under it that we want to evaluate because it can tell us how well our model really works. In the context of predicting acute coronary syndrome this area can be interpreted as the probability of our classifier giving a randomly chosen patient from the ACS group a larger predicted risk than a randomly chosen non-ACS patient. An optimal classifier that always delivers good predictions will have an ROC area of 1. An ROC area of 0.5 means that our classifier is no better than random guessing, and we might as well just toss a coin. If the number drops below 0.5 we've managed to do something rather strange since the classifier managed to model the wrong class. Whether we model the right class or not, the area under the ROC curve is still a brilliant tool for evaluating a classifier's predictive ability independent of class distributions.

i.4 Medical decision support systems

Life's like a box o chocolates, you neva know what you gonna get.
— Forrest Gump

A whole new field of research has grown from the machine learning community. This new trend deals with providing support for specialists making difficult decisions. Often by applying one or many of the fancy algorithms already present in the machine learning toolbox. In the medical domain these kind of systems are still in the proof of concept stage, *i.e.*, very few live applications are actually used in the clinical practice. There are numerous reasons for this lack of usage [41]. The most important ones being insufficient speed of the system and poor integration into the clinical work flow. I would say that speed is not a concern with neural networks since once they are trained, prediction is immediate. Really it's a matter of a fraction of a second. The integration with the present clinical work flow is much more complicated but is more related to engineering than science. So what is the problem then? Well, it turns out that one of the major scientific issues with decision support systems is to make them work well on several different hospitals at different times and conditions. So far hardly any work has been addressing these issues seriously since data is hard to come by, and multi-center studies are expensive and time consuming. This is also related to another field of research that deals with storing and assessing the quality of data. I won't even get started on that. To top this off there is another major problem to overcome for neural networks, and that is to explain their reasoning to their users.

i.4.1 Of neural networks and black boxes

Unfortunately for neural networks physicians always want to know why a certain patient has been diagnosed with a specific disease. I say unfortunately since this is the Achilles' heel [42] of neural networks. Due to their powerful pattern finding abilities, where several variables can be linked to the decision in non-linear ways, it is rather difficult to extract the underlying reasoning. This limitation is the reason why neural networks are often called black box methods. You put a question into it, and out comes an answer, but you have no idea how it was processed. So why is nobody looking? Well to tell you the truth, a lot of people have looked and quickly turned their eyes elsewhere. The experience resembles a situation when you are trying to find out why the computer does not start and consequently open it up in order to locate the broken circuit. A scenario that might very well end up with you running away pulling your hair and screaming "God, oh God why?!".

That being said there are actually quite a few people who have persisted [43], and developed tools for extracting knowledge from trained neural networks. There have been many different attempts to attain this knowledge but the mainstream of them can be divided into pedagogical and decompositional methods. The decompositional approaches scrutinize the network from within, measuring weights and activations of hidden nodes etc. This is a powerful way of attacking the problem since you have access to the entire functional structure of the neural network. On the other hand you may not need to know about that structure if all you want to do is to understand what is happening. This is precisely the philosophy behind the pedagogical techniques. Here you accept the network as a black box and try to understand it by modifying the inputs and observing what happens to the output. One of the strengths with the pedagogical way of doing it is that it does not depend on the machine in the box. In other words, it does not even have to be a neural network in there. This of course makes this approach more portable, in a sense.

The machine learning community has put a lot of effort into using these two paradigms in order to extract rules [44–46] describing the neural network. Rule extraction methods basically try to discretize the ANN in order to decompose it into a number of rules. Although it sounds like precisely the thing you are looking for, I will argue that it is not always the best approach. An artificial neural network is often quite complex and will not give it up without a fight, meaning that rule extraction will often result in a lot of rules. To get around this problem, researchers try to prune away rules that do not influence the performance too much. This is often referred to as maintaining the fidelity.

Another way of viewing the problem has been introduced by sensitivity analysis [47–49] where the focus so far has been on extracting global properties from the neural network, *i.e.*, variables that are important for every patient (case) in the data. Traditionally this has been the way statistics has taught us to view risk factors. For instance we might know from previous studies that obesity is highly associated with heart diseases. The sensitivity analysis would typically strive to find this.

But hang on, now that we are able to extract important variables from the data that applies to all patients we are done, right? Not quite. To see why, let's consider the example variable "Age" which is clearly associated with an elevated risk of heart disease. A 25 year old male with chest pain walks into the emergency and our nifty neural network predicts that he is suffering from acute coronary syndrome. Now the physician wants to know why. Reporting "Age" as one of the reasons why this decision was made is nonsense and gives no further information to the physician. At least in this scenario it seems more natural to extract only the core variables that affected the decision for *this* particular patient. This way of approaching the explanation process for a neural network is referred to as case based [50, 51] since the explanation being generated only depends on the current patient alone. In the context of neural network ensembles I believe that this approach is the most promising so far, and most of my research has actually been focused on developing new methods for the generation of these kind of explanations.

i.4.2 The clinical situation today

In the emergency departments around the world today a lot of patients will arrive with chest pain, some of them by ambulance, others on their own. Once in the emergency department an electrocardiogram will be taken and analyzed by the attending physician. Deciding whether a patient is suffering from acute coronary syndrome or not is tough, even for the most experienced physicians. This is partly due to the fact that ECGs are only predictive in 50% of the cases [9–12] in patients with acute myocardial infarction. Approximately 20% of patients with this condition [11, 12, 52], and about 40% of those with unstable angina [52, 53] have completely normal electrocardiograms. Those are scary numbers! So the physicians need more data in order to make informed decisions, usually in the form of biomarkers from blood samples. However, those tests take time and time is not a surplus in the emergency departments so when in doubt physicians often choose to admit the patient to a higher care level unit. In fact, for patients admitted with a suspicion of acute coronary syndrome, some 7 out of 10 prove not to have the disease [52, 54]. Some of these numbers are due to young physicians making erroneous decisions since they lack the long experience of their older peers. To me this suggests that there is room for improvement. One way to improve it might be to introduce a decision support system, powered by neural networks, into the everyday life of the clinicians. In this thesis, I, together with my coauthors, have shown that an artificial neural network ensemble can be at least as good as even the most experienced physician[1] when diagnosing acute coronary syndromes from ECG data alone. If we could support the younger physicians in their decision making with decision support systems, then we would have come at least one bit further in the pursuit of saving more lives in the emergency departments.

i.5 The papers

If every PhD student changed the world,
everyone would get a migraine.
— Andy Hopper

i.5.1 Paper I

In this study we investigated to what extent artificial neural network ensembles and logistic regression models could be used for the early prediction of acute coronary syndrome (ACS) in chest pain patients in an emergency setting. A thorough comparison of the models with respect to performance, calibration, and correlation was performed. We also investigated to what extent the extracted risk factors from each model coincided. Though we had access to both electrocardiograms and clinical data the results clearly showed that building a model on only ECGs gave the best overall performance. The best performance was achieved by the ANN ensemble with an ROC area of about 80%, while the corresponding statistics for logistic regression model was only 76%. Both models were, however, well calibrated, and interestingly enough we also found that they largely agreed on the risk factors present in the data. This suggests that the ANN was able to pick up features in the data that the logistic regression model could not spot. We conclude that a prediction model based on ANN, combined with the judgment of trained emergency department personnel, could be useful for the early discharge of chest pain patients in populations with a low prevalence of ACS.

[1] Yes you have to read the papers in this thesis to learn more about those results.

i.5.2 Paper II

The purpose of this study was to determine which leads in the standard 12-lead ECG are the best for detecting acute coronary syndrome (ACS) among chest pain patients in the emergency department (ED). Each lead combination was evaluated using a 10 x 10-fold cross validation run where we measured the median area under the ROC curve. We found that using two extremity leads was always as good as using all six of them. This is consistent with Einthoven's relations. Adding a precordial lead always increased the performance, and the best combination found was lead III, aVL and V_2. This three lead combination was at least as good as using all 12 leads, indicating that all the information needed from the ECG to predict ACS is present in these three leads. This, however, does not mean that any given physician could work with only three leads since there may be many patterns found by the ANN ensemble in these leads, rendering the rest of the leads redundant, that are typically hidden from a human interpreter. The results could be important for the creation of clinical decision support systems for ECG prediction of ACS.

i.5.3 Paper III

Estimation of the generalization performance for classification within the medical applications domain is always an important task. In this study we focus on artificial neural network ensembles as the machine learning technique. We present a numerical comparison between five common resampling techniques: k-fold cross validation (CV), holdout, using three cutoffs, and bootstrap using five different data sets. The results show that CV together with holdout 0.25 and 0.50 are the best resampling strategies for estimating the true performance of ANN ensembles. The bootstrap, using the .632+ rule, is too optimistic, while the holdout 0.75 underestimates the true performance.

i.5.4 Paper IV

In this paper we wanted to compare different methods for the early prediction of acute coronary syndrome (ACS) in the emergency department using only information from a single ECG. In this study, however, we had access to a larger database than we had in paper I. The methods we tried were (i) traditional ECG criteria, (ii) consensus interpretation of two expert physicians, (iii) logistic regression model, and (iv) artificial neural network ensembles (ANN). The best overall method was the logistic regression model with an ROC area of 88%, quickly followed by our ANN. Both these methods was significantly better than both the traditional ECG criteria and the expert physicians. We conclude that decision support systems have the potential to improve even experienced ECG readers' ability to predict ACS in the emergency department.

i.5.5 Paper V

Papers I to IV were mainly concerned with evaluating how artificial neural network ensembles could be used as a decision support tool in the clinics and how to estimate their performance. In this paper we instead focus on how we can provide simple and practical explanations for the predictions made by these kind of models. Artificial neural network ensembles (ANN) has long suffered from lack of interpretability. This has severely limited the practical usability of ANNs in settings where an erroneous decision can be disastrous. In this study we develop, explore and compare a set of new methods for the explanation process on two artificial data sets (Monks 1 and 3), and one acute coronary syndrome data set consisting of 861 electrocardiograms (ECG) collected retrospectively at the emergency department at Lund University Hospital. Our view on an explanation is simply highlighting the top most important variables for a given prediction. Using this approach our algorithms managed to extract good explanations in more than 84% of the cases. More to the point, the best method provided 99 and 91% good explanations in Monks data 1 and 3 respectively. The algorithms has the potential to be used as an explanatory aid when using ANN ensembles in clinical decision support systems.

i.5.6 Paper VI

In this paper we compare the two best case-based explanation methods from paper V to two trained physicians on the analysis of electrocardiogram (ECG) data from patients with a suspected acute coronary syndrome (ACS). We investigate which variables the algorithms and physicians typically select as explanations for a given ECG. We also investigated how well the explanations given by the algorithms coincided with those given by the physicians. The algorithms explain the predictions by presenting the top five most important variables for each patient. We could quantify the agreement of an algorithm and a physician by asking them to do the same. In other words we asked the physicians to select the five most important variables for each patient. This way we could use the overlap of the variables selected by a physician and an algorithm to analyze their agreement. The median overlap of the top 5 selected variables between the two physicians, and a given physician and a method, were initially low. Using a correlation analysis of the variables the median overlap increased to values typically in the range 3-4. In conclusion, both our case-based methods generate explanations similar to those of trained expert physicians on the problem of diagnosing ACS from ECG data.

Acknowledgments

Friendship is neither inherited nor transitive.
— Bjarne Stroustrup

I believe that even writing this acknowledgment section is a mistake in itself since there is no way I could ever fit all the people I owe my thanks to in these few lines. Therefor any lack of proper thanks to any given person is mainly due to lack of space, sleep or a combination thereof. Not due to shortage of appreciation.

First and foremost I would like to thank my supervisor Mattias who has always helped me along the way of becoming a scientist. Despite the fact that he has virtually no time whatsoever to spare, he always manages to squeeze me in somehow. I would also like to thank all the people in LUNAR for their unwavering efforts to bring decision support systems to the clinics. Even though we always have more ideas than time, I've greatly enjoyed all the scientific discussion we have had over the years. Thank you all!

For Josefin, my friend, lover and partner in life, I believe that no words can express how much you have helped and supported me in all my endeavors. You taught me that there is an alternative to cynicism.

A special thanks goes to my best friend Jennie, who, despite my argumentative nature, has stood by me through thick and thin. I believe we have shown beyond reasonable doubt that agreement is not a requirement for even the deepest of friendships.

My family is not big by any means, but what we lack in numbers we make up for in devotion and a never ending will to help each other through difficult times. Especially I would like to thank my grandmother who made me realize that studying could actually be a life long challenge worth pursuing. She has been the glue tying us all together throughout the years.

During my time at the department I've come to know a lot of interesting people. Theoretical physics is really that kind of place you know. Especially I would like to thank Carl, Henrik, Liwen, and Spring for many interesting discussions about science, careers and even some things that are not suitable for print. I would also like to thank Patrik, Pontus and Simon for the occasional nonsense that has brightened my stay at the department. A warm thanks goes to Anders who introduced me to new levels of physical exertion through St Hans Extreme. Carsten, thank you for always being cheerful and supportive, and above all for the many interesting anecdotes you've provided me with.

I could not have imagined life at theoretical physics without the basement people, you guys really made the difference. Though I lived in the attic, I still believe that I was secretly longing for the crypt where darkness ruled. This may explain my frequent visits. Getting to know Jari, Markus, and Peter is a bit like getting punched in the face[2]. The benefit of their company might not be immediately clear to you, but after a while

[2] Those of you who know them understand what I mean.

you realize that these guys can go through more science during a 20 minute coffee break than most people manage in a week. I've learned a lot from you guys, thank you.

Lastly, I would like to express my gratitude to my friends, both foreign and domestic, for providing excellent company during countless non-office activities.

i References

[1] *The world health report: report of the director-general, changing history*, Geneva, 2004.

[2] P. W. Wilson, *Established risk factors and coronary artery disease: the framingham study.*, American Journal of Hypertension **7**, 7S–12S (1994).

[3] G. K. Hansson, *Inflammation, atherosclerosis, and coronary artery disease*, New England Journal of Medicine **352**, 1685–1695 (2005).

[4] D. M. Small, *Mechanisms of reversed cholesterol transport*, Agents and Actions. Supplements **26**, 135–146 (1988).

[5] S. S. Barold, *Willem einthoven and the birth of clinical electrocardiography a hundred years ago*, Cardiac electrophysiology review **7**, 99–104 (2003).

[6] M. Sejersten, G. S. Wagner, O. Pahlm, J. W. Warren, C. L. Feldman, and B. M. Horácek, *Detection of acute ischemia from the EASI-derived 12-lead electrocardiogram and from the 12-lead electrocardiogram acquired in clinical practice*, Journal of Electrocardiology **40**, 120–126 (2007).

[7] G. Wehr, R. J. Peters, K. Khalifé, A. P. B. V. Kuehlkamp, A. F. Rickards, and U. Sechtem, *A vector-based, 5-electrode, 12-lead monitoring ECG (EASI) is equivalent to conventional 12-lead ECG for diagnosis of acute coronary syndromes.*, Journal of Electrocardiology **39**, 22–28 (2006).

[8] M. Green, M. Ohlsson, J. L. Forberg, J. Björk, L. Edenbrandt, and U. Ekelund, *Best leads in the standard electrocardiogram for the emergency detection of acute coronary syndrome*, Journal of Electrocardiology **40**, 251–256 (2007).

[9] A. M. Hutter, E. A. Amsterdam, and A. S. Jaffe, *31st bethesda conference. emergency cardiac care. task force 2: Acute coronary syndromes: Section 2b–chest discomfort evaluation in the hospital*, Journal of the American College of Cardiology **35**, 853–862 (2000).

[10] B. G. Abbott and F. J. Wackers, *The role of radionuclide imaging in the triage of patients with chest pain in the emergency department*, Portuguese journal of cardiology : an official journal of the Portuguese Society of Cardiology **19 Suppl 1**, I53–I61 (2000).

[11] K. Channer and F. Morris, *Abc of clinical electrocardiography: Myocardial ischaemia*, BMJ **324**, 1023–1026 (2002).

[12] E. B. Sgarbossa, Y. Birnbaum, and J. E. Parrillo, *Electrocardiographic diagnosis of acute myocardial infarction: Current concepts for the clinician*, American heart journal **141**, 507–517 (2001).

[13] U. Ekelund and J. L. Forberg, *New methods for improved evaluation of patients with suspected acute coronary syndrome in the emergency department*, Emergency Medicine Journal **24**, 811–814 (2007).

[14] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[15] W. Baxt and H. White, *Bootstrapping confidence intervals for clinical input variable effects in a network trained to identify the presence of acute myocardial infarction*, Neural Computation **7**, 624–638 (1995).

[16] S. Cross, R. Harrison, and R. Kennedy, *Introduction to neural networks*, Lancet **346**, 1075–1079 (1995).

[17] G. V. Cybenko, *Approximation by superpositions of a sigmoidal function*, Mathematics of Control, Signals and Systems **2**, 303–314 (1989).

[18] K.-I. Funahashi, *On the approximate realization of continuous mappings by neural networks*, Neural Networks **2**, 183–192 (1989).

[19] F. Rosenblatt, *The perceptron: a probabilistic model for information storage and organization in the brain*, Psychological Review **65**, 386–408 (1958).

[20] F. Rosenblatt, *Principles of neurodynamics: Perceptrons and the theory of brain machanisms*, Spartan, 1962.

[21] M. L. Minsky and S. A. Papert, *Perceptrons: expanded edition*, MIT Press, Cambridge, MA, USA, 1988.

[22] P. Y. Simard, D. Steinkraus, and J. Platt, *Best practice for convolutional neural networks applied to visual document analysis*, International Conference on Document Analysis and Recogntion (ICDAR) (Los Alamitos), IEEE Computer Society, 2003, pp. 958–962.

[23] Wikipedia, *Limbo — wikipedia, the free encyclopedia.*

[24] P.-T. de Boer, D. Kroese, S. Mannor, and R. Rubinstein, *A tutorial on the cross-entropy method*, Annals of Operations Research **134**, 19–67(49) (2005).

[25] B. Flannery, W. Press, S. Teukolsky, and W. Vettering, *Numerical recipies in c*, Cambridge University Press, Cambridge UK,, 1992.

[26] W. C. Davidon, *Variable metric method for minimization*, SIAM Journal on Optimization **1**, 1–17 (1991).

[27] C. M. Bishop, *Neural networks for pattern recognition*, Oxford University Press, 1995.

[28] A. Weigend, D. Rumelhart, and B. Huberman, *Generalization by weight-elimination applied to currency exchangerate prediction*, Neural Networks, 1991. 1991 IEEE International Joint Conference on, 2374–2379 (1991).

[29] L. K. Hansen and P. Salamon, *Neural network ensembles*, IEEE Transactions on Pattern Analysis and Machine Intelligence **12**, 993–1001 (1990).

[30] T. Heskes, *Bias/variance decompositions for likelihood-based estimators*, Neural Computation **10**, 1425–1433 (1998).

[31] T. G. Dietterich, *Ensemble methods in machine learning*, Lecture Notes in Computer Science **1857**, 1–15 (2000).

[32] L. Breiman, *Bagging Predictors*, Machine Learning **24**, 123–140 (1996).

[33] M. Green, J. Björk, J. Hansen, U. Ekelund, L. Edenbrandt, and M. Ohlsson, *Detection of acute coronary syndromes in chest pain patients using neural network ensembles*, Second International Conference on Computational Intelligence in Medicine and Healthcare (Lisbon, Portugal) (J. M. Fonseca, ed.), IEE/IEEE, June-July 2005, pp. 182–187.

[34] Z. Zhou, J. Wu, and W. Tang, *Ensembling neural networks: Many could be better than all*, Artificial Intelligence **137**, 239–263 (2002).

[35] R. Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann, 1995, pp. 1137–1145.

[36] B. Efron, *Estimating the error rate of a prediction rule: improvement on cross-validation*, Journal of the American Statistical Association **78**, 316–331 (1983).

[37] B. Efron and R. Tibshirani, *Improvements on cross-validation: The .632+ bootstrap method*, Journal of the American Statistical Association **92**, 548–560 (1997).

[38] R. Wehrens, H. Putter, and L. Buydens, *The bootstrap: A tutorial*, Chemometrics and Intelligent Laboratory Systems **54**, 35–52 (2000).

[39] J. A. Hanley and B. J. McNeil, *The meaning and use of the area under a receiver operating characteristic(ROC) curve*, Radiology **143**, 29–36 (1982).

[40] T. Fawcett, *ROC graphs: Notes and practical considerations for researchers*, Tech. report, 2004.

[41] D. W. Bates, G. J. Kuperman, S. Wang, T. Gandhi, A. Kittler, L. Volk, et al., *Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality.*, Journal of the American Medical Informatics Association **10**, 523–530 (2003).

[42] Wikipedia, *Achilles' heel — wikipedia, the free encyclopedia*.

[43] D. Elizondo and M. Gongora, *Current trends on knowledge extraction and neural networks*, Lecture Notes in Computer Science (W. D. et al., ed.), vol. 3697, Springer-Verlag, September 2005, p. 485–490.

[44] T. A. Etchells and P. J. G. Lisboa, *Orthogonal search-based rule extraction (OSRE) for trained neural networks: a practical and efficient approach*, IEEE transactions on neural networks **17**, 374–384 (2006).

[45] E. Kolman and M. Margaliot, *Are artificial neural networks white boxes?*, IEEE Transactions on Neural Networks **16**, 844–852 (2005).

[46] M. Craven and J. W. Shavlik, *Using sampling and queries to extract rules from trained neural networks*, International Conference on Machine Learning, 1994, pp. 37–45.

[47] J. J. Montaño and A. Palmer, *Numeric sensitivity analysis applied to feedforward neural networks*, Neural Computing & Applications **12**, 119–125 (2003).

[48] W. Wang, P. Jones, and D. Partridge, *Assessing the impact of input features in a feedforward neural network*, Neural Computing & Applications **9**, 101–112 (2004).

[49] T. Tchaban, M. J. Taylor, and J. P. Griffin, *Establishing impacts of the inputs in a feedforward neural network*, Neural Computing & Applications **7**, 309–317 (1998).

[50] R. Wall, P. Cunningham, P. Walsh, and S. Byrne, *Explaining the output of ensembles in medical decision support on a case by case basis*, Artificial intelligence in medicine **28**, 191–206 (2003).

[51] R. Caruana, *Case-based explanation for artificial neural nets*, Proceedings of Artificial Neural Networks in Medicine and Biology Conference (Göteborg, Sweden) (H. Malmgren, M. Borga, and L. Niklasson, eds.), 2000, pp. 303–308.

[52] J. Pope, R. Ruthazer, J. Beshansky, J. Griffith, and H. Selker, *Clinical features of emergency department patients presenting with symptoms suggestive of acute cardiac ischemia: A multicenter study*, Journal of Thrombosis and Thrombolysis **6**, 63–74 (1998), Center for Cardiovascular Health Services Research, Division of Clinical Care Research, Department of Medicine, New England Medical Center and Baystate Medical Center, Tufts University School of Medicine, Boston, Massachusetts.

[53] J. H. Pope and H. P. Selker, *Diagnosis of acute cardiac ischemia*, Emergency medicine clinics of North America **21**, 27–59 (2003).

[54] U. Ekelund, H.-J. Nilsson, A. Frigyesi, and O. Torffvit, *Patients with suspected acute coronary syndrome in a university hospital emergency department: an observational study*, BMC Emergency Medicine **2**, 1 (2002).

I

Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room

Michael Green[1], Jonas Björk[2], Jakob Lundager Forberg[3], Ulf Ekelund[3], Lars Edenbrandt[4] and Mattias Ohlsson[1]

[1]Computational Biology & Biological Physics, Lund University, Sölvegatan 14A, SE-22362 Lund, Sweden

[2]Competence Centre for Clinical Research, Lund University Hospital, SE-22185 Lund, Sweden

[3]Department of Emergency Medicine, Lund University Hospital, SE-22185 Lund, Sweden

[4]Department of Clinical Physiology, Malmö University Hospital, SE-20502 Malmö, Sweden

Artificial Intelligence in Medicine **38**, 305–318 (2006).

Objective: Patients with suspicion of acute coronary syndrome (ACS) are difficult to diagnose and they represent a very heterogeneous group. Some require immediate treatment while others, with only minor disorders, may be sent home. Detecting ACS patients using a machine learning approach would be advantageous in many situations.

Methods and material: Artificial neural network (ANN) ensembles and logistic regression models were trained on data from 634 patients presenting an emergency department with chest pain. Only data immediately available at patient presentation were used, including electrocardiogram (ECG) data. The models were analyzed using receiver operating characteristics (ROC) curve analysis, calibration assessments, inter- and intra-method variations. Effective odds ratios for the ANN ensembles were compared with the odds ratios obtained from the logistic model.

Results: The ANN ensemble approach together with ECG data preprocessed using principal component analysis resulted in an area under the ROC curve of 80%. At the sensitivity of 95% the specificity was 41%, corresponding to a negative predictive value of 97%, given the ACS prevalence of 21%. Adding clinical data available at presentation did not improve the ANN ensemble performance. Using the area under the ROC curve and model calibration as measures of performance we found an advantage using the ANN ensemble models compared to the logistic regression models.

Conclusion: Clinically, a prediction model of the present type, combined with the judgment of trained emergency department personnel, could be useful for the early discharge of chest pain patients in populations with a low prevalence of ACS.

I.1 Introduction

Patients who present at the emergency department (ED) with chest pain or other symptoms suspicious of myocardial infarction (AMI) or unstable angina pectoris (i.e. acute coronary syndrome, ACS) are common and represent a heterogeneous group. Some have an AMI with a high risk of life-threatening complications whereas others have completely benign disorders which may safely be evaluated on an out-patient basis. Since our ability to diagnose ACS in the ED remains poor, and since the consequences of a missed ACS can be disastrous, there is a large overadmission to in-hospital care; some 7 out of 10 patients admitted with a suspicion of ACS prove not have it [1, 2].

A number of methods have been developed to support the physicians in their decision making regarding patients presenting to the ED with chest pain [3–9]. Goldman et al. [3] developed a statistical model to estimate the relative risk of major events within 72 hours after arrival at the ED. The independent variables used included age, gender and electrocardiographic (ECG) findings, all available at presentation. Another model, the ACI-TIPI [4] was developed to assist triage decisions regarding patients with symptoms of acute cardiac ischemia. This model, using only a few factors (both clinical and ECG), was able to significantly reduce hospitalizations for ED patients without acute cardiac ischemia. In a recent study by Harrison et al. [7] approximately 3000 ACS patients from three different hospitals were analyzed with very good results, using as few as 8 features. They obtained an area under the receiver operating characteristics (ROC) curve as high as 98%. An example of ACS prediction can also be found in the work of Xue et al. [6] where a hybrid machine learning approach was used, combining artificial neural networks (ANN) and decision trees. There are also a number of approaches that have been developed to predict the presence of AMI based on a full range of clinical data [10–13] and data limited to the 12-lead ECG only [14, 15]. Many of these methods used ANN as the classification tool. The performance is usually good compared to interpretation made by experienced physicians.

ANN represents a machine learning tool that has turned out to be useful for complex pattern recognition problems. ANN is also widely used for medical applications (see e.g. [16]). Ensemble learning for ANN is standard procedure to increase the generalization performance by combining several individual networks trained on the same task. The ensemble approach has been justified both theoretically [17, 18] and empirically [19]. Combining the outputs is clearly only relevant when they disagree on some or several of the samples. The most simple method for creating diverse ensemble members is to train each network using randomly initialized weights (also known as injecting randomness). A more elaborate approach is to train the different networks on different subsets of the training set. An example is Bagging [20] where each training set is created by resampling (with replacement) the original one, with uniform probability. Cross splitting [18] is another ensemble creation technique that has performed well in connection with ACS prediction [8].

Comparing ANN models with standard statistical generalized linear models such as logistic regression is an important step in the development procedure. If the results show that the gain of using a non-linear model, such as the ANN, is limited, one should usually go for the less complicated model. Logistic regression always has the nice property of being fully interpretable which can be used to provide feed-back to the user. When performing this comparison it is always important to use more than one measure of performance, since there are several aspects of what is good performance [21].

The aims in this study were twofold. The first aim was to construct an ACS prediction model for our study population and explore to what extent we can confirm previous results obtained for other ACS study populations. Part of this aim was also to identify relevant clinical input factors for the ACS prediction models using an effective odds ratio approach. The second aim was to conduct a detailed comparison between ANN and logistic regression models. In this comparison we used two common techniques for ANN ensemble training together with a single ANN approach. The measures of performance were area under the ROC curve, χ^2 calibration statistics and Person correlations for intra- and inter method variations.

