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

Michael Green MICHAEL@THEP.LU.SE
Computational Biology & Biological Physics, Department of Theoretical Physics, Lund University, Lund, Sweden

Ulf Ekelund ULF.EKELUND@MED.LU.SE
Department of Clinical Sciences, Section for Emergency Medicine, Lund University Hospital, Lund, Sweden

Lars Edenbrandt LARS.EDENBRANDT@MED.LU.SE
Department of Clinical Physiology, Malmö University Hospital, Malmö, Sweden

Jonas Björk JONAS.BJORK@SKANE.SE
Competence Center for Clinical Research, Lund University Hospital, Lund, Sweden

Jakob Lundager Forberg JAKOB@LUNDAGER.EU
Department of Clinical Sciences, Section for Emergency Medicine, Lund University Hospital, Lund, Sweden

Mattias Ohlsson MATTIAS@THEP.LU.SE
Computational Biology & Biological Physics, Department of Theoretical Physics, Lund University, Lund, Sweden

Abstract

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 2-3. In conclusion, both our case-based methods generate explanations somewhat similar to those of trained expert physicians on the problem of diagnosing ACS from ECG data.

Appearing in the Proceedings of the ICML/UAI/COLT 2008 Workshop on Machine Learning for Health-Care Applications, Helsinki, Finland, 2008.

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 (Harrison & Kennedy, 2005; Goldman et al., 1996; Baxt et al., 2002; Green et al., 2006; Green et al., 2005; Kennedy & Harrison, 2006; Lisboa, 2002). Still, as of today, there are very few live applications in use at the clinics. Though the reasons for this low usage are numerous (Bates et al., 2003), one major drawback is the lack of interpretability of the decisions provided by an ANN (Lisboa, 2002).

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* (Kolman & Margaliot, 2005) 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 & Margaliot, 2005) 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 be-

ing intuitively attractive this approach lead to a large number of rules that had to be reduced.

The second one known as the *pedagogical* (Saad & Wunsch, 2007; Etchells & Lisboa, 2006) 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 (Etchells & Lisboa, 2006) the pedagogical approach was used when developing the orthogonal search based rule extraction (OSRE) method that successfully extracted the exact rules for the Monks (Thrun et al., 1991) 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 (Montaño & Palmer, 2003; Tchaban et al., 1998; Wang et al., 2004).

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 (Haraldsson et al., 2004) 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 (Wall et al., 2003) 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 (Caruana, 2000) 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.

2. Methods

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 (Green et al., 2005; Kennedy & Harrison, 2006). 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 (Green et al., 2006; Björk et al., 2006) 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 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.

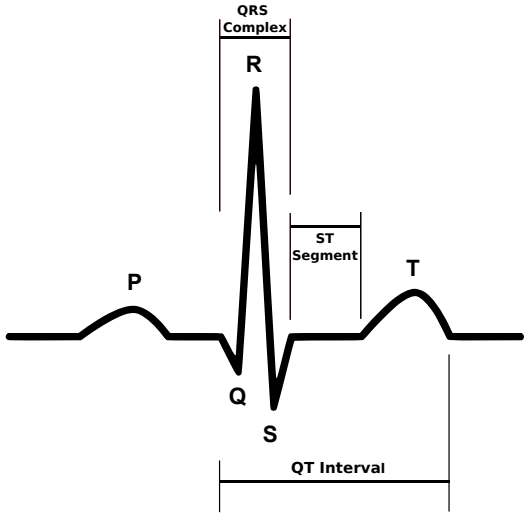

Figure 1. A illustration of an ECG showing different parts of the curve. All amplitudes are measured from the baseline, except for QRS peak to peak which measures the total height of the QRS complex.

2.2. Artificial neural network ensembles

The generalization performance of the ANN ensemble was evaluated in a 10 fold cross validation (Baumann, 2003; Kohavi, 1995) 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 (Hanley & McNeil, 1982).

The ensemble (Krogh & Vedelsby, 1995; Dietterich, 2000) of networks was built by resampling the data using a bagging (Breiman, 1996) 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 (Opitz & Maclin, 1999). Since we were training our ensemble for classification purposes we used a cross-entropy error function (Simard et al., 2003) 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.

2.3. Explanatory models

We decided, together with experienced ECG readers, that a good explanation model for the ECG prediciton is simply to highlight 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 and where the risk factors are standard measurements easy to interpret. The two methods described in this section work 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.

2.3.1. 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

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

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.

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

Algorithm 1 Input sensitivity

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

2.3.2. 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 (Saad & Wunsch, 2007). 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.

Algorithm 2 Euclidean distance

input data $\mathbf{x}$, ensemble *net*, input size L, target y
  Calculate $Err_0 = E(\mathbf{x}) = (y - net(\mathbf{x}, \omega))^2$
  Set $\mathbf{p} = \mathbf{x}$ and $\eta_0 = 0.2$
  repeat
    $\mathbf{p} = \mathbf{p} - \eta_t \frac{\partial E(\mathbf{p})}{\partial \mathbf{x}}$
    Calculate $Err_{t+1} = E(\mathbf{p})$
    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 $\mathbf{d} = \mathbf{x} - \mathbf{p}$
  Calculate $R_l = Rank(|d_l|)\ \forall\ l \in\ [1..L]$
output $R = \{l : R_l \leq 5\}$

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

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

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.

3. Results and discussion

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.

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 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.

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.