I.2 Materials and methods

I.2.1 Study population

This study is based on patients with chest pain attending the ED of Lund University Hospital, Sweden, from July 1 to November 20 1997. Six hundred sixty-five consecutive visits for which electronic ECG data could be retrieved were included. To have as independent data as possible, some visits were removed such that a criterion of at least 20 days between two consecutive visits, for a given patient, was fulfilled. This reduced the dataset to 634 visits, where 130 patients were diagnosed with ACS and 504 with no ACS. ECG data comprised the 12-lead ECG, recorded using computerized electrocardiographs (Siemens-Elema AB, Solna, Sweden). Table I.1 shows the clinical variables used in this study. Missing values were substituted by the most common category for categorical variables and the mean value for continuous variables.

Table I.1: Characteristics of the independent variables used to train the ACS prediction models. There are 130 cases of ACS and 504 cases without ACS. The second column shows the number of missing values for each variable, where '-' indicates no missing value. The last two columns shows the number of patients (percentage) in each category. For continuous variables the mean (standard deviation) is presented.

Input variable	No Miss.	ACS	No ACS
	n	n (%)	n (%)
Age	-	70.1* (13.2)†	61.3* (18.0)†‡
Gender	-		
Male		83 (63.8)	279 (55.4) ‡

Table I.1: (continued)

Input variable	No Miss. n	ACS n (%)	No ACS n (%)
Female		47 (36.2)	225 (44.6)
Diastolic blood pressure	15	83.9* (14.9)†	82.7* (12.4)†‡
Systolic blood pressure	8	148.5* (29.6)†	142.2* (24.0)†
Heart rate	2	79.4* (22.0)†	78.1* (18.1)†
Smoking status	-		
Current		29 (22.3)	98 (19.4)
Not Current/Unknown		101 (77.7)	406 (80.6)
Hypertension	-		
Yes		47 (36.2)	114 (22.6) ‡
No/Unknown		83 (63.8)	390 (77.4)
Diabetes	-		
Yes		19 (14.6)	57 (11.3)
No		111 (85.4)	447 (88.7)
Medication	-		
Yes		82 (63.1)	263 (52.2)
No		48 (36.9)	241 (47.8)
Angina pectoris	2		
Yes, ≤ 1 month		4 (3.1)	5 (1.0) ‡
Yes, > 1 month		56 (43.8)	174 (34.5)
No		68 (53.1)	325 (64.5)
Congestive heart failure	-		
Yes		20 (15.4)	79 (15.7) ‡
No		110 (84.6)	425 (84.3)
Chest discomfort at presentation	-		
Yes		85 (65.4)	238 (47.2) ‡
No		45 (34.6)	266 (52.8)
Symptom duration	2		
0-6 hours		100 (76.9)	263 (52.2) ‡
7-12 hours		16 (12.3)	59 (11.7) ‡
13-24 hours		4 (3.1)	42 (8.3)
> 24 hours		10 (7.7)	140 (27.8)
Tachypnea	-		
Yes		13 (10.0)	27 (5.4)
No		117 (90.0)	477 (94.6)
Lung rales	-		
Yes		12 (9.2)	23 (4.6)
No		118 (90.8)	481 (95.4)

Table 1.1: (continued)

Input variable	No Miss. n	ACS n (%)	No ACS n (%)
Previous myocardial infarction	-		
Yes, ≤ 6 months		13 (10.0)	19 (3.8) ‡
Yes, > 6 months		37 (28.5)	107 (21.2) ‡
No		80 (61.5)	378 (75.0)
Previous PTCA	-		
Yes		4 (3.1)	21 (4.2) ‡
No		126 (96.9)	483 (95.8)
Previous CABG	-		
Yes		10 (7.7)	55 (10.9) ‡
No		120 (92.3)	449 (89.1)

* Mean.

† Standard deviation.

‡ Clinical variables used in the simplified logistic regression model.

ECG data were reduced to smaller sets of more effective variables before entered into the classification models. The reduction was accomplished using principal component analysis (PCA). Prior to this analysis the measurements were grouped into the following 6 sets of measurements namely: QRS area (total area of the QRS complex), QRS duration, QRS amplitudes, ST amplitudes (ST-amp, ST-amp 2/8 and ST-amp 3/8), ST slope (the slope at the beginning of the ST segment) and positive/negative T amplitudes. The ST amplitudes 2/8 and 3/8 were obtained by dividing the interval between ST-J point and the end of the T wave into eight parts of equal duration. The amplitudes at the end of the second and third interval were denoted ST amplitude 2/8 and 3/8, respectively. Each of these 6 sets were then subject to a principal component analysis reduction, e.g. the 12 ST slope variables (one from each lead) were reduced to 2 variables. The final ECG data set, to be used for the ANN training, consisted of a selection [15] of 16 PCA variables.

The diagnosis of ACS is defined as one of the following discharge diagnoses for the patient: AMI and unstable angina pectoris. The discharge diagnoses were made by the attending senior ward physicians and also reviewed by an experienced research nurse. AMI was defined by the WHO criteria [22] where the biochemical criterion was at least one measurement of CK-MB>10 μg/l or Troponin T>0.1 μg/l. The criteria for unstable angina were (i) observed with (ii) and/or (iii):

(i) Ischemic symptoms: chest pain >15 min., syncope, acute heart failure or pulmonary oedema

(ii) Electrocardiogram (ECG) changes: transient or persisting ST segment depression (≥1 mm) and/or T-wave inversion (≥1 mm) without developing Q waves or loss of R wave height.

(iii) Biochemical markers: CK-MB 5-10 μg/l or Troponin T0.05 – 0.1μg/l.

The non ACS cases consisted of patients with the diagnosis of stable and suspected angina pectoris, together with the category "other diagnosis". Out of the 504 non ACS cases, 271 had discharge diagnoses other than stable or suspected angina pectoris. Table I.2 shows common ECG characteristics for both the ACS cases and the non-ACS cases, obtained by the lead measurements.

Table I.2: Characteristics of the ECGs recorded on the patients. There are 130 cases of ACS and 504 cases without ACS. ST-elevation was defined as ST amplitude ≥ 1mm in two or more contiguous leads, whereas ST-depression was defined as a negative ST amplitude ≥ 1mm in any lead. T-wave depression was defined as a negative T-wave (≥ 1mm) with a predominant R-wave.

ECG finding	ACS	No ACS
	n (%)	n (%)
ST-elevation	52 (40.0)	80 (15.9)
ST-depression	52 (40.0)	59 (11.7)
T-wave inversion	74 (56.9)	189 (37.5)

I.2.2 Artificial neural networks

We considered ANN in the form of feed-forward multilayer perceptrons (MLP) with one hidden layer and no direct input-output connections. The hidden unit activation function was the hyperbolic tangents and the output activation function was the standard logistic function. We used the cross-entropy error function for two classes. In addition we introduced a weight elimination term E_{reg} [23], controlled by a tunable parameter λ, to possibly regularize the network.

$$E_{\text{reg}} = \lambda \sum_i \frac{\beta_i^2}{1+\beta_i^2} ,$$

where the sum runs over all weights in the MLP, except threshold weights. The total error is the sum of the cross-entropy part and E_{reg} for the case when using regularized MLPs. The minimization of the error function was accomplished using the gradient descent method.

Among several existing methods for constructing ensembles, such as voting and boosting (see e.g. [24]) we have used two methods; the common Bagging method [20] and S-fold cross-splitting [8,18]. In Bagging one starts with a given training set and then creates new training sets by resampling, with replacement, the original one. Thus, the

Bagging ensemble contains MLPs trained on *bootstrap* samples of the original training set. The ensemble output t^{ens} is simply computed as the mean of the individual ensemble members, i.e.

$$t^{\text{ens}} = \frac{1}{C} \sum_{n=C}^{C} t_n \ , \tag{I.1}$$

where t_n is the output of the n:th MLP in the ensemble and C is the Bagging ensemble size.

Another way to create diverse training sets is to randomly partition the dataset into S bins. One can then create S slightly different training sets by excluding one of the parts each time. This procedure can be repeated N times to create NxS different but similar training sets. By training an MLP on each of these training sets we can create a pool of MLPs that can be combined into a NxS cross-splitting ensemble. As for Bagging the ensemble output is computed as the mean over the NxS MLP outputs (see Eq. (I.1)). Clearly, the difference between the training sets will increase if fewer bins are used, as a larger fraction of the original training set is excluded each time. For the efficiency of the ensemble we therefore used $S = 2$, supported by the findings in Green et al. [8]. This approach to ensemble creation can be found in the work of Krogh et al. [18], but used in a different context.

The ensemble size, C for Bagging and NxS for cross-splitting, influences the performance of the ensemble method compared to single MLP classifiers. In this study we used an ensemble size of 25 for the model selection and 50 for the final test runs. Both sizes are reasonable according to numerical studies (see e.g. [19, 25]).

I.2.3 Ensemble model selection

Even though the use of ensembles decreases the usual negative effect of overtraining, one must perform model selection for the ensemble. We use the standard K-fold cross-validation procedure to estimate the generalization performance. However, to actually validate the ensemble, each training group in the K-fold cross-validation procedure is used to train an ensemble with either Bagging or S-fold cross-splitting. Figure I.1 summarizes the procedure used for performing ensemble model selection. Model selection is performed, using a grid search, over parameters λ and the number of hidden units in the ANN. Alternative procedures can be used with the S-fold cross-splitting ensemble, which combines both the cross-validation and the ensembles creation [8]. However to accurately validate both the bagging and the S-fold cross-splitting ensemble we used the above procedure even though it is more costly in terms of CPU-time.

I.2.4 Multiple logistic regression

Multiple logistic regression [26] was also used to predict the probability of ACS. Both full logistic regression models, using the same inputs as the ANN models, and a sim-

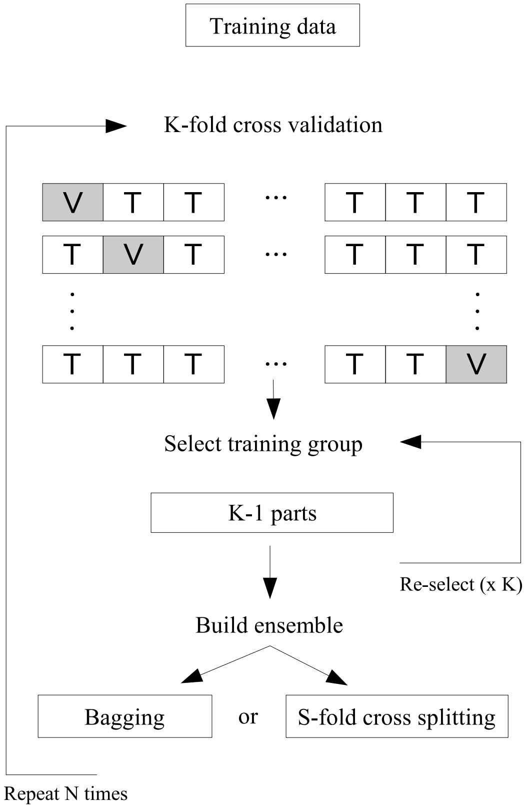

Figure I.1: Ensemble model selection procedure. A given training data set was split into several training/validation parts using K-fold cross-validation. Each of these smaller training sets (T) were then used to create an ANN ensemble and the corresponding validation set (V) was used for validation. For each K-fold cross-validation split, K ensembles were created which resulted in K validation results. The whole procedure was repeated N times with different random K-fold cross-validation splits.

plified model using only clinical input data were trained. The clinical input variables used for the simplified logistic regression model can be found in Table I.1.

The optimization procedure for the simplified logistic regression model was as follows; starting with the full multivariate model with all independent variables included, we excluded one insignificant independent variable at a time, starting with the variable with highest p-value, until only significant and important predictors remained. Categorical variables with more than two categories were kept in the model if the odds ratio associated with any of the categories was significant. The statistical power to detect associations between some of the rare but possibly important clinical characteristics was low. Thus, variables with estimated odds ratio of at least 2.5 (or, equivalently, at most 0.4) were considered as important predictors and kept in the model even if they were not statistically significant. In order to simplify the final model, categories with odds ratios close to one were collapsed with the reference category for that variable. Similarly, unknown response to one of the variables (hypertension) was also added to the reference category.

I.2.5 Statistical analysis

Effective odds ratios

To discern the information content in each of the ANN input features we considered *effective odds ratios*. Odds ratio is the ratio between the odds for an event when a feature is present and the odds for an event when that feature is absent. Odds ratios are well known in the statistical community but cannot be used in conjunction with ANN since the output of an ANN is a non-linear function of the inputs. Odds ratios are defined as:

$$OR = \frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0} = \frac{p_1(1-p_0)}{p_0(1-p_1)} \qquad (\text{I.2})$$

where p_1 is the risk of an event for a patient with a certain feature and p_0 is the risk for the patient without that certain feature. In generalized linear models, such as the logistic regression model used in this study, the odds ratio for a particular feature is e^w where w is the weight for this particular feature. In an ANN we have a non-linear function in the exponent which depends on all other input features in the ANN. However, it is possible to calculate an effective odds ratio by averaging expression (I.2) over all patients [27].

For the logistic regression model there is an alternative interpretation of the odds ratio for a specific feature. The logistic standard bare model can be described by the following relation

$$y = \sum_{i=1}^{m} x_i \omega_i + \omega_0 \ ,$$

where y is the log odds of an event, given the input $(x_1, x_2, ..., x_m)$. If we take the derivative of this relation with respect to a certain feature x_i we end up with:

$$\frac{\partial y}{\partial x_i} = \omega_i = \log\left(OR_{x_i}\right) \qquad (\text{I.3})$$

In other words, we can interpret the derivative with respect to a feature x_i as the log odds ratio for that feature. We can easily generalize this measure to the ANN case. However, the resulting expression will depend on the other input features via the hidden layer function. We can consider odds ratios for an ANN as either the effective odds ratio where we average expression (I.2) over all patients, or we can use the derivative interpretation, by averaging expression (I.3). It is not obvious which one provides the best approximation of odds ratios for the ANN. In this study we used the former approach.

Model calibration

Model calibration, which is a comparison between the observed and predicted ACS risk, was evaluated using the Hosmer-Lemeshow goodness-of-fit test [28], which is given by,

$$\chi^2 = \sum_{j=1}^{G} \frac{(o_j - n_j \bar{\pi}_j)^2}{n_j \bar{\pi}_j (1 - \bar{\pi}_j)} \ .$$

In this expression o_j is the number of observed ACS cases in bin j, and $\bar{\pi}_j$ is the mean average predicted ACS risk in bin j. G is the number of bins meanwhile n_j is the number of samples in the bin. This test follows the χ^2 statistics with $(G - 2)$ degrees of freedom. In this study we have used 10 bins of equal size. The resulting χ^2 statistic is used to indicate non-significant differences ($p > 0.05$) between observed and predicted ACS.

I.2.6 Performance estimation

In addition to the calibration assessment we also constructed ROC curves for all methods. The area under the ROC curve provides yet another (popular) measure of performance. It has the interpretation of the probability that a randomly chosen patient with ACS has a larger predicted ACS risk than a randomly chosen patient without ACS (see e.g. [29]). From the ROC curve we also accessed the specificity at a level of 95% sensitivity. This somewhat arbitrary level was chosen because with current standard evaluation, some 2-5% of the ACS patients are erroneously discharged from the ED, which implies a sensitivity of at least 95% for the routine ED work-up.

To estimate the generalization performance of the tested models we used a 5-fold cross-testing procedure, repeated 20 times, resulting in 100 test sets on which the area under the ROC curve was calculated. The procedure is similar to the cross-validation method used for model selection and is accomplished by dividing the data set into 5 parts of (approximately) equal size. An ACS prediction model is constructed on all parts except one, which is used as the independent test set. The median of the 100 ROC areas is used as the test performance for a given model and selection of independent variables.

An alternative approach to measure the generalization performance is to make an ensemble of the test ACS predictions. This is accomplished by computing the average ACS probability for each patient taken over the 20 cross splittings defined above. The end result is a single list of test ACS probabilities, comprising the full data set, and its corresponding ROC curve. The 100 test set predictions, for a given particular model, is thus transformed into one set of test predictions, defined as the *full test ensemble*.

One would expect this approach to produce an estimation of the generalization performance that is above the one given by the median of the 100 single test results since there is yet another ensemble effect to account for. Furthermore, using the full test ensemble enables a straightforward statistical comparison between different ROC curves and their areas. Associated p-values for ROC area differences using the full test ensemble were calculated using a permutation test (see e.g. [30]).

I.2.7 Software

In this study we used the SAS system to build and develop the logistic regression models meanwhile a C++ based software package was used to build the ANN models. The statistical comparisons were conducted using custom made Perl scripts.

I.3 Results

The test ROC areas obtained for the different methods and different combinations of independent variables are summarized in Table I.3. For each method the ROC area is given both as the median area of the 100 test sets and as the single area of the full test set ensemble.

Table I.3: Test ROC areas obtained from the different methods. For each method two estimations of the generalization performance are presented. The first line corresponds to the median (2.5, 97.5 percentiles) over the 100 test sets defined by the cross-testing procedure. The second line is the ROC area (95% confidence bounds) from the full test set ensemble.

Model	Number of variables (categories†+ continuous)	Test ROC area (%)
ANN bagging ensemble		
Clinical + ECG data	38	79.1 (69.2 86.2)
		80.1 (76.2 84.2)
ECG data	16	79.8 (69.2 88.5)
		81.1 (77.1 85.2)
Clinical data	22	75.3 (67.2 83.0)
		76.0 (71.8 80.4)
ANN cross-splitting ensemble		
Clinical + ECG data	38	78.7 (68.6 86.5)
		80.0 (76.1 84.0)
ECG data	16	80.2 (70.7 89.2)
		81.0 (77.1 85.2)
Clinical data	22	75.1 (67.0 82.6)
		75.3 (70.9 79.8)
ANN single MLP		
Clinical + ECG data	38	76.3 (65.3 83.7)

Table I.3: (continued)

Model	Number of variables (categories†+ continuous)	Test ROC area (%)
		77.1 (72.7 81.6)
ECG data	16	76.0 (60.0 87.1)
		80.0 (76.0 84.2)
Clinical data	22	72.6 (64.9 80.7)
		73.3 (68.6 78.1)
Multiple Logistic Regression, no interaction		
Clinical + ECG data	38	75.7 (63.5 84.2)
		76.4 (71.8 80.9)
ECG data	16	70.5 (54.2 81.2)
		71.0 (65.8 76.2)
Clinical data	22	72.5 (64.6 81.7)
		73.1 (68.4 78.0)
Multiple Logistic Regression, simplified		
Clinical data	13	75.2 (66.4 82.8)
		75.1 (70.7 79.7)

† The base categories are not counted.

The best areas were obtained using the ANN ensemble approach with ECG data, 79.8% and 80.2% (median values) for the bagging and the cross-splitting ensemble, respectively. Adding clinical data to the ANN models did not improve the performance, there was actually a slight decrease of the performance (79.1% and 78.7%), although not significant. Comparing the two ANN ensemble creation methods, it is apparent the both methods yielded similar results. The logistic regression model using both ECG and clinical data received an area of 75.7%. Using only ECG data in the logistic model the results dropped to only 70.5%, indicating the presence non-linearities in the ECG data that the logistic regression model could not capture. Comparing the logistic regression models, built on clinical data alone, the simplified model, using feature selection, and the normal model, with all features present, received an ROC area of 75.2% and 72.5% respectively.

Using the full test ensemble when measuring the performance allows for a (statistical) comparison of two ROC curves. As can be seen in Table I.3 there was an overall increase of the performance using the full test ensemble (except for the simplified logistic model) and this is most certainly due to the ensemble-averaging effect. The difference was significant ($p = 0.05$) when comparing the ANN bagging ensemble trained with clinical data only (76.0%) and ECG data only (81.1%). For the cross-splitting ensemble the corresponding (significant different) areas were 75.3% and 81.0% ($p = 0.03$). Using the simplified logistic regression model, where each non-significant input feature was removed, resulted in an ROC area of 75.1%. The logistic regression model with all features present performed worse, receiving an ROC area of 73.1% ($p = 0.02$). Also

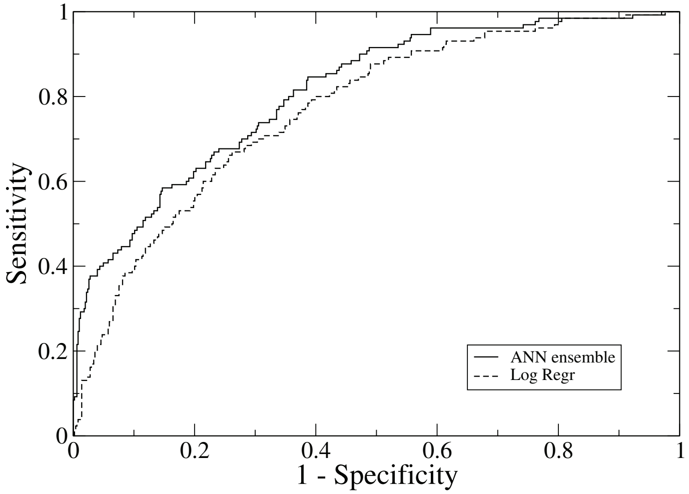

Figure I.2: The ROC curves for the best ANN ensemble and the best logistic regression model using the full test ensemble. The areas under the curves were 81.1% and 76.4%, respectively. The difference was significant (p = 0.03).

including ECG data in the logistic regression model did not significantly improve the performance compared to the simplified model based on clinical data only. It is also interesting to compare sensitivity and specificity values for the different methods. Figure I.2 shows the ROC curve for the full test ensemble using the ANN bagging ensemble and the logistic regression method. At the sensitivity level of 95% we obtained a specificity of 41.1% and 33.7% for the ANN and the logistic model, respectively. With the prevalence of 20.5% ACS in this study population this corresponds to a negative predictive value of 97.2% (96.1%) and a positive predictive value of 29.5% (25.8%) for the ANN ensemble (logistic regression) method.

I.3.1 Calibration comparison

The degree of calibration for the different methods was quantified using the Hosmer-Lemeshow goodness-of-fit test [28]. The results are presented in Table I.4. Comparing the best models (cross-splitting ensemble and logistic regression) we obtained χ^2 values of 11.8 and 24.8, respectively. Both values, taken as the median over the 100 test sets, corresponds to p-values of 0.16 and 0.002. We thus conclude that the best logistic regression model was not calibrated, meanwhile the ANN model was. Moreover, we see that the most calibrated model was the single MLP with a χ^2 and a p-value of 11.5 and 0.17 respectively. Generally models trained with only clinical data received the best calibration scores. The overall worse calibrated model was the single MLP model trained using only ECG data (χ^2 = 40.2).

Table I.4: Test χ^2 calibration and intra Pearson correlation values obtained from the different methods. The values are presented as median (2.5, 97.5 percentiles) over the 100 test sets defined by the cross-testing procedure for the calibration assessment. Pearson correlation values are median (2.5, 97.5 percentiles) over all full test split pairs.

Model	Calibration (χ^2)	Pearson correlation
ANN bagging ensemble		
Clinical + ECG data	14.5 (3.5 58.8)	0.88 (0.85 0.90)
ECG data	12.5 (3.2 47.6)	0.85 (0.81 0.88)
Clinical data	11.7 (4.1 35.3)	0.92 (0.90 0.93)
ANN cross-splitting ensemble		
Clinical + ECG data	13.6 (4.4 65.3)	0.89 (0.86 0.91)
ECG data	11.8 (3.6 24.9)	0.85 (0.82 0.88)
Clinical data	11.6 (3.2 40.8)	0.93 (0.91 0.94)
ANN single MLP		
Clinical + ECG data	15.7 (4.2 65.2)	0.88 (0.85 0.91)
ECG data	40.2 (7.3 436.5)	0.69 (0.59 0.78)
Clinical data	11.5 (3.5 44.1)	0.93 (0.87 0.95)
Multiple Logistic Regression		
Clinical + ECG data	24.8 (6.9 93.6)	0.88 (0.84 0.90)
ECG data	17.1 (3.9 67.2)	0.85 (0.80 0.89)
Clinical data	12.8 (4.5 45.3)	0.93 (0.91 0.95)
Multiple Logistic Regression, simplified		
Clinical data	11.7 (3.6 39.6)	0.96 (0.94 0.97)

An illustration of the degree of calibration in the full test ensemble is presented in Figure I.3 where the solid bars represent the predicted fraction of ACS meanwhile the textured bars represents the true fraction of ACS.

I.3.2 Scatter plots

Although the ROC area and the calibration comparison may reveal differences between the logistic regression and the ANN ensemble model, they are not useful for detecting differences on a patient per patient basis. It is therefore interesting to look at ordinary scatter plots, both for intra- and inter-method comparisons. To quantify the degree of correlation in the scatter plots we used the Pearson correlation coefficient. Results for the intra-method correlations can be found in Table I.4. The simplified logistic regression model obtained the largest correlation coefficient (0.96). Generally methods trained with only clinical data had smaller intra variations compared to method trained with ECG information. Comparing the best ANN and logistic regression model ac-

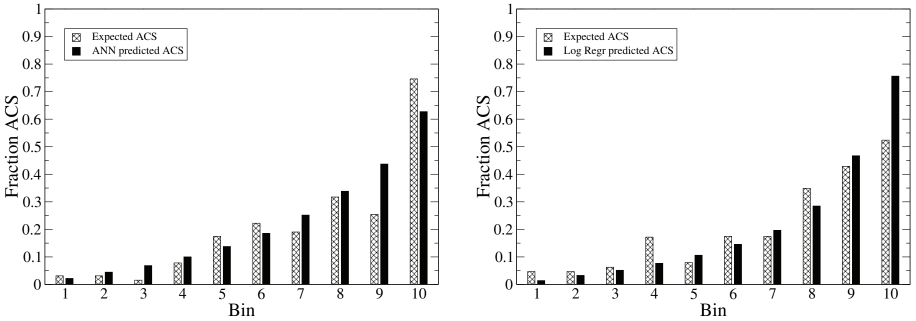

Figure 1.3: This figure shows the expected and the predicted fraction of ACS for patients in the full test ensemble. Left and right figure are the ANN ensemble, trained on ECG data only, and the logistic regression model, trained on both ECG and clinical data, respectively.

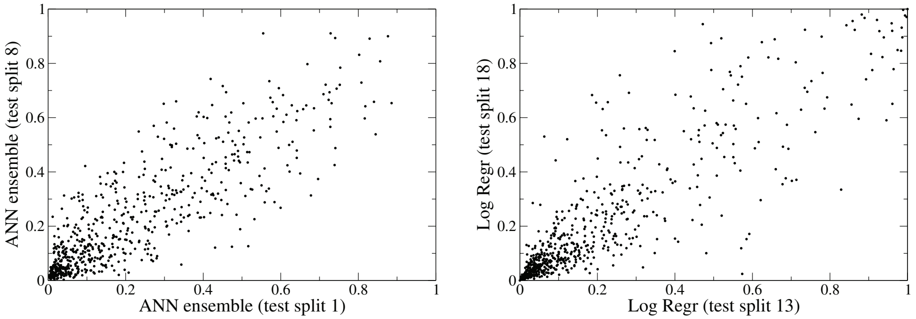

Figure 1.4: Intra-method scatter plots. The left figure shows the ANN cross-splitting ensemble ACS predictions for patients in test splits 1 and 8. The right figure are the corresponding ACS predictions for logistic regression model (test split 13 and 18). The ANN ensemble was trained on ECG data meanwhile the logistic regression model used both ECG and clinical data.

cording to Table 1.3 we can conclude that the ANN had larger intra-method variations (0.85 compared to 0.88 for the logistic regression model). Figure 1.4 shows the scatter plots for these two models, where the test splits are chosen as to correspond to median Pearson correlation values. Thus, the scatter plots in Figure 1.4 represents typical intra-variations in the 20x5-fold cross-testing scheme for the two models.

For inter-method comparisons we first looked the best ANN model and the best logistic regression model according to the ROC area (see Table 1.3). The median Pearson correlation coefficient for all inter-method test split pairs was 0.59 and Figure 1.5 (left part) shows a corresponding scatter plot. Since there was an ROC area difference

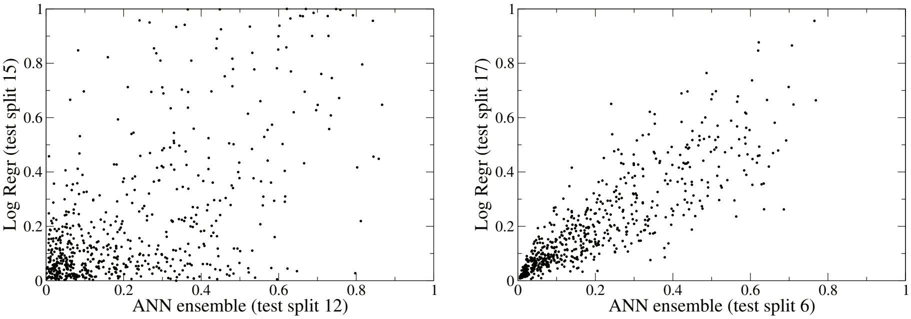

Figure I.5: Inter-method scatter plots. The left figure shows ACS predictions for the ANN cross-splitting ensemble (ECG data) versus the logistic regression model (all input features), using test split 12 and 15, respectively. The right figure corresponds to the bagging ensemble (clinical data) and the simplified logistic regression model, using test split 6 and 17.

of 4.5% between the two models (80.2% compared to 75.7%) one would expect some inter-method differences, but the scatter plot shows a large variation for many patients.

It is also interesting to compare ANN and logistic regression models that had almost the same ROC area and calibration statistics. The bagging ensemble trained on clinical data obtained an ROC area of 75.3% and calibration χ^2 of 11.7. The corresponding numbers for the simplified logistic regression model was 75.2% and 11.7, respectively. The median Pearson correlation coefficient for this comparison was 0.85 and the corresponding scatter plot is shown in Figure I.5 (right part). Although there were no differences in performance and calibration between these two models, there were still significant ACS prediction differences for specific patients. To further analyze the differences we looked at the 10 patients that had the largest ACS prediction differences in this scatter plot. The absolute differences ranged from 0.42 to 0.28. Four ACS patients was part of this subset and the ANN ensemble was correct in 3 cases. Among the remaining 6 non-ACS patients the ANN ensemble correctly classified 4 of them.

I.3.3 Comparing risk factors

For the logistic regression method one can easily compute odds ratios for each of the independent variables. Using odds ratios one can compare the different "predictor" variables. For the ANN ensemble one has to compute *effective* odds ratios because of the non-linearity in the model (see section I.2.5). Odds ratios for the logistic regression model and effective odds ratios for the ANN bagging ensemble are shown in Table I.5. Both models were trained using only clinical data. For the ANN ensemble standard deviations were computed across patients. For both the logistic and the ANN ensemble model the odds ratios were computed using the full data set. For the ANN model this implied training an ANN ensemble on the full data set followed by the effective odds ratio calculation. For the logistic regression model odds ratios were calculated from the weights estimated using the full data set.

Table 1.5: Odds ratios and effective odds ratios for the logistic regression model and the ANN bagging ensemble. These models were trained using clinical data only. For the ANN ensemble the figures in parenthesis are standard deviations computed across patients.

Variable	Logistic regression	ANN
Age	1.04	1.03 (0.01)
Gender		
Male	1.47	1.57 (0.42)
Diastolic blood pressure	1	0.99 (0.01)
Systolic blood pressure	1	1 (0.01)
Heart rate	1	1 (0.01)
Smoking status		
Current	1.59	1.37 (0.16)
Hypertension		
Yes	1.6	1.41 (0.18)
Diabetes		
Yes	1.15	1.07 (0.07)
Medication		
Yes	0.8	0.96 (0.13)
Angina pectoris		
Yes, ≤ 1 month	2.63	2.38 (0.58)
Yes, > 1 month	0.84	1.06 (0.3)
Congestive heart failure		
Yes	0.59	0.65 (0.1)
Chest discomfort at presentation		
Yes	2.14	2.2 (0.49)
Symptom duration		
0-6 hours	5.12	3.79 (0.77)
7-12 hours	3.8	2.67 (0.54)
13-24 hours	1.33	1.02 (0.1)
Tachypnea		
Yes	1.01	1.15 (0.19)
Lung rales		
Yes	1.78	1.55 (0.15)
Previous myocardial infarction		
Yes, ≤ 6 months	3.19	2.94 (0.63)
Yes, > 6 months	1.86	1.97 (0.42)
Previous PTCA		
Yes	0.5	0.58 (0.11)

Table I.5: (continued)

Variable	Logistic regression	ANN
Previous CABG		
Yes	0.41	0.47 (0.11)

There was an overall good agreement between the odds ratios from the logistic regression model and the effective odds ratios obtained from the ANN bagging ensemble. Categorical factors with the largest odds ratios were symptom duration, angina pectoris, previous myocardial infarction and chest discomfort at presentation. It appears that the logistic regression model gave higher weight to "Symptom duration" and that an "Angina pectoris" event that occurred > 1 month ago was not associated with a decrease in ACS risk, as in the logistic regression model. Neither of the models found the factors diastolic and systolic blood pressure, heart rate to be associated with any change of ACS risk.

I.4 Discussion

Part of the aim of this study was to construct a model for ACS prediction at the ED, only using data that are immediately available at presentation. The model was developed using data from chest pain patients at the ED of a university hospital and included clinical and ECG data. The best model was found to be an ANN cross-splitting ensemble, trained on ECG data only, with an area under the ROC curve of about 80%. The model was also well calibrated. There is a general consensus that ECG is one of the most important factors predicting ACS early at the ED. This is confirmed in this study since the best performance was obtained using only the ECG. Adding clinical information did not improve the performance for our study population. The obtained results did not confirm the high levels of ROC areas (> 95%) found in other recent studies (e.g. [5, 7, 9]). One limiting factor in our study was the relatively small study population, however, this cannot be the only explanation. The prevalence of ACS was larger in the work of Kennedy and Harrison [7, 9], ranging from 37%-55% compared to a 21% prevalence of ACS in our study, which we believe is a more realistic number for an ordinary ED [31]. The prevalence of ACS in Baxt et al. [5] was as low as 16%. Furthermore, the presence of ST-elevation, ST-depression or T-wave inversion ECGs, in our population (see Table I.2), was different compared to the cohorts of Kennedy and Harrison, where their training ACS (non-ACS) cases had 32% (1%) ST-elevation, 51% (1%) ST-depression and 44% (4%) T-wave inversion. It is apparent that ECG changes of this kind is very indicative of ACS and may therefore explain why ACS prediction was more difficult in our study population. Baxt et al. [5] obtained an ROC area of 90% with their ANN model, but this included a set of early chemical markers that was not part of our data, since we only included patient data immediately available at presentation. The ECG data used in our model was derived from measurements of the 12-lead ECGs and not from interpretations made by ED staff. The fact that our best model only used such ECG data is interesting since that would allow for a prediction model that is fully automatic without any manual intervention.

Part of this study was also to compare models based on ANN with logistic regression models. Since there are several aspects of how to measure the performance of a given prediction method, we used more than one measurement. The area under the ROC curve is a very popular performance measure in medical applications, but will of course not reveal differences for specific points along the ROC curve. Furthermore, the ROC curve is invariant under any transformation of the ACS predictions as long as the order of the individual ACS predictions is not changed. In a clinical setting however, it is important that the output value of the model can be interpreted as ACS predictions, i.e. we want a good calibration. One approach to measure the degree of calibration for the ACS predictions is the Hosmer-Lemeshow goodness-of-fit test [28]. Comparing models using the area under the ROC curve as performance measure we found an advantage using ANN ensembles compared to both single MLPs and logistic regression. The two different ensemble models tested, bagging and cross-splitting ensemble, obtained comparable ROC areas for the different sets of variables used. It is also apparent that using ensemble averaging increases the performance compared to the single MLP models. Using only clinical data, and no ECG data, there were no significant differences between logistic regression and ANN ensembles. Using only ECG data the performance was better for the ANN ensembles compared to the logistic regression model, indicating non-linear effects not captured by the linear model.

Comparing models using the Hosmer-Lemeshow test we found most ANN ensembles to be well calibrated with χ^2 values ranging from 11.6 to 14.5 with the corresponding p-value range of 0.17 to 0.07. For the logistic regression models the variation was larger ranging from 11.7 to 24.8 for the χ^2. Although the single MLP model using only ECG data obtained a larger ROC area compared to the corresponding logistic regression model, the calibration was much worse. It is obvious that there is no one-to-one correspondence between ROC area and calibration using the Hosmer-Lemeshow test, indicating that it is important to use both measurements for the final model selection. To continue the comparison between models we also looked at intra- and inter-method scatter plots, and the associated Pearson correlation coefficients, to reveal differences on a patient per patient basis. When comparing two models with the same ROC area and calibration statistics large differences for individual ACS predictions was found (see Figure 1.5). An individual patient could be classified as having ACS using one method but with the other one the same patient would be at low risk.

The final choice of ACS prediction model, or even a combination of more than one model, has to be further analyzed and validated in properly designed prospective studies. A hybrid model consisting of both ANN ensembles and logistic regression models, each optimized using different input data, may turn out to be the overall best model.

I.4.1 Clinical implications

Because of possibly disastrous consequences of a missed case of ACS, the evaluation of patients with suspected ACS is very important. The quality of the current standard ED assessment is, however, insufficient. A large number of patients with suspected ACS are incorrectly hospitalized [1, 2, 32] and many patients with ACS are diagnosed only after lengthy (up to 12 hours) observation, with a resulting delay in therapy and an impaired prognosis. At the same time, as many as 5% of those with ACS are erroneously sent home from the ED [31, 33]. Thus, there is a great need for methods to improve ED evaluation. One such method is a decision support system based on ACS prediction models.

The best model developed in this study had a specificity of 41% at the sensitivity level of 95%. For our ACS prevalence of 21%, this corresponds to a positive predictive value of about 30% and a negative predictive value of 97%. The positive predictive value may seem low, but it is likely comparable to that of the ED physician's decision after current standard ED assessment, where some 70% of those admitted for suspected ACS prove not to have it [1, 2, 32]. We have been unable to find any published data on the positive predictive value of standard ED assessment for possible ACS.

Models for ACS prediction based on ECG and clinical characteristics can probably be applied in many different healthcare settings. For the present ACS prediction methods, it seems wise to exploit the reasonably high negative predictive value. Our models are thus probably best used as support for discharging a patient in healthcare settings where ACS prevalence is low, e.g. in primary care, in the initial ED triage or in telemedicine situations where information is limited. Adding the clinical judgment of a physician would probably increase the negative predictive value to close to 100%.

Whatever the use of our models, the limited number of variables imply a small need for manual input, and an increased likelihood that the model will actually be used in a busy environment. With the exception of the ACI-TIPI [4], the need for a time-consuming large input has been a weak point of several previous prediction models, e.g. [5], where up to 40 questions need to be answered before the model gives decision support.

I.4.2 Limitations and future work

The patients included in the present model were retrospectively collected and from one center only. Furthermore the size of the collected dataset has an effect on the performance of the models and increasing the number of patients would probably lead to an increased performance. Before clinical implementation, the model clearly needs to be validated prospectively, preferably at multiple centers. To fully explore the use of ANN ensembles other techniques such as boosting or voting should be tested. Also the observed diversity between between logistic regression models and the ANN models could be utilized using a hybrid approach. The ECG representation using PCA may not be optimal and should be further investigated.

I.4.3 Conclusions

We have found that ANN ensembles, using ECG data only, can predict ACS at the ED with an area under the ROC curve of about 80%. No significant increase in performance was obtained adding clinical data available at presentation. Also, no significant differences were found between the bagging and the cross-splitting ensemble techniques. Comparing ANN ensembles with logistic regression models we found the former approach to be better in terms of ROC area and calibration assessments. Both ANN and logistic regression models showed intra-method variations, as a result of training the models with different parts of the study population. This variation was larger for the ANN ensemble models.

I.5 Acknowledgments

This study is in part supported by the Swedish Knowledge Foundation. We are indebted to research nurse Ewa Mattson for collecting and arranging clinical data and for reviewing the final diagnose of each patient.

I References

[1] J. Pope, R. Ruthazer, J. Beshansky, J. Griffith, and H. Selker, *Clinical features of emergency department patients presenting with symptoms suggestive of acute cardiac ischemia: A multicenter study*, Journal of Thrombosis and Thrombolysis **6**, 63–74 (1998), Center for Cardiovascular Health Services Research, Division of Clinical Care Research, Department of Medicine, New England Medical Center and Baystate Medical Center, Tufts University School of Medicine, Boston, Massachusetts.

[2] U. Ekelund, H.-J. Nilsson, A. Frigyesi, and O. Torffvit, *Patients with suspected acute coronary syndrome in a university hospital emergency department: an observational study*, BMC Emergency Medicine **2**, 1 (2002).

[3] L. Goldman, E. F. Cook, P. A. Johnson, D. A. Brand, G. W. Rouan, and T. H. Lee, *Prediction of the need for intensive care in patients who come to the emergency departments with acute chest pain.*, New England Journal of Medicine **334**, 1498–1504 (1996).

[4] H. P. Selker, J. R. Beshansky, J. L. Griffith, T. P. Aufderheide, D. S. Ballin, S. A. Bernard, et al., *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**, 845–855 (1998).

[5] W. Baxt, F. Shofer, F. Sites, and J. Hollander, *A neural network aid for the early diagnosis of cardiac ischemia in patients presenting to the emergency department with chest pain*, Annals of Emergency Medicine **40**, 575–583 (2002).

[6] J. Xue, T. Aufderheide, R. S. Wright, J. Klein, R. Farrell, I. Rowlandson, et al., *Added value of new acute coronary syndrome computer algorithm for interpretation of prehospital electrocardiograms.*, Journal of Electrocardiology **37 Suppl**, 233–239 (2004).

[7] R. Harrison and R. Kennedy, *Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation*, Annals of Emergency Medicine **46**, 431–439 (2005).

[8] M. Green, J. Björk, J. Hansen, U. Ekelund, L. Edenbrandt, and M. Ohlsson, *Detection of acute coronary syndromes in chest pain patients using neural network ensembles*, Second International Conference on Computational Intelligence in Medicine and Healthcare (Lisbon, Portugal) (J. M. Fonseca, ed.), IEE/IEEE, June-July 2005, pp. 182–187.

[9] R. L. Kennedy and R. F. Harrison, *Identification of patients with evolving coronary syndromes by using statistical models with data from the time of presentation.*, Heart **92**, 183–189 (2006).

[10] W. G. Baxt, *Use of an artificial neural network for the diagnosis of myocardial infarction.*, Annals of internal medicine **115**, 843–848 (1991).

[11] W. G. Baxt and J. Skora, *Prospective validation of artificial neural network trained to identify acute myocardial infarction.*, Lancet **347**, 12–15 (1996).

[12] R. L. Kennedy, A. M. Burton, H. S. Fraser, L. N. McStay, and R. F. Harrison, *Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models.*, European Heart Journal **17**, 1181–1191 (1996).

[13] W. Baxt, F. Shofer, F. Sites, and J. Hollander, *A neural computational aid to the diagnosis of acute myocardial infarction*, Annals of Emergency Medicine **34**, 366–373 (2002).

[14] B. Heden, H. Öhlin, R. Rittner, and L. Edenbrandt, *Acute myocardial infarction detected in the 12-Lead ECG by artificial neural networks*, Circulation **96**, 1798–1802 (1997).

[15] M. Ohlsson, H. Ohlin, S. M. Wallerstedt, and L. Edenbrandt, *Usefulness of serial electrocardiograms for diagnosis of acute myocardial infarction.*, The American journal of cardiology **88**, 478–481 (2001).

[16] P. Lisboa, E. Ifeachor, and P. Szczepaniak (eds.), *Artificial neural networks in biomedicine*, Springer-Verlag, London, 2000.

[17] L. K. Hansen and P. Salamon, *Neural network ensembles*, IEEE Transactions on Pattern Analysis and Machine Intelligence **12**, 993–1001 (1990).

[18] A. Krogh and J. Vedelsby, *Neural network ensembles, cross validation, and active learning*, Advances in Neural Information Processing Systems (San Mateo, CA) (G. Tesauro, D. Touretzky, and T. Leen, eds.), vol. 2, Morgan Kaufman, 1995, pp. 650–659.

[19] D. Opitz and R. Maclin, *Popular ensemble methods: An empirical study*, Journal of Artificial Intelligence Research **11**, 169–198 (1999).

[20] L. Breiman, *Bagging Predictors*, Machine Learning **24**, 123–140 (1996).

[21] A. Niculescu-Mizil and R. Caruana, *Predicting good probabilities with supervised learning*, Proceedings of the 22nd International Conference on Machine Learning (Bonn, Germany) (L. D. Raedt and S. Wrobel, eds.), ACM Press, 2005.

[22] H. Tunstall-Pedoe, K. Kuulasmaa, P. Amouyel, D. A. A. M. Rajakangas, and A. Pajak, *Myocardial infarction and coronary deaths in the world health organization-monica project. registration procedures, event rates, and case-fatalityrates in 38 populations from 21 countries in four continents*, Circulation **90**, 583–612 (1994).

[23] S. J. Hanson and L. Y. Pratt, *Comparing biases for minimal network construction with back–propagation*, Advances in Neural Information Processing Systems 1 (D. S. Touretzky, ed.), Morgan Kaufmann, 1989, pp. 177–185.

[24] T. G. Dietterich, *Ensemble methods in machine learning*, Lecture Notes in Computer Science **1857**, 1–15 (2000).

[25] D. West, P. Mangiameli, R. Rampal, and V. West, *Ensemble strategies for a medical diagnostic decision support system:a breast cancer diagnosis application*, European Journal of Operational Research **162**, 532–551 (2005).

[26] D. Hosmer and S. Lemeshow, *Applied logistic regression*, Wiley, New York, 1989.

[27] R. Lippman and D. Shahian, *Coronary artery bypass risk prediction using neural networks*, The Annals of Thoracic Surgery **63**, 1635–1643 (1997).

[28] D. W. Hosmer, T. Hosmer, S. L. Cessie, and S. Lemeshow, *A comparison of goodness-of-fit tests for the logistic regression model.*, Statistics in Medicine **16**, 965–980 (1997).

[29] J. A. Hanley and B. J. McNeil, *The meaning and use of the area under a receiver operating characteristic(ROC) curve*, Radiology **143**, 29–36 (1982).

[30] R. Wehrens, H. Putter, and L. Buydens, *The bootstrap: A tutorial*, Chemometrics and Intelligent Laboratory Systems **54**, 35–52 (2000).

[31] J. Pope, T. Aufderheide, R. Ruthazer, R. Woolard, J. Feldman, J. Beshansky, et al., *Missed diagnoses of acute cardiac ischemia in the emergency department*, The New England Journal of Medicine **342**, 1163–1170 (2000), Center for Cardiovascular Health Services Research, Department of Medicine, New England Medical Center, Boston, Mass 02111, USA.

[32] B. W. Karlson, J. Herlitz, O. Wiklund, A. Richter, and A. Hjalmarson, *Early prediction of acute myocardial infarction from clinical history, examination and electrocardiogram in the emergency room.*, The American journal of cardiology **68**, 171–175 (1991).

[33] T. H. Lee, G. W. Rouan, M. C. Weisberg, D. A. Brand, D. Acampora, C. Stasiulewicz, et al., *Clinical characteristics and natural history of patients with acute myocardial infarction sent home from the emergency room.*, The American journal of cardiology **60**, 219–224 (1987).

II

Best leads in the standard electrocardiogram for the emergency detection of acute coronary syndrome

Michael Green[1], Mattias Ohlsson[1], Jakob Lundager Forberg[2], Jonas Björk[3], Lars Edenbrandt[4] and Ulf Ekelund[2]

[1]Computational Biology & Biological Physics, Lund University, Sölvegatan 14A, SE-22362 Lund, Sweden

[2]Department of Emergency Medicine, Lund University Hospital, SE-22185 Lund, Sweden

[3]Competence Centre for Clinical Research, Lund University Hospital, SE-22185 Lund, Sweden

[4]Department of Clinical Physiology, Malmö University Hospital, SE-20502 Malmö, Sweden

Journal of Electrocardiology **40**, 251–256 (2007).

Background and Purpose

The purpose of this study was to determine which leads in the standard 12-lead ECG are the best for detecting acute coronary syndrome (ACS) among chest pain patients in the emergency department (ED).

Methods

Neural network classifiers were used to determine the predictive capability of individual leads and combinations of leads from 862 ECGs from chest pain patients in the ED at Lund University Hospital.

Results

The best individual lead was aVL with an area under the receiver operating characteristics (ROC) curve of 75.5%. The best 3-lead combination was III, aVL and V_2 with an ROC area of 82.0%, compared to the 12-lead ECG performance of 80.5%.

Conclusions

Our results indicate that leads III, aVL and V_2 are sufficient for computerized prediction of ACS. The present results are likely important in situations where the 12-lead ECG is impractical, and for the creation of clinical decision support systems for ECG prediction of ACS.

II.1 Introduction

In the emergency department (ED), the ECG is crucial in the evaluation of a possible acute myocardial infarction (AMI) or unstable angina pectoris, i.e. acute coronary syndrome (ACS). The standard 12-lead ECG may in this situation convey as much diagnostic information as all other clinical data taken together [1]. For the ED diagnosis of ACS, it is conceivable that all of the standard 12 leads are not equally important. Myocardial ischemia and infarction are more frequent in some parts of the heart, and there are also “blind spots” in the standard ECG for certain regions of the heart, e.g. that supplied by the left circumflex artery [2]. If a few leads, or combinations of leads, would have as good or almost as good performance for ACS as the complete standard 12-lead ECG, this would be of interest both in situations where the 12-lead ECG is impractical, as in prehospital triage or in ECG monitoring of possible ACS, and for the creation of ECG decision support software. Selection of the best leads from a 12-lead ECG has previously been attempted for detection of coronary artery disease [3] and for the assessment of QT prolongation [4].

Artificial neural networks represents a machine learning tool that has proved useful for complex pattern recognition problems, and is widely used for medical applications (see e.g. [5]). The networks learn by associating different ECG patterns with the desired classification, not by being fed a set of predefined diagnostic criteria. Data from a large group of observations are presented to the networks, together with the desired classification, during a so-called training session. Neural networks have already been applied to different aspects of automated interpretation of ECGs, for example in the diagnosis of myocardial infarction [6–8]. These studies have demonstrated a significantly improved performance over both conventional ECG criteria and experienced ECG readers. Neural networks have also been used for ACS prediction in acute chest pain patients [9–11] and have been compared to standard statistical methods such as multiple logistic regression [12]. These studies indicate that networks are well suited as a tool for analyzing ECGs in suspected ACS patients.

The aim of this study was to elucidate, with the use of neural networks, which of the standard ECG leads, or which combination of these leads, have the largest predictive capability for the emergency diagnosis of ACS when being used together with a machine learning tool.

II.2 Methods

II.2.1 Study population

This retrospective study was based on the first ECGs recorded in the ED of Lund University Hospital on patients with a principal complaint of chest pain, from July 1997 to March 1999. ECGs were recorded 5 min to 1 h after the patient arrived at the ED. Only ECGs for which the electronic ECG data could be retrieved were included, excluding

ECGs with severe technical deficiencies and ECGs from pacemaker patients. Each ECG was classified as either "ACS" or "non ACS", depending on the hospital discharge diagnosis of the patient. A diagnosis of ACS was defined as a discharge diagnosis of AMI or unstable angina pectoris, and the criteria for these diagnoses were the ones used during the ECG recording period. AMI was defined by the WHO criteria [13] where the biochemical criterion was at least one measurement of CK-MB > 10 μg/l or Troponin T > 0.1 μg/l. The criteria for unstable angina were ischemic symptoms (chest pain >15 min., syncope, acute heart failure or pulmonary edema) together with at least one of the following: a) Electrocardiogram (ECG) changes: transient or persisting ST segment depression (≥ 1 mm) and/or T-wave inversion (≥ 1 mm) without developing Q waves or loss of R wave height, or b) Biochemical markers: CK-MB 5-10 μg/l or Troponin T 0.05-0.1 μg/l.

All discharge diagnoses were made by the senior ward physician or the ED physician (in cases discharged from the ED), reviewed by a senior research nurse, and when ambiguous, further reviewed by a senior cardiologist. In the review of diagnoses for cases discharged from the ED, available data from the patient records indicated that the rate of missed diagnosis of ACS, compared to the above described criteria, was low (not more than 2%).

The final dataset consisted of 862 patients, 345 with diagnosis ACS and 517 with diagnosis no ACS. Among the non-ACS cases 123 patients were diagnosed as stable angina pectoris, 114 as suspected angina pectoris and the remaining 280 patients belonged to the category "other diagnoses". The mean age within the ACS and non-ACS group was 69 (13) and 62 (18), respectively, and where the numbers in parenthesis are standard deviations. Additionally, the ACS group consisted of 227 men and 118 women and the corresponding numbers for the non-ACS group were 291 and 226.

This study was approved by the Lund University Research Ethics Committee.

II.2.2 Electrocardiography

The 12-lead ECGs were recorded by the use of computerized electrocardiographs (from Siemens-Elema AB, Solna, Sweden), and the following 12 measurements taken from each of the 12 leads were selected for further analysis: QRS duration, QRS area, Q duration, Q amplitude, R duration, R amplitude, ST-J amplitude, ST slope (the slope at the beginning of the ST segment), ST amplitude 2/8, ST amplitude 3/8, positive T amplitude and negative T amplitude. All durations and amplitudes are measured in milliseconds and micro-volts, respectively. The ST amplitude 2/8 and ST amplitude 3/8 were obtained by dividing the interval between ST-J point and the end of the T wave into 8 parts of equal duration. The amplitudes at the end of the second and the third intervals were denoted ST amplitude 2/8 and ST amplitude 3/8. In total 144 measurements from each 12-lead ECG were collected. In order to reduce the number of input measurements for the neural networks a principal component analysis (PCA) [14] on the 12 measurements within each lead was used. Using only the first six principal components in each lead resulted in a total of 72 measurements when considering all 12 leads.

The number of selected principal components was chosen as to include at least 90% of the variance in each lead. The variance captured in each lead varied within a range of 91.1-94.9%. The PCA analysis was based on the correlation matrix.

II.2.3 Artificial neural networks

In this work we built ACS prediction classifiers using neural network ensembles with the bagging technique [15]. A general presentation of artificial neural networks can be found in the work of Cross et al. [16]. An ensemble size of 50 was chosen which has been found to be sufficient in numerical studies [17]. The ensemble prediction was computed as the average over the output of each of the individual networks. All six principal components from the PCA step was fed to the ANN as continuous variables.

The model selection [11] consisted of selecting the best architecture and regularization parameter for each neural network ensemble with respect to the area under the ROC curve [18]. The ROC area is commonly used as a performance measure and can be interpreted as the probability that a randomly chosen patient with ACS has a higher risk output than a randomly chosen patient without ACS. We used K-fold cross validation [19] to estimate the best ensemble parameters. To accomplish this the training data was split into K random equally sized disjoint parts. One part was selected for the validation of the neural network ensemble which was constructed on the other K-1 parts. This procedure was repeated for all K parts. The K-fold cross validation was repeated N times, and the total validation result was taken as the mean of the N x K validation results. We used N=10 and K=5 for the model selection.

In order to estimate the generalization performance of the neural network ensemble an outer cross validation loop was used. The data was randomly split into 5 disjoint parts. Each part was selected as a test set with the rest of the parts as the corresponding training set. The outer cross validation loop was repeated 20 times resulting in 100 training and test sets. The total test result was evaluated as the median over the 100 test results.

II.2.4 Statistics

We used the area under the ROC curve to assess the performance of the neural networks. When comparing two different neural network classifiers, on a given test set, we used their corresponding outputs to evaluate whether they produced significantly different ROC areas or not. Statistical significance was evaluated using a permutation test [20] where we considered a p-value < 0.05 as statistically significant.

II.3 Results

All results are presented as medians over the 100 ROC areas produced by the outer cross validation loop. The results for the neural network classifiers fed with single leads as

Table II.1: The test ROC areas for the individual leads. The ROC area is presented as median (2.5, 97.5 percentiles) over the 100 test sets.

Selected ECG lead	Test ROC area (%)
I	74.1 (67.9, 81.8)
II	68.6 (61.6, 76.2)
III	75.0 (68.2, 80.6)
aVR	67.9 (62.2, 75.3)
aVL	75.5 (65.8, 82.6)
aVF	72.0 (63.7, 78.1)
V_1	67.8 (60.6, 75.7)
V_2	74.3 (67.7, 82.5)
V_3	73.7 (65.4, 81.3)
V_4	72.3 (66.1, 79.6)
V_5	71.5 (65.6, 79.3)
V_6	73.7 (65.1, 81.6)

input are presented in Table II.1. The three best limb leads I, III and aVL had similar performance with ROC areas of 74.1%, 75.0% and 75.5%, respectively. Leads II, aVR and aVF did not match that performance. For the precordial leads, the best performance was obtained using lead V_2 with an ROC area of 74.3%. However, leads V_3 and V_6 were almost as good with ROC areas of 73.7%. Statistical evaluations showed that a significant difference between the best (aVL) and the worst (V_1) performing leads was found in 36 out of the 100 test sets.