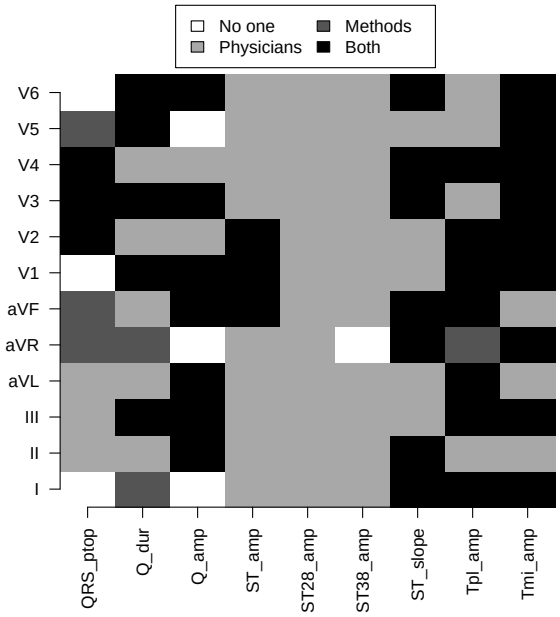

Figure 2. The set of features that was considered important in one or more patients over the entire ECG data set. The x-axis shows the measurements we extracted from every ECG. The y-axis represents the 12 different leads. A feature is thus a measurement in a lead. Each feature is color coded depending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

In Table 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 1).

Table 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

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 3 and Table 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 3 (left image), with a median of 5 out of 5 possible.

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. Typically 25 features remained after the reduction. This vastly improved the agreement,

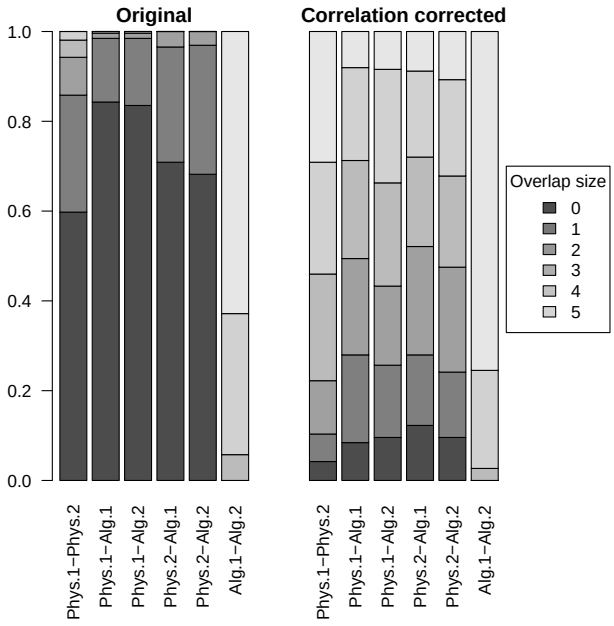

Figure 3. Illustration of the relative frequencies of the overlaps for each pair of evaluators with (right image) and without (left image) correlation correction.

Table 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

in both comparisons between physicians and comparison between a given physician and method (see right image in Figure 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 (Olsson et al., 2006; Forberg et al., 2008). 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.

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 somewhat different explanations, even after a correlation analysis. One interpretation of this result is that the network ensemble finds important information in the ECG that is typically hidden from the human experts.

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 (Wennberg et al., 1982) 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 (Lisboa, 2002). 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.

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.

References

Bates, D. W., Kuperman, G. J., Wang, S., Gandhi, T., Kittler, A., Volk, L., Spurr, C., Khorasani, R., Tanasijevic, M., & Middleton, B. (2003). 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.

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

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

Björk, J., Forberg, J. L., Ohlsson, M., Edenbrandt, L., Öhlin, H., & Ekelund, U. (2006). 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.

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

Caruana, R. (2000). Case-based explanation for artificial neural nets. *Proceedings of Artificial Neural Networks in Medicine and Biology Conference* (pp. 303–308). Göteborg, Sweden.

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

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

Forberg, J., Green, M., Björk, J., Ohlsson, M., Edenbrandt, L., Öhlin, H., & Ekelund, U. (2008). In search of the best method to predict acute coronary syndrome using only the ecg from the emergency department. submitted.

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

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

Green, M., Björk, J., Hansen, J., Ekelund, U., Edenbrandt, L., & Ohlsson, M. (2005). Detection of acute coronary syndromes in chest pain patients using neural network ensembles. *Proceedings of the 2nd International Conference on Computational Intelligence in Medicine and Healthcare* (pp. 182–187). Lisbon, Portugal.

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

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

Harrison, R., & Kennedy, R. (2005). 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.

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

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

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

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems* (pp. 650–659). San Mateo, CA: Morgan Kaufman.

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

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

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

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

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

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

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

Thrun, S. B., Bala, J., Bloedorn, E., Bratko, I., Cestnik, B., Cheng, J., Jong, K. D., Dzeroski, S., Fahlman, S. E., Fisher, D., Hamann, R., Kaufman, K., Keller, S., Kononenko, I., Kreuziger, J., Mitchell, T., Pachowicz, P., Reich, Y., Vafaie, H., Welde, W. V. D., Wenzel, W., & Wnek, J. (1991). *The MONK's problems: A performance comparison of different learning algorithms* (Technical Report CS-91-197). Carnegie Mellon University, Pittsburgh, PA.

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

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

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