The performance of the neural networks classifiers fed with inputs from different combinations of leads are presented in Table II.2. The two (III and aVL) best individual leads were combined and this combination obtained an ROC area of 78.9%. Any two lead combination of the six limb leads resulted in similar ROC areas with a median area of 77.9% (range 74.5% – 78.9%). Adding one precordial lead to the best two lead combination almost always increased the performance (see Table II.2). The best three lead combination was III, aVL and V_2 with an area under the ROC curve of 82%.

Table II.2 also shows the results for the combination of all limb leads (denoted 6-lead ECG), the two best combinations of the 6-lead ECG and one precordial lead, and the full 12-lead ECG. The performance of the neural network when using the 12-lead ECG was 80.5%. A statistical comparison of the best 3-lead combination (III-aVL-V_2) and the full 12-lead ECG resulted in only 10 of the 100 test splits being significantly different, indicating that performance of these two combinations of leads are comparable.

The ROC curves for the best single lead, the best 3-lead combination and for the 12-lead ECG are shown in Figure II.1. A comparison with traditional ECG criteria for AMI detection resulted in a specificity and sensitivity of 95.6% and 24.3%, respectively. The sensitivity of the AMI subgroup was 34.1% and the corresponding result for the

Table II.2: The test ROC areas for combination of leads. The 6-lead ECG refers to the combination of all limb leads. The ROC areas are presented as median (2.5, 97.5 percentiles) over the 100 test sets.

Selected ECG leads	Test ROC area (%)
III-aVL	78.9 (71.1, 83.9)
III-aVL-V_1	78.9 (70.9, 84.9)
III-aVL-V_2	82.0 (74.2, 87.7)
III-aVL-V_3	81.1 (74.0, 86.9)
III-aVL-V_4	81.1 (73.3, 87.1)
III-aVL-V_5	80.9 (72.8, 85.8)
III-aVL-V_6	80.6 (72.0, 87.0)
6-lead ECG	78.0 (68.7, 81.6)
6-lead ECG + V_2	80.2 (73.6, 86.5)
6-lead ECG + V_3	80.7 (73.6, 86.7)
12-lead ECG	80.5 (72.8, 86.2)

unstable angina subgroup was 5.2%.

II.4 Discussion

In the present study we attempted to establish the best lead, or combination of leads, for the ED diagnosis of ACS. The results showed that the best individual lead was aVL (ROC area of 75.5%), and that the six limb leads together with either V_2 (80.2%) or V_3 (80.7%) had principally the same performance for ACS as the complete 12-lead ECG (80.5%). Somewhat surprisingly, using only leads III, aVL and V_2 gave similar discriminatory power for ACS (82.0%). It thus seems that these three leads together contain all the ACS predicting information present in the standard 12-lead ECG, at least in the present patient material. This can partially be explained by the fact that any two limb leads can be used to derive the other four limb leads when using the raw ECG lead recording. Thus, given that our representation of the ECG is good enough, the ANN will be able to extract information about all six limb leads even though only two of them are fed to the network as inputs.

The present results are compatible with previous studies on optimal leads for detection of ST segment deviations in acute myocardial ischemia. During coronary occlusion induced by balloon angioplasty, the largest ST changes have been observed in leads V_2-V_4 (occlusions of the left anterior descending or circumflex arteries) and in leads III and aVF (right coronary artery) [21–23], and these leads have therefore been suggested to be optimal for ischemia detection during balloon angioplasty. For identification of ST changes in established AMI, leads III and V_2 have been suggested to be

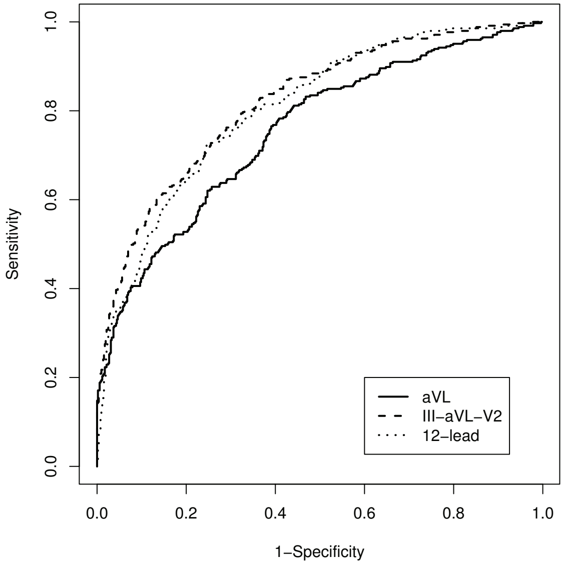

Figure II.1: The ROC curves for the best single-lead, 3-lead, and 12-lead ECG using the respective median test split. The ROC curves were produced by concatenating all 5 test results from the outer cross-validation split, with an ROC area most similar to the median test results, as presented in Tables II.1 and II.2.

optimal [24]. However, these results are not immediately applicable to ED patients with suspected ACS. First, many ED patients with ACS do not have ST segment changes at all, but rather T-wave inversions, new Q-waves or no ECG changes at all, and the ECG changes may in turn be due to subtotal and varying occlusion of branches of the large coronary arteries. Since we considered not only the ST segment but several other ECG variables (QRS duration, QRS area, Q duration, Q amplitude, R duration and R and T amplitudes), it is not surprising that our results differ from those in studies focusing only on the ST segment. For instance, aVL was the single best lead for ACS prediction in our study, whereas during balloon angioplasty [22,25] ΔST in aVL was too low to be of any use for ischemia detection. Second, in the present study only one ECG from each patient was considered. The neural networks thus only had access to absolute measures in the ECG, and not to any relative changes induced by ischemia in the ACS patients. It may be that preexisting ECG changes unrelated to current ischemia contributed to ACS detection by the neural networks in our patients.

In the present results, good ACS discriminating power with only three leads was observed. ECG registration with reduced lead sets is practical for many reasons. Few leads interfere less with the everyday care of the patient, with diagnostic tests such as echocardiography, and with emergency procedures such as defibrillation. To detect acute ischemia by ST deviation, however, current consensus is that all 12 leads of the standard ECG are necessary [21]. Indeed, many ischemic events were missed when only the usual telemetry leads (V_1 and II) [26] were used, or even the three single best leads for detection of ST deviation [27]. Ischemia detection with reduced lead sets have in fact so far only been successful when the omitted leads have been calculated, or when a derived 12-lead ECG has been used [28–30]. Thus, in reduced lead sets, it seems that ischemia detection will not be satisfactory if only the ST segment is monitored. To our knowledge, detection of ischemia using multiple ECG variables in reduced lead sets has not previously been tested. Our finding that leads III, aVL and V_2 together predicted ACS as well as the standard 12-lead ECG may thus be explained by the fact that we included several ECG measurements in addition to the ST segment. We have not, however, investigated the relative importance among the ECG measurements within each included lead.

In this study we used neural networks as the method for ACS prediction with a varying number of input leads. This choice of classification method was guided by previous work (e.g. [7, 9, 11]) where neural networks has proved to be useful for ACS and AMI prediction. Standard linear statistical methods, such as multiple logistic regression, would not have been sufficient, since there are nonlinear relationships among the lead measurements, utilized by the networks, that are important for predicting ACS [11]. The PCA preprocessing of the ECGs has been used previously [31] and can be motivated by the fact that measurements for each lead showed large correlations. Furthermore it is always advantageous to keep the number of inputs to the network models as low as possible since the problem of overtraining usually increases with an increasing number of inputs. Using PCA for this reduction is a commonly used method. Care was taken to obtain as reliable estimates as possible of the generalization performance for each lead selection. Even though the study population was relatively small which may have influenced the absolute values for the ROC areas, we believe that the obtained selection of important leads is valid.

II.4.1 Clinical implications

The present results have their main implications for the creation of future clinical decision support systems (CDSS) for ECG interpretation. For a CDSS to produce as robust ACS predictions as possible, it is essential that it is allowed to work only with the ECG elements crucial for ACS prediction, and that other information is left out. With more robust ACS predictions, the CDSS will of course be more valuable to the patients and the physicians using them. The identified leads III, aVL and V_2, together with clinical patient data such as chest pain history and blood pressure, could be used to develop a neural network based CDSS that would potentially be useful in situations where the standard 12-lead ECG is impractical, as in e.g. prehospital triage or in telemedicine settings. For true clinical usefulness, such a CDSS should also include an ANN able to detect ST elevation myocardial infarction in need of urgent reperfusion therapy, e.g. [32]. Before clinical implementation, the CDSS would of course need to be validated prospectively, preferably at multiple centers.

II.4.2 Limitations of the study

The results from this study are probably not applicable to the manual interpretation of ECGs by physicians. It is not at all evident, and perhaps even unlikely, that leads III, aVL and V_2 together would be as useful as the 12-lead ECG to the physician trying to establish whether the patient has ACS or not. Some of the variables used in the present study are not part of the standard ECG interpretation routine, and are not easy to appreciate by eye. Further, if ANNs such as those in the present study are to be used in CDSS for physicians, a problem is that the ANN is unable to explain to the user the reasons for the suggested decisions. Current research is trying to overcome this problem [33, 34].

The results were obtained using ECG data collected from a limited number of patients during a limited time period, and at one center only. Other populations might of course produce different results. Likewise, we cannot exclude the possibility that another set of ECG variables than the ones chosen would produce other results. However, we believe it is unlikely that results in other populations or with other variables would differ substantially, since only the relative performances of the different leads and combinations of leads were analyzed in this study.

The ECGs in the present study were collected in the late 1990's, and old definitions of AMI and unstable angina were used. More recent definitions of AMI have lower cut-off values for biochemical markers [35], and for the diagnosis of unstable angina no marker elevation is currently needed. A few patients classified as non-ACS in the present study may thus be classified as having an ACS with current diagnostic criteria.

II.5 Conclusions

The aim of this study was not to find the best neural network classifiers for prediction of ACS, but rather to compare the information content of the different leads, and of the different combinations of leads. We found that the lead aVL was the single best lead for ACS detection, and that the leads III, aVL and V_2 together yielded similar performance as the full 12-lead ECG for predicting ACS. It thus seems that these three leads together contain all the ACS predicting information present in the standard 12-lead ECG, at least in our patient population. These findings may be useful for the creation of ECG decision support software to be used in situations where the 12-lead ECG is impractical.

II.6 Acknowledgments

This study was supported by grants from the Swedish Knowledge Foundation (2208/04-504), the Swedish Heart-Lung Foundation, Region Skåne and The Leardal foundation for Acute Medicine. This work has also been supported by the Swedish Knowledge Foundation through the Industrial PhD program in Medical Bioinformatics at the Centre for Medical Innovations (CMI) at the Karolinska Institute.

II References

[1] L. Goldman and A. J. Kirtane, *Triage of patients with acute chest pain and possible cardiac ischemia: the elusive search for diagnostic perfection.*, Annals of internal medicine **139**, 987–995 (2003).

[2] E. B. Sgarbossa, Y. Birnbaum, and J. E. Parrillo, *Electrocardiographic diagnosis of acute myocardial infarction: Current concepts for the clinician*, American heart journal **141**, 507–517 (2001).

[3] J. Viik, R. Lehtinen, and J. Malmivuo, *Detection of coronary artery disease using maximum value of ST/HR hysteresis over different number of leads.*, Journal of Electrocardiology **32 Suppl**, 70–75 (1999).

[4] T. Sadanaga, F. Sadanaga, H. Yao, and M. Fujishima, *An evaluation of ECG leads used to assess QT prolongation.*, Cardiology **105**, 149–154 (2006).

[5] P. Lisboa, E. Ifeachor, and P. Szczepaniak (eds.), *Artificial neural networks in biomedicine*, Springer-Verlag, London, 2000.

[6] B. Hedén, M. Ohlsson, R. Rittner, O. Pahlm, W. K. Haisty, C. Peterson, et al., *Agreement between artificial neural networks and experienced electrocardiographer on electrocardiographic diagnosis of healed myocardial infarction.*, Journal of the American College of Cardiology **28**, 1012–1016 (1996).

[7] B. Heden, H. Öhlin, R. Rittner, and L. Edenbrandt, *Acute myocardial infarction detected in the 12-Lead ECG by artificial neural networks*, Circulation **96**, 1798–1802 (1997).

[8] W. G. Baxt and J. Skora, *Prospective validation of artificial neural network trained to identify acute myocardial infarction.*, Lancet **347**, 12–15 (1996).

[9] W. Baxt, F. Shofer, F. Sites, and J. Hollander, *A neural network aid for the early diagnosis of cardiac ischemia in patients presenting to the emergency department with chest pain*, Annals of Emergency Medicine **40**, 575–583 (2002).

[10] R. Harrison and R. Kennedy, *Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation*, Annals of Emergency Medicine **46**, 431–439 (2005).

[11] M. Green, J. Björk, J. Hansen, U. Ekelund, L. Edenbrandt, and M. Ohlsson, *Detection of acute coronary syndromes in chest pain patients using neural network ensembles*, Second International Conference on Computational Intelligence in Medicine and Healthcare (Lisbon, Portugal) (J. M. Fonseca, ed.), IEE/IEEE, June-July 2005, pp. 182–187.

[12] M. Green, J. Björk, J. Forberg, U. Ekelund, L. Edenbrandt, and M. Ohlsson, *Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room*, Artificial Intelligence in Medicine **38**, 305–318 (2006).

[13] H. Tunstall-Pedoe, K. Kuulasmaa, P. Amouyel, D. A. A. M. Rajakangas, and A. Pajak, *Myocardial infarction and coronary deaths in the world health organization-monica project. registration procedures, event rates, and case-fatalityrates in 38 populations from 21 countries in four continents*, Circulation **90**, 583–612 (1994).

[14] M. J. O'Connell, *Search program for significant variables*, Computer Physics Communications **8**, 49–55 (1974).

[15] L. Breiman, *Bagging Predictors*, Machine Learning **24**, 123–140 (1996).

[16] S. Cross, R. Harrison, and R. Kennedy, *Introduction to neural networks*, Lancet **346**, 1075–1079 (1995).

[17] D. Opitz and R. Maclin, *Popular ensemble methods: An empirical study*, Journal of Artificial Intelligence Research **11**, 169–198 (1999).

[18] J. A. Hanley and B. J. McNeil, *The meaning and use of the area under a receiver operating characteristic(ROC) curve*, Radiology **143**, 29–36 (1982).

[19] K. Baumann, *Cross-validation as the objective function for variable-selection techniques*, Trends in Analytical Chemistry **22**, 395–406 (2003).

[20] R. Wehrens, H. Putter, and L. Buydens, *The bootstrap: A tutorial*, Chemometrics and Intelligent Laboratory Systems **54**, 35–52 (2000).

[21] B. J. Drew and M. W. Krucoff, *Multilead ST-segment monitoring in patients with acute coronary syndromes:a consensus statement for healthcare professionals. ST-segment monitoringpractice guideline international working group*, American journal of critical care **8**, 372–386 (1999).

[22] H. S. Bush, J. J. Ferguson, P. Angelini, and J. T. Willerson, *Twelve-lead electrocardiographic evaluation of ischemia during percutaneous transluminal coronary angioplasty and its correlation with acute reocclusion.*, American heart journal **121**, 1591–1599 (1991).

[23] E. Persson, J. Pettersson, M. Ringborn, L. Sörnmo, S. G. Warren, G. S. Wagner, et al., *Comparison of ST-segment deviation to scintigraphically quantified myocardial ischemia during acute coronary occlusion induced by percutaneous transluminal coronary angioplasty.*, The American journal of cardiology **97**, 295–300 (2006).

[24] H. R. Aldrich, N. B. Hindman, T. Hinohara, M. G. Jones, J. Boswick, K. L. Lee, et al., *Identification of the optimal electrocardiographic leads for detecting acute epicardial injury in acute myocardial infarction.*, The American journal of cardiology **59**, 20–23 (1987).

[25] B. M. Horácek, J. W. Warren, C. J. Penney, R. S. MacLeod, L. M. Title, M. J. Gardner, et al., *Optimal electrocardiographic leads for detecting acute myocardial ischemia.*, Journal of Electrocardiology **34 Suppl**, 97–111 (2001).

[26] B. J. Drew, M. M. Pelter, M. G. Adams, S. F. Wung, T. M. Chou, and C. L. Wolfe, *12-lead ST-segment monitoring vs single-lead maximum ST-segment monitoring for detecting ongoing ischemia in patients with unstable coronary syndromes.*, American journal of critical care **7**, 355–363 (1998).

[27] P. Klootwijk, S. Meij, G. A. von Es, E. J. Müller, V. A. Umans, T. Lenderink, et al., *Comparison of usefulness of computer assisted continuous 48-h 3-lead with 12-lead ECG ischaemia monitoring for detection and quantitation of ischaemia in patients with unstable angina.*, European Heart Journal **18**, 931–940 (1997).

[28] B. J. Drew, M. M. Pelter, D. E. Brodnick, A. V. Yadav, D. Dempel, and M. G. Adams, *Comparison of a new reduced lead set ECG with the standard ECG for diagnosing cardiac arrhythmias and myocardial ischemia.*, Journal of Electrocardiology **35 Suppl**, 13–21 (2002).

[29] B. J. Drew, M. G. Adams, M. M. Pelter, S. F. Wung, and M. A. Caldwell, *Comparison of standard and derived 12-lead electrocardiograms for diagnosis of coronary angioplasty-induced myocardial ischemia.*, The American journal of cardiology **79**, 639–644 (1997).

[30] G. Wehr, R. J. Peters, K. Khalifé, A. P. B. V. Kuehlkamp, A. F. Rickards, and U. Sechtem, *A vector-based, 5-electrode, 12-lead monitoring ECG (EASI) is equivalent to conventional 12-lead ECG for diagnosis of acute coronary syndromes.*, Journal of Electrocardiology **39**, 22–28 (2006).

[31] M. Ohlsson, H. Ohlin, S. M. Wallerstedt, and L. Edenbrandt, *Usefulness of serial electrocardiograms for diagnosis of acute myocardial infarction.*, The American journal of cardiology **88**, 478–481 (2001).

[32] S.-E. Olsson, M. Ohlsson, H. Ohlin, S. Dzaferagic, M.-L. Nilsson, P. Sandkull, et al., *Decision support for the initial triage of patients with acute coronary syndromes.*, Clinical physiology and functional imaging **26**, 151–156 (2006).

[33] H. Haraldsson, L. Edenbrandt, and M. Ohlsson, *Detecting acute myocardial infarction in the 12-lead ECG using Hermite expansions and neural networks*, Artificial Intelligence in Medicine **32**, 127–136 (2004).

[34] L. Yang, P. Wang, Y. Jiang, and J. Chen, *Studying the explanatory capacity of artificial neural networks for understanding environmental chemical quantitative structure-activity relationship models.*, Journal of Chemical Information and Modeling **45**, 1804–1811 (2005).

[35] J. Trevelyan, E. W. A. Needham, S. C. H. Smith, and R. K. Mattu, *Impact of the recommendations for the redefinition of myocardial infarction on diagnosis and prognosis in an unselected united kingdom cohort with suspected cardiac chest pain.*, The American journal of cardiology **93**, 817–821 (2004).

III

Comparison of standard resampling methods for performance estimation of artificial neural network ensembles

Michael Green and Mattias Ohlsson

Computational Biology & Biological Physics,
Department of Theoretical Physics,
Lund University, Sölvegatan 14A, SE–223 62 Lund, Sweden

In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare*, E. Ifeachor (ed.) (2007).

Estimation of the generalization performance for classification within the medical applications domain is always an important task. In this study we focus on artificial neural network ensembles as the machine learning technique. We present a numerical comparison between five common resampling techniques: k-fold cross validation (CV), holdout, using three cutoffs, and bootstrap using five different data sets. The results show that CV together with holdout 0.25 and 0.50 are the best resampling strategies for estimating the true performance of ANN ensembles. The bootstrap, using the .632+ rule, is too optimistic, while the holdout 0.75 underestimates the true performance.

III.1 Introduction

Machine learning applications for classification in medicine is developing rapidly today and the question of how to best evaluate them has been addressed by many scientists. In the machine learning community it is well known that when training a classifier one should set aside a portion of the data for testing. Preferably this procedure should be repeated a number of times to collect statistics. Methods such as K-fold cross validation (CV), bootstrap [1] and holdout methods have been developed for dividing data into a training and test set. Rigorous resampling procedures are especially important when dealing with unstable learners such as Artificial Neural Networks (ANN) [2]. This machine learning concept has been used extensively over the years in many different areas of pattern recognition.

It is common knowledge that CV gives a nearly unbiased estimate of the performance of a classifier. However, this only applies if all aspects of model training is carried out within the CV loop [3]. Also CV often pay for this low bias in terms of large variance. In the late 90's Efron et. al. [1] introduced the .632+ bootstrap method as an improvement over CV. This method maintained low variance. There are, however, also reports that the .632+ rule can give large bias. Molinaro et. al. [4] found the .632+ method to be severely biased when dealing with high dimensional genomic data, and that CV, despite its large variance, was better at estimating the true performance of a classifier. Few comparisons of standard resampling methods for performance estimation has been conducted as of today [4, 5] and there is currently, to our knowledge, no study focusing on ANN ensembles. Furthermore, when using ANN ensembles it is important to incorporate all model training and model selection procedures within the performance estimation loop to avoid information leakage that would otherwise bias the estimation.

The aim of this study was to compare five common resampling methods for estimating the generalization performance of an ANN ensemble on five datasets. All of them were binary classification problems. First we tried to numerically establish which of the resampling methods that came closest to the true performance. Second we investigated whether the choice of resampling method for model selection had any influence on the true performance as estimated by the outer resampling method. Two common ensemble creation techniques were used, bagging [6] and the cross validation ensemble [7].

III.2 Methods

III.2.1 Datasets

Five datasets were used in this study. Three real world and two simulated datasets. The first real world dataset contained 12-lead electrocardiogram (ECG) data extracted from chest pain patients suspected of having transmural infarction (TMI) [8]. We used 18 features from the ECG and the training and test set consisted of 1000 and 3000 data points respectively. The second real world dataset was the Wisconsin Breast Cancer Database [9]. This breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. The database contained 699 patients of which 458 was diagnosed benign and 241 as malignant. In total 10 features was collected from each patient. Real world dataset number three was collected in 1997 and comes from 862 consecutive patients attending the Lund University emergency department with a principal complaint of chest pain [10, 11]. The diagnosis was either Acute Coronary Syndrome (ACS) or non-ACS. We used 16 PCA components extracted from 12-lead ECG recordings. Simulated dataset 1 contained data drawn from two multivariate Gaussian distributions with equal mean but with different covariance matrices. Specifically the two classes were generated from

$$p(\boldsymbol{x}|\mathcal{C}_k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \sigma_k^2 I)$$

where I is the identity matrix, and σ_k^2 the variance for class $\mathcal{C}_k$. The input vector $\boldsymbol{x}$ is eight dimensional and the size of the training and test set was 600 and 10000 respectively.

The second simulated dataset was acquired from two overlapping multivariate Gaussian distributions with equal covariance matrices but differing mean, as

$$p(\boldsymbol{x}|\mathcal{C}_k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \sigma^2 I)$$

where I is the identity matrix, and $\boldsymbol{\mu}_k$ the mean vector for class $\mathcal{C}_k$. The number of dimensions and the number of samples in the training and test set was the same as the first simulated dataset.

III.2.2 Artificial neural networks ensembles

We used ANN in the context of bagging [6] or CV ensembles [7] of size 25 and 24, respectively, which has been found to be sufficient in numerical studies [12]. The individual ensemble member ANNs were implemented as fully connected feedforward multilayer perceptrons (MLP) with no direct input-output connections. Only one hidden layer with five hidden units was used for all datasets. Each individual ANN in the ensemble was trained using a Quasi-Newton algorithm with the kullback-leibler error function for two classes

$$E = \sum_n (t_n \ln y_n + (1 - t_n) \ln(1 - y_n)) + \alpha E_{reg}$$

featuring a weight-elimination term

$$E_{reg} = \sum_i \frac{\omega_i^2}{\omega_0^2 + \omega_i^2}$$

to possibly regularize the network. The sum runs over all the weights in the network except the biases. The reason for excluding the biases from the penalty term is that we do not wish to force the decision boundary to pass through the origin in input space.

The CV ensemble method was used as follows: The training set was randomly divided into two parts of equal size. Two ensemble members were created by training one MLP on each of the two parts. This procedure was repeated 12 times, with a new random division each time. The resulting CV ensemble consisted of 24 MLPs.

III.2.3 Performance estimation

We used the area under the receiver operating characteristic curve (AUC) as a performance measure for a given ANN ensemble. The AUC can be interpreted as the probability that a randomly chosen data point from class $\mathcal{C}_1$ has a higher output value than a randomly chosen data point from class $\mathcal{C}_2$ [13]. The choice of AUC as the performance measure was mostly governed by its popularity, but also because it is independent of any cut on the output value.

In every dataset used in this study we put aside a large fraction (approximately 70%) to be used as an independent test set. The remaining data was used to estimate the performance of the ANN ensemble using five different resampling methods. The *true* performance of the ANN ensemble was evaluated by training an optimal ANN ensemble on the remaining data and testing on many bootstrap samples of the test set. The optimal model was chosen by a model selection procedure described later. In other words we used the performance of the ANN ensemble on the test set as a baseline for comparing the capability of the different resampling methods for correctly estimating the true performance. The whole procedure is illustrated in Figure III.1.

Five resampling methods was investigated; 5x5 fold CV, 25 fold bootstrap and 25 fold holdout using three cutoffs (0.25, 0.50 and 0.75). Thus, each method produced 25 new test and training data sets, labeled *TstP* and *TrnP* in Figure III.1, from the original training data. We built an optimized ANN ensemble for every training set (*TrnP*) using a model selection procedure described in the next section. The best model was then tested on the corresponding test data (*TstP*). This resulted in 25 training and test results that we used to estimate the performance of the ANN ensemble for each method. We used the mean of the 25 test AUCs for the CV and the holdout techniques, meanwhile the .632+ rule was used when evaluating the bootstrap method. This rule is less biased than its predecessor since it corrects, to some extent, for overfitting.

III.2.4 Model selection

The model selection consisted of a grid search for the optimal weight elimination parameter α. For each value of α an inner resampling session using bootstrap or CV was carried out on the training data. This process is also illustrated in Figure III.1. We used 25 resamples for the inner loop, i.e., a 5x5 fold CV or a 25 fold bootstrap. A full ANN

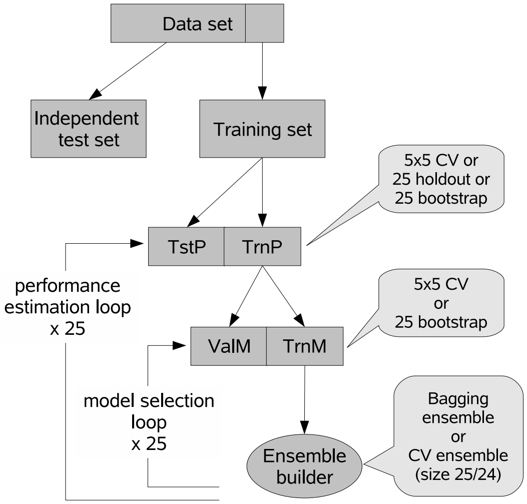
Data set
Independent test set
Training set
5x5 CV or 25 holdout or 25 bootstrap
TstP
TrnP
performance estimation loop x 25
5x5 CV or 25 bootstrap
ValM
TrnM
model selection loop x 25
Bagging ensemble or CV ensemble (size 25/24)
Ensemble builder

Figure III.1: Illustration of the performance estimation procedure. The data set is split into an independent test set and a training set. The latter is further divided into a *TstP* and a *TrnP* set used for the performance evaluation. The *TrnP* is finally split into a *ValM* and a *TrnM* set for the model selection. The methods used to split the data sets are indicated in the balloons.

ensemble was built from each resample (*TrnM*) using bagging [6] or the CV ensemble [7]. The α receiving the best AUC from the inner loop was used to construct an ANN bagging ensemble on the whole training set (*TrnP*).

III.2.5 Optimal Bayes classifier for simulated data

The two artificial datasets were generated from variants of the multivariate Gaussian distribution. Knowing the generating distribution allows us to derive the optimal Bayes classifier, that is, we can evaluate the posterior probability for class $\mathcal{C}_1$ given the data using Bayes' theorem. Following Bishop [2] and taking the functional form of the posterior to be sigmoid we set

$$p(\mathcal{C}_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} = \frac{1}{1+e^{-a}}$$

so that

$$a = \ln\frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} = \ln\frac{p(\boldsymbol{x}|\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)}$$

using the fact that we have $p(\mathcal{C}_1) = p(\mathcal{C}_2)$. Setting $a = 0$ gives us the decision boundary for our problem, corresponding to a cut of 0.5. For simulated data 1 this can be interpreted as a hypersphere with radius

$$r^2 = \frac{2}{1/\sigma_1^2 - 1/\sigma_2^2} \cdot \ln \frac{\sigma_2}{\sigma_1}.$$

The corresponding interpretation for simulated data 2 is a hyperplane defined by

$$\left(\frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{\sigma^2}\right)^T \boldsymbol{x} + \frac{\boldsymbol{\mu}_2^2 - \boldsymbol{\mu}_1^2}{2\sigma^2} = 0.$$

The performance of these classifiers was estimated, using AUC, by evaluating them on one hundred thousand samples from each class. Their performance should serve as an upper bound for the ANN ensemble since the ANN is known to estimate the Bayesian posterior probability [14].

III.3 Results

The results for the five data sets using bagging ensembles and CV in the model selection are presented in Table III.1 and III.2. The CV and holdout using cuts 0.25 and 0.50 had similar performance for all datasets. They differed with at most three percent from each other. See Figure III.3 and III.2. These three methods also did a good job estimating the true performance with differences ranging from one to three percent. The holdout 0.75 and bootstrap methods were strongly biased for the majority of the datasets, and rarely performed well. The bootstrap constantly overestimated the true performance meanwhile the holdout 0.75 remained rather pessimistic in its estimation.

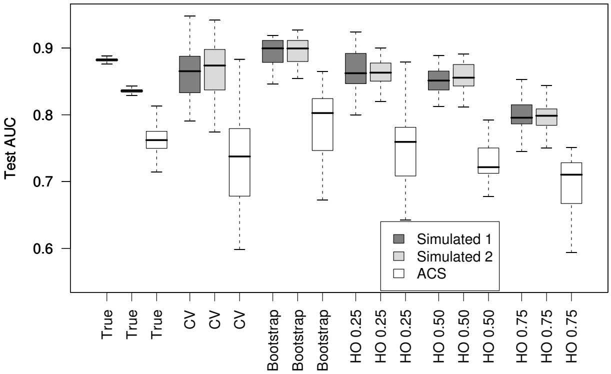

Figure III.2: Boxplots for ACS data and for Simulated data 1 and 2 using bagging ensembles and CV for the model selection.

Table III.1: Results for all five resampling methods on ACS and both simulated datasets using bagging ensembles and CV for the model selection. The results are presented as mean AUC, except for the bootstrap method where the .632+ estimator was used.

	Training	Validation	Test
		Simulation 1	
CV	0.97	0.87	0.86
Bootstrap	0.99	0.94	0.89
HO 0.25	0.97	0.86	0.86
HO 0.50	0.99	0.85	0.85
HO 0.75	1.00	0.79	0.80
True	0.96	0.88	0.88
		Simulation 2	
CV	0.97	0.87	0.87
Bootstrap	0.99	0.94	0.90
HO 0.25	0.98	0.86	0.86
HO 0.50	0.99	0.84	0.86
HO 0.75	1.00	0.79	0.80
True	0.88	0.86	0.84
		ACS	
CV	0.91	0.76	0.73
Bootstrap	1.00	0.90	0.79
HO 0.25	0.93	0.76	0.75
HO 0.50	0.96	0.75	0.73
HO 0.75	0.99	0.71	0.70
True	0.86	0.77	0.76

The true validation result was defined as the largest AUC during the model selection procedure on the whole training set. A closer look at the validation results revealed a bias for the holdout 0.75 method. It underestimated the validation performance with 4 to 12 percent. The opposite was true for the bootstrap method. It overshot the true validation performance by magnitudes ranging from 1 to 14 percent. The CV, holdout 0.50 and 0.25 methods only differed slightly from the true validation performance.

Turning our eyes to the CV ensembles in Table III.3 and III.4 we see that the results closely resembles the results for the bagging ensembles. Comparing the box plots in Figure III.5 and III.4 with Figure III.3 and III.2, no obvious differences could be found between CV and bagging ensembles for any of the data sets.

Training a single MLP instead of an entire ensemble resulted in a downward bias for all datasets. All methods underestimated the true performance to a larger extent than when using ensembles indicating that the single MLP was not able to generalize as well from the data. Boxplots for the single MLP are shown in Figures III.6 and III.7.

Using the .632+ bootstrap estimator during the model selection produced the same

Table III.2: Results for all five resampling methods on Breast cancer and TMI data using bagging ensembles and CV for the model selection. The results are presented as mean AUC, except for the bootstrap method where the .632+ estimator was used.

	Training	Validation	Test
	Breast cancer		
CV	1.00	0.99	0.99
Bootstrap	1.00	1.00	0.99
HO 0.25	1.00	0.99	0.99
HO 0.50	1.00	1.00	0.99
HO 0.75	1.00	0.87	0.99
True	1.00	0.99	0.99
	TMI		
CV	0.99	0.94	0.93
Bootstrap	1.00	0.97	0.95
HO 0.25	0.99	0.92	0.94
HO 0.50	0.99	0.91	0.92
HO 0.75	0.99	0.87	0.88
True	0.99	0.94	0.93

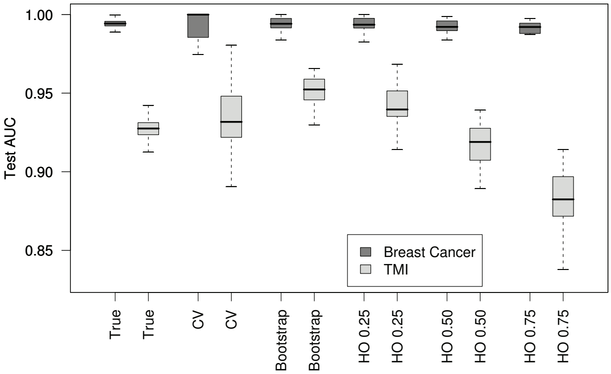

Figure III.3: Boxplots for the Breast cancer data and the TMI data using bagging ensembles and CV for the model selection.

estimation of the true performance as the CV. However, the models selected by the two different strategies differed as well as the corresponding AUCs. Looking closer into the values of the regularization parameter α selected by the two different model selection methods we found that most α's were small, indicating that no or very little

Table III.3: Results for all five resampling methods on ACS and both simulated datasets using CV ensembles and CV for the model selection. The results are presented as mean AUC, except for the bootstrap method where the .632+ estimator was used.

	Training	Validation	Test
		Simulation 1	
CV	0.97	0.87	0.87
Bootstrap	0.99	0.94	0.90
HO 0.25	0.97	0.87	0.87
HO 0.50	0.98	0.84	0.85
HO 0.75	0.99	0.78	0.80
True	0.95	0.88	0.88
		Simulation 2	
CV	0.97	0.87	0.87
Bootstrap	0.99	0.94	0.90
HO 0.25	0.98	0.87	0.87
HO 0.50	0.99	0.84	0.85
HO 0.75	0.99	0.79	0.80
True	0.88	0.86	0.83
		ACS	
CV	0.92	0.76	0.72
Bootstrap	0.99	0.90	0.79
HO 0.25	0.94	0.76	0.75
HO 0.50	0.97	0.76	0.73
HO 0.75	0.99	0.72	0.70
True	0.92	0.76	0.76

regularization were optimal for the ensembles.

III.4 Discussion and Conclusion

In this paper we examine five common resampling methods for the purpose of estimating the generalization performance using ANN classification ensembles. The process of training an ANN ensemble also includes resampling methods for creating the ensemble and resampling methods for the model selection part. To limit the number of combinations of resampling methods to test, the ensemble creation was limited to the bagging and cross validation ensemble. Furthermore, in the model selection part only two resampling methods were tested, CV and bootstrap. Although CV and bootstrap gave different estimations of the true test performance, no difference was found when using them in the model selection part. The reason for this is probably because the purpose of the model selection is to determine the regularization parameter α. Now

Table III.4: Results for all five resampling methods on Breast cancer and TMI data using CV ensembles and CV for the model selection. The results are presented as mean AUC, except for the bootstrap method where the .632+ estimator was used.

	Training	Validation	Test
		Breast cancer	
CV	1.00	0.99	0.99
Bootstrap	1.00	1.00	0.99
HO 0.25	1.00	0.99	0.99
HO 0.50	1.00	1.00	0.99
HO 0.75	1.00	0.95	0.99
True	1.00	0.99	0.99
		TMI	
CV	0.99	0.92	0.93
Bootstrap	1.00	0.97	0.95
HO 0.25	0.99	0.92	0.92
HO 0.50	0.99	0.91	0.92
HO 0.75	0.98	0.87	0.87
True	0.98	0.94	0.92

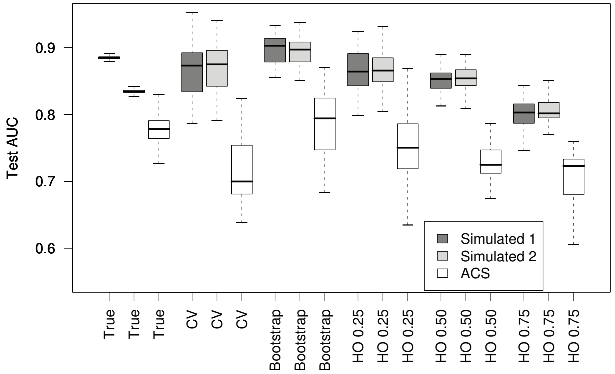

Figure III.4: Boxplots for ACS data and for Simulated data 1 and 2 using CV ensembles and CV for the model selection.

for ANN ensembles in general one expects little or no regularization at all, and this was confirmed in our results since the selected models had overall small α's. The model selection part is therefore not crucial, hence no difference between CV and bootstrap. In this study no feature selection was performed since the input variables were prede-

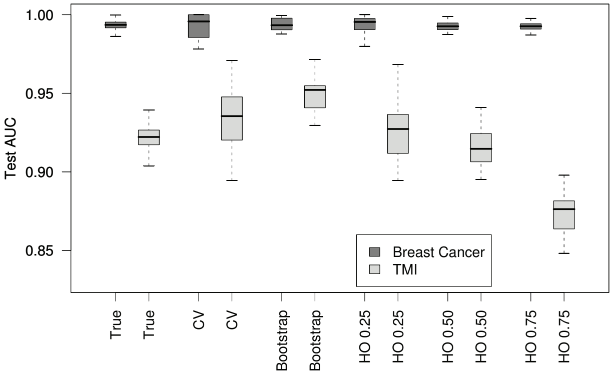

Figure III.5: Boxplots for the Breast cancer data and the TMI data using CV ensembles and CV for the model selection.

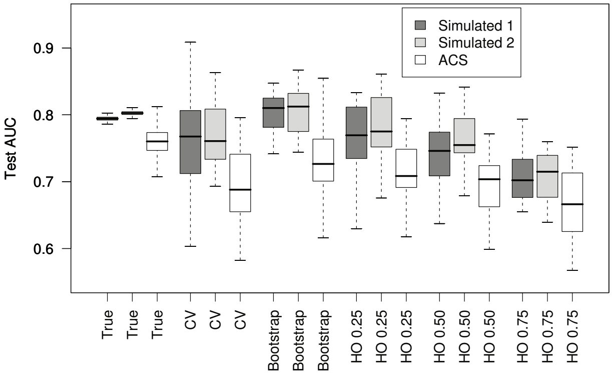

Figure III.6: Boxplots for ACS data and for Simulated data 1 and 2 using a single MLP and CV for the model selection.

fined. When including feature selection in the model selection process it may turn out that different resampling methods for the model selection will give different results.

Turning to the true performance estimation results we found that CV, holdout 0.25 and 0.50 performed equally well. The bootstrap method, using the .632+ rule, was constantly overestimating the test performance. Although the .632+ rule should compensate for possible overfitting, which is the case for the individual members of the

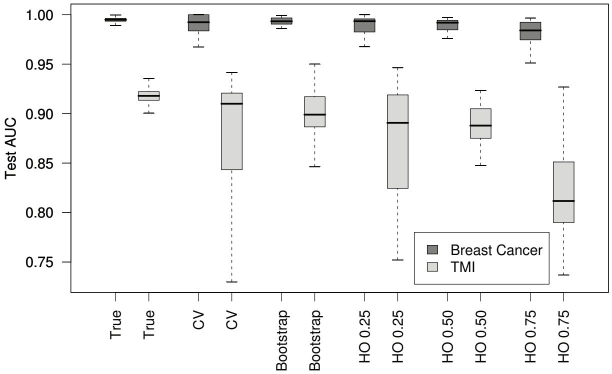

Figure III.7: Boxplots for the Breast cancer data and the TMI data using a single MLP and CV for the model selection.

ensemble, it is still biased. The holdout 0.75 resampling method was on the other hand constantly underestimating the true performance. This is probably due to the low fraction of data used to construct the ANN ensemble, hence a very inaccurate model.

Lingering on the true performance we found that the ANN ensemble succeeded to reach the optimal Bayes estimate using the second artificial data set. The first artificial problem was much more difficult and the true performance of the ANN ensemble did not match the Bayes estimate of 0.93. However, this was mainly an effect of undersampling since only 600 data points were used to construct the ANN ensemble. Increasing the flexibility of the networks as well as the amount of data available to the construction of the ANN ensemble alleviated this problem, indicating that our definition of *truth* made sense.

In this study we tested five different data sets, originating from three medical classification problems and two artificial ones. The medical applications ranged from being difficult to very easy. For the simulated data sets, one was linear and the second one required nonlinearity for the optimal solution. The advantage of using simulated data is of course the unlimited amount of test data. Although only a small number of data sets were used we believe that they represents suitable mix of different classification problems.

In conclusion we found, for our choice of data sets and training procedures, the best resampling strategies for estimating the true performance of an ANN ensemble to be the CV and holdout, using cutoff 0.25 and 0.50, methods. The .632+ bootstrap did not match this performance but still gave a much more accurate estimation than holdout 0.75. The choice of resampling technique in the model selection did not influence the final estimation. We can also confirm the well known advantage of using

ANN ensembles compared to single ANNs.

III.5 Acknowledgments

This work has been supported by the Swedish Knowledge Foundation through the Industrial PhD program in Medical Bioinformatics at the Strategy and Development Office (SDO) at the Karolinska Institute.

III References

[1] B. Efron and R. Tibshirani, *Improvements on cross-validation: The .632+ bootstrap method*, Journal of the American Statistical Association **92**, 548–560 (1997).

[2] C. M. Bishop, *Neural networks for pattern recognition*, Oxford University Press, 1995.

[3] S. Varma and R. Simon, *Bias in error estimation when using cross-validation for model selection*, BMC Bioinformatics **7**, 91 (2006).

[4] A. M. Molinaro, R. Simon, and R. M. Pfeiffer, *Prediction error estimation: a comparison of resampling methods.*, Bioinformatics **21**, 3301–3307 (2005).

[5] R. Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann, 1995, pp. 1137–1145.

[6] L. Breiman, *Bagging Predictors*, Machine Learning **24**, 123–140 (1996).

[7] M. Green, J. Björk, J. Hansen, U. Ekelund, L. Edenbrandt, and M. Ohlsson, *Detection of acute coronary syndromes in chest pain patients using neural network ensembles*, Second International Conference on Computational Intelligence in Medicine and Healthcare (Lisbon, Portugal) (J. M. Fonseca, ed.), IEE/IEEE, June-July 2005, pp. 182–187.

[8] S.-E. Olsson, M. Ohlsson, H. Ohlin, S. Dzaferagic, M.-L. Nilsson, P. Sandkull, et al., *Decision support for the initial triage of patients with acute coronary syndromes.*, Clinical physiology and functional imaging **26**, 151–156 (2006).

[9] O. L. Mangasarian and W. H. Wolberg, *Cancer diagnosis via linear programming*, SIAM News **23**, 1,18 (1990).

[10] M. Green, J. Björk, J. Forberg, U. Ekelund, L. Edenbrandt, and M. Ohlsson, *Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room*, Artificial Intelligence in Medicine **38**, 305–318 (2006).

[11] J. Björk, J. L. Forberg, M. Ohlsson, L. Edenbrandt, H. Ohlin, and U. Ekelund, *A simple statistical model for prediction of acute coronary syndrome in chest pain patients in the emergency department.*, BMC medical informatics and decision making **6**, 28 (2006).

[12] D. Opitz and R. Maclin, *Popular ensemble methods: An empirical study*, Journal of Artificial Intelligence Research **11**, 169–198 (1999).

[13] J. A. Hanley and B. J. McNeil, *The meaning and use of the area under a receiver operating characteristic(ROC) curve*, Radiology **143**, 29–36 (1982).

[14] M. D. Richard and R. P. Lippmann, *Neural network classifiers estimate bayesian a posteriori probabilities*, Neural Computation **3**, 461–483 (1991).

IV

In search of the best method to predict acute coronary syndrome using only the electrocardiogram from the emergency department

Jakob Lundager Forberg[1], Michael Green[2], Jonas Björk[3], Mattias Ohlsson[2], Lars Edenbrandt[4,5], Hans Öhlin[6] and Ulf Ekelund[1]

[1]Department of Clinical Sciences, Section for Emergency Medicine, Lund University Hospital, SE-22185 Lund, Sweden

[2]Computational Biology & Biological Physics, Lund University, Sölvegatan 14A, SE-22362 Lund, Sweden

[3]Competence Center for Clinical Research, Lund University Hospital, SE-22185 Lund, Sweden

[4]Department of Clinical Physiology, Malmö University Hospital, SE-20502 Malmö, Sweden

[5]Department of Clinical Physiology, Sahlgrenska University Hospital, SE-41345 Gothenburg, Sweden

[6]Department of Cardiology, Lund University Hospital, SE-22185 Lund, Sweden

LU TP 07-42 (submitted)

Introduction
The aim of this study was to compare different methods to predict acute coronary syndrome (ACS) using only data from a single ECG in the emergency department (ED).
Method
We compared the ACS prediction abilities of classical ECG criteria, human expert ECG interpretation, a logistic regression model and an artificial neural network ensemble (ANN). The ED ECG and discharge diagnoses were retrieved for 861 patient visits to the ED for chest pain. Cross-validation was used to estimate the generalization performance of the logistic regression and the ANN model.
Results
The logistic regression model had the overall best performance in predicting ACS with an area under the receiver operating characteristic curve of 0.88. The sensitivities of logistic regression, ANN, expert physicians and classical ECG criteria were 95, 95, 82 and 75% respectively, and the specificities were 54, 44, 63 and 69%.
Conclusion
Our logistic regression model was the best overall method to predict ACS, followed by our ANN. Decision support models have the potential to improve even experienced ECG readers' ability to predict ACS in the ED.

IV.1 Introduction

The correct identification of acute coronary syndrome (ACS; acute myocardial infarction or unstable angina pectoris) in patients with chest pain is a major challenge in the emergency department (ED). No diagnostic strategy in the ED has yet been able to reliably identify all cases of ACS, and hence admission to rule out ACS is often used. Among patients admitted for suspected ACS, 7 out of 10 prove not to have it [1–3]. Despite this marked over admission, some 2-5% of the patients with ACS are erroneously discharged from the ED [4]. Improved diagnostic methods are therefore needed.

The ECG is the single most important method to predict ACS in the ED [5,6]. ECG changes traditionally considered to indicate ACS [7,8] are ST-segment elevation in two or more contiguous leads ≥ 2 mm (leads V_1, V_2 and V_3) or ≥ 1 mm (other leads), ST-segment depression > 1 mm in two or more contiguous leads, and inverted T waves (>1 mm) in leads with predominant R-waves. Transient ST-segment changes >0.5 mm during symptoms also indicate ACS [9]. However, the sensitivity and specificity of a single ED ECG for acute myocardial infarction (AMI) is only 30-70% and 70-95% [10–13], and less for unstable angina. Attempts have therefore been made to improve the diagnostic yield of the ED ECG by ACS prediction models based on statistical methods [14] or artificial neural networks (ANN) [15,16]. These models have shown that ECG variables other than the above, and also interaction terms between variables, can be used as good predictors of ACS.

In the present study, we analyzed the ability of four different methods to predict ACS using only the first ECG recorded in the ED. A comparison was made between the classical ECG criteria, human expert ECG interpretation, a logistic regression model and an ANN.

IV.2 Methods

IV.2.1 Setting and patient material

The ED at Lund University Hospital has about 60,000 visits per year and serves a population of about 250,000 inhabitants. We retrospectively included 861 ED visits for chest pain where the first ECG could be electronically retrieved (Table IV.1). 643 of the visits were consecutive patients presenting at the ED from July 1 to November 20, 1997, and 218 visits were patients with a final diagnosis of ACS presenting at the ED from January 1, 1998 to March 16, 1999.

IV.2.2 Reference standard

Discharge diagnoses were made by the ED physician if patients were sent home from the ED. If patients were admitted the senior ward physician made the diagnosis after blood test (incl. biomarkers) and other investigations (exercise testing, echocardio-

Table IV.1: Selected patient and ECG characteristics.

	ACS (n = 344)	Non-ACS (n = 517)	Total (n = 861)
Males, n(%)	226(66)	291(56)	517(60)
Age, median (Q1-Q3)	70(60-78)	63(49-76)	66(54-77)
ST elevation, n(%)	96(28)	29(6)	125(15)
ST depression, n(%)	94(27)	29(6)	123(14)
T inversion, n(%)	223(65)	152(29)	375(44)

graphy etc.) were available. The diagnosis was reviewed by a senior research nurse according to the criteria given below, and when ambiguous, further reviewed by senior cardiologist. In the review of diagnoses for cases discharged from the ED, available data from the patient records indicated that the rate of missed diagnosis of ACS was low (not more than 2%). AMI was defined by the WHO criteria where the biochemical criterion was at least one measurement of CK-MB>10 $\mu g/L$ or Troponin T>0.1 $\mu g/L$. The criteria for unstable angina were ischemic symptoms (chest pain >15 min., syncope, acute heart failure or pulmonary edema) together with at least one of the following: a) Electrocardiogram (ECG) changes: transient or persisting ST segment depression (≥1 mm) and/or T-wave inversion (≥1 mm) without developing Q waves or loss of R wave height, or b) Biochemical markers: CK-MB 5-10 $\mu g/L$ or Troponin T 0.05-0.1 $\mu g/L$. This study was approved by the regional ethics committee at Lund.

IV.2.3 Electrocardiogram

The 12-lead ECGs were recorded and ECG measurements obtained by the use of computerized ECG recorders (Siemens-Elema AB, Solna, Sweden). For each of the 12 leads, the following 14 measurements were obtained: amplitudes of Q, R, S, ST, T+, T-, ST 2/8, and ST 3/8, duration of Q, QRS, R, and S, area of QRS, and slope of ST. The ST amplitudes 2/8 and 3/8 were the amplitude at the end of the second and third interval when the interval between ST-J point and the end of the T wave was divided into eight parts of equal duration.

IV.2.4 Statistical ECG predicting model

The statistical analyses were conducted using SPSS release 12.0.1 (SPSS Inc, Chicago, U.S.). All $12 \cdot 14 = 168$ obtained ECG measurements were used as continuous variables. The ECG measurements that had both positive and negative values, i.e. QRS area, slope of ST, and the amplitudes of ST, ST 2/8, and ST 3/8, were replaced by two distinct variables in the statistical analyses, yielding another $5 \cdot 12 = 60$ variables. Thus, we had in total $168 + 60 = 228$ ECG variables.

The statistical model was established in two steps. In the first explorative step, possibly important ECG variables were identified using discriminant analysis, with stepwise forward selection of variables based on Wilk's lambda [17]; $p < 0.05$ for inclusion and

$p > 0.10$ for exclusion). In the second step, identified ECG variables from the discriminant analysis were entered into a logistic regression model that estimates the probability of ACS for each patient [18]. We also entered three additional variables, derived from the original ECG measurements and regarded as important by two experienced ED physicians (JF and UE). We then excluded insignificant variables, one at a time. For the final set of significant variables (Table iv.2), significant two-way interaction terms were added to the model.

Table iv.2: Descriptive statistics for the significant determinants of acute coronary syndrome (ACS) in the logistic regression modeling. The numbers under ACS and No ACS are given as median (Q1, Q3). Also the symbol † means that the variable was derived from the original ECG measurements and regarded as important by two experienced ED physicians (JG and UE).

Variable	Lead	Effect	ACS	No ACS
Q amplitude	V_3	-	0 (0, 0)	0 (-26, 0)
	aVL	-	-36 (-88, 0)	-28 (-84, 0)
	aVF	-	0 (-76, 0)	0 (-98, 0)
	III	-	0 (-119, 0)	0 (-234, 0)
QRS area	aVF	-	0 (-105, 0)	0 (-336, 0)
R amplitude	III	+	240 (90, 560)	192 (83, 466)
R duration if at least 0.12 *sec*†	I	+	0 (0, 0)	0 (0, 0)
ST amplitude	V_3	-	0 (-8, 0)	0 (-47, 0)
	V_5	+	0 (0, 16)	0 (0, 26)
No. of ST amplitudes ≥ 1 *mm*†	V_1 - V_6	+	0 (0, 1)	0 (0, 2)
No. of ST amplitudes ≥ 0.5 *mm*†	I - III	+	0 (0, 0)	1 (0, 2)
ST amplitude 3/8	V_3	-	0 (0, 0)	0 (0, 0)
ST amplitude 2/8	V_4	-	0 (-3, 0)	0 (-56, 0)
ST slope	V_1	-	0 (0, 0)	0 (0, 0)
	I	+	5 (0, 10)	0 (0, 6)
T amplitude	V_1	+	81 (29, 160)	118 (47, 264)
	V_2	-	0 (0, 0)	0 (-72, 0)
	I	-	0 (-32, 0)	-30 (-124, 0)
	aVL	-	0 (-60, 0)	-68 (-175, 0)
	III	-	0 (-49, 0)	0 (-78, 0)

IV.2.5 Artificial neural network

One of the more powerful classification methods today is artificial neural networks [19], and they have often been used in medical studies [20] over the years. In this work we created ensembles [21] of artificial neural networks where each ensemble member had one hidden layer with 10 units. This many hidden nodes allow the networks to become fairly complex. Thus, in order for the networks to generalize well, we introduced a regularization framework to tune the complexity further. A gentle introduction to neural networks can be found in Cross et. al [19]. The 168 continuous variables available to the network was reduced down to 54 by a backward feature selection procedure. Each feature subset was subjected to a cross-validation run, and a feature was removed only if it had no apparent effect on the performance of the ensemble. All the networks were created using custom-made C++ software.

IV.2.6 Expert consensus interpretation

Two physicians highly experienced in ECG reading separately classified all 861 ECGs into one of the following three ordinal groups: 1. No signs of ACS, 2. Possible ACS or 3. ACS. Classical criteria ECG criteria for ACS were not strictly used. Instead, also the configuration of the ST segment and the shape of the QRS complex were considered, i.e. a pattern recognition analysis was applied as in the clinical routine interpretation of ECGs. The experts made the same primary classification in 520 of the 861 ECGs. For the discrepant cases, a consensus classification was done.

IV.2.7 Classical ECG criteria

A regression model was established using only the classical ECG criteria: ST-segment elevation in two or more contiguous leads ≥ 2 mm (leads V_1, V_2 and V_3) or ≥ 1 mm (other leads), ST-segment depression >1 mm in two or more contiguous leads, and inverted T waves (>1 mm) in leads with predominant R-waves.

IV.2.8 Performance evaluation and cross-validation of the statistical and the ANN model

The area under the receiver-operating-characteristic (ROC) curve was used as an overall measure of the prediction abilities of the different models. For calculation of specificity and predictive values for the models, the sensitivity was set to 95%. This somewhat arbitrary level was chosen because with current standard evaluation, some 2-5% of the ACS patients are erroneously discharged from the ED, which implies a sensitivity of at least 95% for the routine ED work-up. We used McNemar's exact test for correlated proportions to test differences in sensitivity and specificity between the different models.

We used a 10-fold cross-validation procedure to estimate the generalization performance of the statistical and the ANN model. A total of 100 evaluation sets were established randomly, each with 90% of the patients used for training and the remaining 10% used for validation. Median values of ROC and the mean probabilities from 100 generated validation sets were used to assess the generalization ability.

IV.3 Results

IV.3.1 Statistical model

The cross-validated median ROC-area of the statistical model was 88% (2.5 - 97.5 percentiles 79 - 94%). The specificity was 54% at 95 % sensitivity when the mean probabilities from the generated validation sets were used .

IV.3.2 Artificial neural network

The ANN model yielded a cross-validated median area under ROC of 86% in (Figure IV.1; 2.5 - 97.5 percentiles 78 - 93%). At a set sensitivity of 95 % the median specificity was 44% , which was significantly lower than for the statistical model ($p < 0.001$).

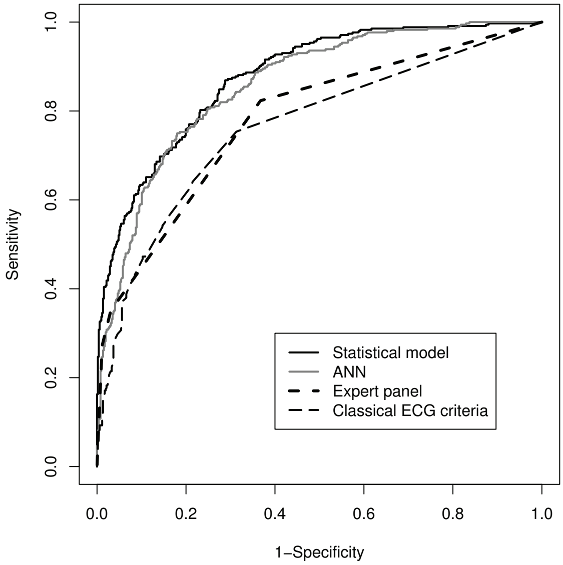

Figure IV.1: Receiver-operating-characteristic (ROC) curve for prediction of acute coronary syndrome among 861 patients based on i) the statistical model with interaction terms (ROC-area=88%, black solid curve), ii) the ANN (ROC-area=86%, grey solid curve), iii), the expert panel (ROC-area=78%, dotted curve), iv) classical ECG criteria (ROC-area=76%, dashed curve). Both i) and ii) are based on the mean probabilities from the generated validation sets.

IV.3.3 Expert consensus interpretation

The performance of the expert classifications was also below that of the statistical model (area of the ROC-curve = 78%; Figure IV.1). If all patients classified as "possible ACS" or "ACS" were regarded as ACS, then this would yield a sensitivity of 82% and a specificity of 63% (Table IV.3). A comparison between the expert classifications and the statistical model, using mean ACS probabilities from the validation sets, showed that the latter was able to identify a larger proportion of true ACS (Table IV.3). Thus, the statistical model had significantly higher sensitivity (95% vs. 82%) than the expert classifications ($p < 0.001$; expert classifications dichotomized as above). On the other hand, the statistical model had lower specificity than the expert classifications (54% vs. 63%; $p < 0.001$). Similarly, the mean ACS probabilities for the ANN from the validation sets yielded significantly higher sensitivity (95%) than the expert classifications, but significantly lower specificity (43%, $p < 0.001$ for both comparisons). However, looking at the ROC curve (Figure IV.1) it is apparent that for any given sensitivity, both the statistical and the ANN model are always more specific then the expert classifications.

Table IV.3: Comparison of the predictions of the statistical model, using mean probabilities from the generated validation sets, with the consensus classifications of the two experts for the 861 patients with complete ECG data. All numerals are given as numbers (with percentages in brackets for marginal totals). In the comparison with the statistical model all patients with an estimated probability of at least 0.13 were classified as ACS.

		Expert classification			Total
		ACS	Possible ACS	No signs of ACS	
True ACS		121(35)	162(47)	61(18)	344(100)
Statistical model	ACS	117	157	53	327(95)
	No ACS	4	4	8	17(5)
True Non-ACS		16(3)	174(34)	327(63)	517(100)
Statistical model	ACS	14	110	114	238(46)
	No ACS	2	64	213	279(54)

IV.3.4 Using the classical ECG criteria to predict ACS

The area under of the ROC-curve was 76%, which is an overall performance below that of the statistical and the ANN models. The ROC-curve could e.g. yield 75% sensitivity and 69% specificity, if all patients with ST-segment elevation/depression or with inverted T-waves according to the classical criteria (see Methods) were classified as ACS.

A summary of the performances of all four methods to predict ACS is shown in Table IV.4.

Table IV.4: Performance of methods to predict ACS. The predictive values are based on an ACS prevalence of 21% as in the consecutive ED subgroup. For the statistical method and ANN, sensitivities are set to 95%.

	Sensitivity	Specificity	PPV	NPV
Classical criteria	0.75	0.69	0.39	0.91
Expert interpretation	0.82	0.63	0.37	0.93
Statistical model	0.95	0.54	0.35	0.98
Artificial neural network	0.95	0.44	0.53	0.94

IV.4 Discussion

In the present study, our statistical model based on multivariable logistic regression showed the best overall performance in predicting ACS from the ED ECG. The few previously published statistical prediction models for ACS have been using both the ECG and clinical data, and have produced ROC-areas of between 81 and 89%. A model with seemingly better performance (ROC area 96%) was reported by Kennedy & Harrison, but their model was both developed and tested in a chest pain population with a clearly higher prevalence of significant ECG changes among the ACS patients. Our statistical model, using only the ECG, can therefore be considered to perform at the same level as previous models based on both clinical and ECG variables. This further supports the notion that most of the ACS predictive information currently available in the ED can be found in the ECG [5, 6].

The present ANN performed well in predicting ACS, although significantly less well than our statistical model. In previous studies, logistic regression and ANN models have been equally good at predicting acute cardiac ischemia when using identical ECG and clinical variables [22–24], and it has been suggested that logistic regression should be the method of choice because of easier use [24]. However, when ANNs and statistical models are established on exactly the same input variable definitions, ANNs may in some settings perform better [25]. One reason for the decrease in performance for the ANN in this study might be that no additional features were added before hand. The performance of the present ANN to predict ACS was similar to previous ANN models predicting AMI using ECG data only [15].

Both the statistical model and the ANN were superior to the expert ECG readings. This confirms and extends previous findings that ANN were superior in AMI prediction than the experienced physician [15]. We therefore believe that statistical models and ANN in general are better than human expert reading in predicting ACS from the ECG and that they have the potential to improve doctors' ECG interpretation in clinical practice [16].

As might have been expected, the model based on classical ECG criteria for ACS was the least powerful method to identify ACS. It is well known that many ACS patients lack the traditional ECG changes. For instance, the ED ECG is considered to be diagnostic in only some 50% of AMI patients [26], and as much as 40% of unstable angina patients have normal ECGs in the ED [27]. The markedly better performances by the expert physicians, the ANN and the statistical model indicate that a great amount of diagnostic information is not present in the classical ECG criteria for ACS. Some of this information seems to be "hidden" also from expert physicians, since both the ANN and the statistical model performed better than the experts.

Before it can be used in clinical practice, our statistical model has to be prospectively validated, preferably at other institutions. Should the model retain positive and negative predictive values of 35 and 98% after such validation, we believe it could be used as decision support in clinical practice, perhaps even for experienced physicians. However, the model in itself is not powerful enough to reliably indicate which patients should be admitted or discharged, or which patients should be given immediate treatment for suspected ACS. Adding results of blood samples and physician judgment is therefore imperative, and would probably increase the predictive ability to very high levels. A comparison with a previous ECG would probably also increase the predictive performance, since 20% of ECG changes in the ED are old [28] and the availability of a previous ECG has been shown to increase diagnostic specificity in cases of suspected ACS [29]. Future statistical models for ACS prediction should therefore probably include a comparison with previous ECGs, if they are available.

In summary we found that using data from a single ECG in patients seeking the ED because of chest pain, our logistic regression model and our ANN model were both superior to expert physicians and classical ECG criteria in predicting ACS. It can be concluded that decision support models have the potential to improve ACS prediction even by experienced ECG readers in the ED.

IV.4.1 Limitations of the study

The patients included in the present study were retrospectively collected and from one center only. The diagnoses of the patients discharged from the ED were not tested with routine post-discharge ECG or blood samples for cardiac markers. An old definition of AMI was used. Newer definitions of AMI have lower cut-off values for biochemical markers, and some of the unstable angina diagnoses in this study would currently be classified as AMI. However, the total number of ACS cases would probably be little changed. Since it is impossible to analyze all information with a potential impact on the likelihood of ACS, it is possible that variables in addition to those included in our model could be important for ACS prediction. The size of the sample might also have limited our abilities to detect ECG characteristics of low prevalence that nevertheless are important for ACS classification. A cross-validation procedure was used to obtain an accurate estimate of the ROC performance. However, the feature selection was not performed within the cross-validation loop, instead the selected features, for both methods, was found using the full patient material. This can result in a too optimistic estimate of the ROC performance.

IV.5 Financial support

This study was supported by grants from the Skåne Region, the Laerdal Foundation for Acute Medicine and by the Swedish Knowledge Foundation through the Industrial PhD program in Medical Bioinformatics at the Strategy and Development Office at the Karolinska Institute.

IV References

[1] B. W. Karlson, J. Herlitz, O. Wiklund, A. Richter, and A. Hjalmarson, *Early prediction of acute myocardial infarction from clinical history, examination and electrocardiogram in the emergency room.*, The American journal of cardiology **68**, 171–175 (1991).

[2] J. L. Forberg, L. S. Henriksen, L. Edenbrandt, and U. Ekelund, *Direct hospital costs of chest pain patients attending the emergency department: a retrospective study*, BMC Emerg Med **6**, 6 (2006).

[3] U. Ekelund, H.-J. Nilsson, A. Frigyesi, and O. Torffvit, *Patients with suspected acute coronary syndrome in a university hospital emergency department: an observational study*, BMC Emergency Medicine **2**, 1 (2002).

[4] J. Pope, T. Aufderheide, R. Ruthazer, R. Woolard, J. Feldman, J. Beshansky, et al., *Missed diagnoses of acute cardiac ischemia in the emergency department*, The New England Journal of Medicine **342**, 1163–1170 (2000), Center for Cardiovascular Health Services Research, Department of Medicine, New England Medical Center, Boston, Mass 02111, USA.

[5] L. Goldman and A. J. Kirtane, *Triage of patients with acute chest pain and possible cardiac ischemia: the elusive search for diagnostic perfection.*, Annals of internal medicine **139**, 987–995 (2003).

[6] S. Yusuf, M. Pearson, H. Sterry, S. Parish, D. Ramsdale, P. Rossi, et al., *The entry ECG in the early diagnosis and prognostic stratification of patients with suspected acute myocardial infarction.*, European Heart Journal **5**, 690–696 (1984).

[7] M. E. Bertrand, M. L. Simoons, K. A. A. Fox, L. C. Wallentin, C. W. Hamm, E. McFadden, et al., *Management of acute coronary syndromes in patients presenting without persistent ST-segment elevation.*, European Heart Journal **23**, 1809–1840 (2002).

[8] J. S. Alpert, K. Thygesen, E. Antman, and J. P. Bassand, *Myocardial infarction redefined–a consensus document of the joint european society of cardiology/american college of cardiology committee for the redefinition of myocardial infarction*, Journal of the American College of Cardiology **36**, 959–969 (2000).

[9] E. Braunwald, E. M. Antman, J. W. Beasley, R. M. Califf, M. D. Cheitlin, J. S. Hochman, et al., *Acc/aha guideline update for the management of patients with unstable angina and non-ST-segment elevation myocardial infarction–2002: summary article: a report of the american college of cardiology/american heart association task force on practice guidelines (committee on the management of patients with unstable angina).*, Circulation **106**, 1893–1900 (2002).

[10] F. M. Fesmire, R. F. Percy, J. B. Bardoner, D. R. Wharton, and F. B. Calhoun, *Usefulness of automated serial 12-lead ECG monitoring during the initial emergency department evaluation of patients with chest pain.*, Annals of emergency medicine **31**, 3–11 (1998).

[11] G. P. Young and T. R. Green, *The role of single ECG, creatinine kinase, and ckmb in diagnosing patients with acute chest pain.*, The American journal of emergency medicine **11**, 444–449 (1993).

[12] P. J. Kudenchuk, C. Maynard, L. A. Cobb, M. Wirkus, J. S. Martin, J. W. Kennedy, et al., *Utility of the prehospital electrocardiogram in diagnosing acute coronary syndromes: the myocardial infarction triage and intervention (MITI) project*, Journal of the American College of Cardiology **32**, 17–27 (1998).

[13] R. L. Jesse and M. C. Kontos, *Evaluation of chest pain in the emergency department.*, Current problems in cardiology **22**, 149–236 (1997).

[14] J. Björk, J. L. Forberg, M. Ohlsson, L. Edenbrandt, H. Ohlin, and U. Ekelund, *A simple statistical model for prediction of acute coronary syndrome in chest pain patients in the emergency department.*, BMC medical informatics and decision making **6**, 28 (2006).

[15] B. Heden, H. Öhlin, R. Rittner, and L. Edenbrandt, *Acute myocardial infarction detected in the 12-Lead ECG by artificial neural networks*, Circulation **96**, 1798–1802 (1997).

[16] S.-E. Olsson, M. Ohlsson, H. Ohlin, S. Dzaferagic, M.-L. Nilsson, P. Sandkull, et al., *Decision support for the initial triage of patients with acute coronary syndromes.*, Clinical physiology and functional imaging **26**, 151–156 (2006).

[17] R. Johnson and D. Wichern, *Applied multivariate statistical analysis*, 2 ed., Prentice Hall Inc., 1998.

[18] D. Hosmer and S. Lemeshow, *Applied logistic regression*, Wiley, New York, 2000.

[19] S. Cross, R. Harrison, and R. Kennedy, *Introduction to neural networks*, Lancet **346**, 1075–1079 (1995).

[20] P. J. G. Lisboa, *A review of evidence of health benefit from artificial neural networks in medical intervention*, Neural Networks **15**, 11–39 (2002).

[21] A. Krogh and J. Vedelsby, *Neural network ensembles, cross validation, and active learning*, Advances in Neural Information Processing Systems (San Mateo, CA) (G. Tesauro, D. Touretzky, and T. Leen, eds.), vol. 2, Morgan Kaufman, 1995, pp. 650–659.

[22] H. P. Selker, J. L. Griffith, S. Patil, W. J. Long, and R. B. D'Agostino, *A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients*, Journal of investigative medicine **43**, 468–476 (1995).

[23] W. Baxt, F. Shofer, F. Sites, and J. Hollander, *A neural computational aid to the diagnosis of acute myocardial infarction*, Annals of Emergency Medicine **34**, 366–373 (2002).

[24] R. Harrison and R. Kennedy, *Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation*, Annals of Emergency Medicine **46**, 431–439 (2005).

[25] M. Green, J. Björk, J. Forberg, U. Ekelund, L. Edenbrandt, and M. Ohlsson, *Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room*, Artificial Intelligence in Medicine **38**, 305–318 (2006).

[26] E. B. Sgarbossa, Y. Birnbaum, and J. E. Parrillo, *Electrocardiographic diagnosis of acute myocardial infarction: Current concepts for the clinician*, American heart journal **141**, 507–517 (2001).

[27] J. Pope, R. Ruthazer, J. Beshansky, J. Griffith, and H. Selker, *Clinical features of emergency department patients presenting with symptoms suggestive of acute cardiac ischemia: A multicenter study*, Journal of Thrombosis and Thrombolysis **6**, 63–74 (1998), Center for Cardiovascular Health Services Research, Division of Clinical Care Research, Department of Medicine, New England Medical Center and Baystate Medical Center, Tufts University School of Medicine, Boston, Massachusetts.

[28] C. J. Lindsell, V. Anantharaman, D. Diercks, J. H. Han, J. W. Hoekstra, J. E. Hollander, et al., *The internet tracking registry of acute coronary syndromes (i*tracs): a multicenter registry of patients with suspicion of acute coronary syndromes reported using the standardized reporting guidelines for emergency department chest pain studies*, Annals of emergency medicine **48**, 666–77, 677.e1–9 (2006).

[29] T. H. Lee, E. F. Cook, M. C. Weisberg, G. W. Rouan, D. A. Brand, and L. Goldman, *Impact of the availability of a prior electrocardiogram on the triage of the patient with acute chest pain*, Journal of general internal medicine **5**, 381–388 (1990).

V

Exploring new possibilities for case-based explanation of artificial neural network ensembles

Michael Green[1], Ulf Ekelund[2], Lars Edenbrandt[3,4], Jonas Björk[5], Jakob Lundager Forberg[2] and Mattias Ohlsson[1]

[1]Computational Biology & Biological Physics, Lund University, Sölvegatan 14A, SE-22362 Lund, Sweden

[2]Department of Clinical Sciences, Section for Emergency Medicine, Lund University Hospital, SE-22185 Lund, Sweden

[3]Department of Clinical Physiology, Malmö University Hospital, SE-20502 Malmö, Sweden

[4]Department of Clinical Physiology, Sahlgrenska University Hospital, SE-41345 Gothenburg, Sweden

[5]Competence Center for Clinical Research, Lund University Hospital, SE-22185 Lund, Sweden

LU TP 07-43 (submitted)

Artificial neural network (ANN) ensembles has long suffered from lack of interpretability. This has severely limited the practical usability of ANNs in settings where an erroneous decision can be disastrous. Several attempts have been made to alleviate this problem. Many of them are based on decomposing the decision boundary of the ANN into a set of rules.

We explore and compare a set of new methods for this explanation process on two artificial data sets (Monks 1 and 3), and one acute coronary syndrome data set consisting of 861 electrocardiograms (ECG) collected retrospectively at the emergency department at Lund University Hospital.

The algorithms managed to extract good explanations in more than 84% of the cases. More to the point, the best method provided 99 and 91% good explanations in Monks data 1 and 3 respectively. Also there was a significant overlap between the algorithms. Furthermore, when explaining a given ECG, the overlap between this method and one of the physicians was the same as the one between the two physicians in this study. Still the physicians were significantly, p-value < 0.001, more similar to each other than to any of the methods.

The algorithms has the potential to be used as an explanatory aid when using ANN ensembles in clinical decision support systems.

V.1 Introduction

Artificial neural networks (ANN) has been gaining interest in the medical community for quite some time now, and has proven useful for many clinical decision problems [1, 2]. Still, as of today, there are very few live applications in use at the clinics. Though the reasons for this low usage are numerous, one major drawback is the lack of interpretability of the decisions provided by an ANN [2].

Most efforts of making sense out of an ANN decision is based on rule extraction methods where the decision boundary is discretized into segments. There are basically two ways of attacking this problem in neural networks. The first is the *decompositional* [3] approach where the network is scrutinized from within in order to extract useful information about a decision. This is usually done by analyzing the activations of individual nodes in the network as well as the weights leading into them. This methodology was used by Kolman et. al. [3] where they demonstrated that an ANN is mathematically equivalent to an all permutation fuzzy rule base. Their work provided an explicit way of transforming an ANN into a set of IF THEN rules. Despite being intuitively attractive this approach lead to a large number of rules that had to be reduced.

The second one known as the *pedagogical* [4, 5] approach treats the network as a black box. Here the analysis in based on examining the relationship between what is fed into the network with what is returned as output. In a recent paper [5] the pedagogical approach was used when developing the orthogonal search based rule extraction (OSRE) method that successfully extracted the exact rules for the Monks [6] data. They also point out that, in the presence of large node output weights, the decompositional approach may fail to accurately describe the logic of the network.

Another way to analyse a neural network is by sensitivity analysis where the main focus has been on extracting global properties. Usually this has been accomplished by analysing the weights in the network on a pattern by pattern basis. Interestingly enough this has been considered a drawback by several authors [7–9].

From a medical application point of view it is often necessary to provide an explanation underlying a given decision. If the decision support is to function in a clinical stressful setting (e.g. an emergency department) then it is required to provide a fast explanation for each case, easily interpretable by the operator. This case-based feed-back requirement is lacking in most methods for analyzing the operation of a neural network ensemble. We believe this has severely limited the full potential of using neural networks in a clinical decision support system. The idea of using the specific case at hand as the basis for the feed-back algorithm is not new. In [10] a specific method was developed for electrocardiogram curves, where the case-based feed-back was presented as modified curves representing changes towards being more healthy or non-healthy. In [11] rules were extracted and later ranked depending on the prediction of the case. The idea was that more complex rules should be presented when the decision support system classified a patient as healthy. Conversely if a patient were classified as non-healthy, less complex rules were given as feed-back. Another approach to case-based explanation can be found in [12] where the reasoning behind the neural network was presented as showing a set of similar cases.

When providing feedback to a physician in a clinical situation we need to make sure that only the core of the driving forces behind a classification is presented. This means that a rule based approach, where possibly more than 10 rules are presented per case, will be difficult to use in practise. Also many of the rules will be non-specific for a given case since the rules are extracted globally from the data set with the aim of approximating the decision boundary of the ANN. To us this suggests that any case-based feedback should be derived from a single case and not the entire data set. Case-based feed back is indeed dependent on the question one is asking. In a clinical setting we often find the important feed-back to simply be the set of variables, most important for the decision. The five approaches described in this study will all result in a ranked list of important variables and the explanation will simply consist of the topmost important ones, for each case.

V.2 Methods

V.2.1 Artificial neural network ensembles

Definition

The ANN ensemble consists of standard multi-layer perceptrons (MLP) combined by simple averaging. Hence, the output of the ANN ensemble of size I, for a data point $\boldsymbol{x}$, is given by

$$y(\boldsymbol{x}, \boldsymbol{\omega}) = \frac{1}{I} \sum_{i=1}^{I} f\left(\sum_{j=1}^{J} \omega_{ij} \cdot g\left(\sum_{k=1}^{K} \tilde{\omega}_{ijk} \cdot x_k + \tilde{\omega}_{ij0} \right) + \omega_{i0} \right) \tag{v.1}$$

where the functions f and g are the logistic sigmoid and hyperbolic tangent, respectively. The size of the MLPs are (J-K-1) and where the set of weights $(\omega_{ij}, \omega_{i0}, \tilde{\omega}_{ijk}, \tilde{\omega}_{ij0})$ follows obvious notation, e.g. $\tilde{\omega}_{ijk}$ is the weight between input k and hidden node j in ensemble member i.

Training procedure

Our ensembles were created by bagging [13], where 25 MLPs were trained on bootstrap [14] samples from the training data. The MLPs were trained by gradient descent using the cross entropy error function for two classes with an added weight elimination term

$$E = \sum_{n=1}^{N} \left(\ln y_n^{t_n} + \ln(1 - y_n)^{1-t_n} \right) + \alpha \sum_{i} \frac{\omega_i^2}{\omega_0^2 + \omega_i^2} \tag{v.2}$$

to possibly regularize the network. The last sum runs over all the weights in the network except the biases. The reason for excluding the biases from the penalty term is that we do not wish to force the decision boundary to pass through the origin in input space. The classes are encoded in the zero-one variable t_n where, for medical applications, $t_n = 1$ usually indicates the event of interest (e.g. not healthy).

The α parameter was tuned by a grid search using the area under the receiver operating characteristics curve (AUC) [15]. In a medical setting the AUC can be interpreted as the probability of a randomly chosen sick patient having a larger predicted risk than a randomly chosen healthy patient. The AUC is frequently used in medical applications as the performance measure and have the advantage of being independent of any cuts imposed on the output value. Figure v.1 illustrates the model selection procedure, where a grid search is performed with respect to the ensemble and not the individual networks.

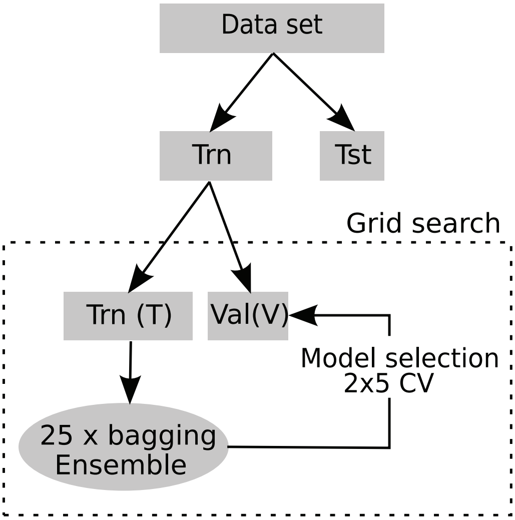

Figure v.1: The model selection scheme. A given training data set was split into several training/validation parts using 5-fold cross validation. Each of these smaller training sets (T) were then used to create an ANN ensemble and the corresponding validation set (V) was used for validation. The whole procedure was repeated 2 times with different random 5-fold cross validation splits.

The grid search consists of a 2x5 fold cross validation run for each value of the regularization parameter α. The α that corresponded to the largest AUC was selected and a final ANN ensemble was created using the original training set.

V.2.2 Explanation methods

The main purpose of the explanation methods presented below is to produce an input variable importance ranking, given a particular data point and a trained ANN ensemble. The ranking list will then be used to provide an explanation for each case.

Input sensitivity

The sensitivity of an ANN ensemble output can be defined by a partial derivative of $y(\boldsymbol{x}, \boldsymbol{\omega})$, Eqn. (v.1), with respect to a variable x_l. The resulting derivative is

$$\frac{\partial y(\boldsymbol{x}, \boldsymbol{\omega})}{\partial x_l} = \frac{1}{I} \sum_{i=1}^{I} f_i' \sum_{j=1}^{J} \omega_{ij} \cdot g_{ij}' \cdot \tilde{\omega}_{ijl} \tag{v.3}$$

where f_i' is the derivative of ensemble member i:s output function. Similarly g_{ij}' is the derivative of the output from hidden node j in ensemble member i. Because $f_i' = f_i \cdot (1 - f_i)$, data points with outputs close to 1 or 0 e.g. confident outputs, will have very low values. To avoid this unfortunate property we simply remove the f_i' and define our input sensitivity measure for input x_l using pattern $\boldsymbol{x}$ as

$$S_l(\boldsymbol{x}) = \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \omega_{ij} \cdot g_{ij}' \cdot \tilde{\omega}_{ijl} \tag{v.4}$$

Removing f_i' from Eqn. (v.3) can also be seen as taking the partial derivative of the argument to the output activation function f in Eqn. (v.1). The input variables are ranked using the absolute value of $S_l(\boldsymbol{x})$, since we do not distinguish between an increase or a decrease of the output. The smaller value of $|S_l(\boldsymbol{x})|$ the lower importance (rank) for input variable l and case $\boldsymbol{x}$.

Generalized odds ratio

The well known concept of odds ratios from statistics can be used with ANN as well. The difference is that the ratio will be data point dependent, which is exactly what we want for case-based explanations. In other words, we observe how the odds of an event varies when we do a unit increase of the variable of interest. More specifically we calculate the ratio of the odds of the event for the two different cases.

Given a data point $\boldsymbol{x}$ we calculate the odds ratio for a variable x_l by

$$OR_l(\boldsymbol{x}) = \frac{y(\boldsymbol{x}^{l,*}, \boldsymbol{\omega})(1 - y(\boldsymbol{x}, \boldsymbol{\omega}))}{y(\boldsymbol{x}, \boldsymbol{\omega})(1 - y(\boldsymbol{x}^{l,*}, \boldsymbol{\omega}))} \tag{v.5}$$

where y is defined in Eqn. (v.1) and $\boldsymbol{x}^{l,*}$ is the same as $\boldsymbol{x}$ with the exception that $x_l^{l,*} = x_l + 0.1\sigma_l$. Thus the increase compensates for distributional differences among the variables. The input variables are ranked using the value $OR_l(\boldsymbol{x})$.

In the case of an ANN ensemble of size 1, expression (v.5) simplifies into

$$OR_l(\boldsymbol{x}) = \exp\left(\sum_{j=1}^{J} \omega_{ij}\left(g_{ij} - g_{ij}^{l,*}\right)\right) \tag{v.6}$$

which is similar to the well known result for the odds ratio in the linear case.

Euclidean distance

Here we pursue a geometrical approach by finding the the point $\boldsymbol{p}$ on the decision boundary closest to a given data point $\boldsymbol{x}$. We define the decision boundary to be the set of input patterns that produce an output of the network ensemble equal to the prevalence of the training data set. We find the boundary point $\boldsymbol{p}$ by network ensemble inversion, see e.g. [4]. This allows us to calculate a distance vector $\boldsymbol{D} = \boldsymbol{x} - \boldsymbol{p}$. The inputs are ranked by the magnitude of the resulting vector's components i.e. by $|D_i|$. Thus features far away from the decision boundary will have the largest impact. The interpretation is that features with large $|D_i|$ has contributed the most in establishing the decision. The rationale is that the further away from the decision boundary our datum is, the more confident the prediction.

The network ensemble inversion starts with the error function

$$E_I(\boldsymbol{x}) = (y_p - y(\boldsymbol{x}, \boldsymbol{\omega}))^2 \tag{v.7}$$

where y is the network ensemble output, Eqn. (v.1), with all weights $\boldsymbol{\omega}$ being fixed. The target value y_p is the prevalence of the data set used. Minimization of E_I is carried out using gradient descent, starting from the given input data point $\boldsymbol{x}$, and augmented with a dynamical learning rate η, according to Eqn. v.8.

$$\eta_{t+1} = \begin{cases} 1.1\eta_t, & Err_{t+1} < Err_t \\ 0.9\eta_t, & \text{otherwise} \end{cases} \tag{v.8}$$

The above network ensemble inversion may not always result in a point $\boldsymbol{p}$ closest to $\boldsymbol{x}$. Such errors are however ignored in this study.

Iterative input clamping

This is similar to the Euclidean distance method in that we once again find the point $\boldsymbol{p}$ on the decision boundary closest to data point $\boldsymbol{x}$, where the decision boundary is defined as for the Euclidean distance method. The method recursively finds the most important features by ranking the effect they have on the ensemble output. The reference point for the comparison is $\boldsymbol{p}$ with the corresponding ensemble output $y(\boldsymbol{p})$. The boundary point $\boldsymbol{p}$ is moved towards the data point $\boldsymbol{x}$ and the corresponding changes of the ensemble output are used to rank the input variables according to the following scheme:

1. Let M be the number of input variables.

2. Find the point $\boldsymbol{p} = (p_1, p_2, \cdots, p_M)$ on the decision boundary closest to the current data point $\boldsymbol{x}_n$ using network ensemble inversion.

3. For $i = 1, 2, \cdots, M$

```
    Define $\tilde{\boldsymbol{p}}_i$ to be the vector $(p_1, \cdots, x_i, \cdots, p_M)$
    Calculate change $e_i = |f(\boldsymbol{p}) - f(\tilde{\boldsymbol{p}}_i)|$
End
```

4. Find the component j resulting in the largest change, i.e. $e_j \geq e_i \ \forall \ i$.

5. Redefine $\boldsymbol{p}$ to be $\tilde{\boldsymbol{p}}_j$.

6. Repeat step (3)-(5) M times until $\boldsymbol{p}$ has become $\boldsymbol{x}$.

7. The *iterative input clamping* ranklist is given by the order in which the different input variables were clamped when $\boldsymbol{p}$ was moved towards $\boldsymbol{x}$.

The non-linearity of the network ensemble can of course result in ranking of the input variables that is misleading. Still, it turns out to provide a reasonable idea of the effect of the different variables for a given data point $\boldsymbol{x}$.

Hybrid method

The hybrid method, for feature l in patient $\boldsymbol{x}$, was defined by

$$H_l(\boldsymbol{x}) = \begin{cases} |S_l(\boldsymbol{x})|, & \frac{p}{2} < y(\boldsymbol{x}, \boldsymbol{\omega}) < \frac{1+p}{2} \\ D_l(\boldsymbol{x}), & otherwise \end{cases} \qquad \text{(v.9)}$$

where $y(\boldsymbol{x}, \boldsymbol{\omega})$ is the output from the ensemble for patient $\boldsymbol{x}$, and p is the prevalence of non-healthy patients. $S_l(\boldsymbol{x})$ and $D_l(\boldsymbol{x})$ is the input sensitivity and Euclidean distance method respectively. The input variables are ranked using the value $H_l(\boldsymbol{x})$. The reason for using a hybrid of input sensitivity and Euclidean distance is that they function differently in different domains in input space. The input sensitivity is designed for analyzing cases near the decision boundary where small changes may have a strong impact on the output from the ensemble. Conversely, the Euclidean distance was designed to work for cases further away from the decision boundary. However, due to the fact that we remove the derivatives f_i' from $S_l(\boldsymbol{x})$, the input sensitivity can function well even when far away from the boundary separating the two classes.

V.2.3 Data sets

We used three different data sets in this study consisting of two artificial ones, and one real-world data set from the medical domain. All three data sets represents binary classification problems. All variables in the three data sets were normalized to Z-scores using the following formula:

$$z_i = \frac{x_i - \bar{X}}{\sigma_X} \tag{v.10}$$

where x_i is the i:th observation of the random variable X. Meanwhile $\bar{X}$ and σ_X is the sample mean and sample standard deviation, respectively.

Monks data

These data sets [6], representing binary classification problems, are generated from three different sets of rules. The data sets are called Monks 1, 2 and 3, corresponding to the three generating rules. Each data set consists of six categorical variables. The valid range for them is $x_1, x_2, x_4 \in \{1, 2, 3\}$, $x_3, x_6 \in \{1, 2\}$ and $x_5 \in \{1, 2, 3, 4\}$. The data set sizes and the generating rules are given in Table v.1.

Table v.1: The cardinality and prevalence of the three Monks data sets. The indicator function $I_1(x)$ is one if $x = 1$, otherwise zero.

Name	Size	Positives	Rule
Monks 1	556	50%	$(x_1 = x_2) \vee (x_5 = 1)$
Monks 2	601	34%	$\sum_{i=1}^{6} I_1(x_i) = 2$
Monks 3	554	52%	$(x_2 \neq 3) \wedge (x_5 \neq 4) \vee (x_4 = 1) \wedge (x_5 = 3)$

Monks 1 and 3 have rules that makes it easy to determine, at a given instance, which variables that are important. The same is not intuitively true for the Monks 2 rule, where it is clear that any variable can at any time influence the classification. Since, giving feedback to a user on this data is indeed non-trivial we chose not to include it in this study.

Electrocardiogram data

This is a real-world data set where one tries to classify electrocardiograms (ECG) as being healthy or not. The background is as follows: Patients who present at the emergency department with chest pain or other symptoms suspicious of myocardial infarction or unstable angina pectoris (i.e. acute coronary syndrome, ACS) are common and represent a heterogeneous group. Some have a myocardial infarction with a high risk of life-threatening complications whereas others have completely benign disorders which may safely be evaluated on an out-patient basis. A number of methods have been developed to support the physicians in their decision making regarding patients presenting to the emergency department with chest pain, see e.g. [16,17], where ANN is a common classification method. One approach to detect ACS as early as possible at the emergency department is based on using only the 12-lead ECG, as this is usually the first type of examination that is performed. This approach was carried out previously [1,18] and the current ECG data set originates from these studies.

The data set was collected in 1997 and comes from 861 patients attending the Lund University emergency department with a principal complaint of chest pain. The diagnosis was either ACS or non-ACS. In total this data set consisted of 14 measurements from 12 ECG leads leading to a total of 168 variables. This list was reduced by experienced physicians in order to get rid of redundant features and facilitate a more straightforward comparison between the physicians and the algorithms. The final ECG variables selected was QRS duration, QRS peak to peak, Q amplitude, ST amplitude, ST slope, T_+ amplitude and T_- amplitude in leads I, III, V_1 and V_6. In addition to these we also added QRS axis, in total 29 variables. An illustration of an ECG can be seen in Fig. v.2.

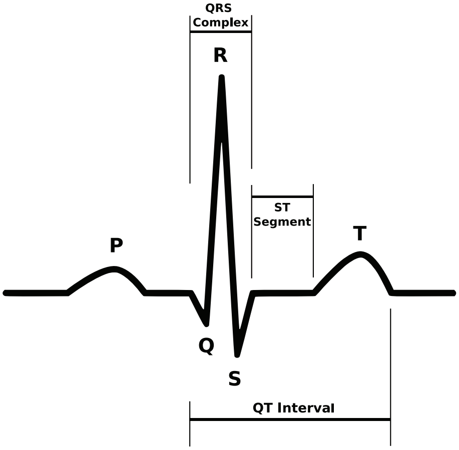

Figure v.2: A illustration of an ECG showing different parts of the curve. All amplitudes are measured from the baseline, except for QRS peak to peak which measures the total height of the QRS complex.

V.2.4 Performance evaluation of explanations

Monks data

In the case of Monks data, we already know the important features for any given instance, since we know the generating rules. This advantage allowed us to discriminate between good and bad explanations. We considered an explanation as good if at least two of the important features were present in the top three ranked features. This is reasonable since, at a given instance, you never need more than two variables to fire one of the Monks rules. We can expect the number of good rules appearing by chance to follow a binomial distribution with a probability $p = \frac{1}{5}$.

Electrocardiogram data

The evaluation of the ECG data is more difficult since there is no real truth here. Instead we used the knowledge of two experts in ECG interpretation as a reference. We randomly selected 50 positive cases out of the possible 344 ACS ECGs, and let two experienced physicians extract the five most important variables for each ECG. The variables selectable by the physicians was described in section v.2.3. From each physician we received a list of the 5 most important variables for 33 of the 50 ECGs. The reason for this reduction was due to the fact that none of the physicians could point out 5 interesting variables for every ECG. Non-ACS cases were not included in this review, since physicians are trained to find signs of ACS and are not comfortable in finding ECG changes that do not indicate ACS. These lists were then compared to the corresponding ones generated by our explanation methods. We looked at both inter and intra overlap between physicians and methods. P-values were calculated from an approximated Normal distribution describing the number of overlaps one might expect on average.

V.3 Results

In order to compare between the five explanational methods, we generated a ranklist for each method and data set. These ranklists were created from test data not previously seen by the ANN ensemble. Thus, all variables were ranked within each data point according to the case-based explanational methods. The AUC for the training and test runs for each of the data sets used in this study is presented in Table v.2. Monks 1 and 3 are easy problems with almost 100% AUC for both test sets, while the ACS prediction is a harder problem. The obtained ROC of 83% is however in line with previous studies on a similar data set [1], even though we used a restricted set of input variables.

Table v.2: The resulting training and test AUC for the three data sets used in this study.

Data set	Training	Test
Monks 1	100%	100%
Monks 3	100%	99%
Electrocardiogram	96%	83%

V.3.1 Monks data

In general the proportion of good explanations for the Monks data, as illustrated in Table v.3 was more than adequate. Input sensitivity performed the best on both Monks 1 and 3 data with 99% and 91% good explanations respectively. The Euclidean distance and the iterative input clamping methods had similar results with an average of 94%

on Monks 1 and 88% on Monks 3. The hybrid method did not improve the result, compared to the two individual methods it was based on. The methods had a significant amount of overlap when looking at the top three ranked features for the Monks data sets. In the Monks 1 data, all the methods had a median overlap of 3. Also, the first and third quantile was 3 as well.

Table v.3: The proportion of good explanations evaluated on the Monks 1 and 3 data. All of these fractions were significant with p-values $< 10^{-100}$.

Method	Monks 1	Monks 3	Average
Input sensitivity	0.99	0.91	0.95
Odds ratio	0.88	0.84	0.86
Euclidean distance	0.93	0.88	0.91
Iterative input clamping	0.95	0.89	0.92
Hybrid	0.93	0.89	0.91

To illustrate the overall importance of the six variables in the Monks 1 and 3 data sets we generated bar plots, showing the distribution of the variables selected at a given rank. Figures v.3 and v.4 shows such bars plots for the input sensitivity method. The other methods had qualitatively similar plots. Looking at Fig. v.3 we see that the variables most frequently ranked 1 and 2 is indeed x_1, x_2 and x_5. Correspondingly, variables x_6, x_4 and x_3 is ranked 5 or 6 most of the time.

For the Monks 3 data set (see Fig. v.4) the input sensitivity method ranked variables x_2, x_4 and x_5 as the top two in the majority of cases. However, the trend for the unimportant variables was not as prominent as for the Monks 1 data. This indicates that a feature present in the generating rule sometimes found itself in the bottom of the rank and unimportant variables sometimes got a high rank.

V.3.2 ECG data

When looking at the intersections for the top five selected features we found that, in general, the median overlap between a method and a physician was typically $1(1,2)$. The numbers in the parenthesis correspond to the first and third quantile. The exceptions were the overlaps between physician 1 and the two methods input sensitivity and Euclidean distance, where the corresponding result was $2(1,2)$. The statistics for the overlap between the physicians was $2(2,3)$. In general the overlap between the physicians was larger than the overlap between a given physician and method. This difference was significant with a p-value < 0.001. The Hybrid method, that consisted of the input sensitivity and the Euclidean distance method, resulted in the same median, first and third quantile as when applying the methods individually.

When observing the intersection between the methods we see that they produce very similar ranks for the top 5 features. The distribution of the overlaps of the different methods are presented in Table v.4. From this it is evident that the distributions

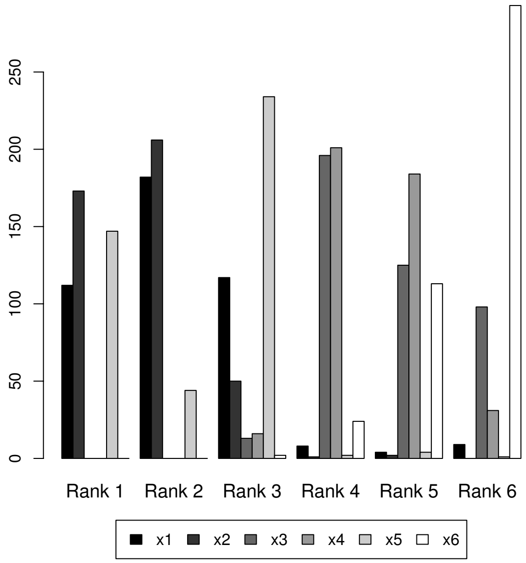

Figure v.3: The rank distribution, as generated by the input sensitivity method, among the six variables from the Monks 1 data set.

Table v.4: The median intersection of the top 5 ranked features between all methods, except the hybrid, on the ECG data. The hybrid method is not presented here since it produces similar results as the input sensitivity method. This table also presents the first and third quantile of the distributions.

Intersection	Median	Q1	Q3
Input sensitivity - Odds ratio	4	3	5
Input sensitivity - Euclidean distance	5	4	5
Input sensitivity - Iterative input clamping	5	4	5
Euclidean distance - Odds ratio	4	3	4
Euclidean distance - Iterative input clamping	5	4	5
Odds ratio - Iterative input clamping	4	4	4

are quite narrow and centered around 4 and 5. Thus, there is a substantial amount of overlap between the methods indicating that little is gained from using more than one for a given case.

Turning our eyes to what features a physician selects versus the ones selected by our input sensitivity method we notice that, for a specific case, the overlap varies to some extent. Figure v.5 and v.6 illustrates a scenario where the method and physician agree

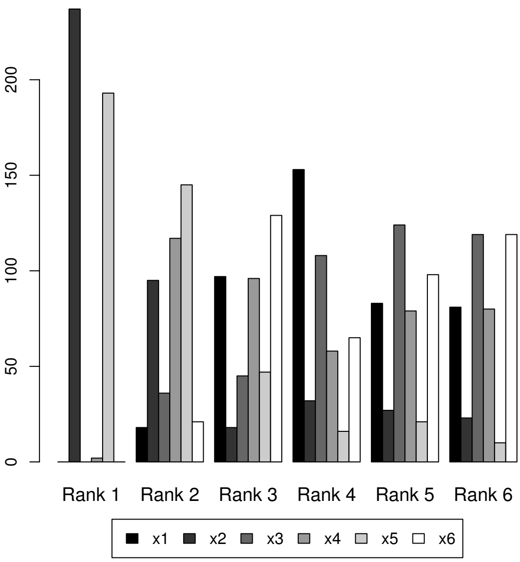

Figure v.4: The rank distribution, as generated by the input sensitivity method, among the six variables from the Monks 3 data set.

and disagree respectively. For the former, with a large overlap (Fig. v.5), the ensemble output was 0.98, indicating a clear ACS case. For the case with no overlap (Fig. v.6) the ensemble output was 0.43, which is near the prevalence cut, indicating a difficult case.

If we stratify the distribution of ensemble output on the number of features, among the top 5, that overlaps between a physician and the input sensitivity method, we see that on average the output gets closer to 0.4 the less overlap we have. This might indicate that the cases with fewer overlaps were indeed the most difficult to classify. The distributions are shown in Figure v.7.

V.4 Discussion

In this work we tried new approaches to explain the results of a complicated algorithm in a simple way to e.g. physicians in a clinical setting. Traditionally this has been tried by extracting variables important to the entire algorithm or by extracting rules describing the operation of the algorithm. In this study we aimed at providing an explanation on a case-by-case level, rather than global operation of the algorithm. Also in this explanational framework we chose not to introduce cutoffs for variables indicating risk. Rather we highlighted the variables that played a significant role in determining the

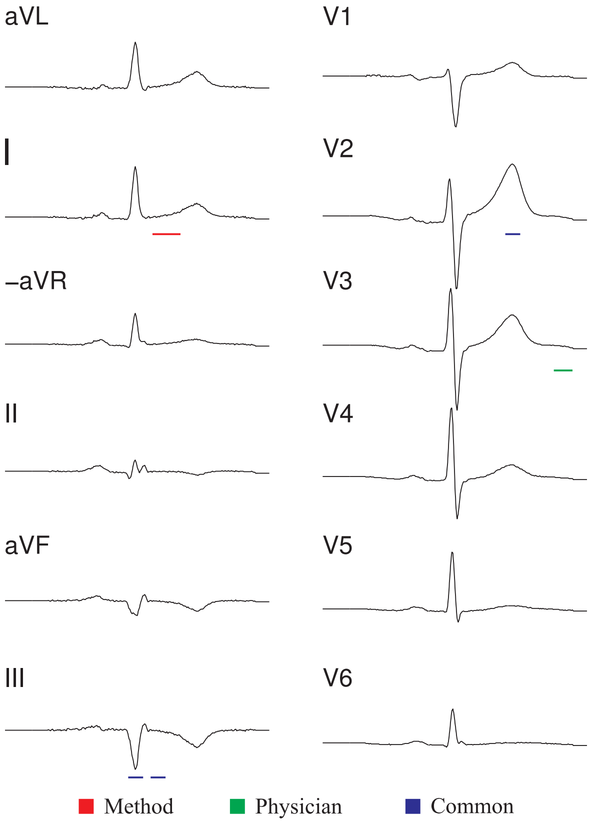

Figure v.5: Example of an ECG where a physician and the input sensitivity method agree on 4 out of 5 of the important features. The important features as decided by the physician and the method are marked by green (T_- amplitude, lead V_2) and red (ST slope, lead I) color respectively. The features that overlap are marked with blue and are Q amplitude and ST amplitude in lead III, T_+ amplitude (lead V_2) and QRS axis (not marked).

outcome. Our main inspiration for this work comes from medical applications, focused on the triage of chest pain patients, where the lack of case-based feed-back is limiting the use of decision support tools.

Although all methods performed rather well on the Monks data sets the results in Table v.3 still favors the input sensitivity method with 95% good explanations on average. The performance of generalized odds ratios was almost 10 percentage points worse. This method however, contains one parameter that needs to be tuned, namely

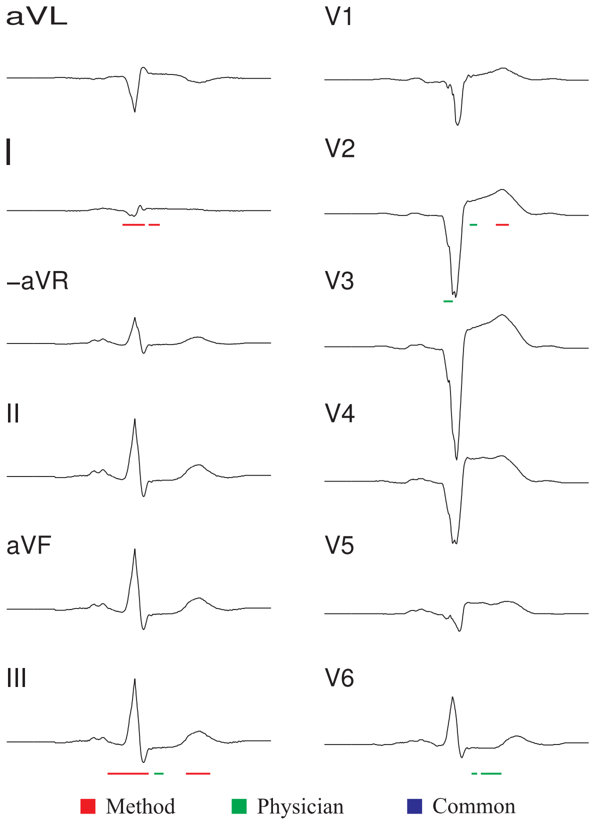

Figure v.6: Example of an ECG where a physician and the input sensitivity method disagree on all 5 of the important features. The important features as decided by the physician are marked by green and are Q amplitude (V_2), ST slope (V_6) and ST amplitude in leads III,V_2 and V_6. The important features from the method are marked by green and are QRS peek-to-peek in leads I and III, ST slope (I) and T_+ in leads III and V_2.

the increase in the variable of interest. In this work we chose to perform a scaled increase for each variable depending on the standard deviation. More precisely we chose the increment to be $0.1\sigma_l$. Changing this increment may produce different results.

The median overlap of a given method and physician was never more than 2. On the other hand the median overlap of the two physicians was 2 as well. This poses a problem since the notion of truth appears difficult to define. Benchmarking against expert physicians may not be the answer to this question since there was no apparent consensus between them either. The reason for this lack of overlap may be due to the

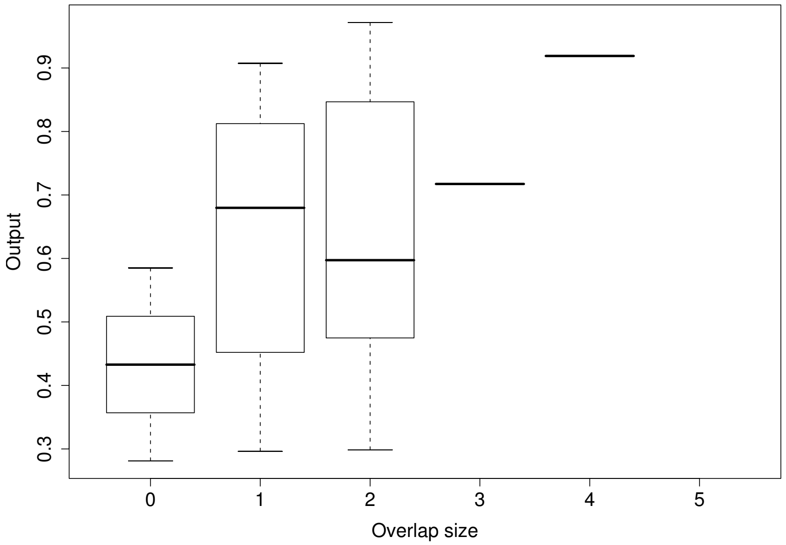

Figure v.7: The output distribution stratified on the number of overlapping features among the top 5 as ranked by a physician and the input sensitivity method.

complexity of predicting ACS from ECG data alone. Furthermore the two physicians may not be trained to interpret ACS ECGs in exactly the same way, due to correlations among the different ECG leads and the corresponding measurements. However, the feedback provided by any explanational method must still make sense to the physicians receiving it in order to be adopted. A comparison of the classification performance between the network ensemble and the physicians would be interesting, given the above result of the overlap between the explanational method and the physicians. Such a comparison is not possible for the limited set of ECG variables used here. However, a comparison using all ECG measurements, on the same data set, can be found in [19]. Their results show that the network ensemble is superior compared to an expert consensus interpretation, with ROC areas of 86% and 78%, respectively.

Though there was no apparent gain from using the hybrid method in our results we still believe that producing a hybrid method is useful since the input sensitivity and Euclidean distance method is designed for intrinsically different domains in input space. Specifically input sensitivity works best close to the decision boundary meanwhile the Euclidean distance method was constructed to work well far away from the boundary. However, since we know that these methods have a significant overlap it is tempting to believe that we need more samples to see the real use of the hybrid approach. This needs to be investigated further.

The observed lack of consensus between experienced physicians when determining important features for a specific case must be investigated further in order to determine how to best validate explanational methods for decision support systems in medicine. For the method based on defining a decision boundary we have not investigated, in depth, the dependence of using different prevalence cuts. Furthermore, this paper was only concerned with the problem of binary classification.

V.5 Conclusion

This work was largely motivated from a medical application point of view where we clearly can identify a need for explanations when using neural network ensembles as decision support tools. This explanation is given on a case-by-case basis, not by providing rules, but rather providing a ranking of importance for the input variables.

We have presented five approaches able to produce case-based feed-back for neural network ensembles. Two methods were based on defining a decision boundary and by network ensemble inversion finding the closest point on the boundary. Ranking lists were then defined by distances in input space (Euclidean distance method) and their effects on the ensemble output value (iterative input clamping). Two other methods (input sensitivity and generalized odds ratio) rank the inputs based in the standard idea of measuring the effect a small input change will have on the ensemble output. We also included a hybrid method consisting of a combination of input sensitivity and Euclidean distance method. When evaluating the methods on three different data sets, with varying difficulty level, we found an advantage using the input sensitivity method, followed by the Euclidean distance approach.

Future work may include classification problems with many classes and even case-based explanation for regression problems. The ranking of input variables based on the input sensitivity method can easily be extended to multiple classes and also work for regression problems. The methods based on defining a decision boundary is less obvious, at least for regression problems.

V References

[1] M. Green, J. Björk, J. Forberg, U. Ekelund, L. Edenbrandt, and M. Ohlsson, *Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room*, Artificial Intelligence in Medicine **38**, 305–318 (2006).

[2] P. J. G. Lisboa, *A review of evidence of health benefit from artificial neural networks in medical intervention*, Neural Networks **15**, 11–39 (2002).

[3] E. Kolman and M. Margaliot, *Are artificial neural networks white boxes?*, IEEE Transactions on Neural Networks **16**, 844–852 (2005).

[4] E. W. Saad and D. C. Wunsch, *Neural network explanation using inversion*, Neural Networks **20**, 78–93 (2007).

[5] T. A. Etchells and P. J. G. Lisboa, *Orthogonal search-based rule extraction (OSRE) for trained neural networks: a practical and efficient approach*, IEEE transactions on neural networks **17**, 374–384 (2006).

[6] S. B. Thrun, J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, et al., *The MONK's problems: A performance comparison of different learning algorithms*, Tech. Report CS-91-197, Carnegie Mellon University, Pittsburgh, PA, 1991.

[7] J. J. Montaño and A. Palmer, *Numeric sensitivity analysis applied to feedforward neural networks*, Neural Computing & Applications **12**, 119–125 (2003).

[8] T. Tchaban, M. J. Taylor, and J. P. Griffin, *Establishing impacts of the inputs in a feedforward neural network*, Neural Computing & Applications **7**, 309–317 (1998).

[9] W. Wang, P. Jones, and D. Partridge, *Assessing the impact of input features in a feedforward neural network*, Neural Computing & Applications **9**, 101–112 (2004).

[10] H. Haraldsson, L. Edenbrandt, and M. Ohlsson, *Detecting acute myocardial infarction in the 12-lead ECG using Hermite expansions and neural networks*, Artificial Intelligence in Medicine **32**, 127–136 (2004).

[11] R. Wall, P. Cunningham, P. Walsh, and S. Byrne, *Explaining the output of ensembles in medical decision support on a case by case basis*, Artificial intelligence in medicine **28**, 191–206 (2003).

[12] R. Caruana, *Case-based explanation for artificial neural nets*, Proceedings of Artificial Neural Networks in Medicine and Biology Conference (Göteborg, Sweden) (H. Malmgren, M. Borga, and L. Niklasson, eds.), 2000, pp. 303–308.

[13] L. Breiman, *Bagging Predictors*, Machine Learning **24**, 123–140 (1996).

[14] R. Wehrens, H. Putter, and L. Buydens, *The bootstrap: A tutorial*, Chemometrics and Intelligent Laboratory Systems **54**, 35–52 (2000).

[15] J. A. Hanley and B. J. McNeil, *The meaning and use of the area under a receiver operating characteristic(ROC) curve*, Radiology **143**, 29–36 (1982).

[16] M. Green, J. Björk, J. Hansen, U. Ekelund, L. Edenbrandt, and M. Ohlsson, *Detection of acute coronary syndromes in chest pain patients using neural network ensembles*, Second International Conference on Computational Intelligence in Medicine and Healthcare (Lisbon, Portugal) (J. M. Fonseca, ed.), IEE/IEEE, June-July 2005, pp. 182–187.

[17] R. L. Kennedy and R. F. Harrison, *Identification of patients with evolving coronary syndromes by using statistical models with data from the time of presentation.*, Heart **92**, 183–189 (2006).

[18] J. Björk, J. L. Forberg, M. Ohlsson, L. Edenbrandt, H. Ohlin, and U. Ekelund, *A simple statistical model for prediction of acute coronary syndrome in chest pain patients in the emergency department.*, BMC medical informatics and decision making **6**, 28 (2006).

[19] J. Forberg, M. Green, J. Björk, M. Ohlsson, L. Edenbrandt, H. Öhlin, et al., *In search of the best method to predict acute coronary syndrome using only the ECG from the emergency department*, (2008), submitted.

VI

Explaining artificial neural network ensembles: A case study with electrocardiograms from chest pain patients

Michael Green[1], Ulf Ekelund[2], Lars Edenbrandt[3,4], Jonas Björk[5], Jakob Lundager Forberg[2] and Mattias Ohlsson[1]

[1]Computational Biology & Biological Physics, Lund University, Sölvegatan 14A, SE-22362 Lund, Sweden

[2]Department of Clinical Sciences, Section for Emergency Medicine, Lund University Hospital, SE-22185 Lund, Sweden

[3]Department of Clinical Physiology, Malmö University Hospital, SE-20502 Malmö, Sweden

[4]Department of Clinical Physiology, Sahlgrenska University Hospital, SE-41345 Gothenburg, Sweden

[5]Competence Center for Clinical Research, Lund University Hospital, SE-22185 Lund, Sweden

LU TP 08-06 (submitted)

Artificial neural networks is one of the most commonly used machine learning algorithms in medical applications. However, they are still not used in practice in the clinics partly due to their lack of explanatory capacity. We compare two case-based explanation methods to two trained physicians on analysis of electrocardiogram (ECG) data from patients with a suspected acute coronary syndrome (ACS). The median overlaps of the top 5 selected features between the two physicians, and a given physician and a method, were initially low. Using a correlation analysis of the features the median overlap increased to values typically in the range 3-4. In conclusion, both our case-based methods generate explanations similar to those of trained expert physicians on the problem of diagnosing ACS from ECG data.

VI.1 Background

Artificial neural networks (ANN) has been gaining interest in the medical community for quite some time now, and has proven useful for many clinical decision problems [1–7]. Still, as of today, there are very few live applications in use at the clinics. Though the reasons for this low usage are numerous [8], one major drawback is the lack of interpretability of the decisions provided by an ANN [7].

Most efforts of making sense out of an ANN decision is based on rule extraction methods where the decision boundary is discretized into segments. There are basically two ways of attacking this problem in neural networks. The first is the *decompositional* [9] approach where the network is scrutinized from within in order to extract useful information about a decision. This is usually done by analyzing the activations of individual nodes in the network as well as the weights leading into them. This methodology was used by [9] where they demonstrated that an ANN is mathematically equivalent to an all permutation fuzzy rule base. Their work provided an explicit way of transforming an ANN into a set of IF THEN rules. Despite being intuitively attractive this approach lead to a large number of rules that had to be reduced.

The second one known as the *pedagogical* [10, 11] approach treats the network as a black box. Here the analysis is based on examining the relationship between what is fed into the network with what is returned as output. In a recent paper by [11] the pedagogical approach was used when developing the orthogonal search based rule extraction (OSRE) method that successfully extracted the exact rules for the Monks [12] data. They also point out that, in the presence of large node output weights, the decompositional approach may fail to accurately describe the logic of the network.

Another way to analyze a neural network is by sensitivity analysis where the main focus has been on extracting global properties. Usually this has been accomplished by analyzing the weights in the network on a pattern by pattern basis. Interestingly enough this has been considered a drawback by several authors [13–15].

From a medical application point of view it is often necessary to provide an explanation underlying a given decision. If the decision support is to function in a stressful clinical setting (e.g. an emergency department) then it is required to provide a fast explanation for each case, easily interpretable by the operator. This case-based feed-back requirement is lacking in most methods for analyzing the operation of a neural network ensemble. We believe this has severely limited the full potential of using neural networks in a clinical decision support system. The idea of using the specific case at hand as the basis for the feed-back algorithm is not new. In [16] a specific method was developed for electrocardiogram curves, where the case-based feed-back was presented as modified curves representing changes towards being more healthy or non-healthy. In [17] rules were extracted and later ranked depending on the prediction of the case. The idea was that more complex rules should be presented when the decision support system classified a patient as healthy. Conversely if a patient were classified as non-healthy, less complex rules were given as feed-back. Another approach to case-based explanation can be found in [18] where the reasoning behind the neural network was presented as showing a set of similar cases.

When providing feedback to a physician in a clinical situation we need to make sure that only the core of the driving forces behind a classification is presented. This means that a rule based approach, where possibly more than 10 rules are presented per case, will be difficult to use in practice. Also many of the rules will be non-specific for a given case since the rules are extracted globally from the data set with the aim of approximating the decision boundary of the ANN. To us this suggests that any case-based feedback should be derived from a single case and not the entire data set. Case-based feed back is indeed dependent on the question one is asking. In a clinical setting we often find the important feed-back to simply be the set of variables, most important for the decision. The two approaches described in this study will both result in a ranked list of important variables and the explanation will simply consist of the topmost important ones, for each case.

In this work a case study was performed where we explored the explanatory power of an ANN ensemble in the context of predicting acute coronary syndromes, in chest pain patients, from electrocardiogram (ECG) data alone. Even though we only investigated this particular medical application, we still believe that the results are transferable to many other medical problems as well.

VI.2 Methods

VI.2.1 Study population

A number of methods have been developed to support the physicians in their decision making regarding patients presenting to the emergency department with chest pain [5, 6]. One approach to detect ACS as early as possible at the emergency department is based on using only the 12-lead ECG, as this is usually the first type of examination that is performed. This approach was carried out in [4, 19] and the current ECG data set originates from these studies.

The data set was collected in 1997 and comes from 861 patients attending the Lund University emergency department with a principal complaint of chest pain. Patients who present at the emergency department with chest pain or other symptoms suspicious of myocardial infarction or unstable angina pectoris (i.e. acute coronary syndromes, ACS) are common and represent a heterogeneous group. Some have a myocardial infarction with a high risk of life-threatening complications whereas others have completely benign disorders which may safely be evaluated on an out-patient basis.

The diagnosis was either ACS or non-ACS. The 12-lead ECGs were recorded by the use of computerized electrocardiographs (Siemens-Elema AB, Solna, Sweden), resulting in 14 measurements from 12 ECG leads leading to a total of 168 variables. This list was reduced by experienced physicians in order to get rid of redundant features and facilitate a more straightforward comparison between the physicians and the algorithms. The final ECG variables selected was QRS peak to peak amplitude, Q duration, Q amplitude, ST amplitude, ST 2/8 amplitude, ST 3/8 amplitude, ST slope, T_+ amplitude and T_- amplitude in all 12 leads. An illustration of an ECG can be seen in Figure vi.1. In addition to these measurements we also added QRS axis and the maximum QRS duration in any lead. In total 110 variables were selected.

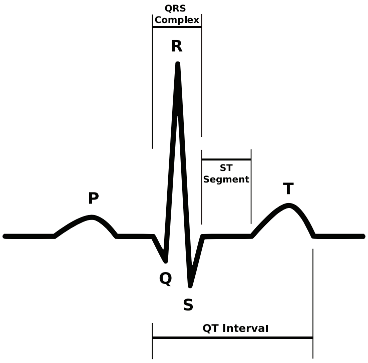

Figure vi.1: A illustration of an ECG showing different parts of the curve. All amplitudes are measured from the baseline, except for QRS peak to peak which measures the total height of the QRS complex.

VI.2.2 Artificial neural network ensembles

The generalization performance of the ANN ensemble was evaluated in a 10 fold cross validation [20,21] loop where the entire data set was split into 10 disjoint parts. Each of these parts served as a test set for an ANN ensemble constructed from the remaining 9 parts. The generalization ability of the ensemble was then evaluated as the median ROC area over these 10 data sets. The ROC area can be interpreted as the probability of a randomly chosen sick patient having a larger predicted risk than a patient chosen at random from the control group [22].

The ensemble [23,24] of networks was built by resampling the data using a bagging [25] procedure, that allowed us to create more diverse ensemble members. We chose an ensemble size of 25 since it has been shown to be enough in numerical studies [26]. Since we were training our ensemble for classification purposes we used a cross-entropy error function [27] with an added weight elimination term that can improve its ability to generalize. The complete error function is shown below

$$E = \sum_{n=1}^{N} \left(\ln y_n^{t_n} + \ln(1 - y_n)^{1-t_n} \right) + \alpha \sum_i \frac{\omega_i^2}{\omega_0^2 + \omega_i^2}$$

where the t_n, y_n, ω_i's, and α is the target, network output, parameters and weight elimination constant respectively. The parameter α effectively controls how much regularization we want to use and it was tuned with respect to the ensemble and not to the individual networks. All the individual networks had a hidden layer with 15 nodes, which in our opinion is rather liberal, since the regularization framework should prevent the ensemble from overfitting the data.

All the models were carefully trained in an internal cross validation loop to make sure that no information leak occurred. In other words, every optimization step was carried out on training data alone.

VI.2.3 Explanatory models

Our view of an explanation is basically highlighting the variables most significant to a given decision. Though this may seem controversial when compared to the traditional way of extracting risk factors from a data set, we consider this approach to be valid. In effect what we are doing is extracting risk factors for a given patient rather than a given data set. The two methods described in this section works as follows:

1. generate a decision for a given patient;
2. rank all input variables according to some measure;
3. select the top five most important variables based on their rank and present them to the physician.

Thus, for each patient we get an individual list of the five variables most important to the decision as given by the network ensemble.

Input sensitivity analysis

This approach is basically a modified partial derivative of the ensemble output with respect to a given input variable. It measures how sensitive the output of the ensemble is to a small perturbation of that particular input variable. This method was mainly developed for use with patients that the network ensemble predicted as uncertain, i.e. patients with predicted risks near the prevalence of the disease in the data. However, the method also works well on patients receiving more certain predictions.

We modify the partial derivative in order to avoid saturation effects that could potentially prevent us from finding important features. An example of this would be when the output of the ensemble is close to either 1 or 0. The problem arises from the sigmoid activation function σ in the output node, since $\frac{\partial\sigma(x)}{\partial x} = \sigma(x)(1 - \sigma(x))$. Thus confident predictions, whose output is near 1 or 0, will never be considered as having a large impact on the ensemble output. We avoid this by defining an input sensitivity function

$$S_l(\boldsymbol{x}) = \left| \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \omega_{ij} \cdot g'_{ij} \cdot \tilde{\omega}_{ijl} \right|$$

which is just the magnitude of the partial derivative of the ensemble output with respect to a variable x_l, where the derivative of the output nodes, from the individual networks, has been removed. The first sum runs over all ensemble members and the second over the hidden nodes in each network. Also ω_{ij} is the weight connecting ensemble members i's output node to its hidden node j, and g'_{ij} is just the partial derivative of the activation function g in that hidden node. Similarly $\tilde{\omega}_{ijl}$ is the weight connecting hidden node j to input l in network i.

$S_l(\boldsymbol{x})$ is used to rank the importance of each variable. The entire procedure is given in Algorithm 1.

Algorithm 1 Input sensitivity

```
input  data $\boldsymbol{x}$, ensemble $net$, input size $L$
  for $l = 1$ to $L$ do
    Calculate $S_l = S_l(\boldsymbol{x},\text{net})$
  end for
  Calculate $R_l = Rank(S_l)\ \forall\ l \in [1..L]$
output  $R = \{l : R_l \leq 5\}$
```

Euclidean distance

The neural network ensemble produces a decision boundary that separates the sick from the healthy in the input space built from the 110 ECG variables. Knowing where this boundary is located is useful since we can then measure the distance, in all 110 variables, to it from a given patient. In order to utilize this distance we need to know where the boundary is located in input space. We find the closest[1] point p on the decision boundary, corresponding to a network output equal to the prevalence of ACS in our material, by network inversion [10]. The inversion proceeds by gradient descent with an added adaptive learning rate. The whole procedure is presented in Algorithm 2.

The idea behind this approach is that the further away the value of a variable is from the decision boundary the more impact it had on the decision. The reason for this assumption lies within the fact that for a variable far away from the decision boundary one would have to make substantial changes to it for it to affect the decision. Thus, the confidence for the decision in this variable is high.

[1] This is only approximately true since a line minimization would be required in order to find it.

Algorithm 2 Euclidean distance

```
input  data $\boldsymbol{x}$, ensemble $net$, input size $L$, prevalence $y$
  Calculate $Err_0 = E(\boldsymbol{x}) = (y - net(\boldsymbol{x}, \boldsymbol{\omega}))^2$
  Set $\boldsymbol{p} = \boldsymbol{x}$ and $\eta_0 = 0.2$
  repeat
    $\boldsymbol{p} = \boldsymbol{p} - \eta_t \frac{\partial E(\boldsymbol{p})}{\partial \boldsymbol{x}}$
    Calculate $Err_{t+1} = E(\boldsymbol{p})$
    if $Err_{t+1} < Err_t$ then
      $\eta_{t+1} = 1.1\eta_t$
    else
      $\eta_{t+1} = 0.9\eta_t$
    end if
  until $Err_t < 10^{-7}$
  Calculate $\boldsymbol{d} = \boldsymbol{x} - \boldsymbol{p}$
  Calculate $R_l = Rank(|d_l|)\ \forall\ l \in [1..L]$
output  $R = \{l : R_l \leq 5\}$
```

VI.2.4 Comparison with physicians

To evaluate the ranked list of features provided by the above methods we asked two physicians to select the most important features for each ECG in a group of patients. Only patients diagnosed with ACS was evaluated during this comparison between the physicians and our methods, since physicians in general have difficulties identifying specific factors indicating health. In summary we handed out 344 ECGs from patients with ACS and asked them to select the top five most important features from the 110 available ones. No priority was given among the five features, i.e. they were all considered as equally important.

Any two feature lists, coming from either a method or from a physician, are then compared to each other by performing the intersection. This is then carried out for each patient, which leaves us with a distribution of intersections for any comparison between two feature lists.

VI.3 Results and discussion

VI.3.1 Performance of the ANN ensemble

The average training and test ROC area (± SD) for the neural network ensemble, over the 10 fold cross validation, was 98.7 (±0.12) and 83.4 (±0.33) respectively. Although the numbers might suggest overfitting, we found no advantage of adding more regularization since the average test ROC area did not increase. This effect can be explained by our use of ensembles, where each MLP in the ensemble might be overfitted. However, since they will be overfitted on different parts of the data set, we get a well performing classification machine when combining their individual predictions. This of course depends on the weighting scheme used for combining the individual predictions.

VI.3.2 Features selected

A list of the features used from the electrocardiograms and the leads in which these were found to be important by the methods and the physicians is shown in Figure vi.2, except QRS-axis and maximum Q_dur which are lead independent. All features deemed as important in at least one patient over the entire dataset was included.

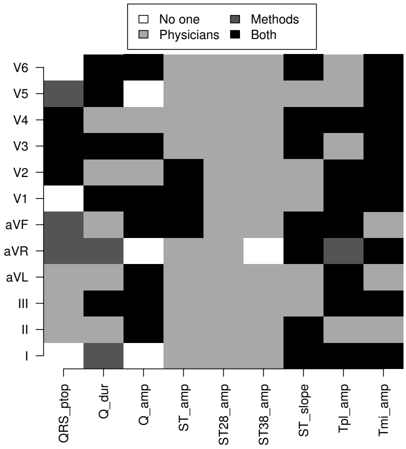

Figure vi.2: The set of features that was considered important in one or more patients over the entire ECG data set. The x-axis shows the measurements we extracted from every ECG. The y-axis represents the 12 different leads. A feature is thus a measurement in a lead. Each feature is color coded depending on which evaluator considered it important.

The figure illustrates an important distinction between the physicians and the methods, namely that the physicians in general chose from a much larger subset of features than the methods did. In effect, the physicians chose from a total of 97 features. The corresponding number for the methods was 47. So it seems as though the methods are more selective when it comes to the features it chooses to present. This reduction in the number of features used by the methods most certainly arises from the high correlation between some of the features (see next section). A fact that will be picked up by the network regularization during the training of the network ensemble. This can to some extent explain why the methods did not find the amplitudes ST 2/8 and ST 3/8 to be important. On the other hand the methods used features from lead aVR somewhat more frequently compared to the physicians. This can be explained by the fact that traditional criteria for detecting ACS almost never use aVR, hence the relatively low frequency among the physicians. Both the methods and the physicians often used T_{-} amplitudes as an explanation for ACS and this is not surprising since negative T-waves is a classical sign of ACS.

In Table VI.1 we looked more closely into the distribution of the number of selected features within a given comparison between two evaluators. In this setting we denote an evaluator to be a given physician or method. The table reveals the number of features i) not chosen by either of the evaluators, ii) chosen by the right evaluator but not the left, iii) chosen by the left evaluator but not the right, and iv) chosen by both evaluators. As earlier stated physicians, in general, considered a larger set of features as important than the methods did. However, looking at the consensus number of important features, within the physicians and the methods we found that the numbers were 59 and 44 respectively. Comparing these numbers to the ones in the previous paragraph it is evident that the larger fraction of features considered as important by the physicians was mainly an effect of them disagreeing. The disagreement between the methods was significantly lower (See Table VI.1).

Table VI.1: Description of the distribution of the selected features for every pair of evaluators. The encoding in the column names refer to the presence (+) and absence (-) of selected features. The sign to the left (right) in each column refers to the first (second) evaluator in the pair. Thus the encoding '- +' refers to the number of features selected by the right evaluator but not by the left one.

Evaluators	- -	- +	+ -	+ +
Phys. 1 - Phys. 2	13	34	4	59
Phys. 1 - Alg. 1	27	20	37	26
Phys. 1 - Alg. 2	27	20	38	25
Phys. 2 - Alg. 1	10	7	54	39
Phys. 2 - Alg. 2	12	5	53	40
Alg. 1 - Alg. 2	63	1	2	44

VI.3.3 Analyzing the overlap

To answer the question of how similar the explanations given by the physicians and the methods are, we compared the list of important features that each of them selected for each patient. We made every possible pairwise comparison between the two physicians and methods. The relative frequencies of the overlaps between two evaluators can be seen in Figure VI.3 and Table VI.2 quantifies the overlaps by listing median, first and third quantile values. We can conclude that the physicians and the methods feature lists do not overlap to a large extent, in fact the median overlap is 0 for any comparison between a physician and a method. To our surprise the overlap between the two physicians was also low, indicating a degree of redundancy when selecting important features. The overlap between the two explanation methods was however large, as seen in Figure VI.3 (left image), with a median of 5 out of 5 possible.

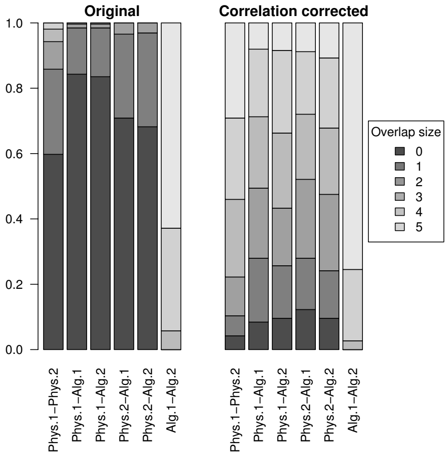

Figure VI.3: Illustration of the relative frequencies of the overlaps for each pair of evaluators with (right image) and without (left image) correlation correction.

There was an overall low degree of agreement of the features selected by the physicians and those highlighted by the methods. This low overlap can be explained by the high degree of correlation among the measurements, which is partly an effect of the fact that any two limb leads (I-III,aVF,aVR,aVL) can be used to derive the other four limb leads when using the raw ECG lead recording. This suggests grouping measurements based on a correlation analysis. When searching for features with a high degree of correlation, defined as a Pearson correlation coefficient larger than 0.5, the feature list was reduced down to a smaller effective set of features. This vastly improved the agreement, in both comparisons between physicians and comparison between a given physician and method (see right image in Figure VI.3). The median overlap between the physicians increased to 4 and almost all comparisons between a physician and a method obtained an overlap of 3. However, after the correlation analysis, the median overlap between a given physician and method is still *significantly* lower than that of the two physicians. There may be several reasons to why this happens. For instance, we know that the neural network ensemble is superior to the physicians when it comes to predicting ACS from ECG data alone [28, 29]. Thus the networks may very well have found a pattern that is typically hidden from human ECG readers. This suggests that there may be a biological interpretation of the ECGs not yet discovered by experienced physicians.

Table VI.2: The median, first and third quantile overlap of the selected features for every pair of evaluators. Values before and after correlation analysis are shown in the upper and lower part, respectively.

Evaluators	Median	Q1	Q3
Phys. 1 - Phys. 2	0	0	1
Phys. 1 - Alg. 1	0	0	0
Phys. 1 - Alg. 2	0	0	0
Phys. 2 - Alg. 1	0	0	1
Phys. 2 - Alg. 2	0	0	1
Alg. 1 - Alg. 2	5	4	5
After correlation analysis			
Phys. 1 - Phys. 2	4	3	5
Phys. 1 - Alg. 1	3	1	4
Phys. 1 - Alg. 2	3	1	4
Phys. 2 - Alg. 1	2	1	4
Phys. 2 - Alg. 2	3	2	4
Alg. 1 - Alg. 2	5	5	5

VI.4 Conclusions

In this work we investigated two methods of explaining the predictions of an artificial neural network ensemble, case by case, for 344 ECGs taken from patients entering the emergency department at Lund University Hospital with a principal complaint of chest pain suspicious of ACS. We compared the feedback given by these methods to two experienced physicians and found that they produced similar explanations.

One of the main strengths of the network ensemble is that it will be consistent in its predictions between different days. This means that if two patients, with the exact same medical condition, walks in to the emergency department on two separate occasions they will get the same diagnosis. The same thing cannot be said about physicians since they may vary in their predictive abilities from day to day [30] depending on a number of factors, e.g. fatigue, stress, illness or lack of motivation. Because most emergency departments are hectic working places, none of these factors is uncommon.

An ensemble of artificial neural networks is a powerful classification tool for medical applications [7]. Despite this promising ability ANN ensembles is not currently used in the clinics, since its reasoning is often complex and consequently difficult to explain to a physician. We believe that case-based feed back is the best way to address this problem, and even though we only considered ECGs from chest pain patients, we believe that the methods presented in this paper are transferable to other medical applications as well.

VI.5 Acknowledgments

This work has been supported by the Swedish Knowledge Foundation through the Industrial PhD program in Medical Bioinformatics at the Strategy and Development Office (SDO) at the Karolinska Institute.

VI References

[1] R. Harrison and R. Kennedy, *Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation*, Annals of Emergency Medicine **46**, 431–439 (2005).

[2] L. Goldman, E. F. Cook, P. A. Johnson, D. A. Brand, G. W. Rouan, and T. H. Lee, *Prediction of the need for intensive care in patients who come to the emergency departments with acute chest pain.*, New England Journal of Medicine **334**, 1498–1504 (1996).

[3] W. Baxt, F. Shofer, F. Sites, and J. Hollander, *A neural computational aid to the diagnosis of acute myocardial infarction*, Annals of Emergency Medicine **34**, 366–373 (2002).

[4] M. Green, J. Björk, J. Forberg, U. Ekelund, L. Edenbrandt, and M. Ohlsson, *Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room*, Artificial Intelligence in Medicine **38**, 305–318 (2006).

[5] M. Green, J. Björk, J. Hansen, U. Ekelund, L. Edenbrandt, and M. Ohlsson, *Detection of acute coronary syndromes in chest pain patients using neural network ensembles*, Second International Conference on Computational Intelligence in Medicine and Healthcare (Lisbon, Portugal) (J. M. Fonseca, ed.), IEE/IEEE, June-July 2005, pp. 182–187.

[6] R. L. Kennedy and R. F. Harrison, *Identification of patients with evolving coronary syndromes by using statistical models with data from the time of presentation.*, Heart **92**, 183–189 (2006).

[7] P. J. G. Lisboa, *A review of evidence of health benefit from artificial neural networks in medical intervention*, Neural Networks **15**, 11–39 (2002).

[8] D. W. Bates, G. J. Kuperman, S. Wang, T. Gandhi, A. Kittler, L. Volk, et al., *Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality.*, Journal of the American Medical Informatics Association **10**, 523–530 (2003).

[9] E. Kolman and M. Margaliot, *Are artificial neural networks white boxes?*, IEEE Transactions on Neural Networks **16**, 844–852 (2005).

[10] E. W. Saad and D. C. Wunsch, *Neural network explanation using inversion*, Neural Networks **20**, 78–93 (2007).

[11] T. A. Etchells and P. J. G. Lisboa, *Orthogonal search-based rule extraction (OSRE) for trained neural networks: a practical and efficient approach*, IEEE transactions on neural networks **17**, 374–384 (2006).

[12] S. B. Thrun, J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, et al., *The MONK's problems: A performance comparison of different learning algorithms*, Tech. Report CS-91-197, Carnegie Mellon University, Pittsburgh, PA, 1991.

[13] J. J. Montaño and A. Palmer, *Numeric sensitivity analysis applied to feedforward neural networks*, Neural Computing & Applications **12**, 119–125 (2003).

[14] T. Tchaban, M. J. Taylor, and J. P. Griffin, *Establishing impacts of the inputs in a feedforward neural network*, Neural Computing & Applications **7**, 309–317 (1998).

[15] W. Wang, P. Jones, and D. Partridge, *Assessing the impact of input features in a feedforward neural network*, Neural Computing & Applications **9**, 101–112 (2004).

[16] H. Haraldsson, L. Edenbrandt, and M. Ohlsson, *Detecting acute myocardial infarction in the 12-lead ECG using Hermite expansions and neural networks*, Artificial Intelligence in Medicine **32**, 127–136 (2004).

[17] R. Wall, P. Cunningham, P. Walsh, and S. Byrne, *Explaining the output of ensembles in medical decision support on a case by case basis*, Artificial intelligence in medicine **28**, 191–206 (2003).

[18] R. Caruana, *Case-based explanation for artificial neural nets*, Proceedings of Artificial Neural Networks in Medicine and Biology Conference (Göteborg, Sweden) (H. Malmgren, M. Borga, and L. Niklasson, eds.), 2000, pp. 303–308.

[19] J. Björk, J. L. Forberg, M. Ohlsson, L. Edenbrandt, H. Ohlin, and U. Ekelund, *A simple statistical model for prediction of acute coronary syndrome in chest pain patients in the emergency department.*, BMC medical informatics and decision making **6**, 28 (2006).

[20] K. Baumann, *Cross-validation as the objective function for variable-selection techniques*, Trends in Analytical Chemistry **22**, 395–406 (2003).

[21] R. Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann, 1995, pp. 1137–1145.

[22] J. A. Hanley and B. J. McNeil, *The meaning and use of the area under a receiver operating characteristic(ROC) curve*, Radiology **143**, 29–36 (1982).

[23] A. Krogh and J. Vedelsby, *Neural network ensembles, cross validation, and active learning*, Advances in Neural Information Processing Systems (San Mateo, CA) (G. Tesauro, D. Touretzky, and T. Leen, eds.), vol. 2, Morgan Kaufman, 1995, pp. 650–659.

[24] T. G. Dietterich, *Ensemble methods in machine learning*, Lecture Notes in Computer Science **1857**, 1–15 (2000).

[25] L. Breiman, *Bagging Predictors*, Machine Learning **24**, 123–140 (1996).

[26] D. Opitz and R. Maclin, *Popular ensemble methods: An empirical study*, Journal of Artificial Intelligence Research **11**, 169–198 (1999).

[27] P. Y. Simard, D. Steinkraus, and J. Platt, *Best practice for convolutional neural networks applied to visual document analysis*, International Conference on Document Analysis and Recogntion (ICDAR) (Los Alamitos), IEEE Computer Society, 2003, pp. 958–962.

[28] S.-E. Olsson, M. Ohlsson, H. Ohlin, S. Dzaferagic, M.-L. Nilsson, P. Sandkull, et al., *Decision support for the initial triage of patients with acute coronary syndromes.*, Clinical physiology and functional imaging **26**, 151–156 (2006).

[29] J. Forberg, M. Green, J. Björk, M. Ohlsson, L. Edenbrandt, H. Öhlin, et al., *In search of the best method to predict acute coronary syndrome using only the ECG from the emergency department*, (2008), submitted.

[30] J. E. Wennberg, B. A. Barnes, and M. Zubkoff, *Professional uncertainty and the problem of supplier-induced demand*, Social Science & Medicine **16**, 811–824 (1982).