Deterministic risk analysis smart health model for heart ailments using neural networks

Kartik Bhanot* and Reshub Kr. Nigam

Information and Communication Technology Department,
Manipal Institute of Technology,
Manipal University,
Karnataka-576104, India
Email: kartik.bhanot1994@gmail.com
Email: reshubnigam4@gmail.com
*Corresponding author

Abstract: Cardiovascular diseases are the leading cause of death all across the globe, much more than all forms of cancer combined. In order to study these heart ailments and stress levels of patients, electrocardiogram (ECG) data is used. The challenge is to develop a risk analysis model that can determine the risk or the possibility of a heart attack based on the current state of ECG data. In the current paper, authors have developed a deterministic risk analysis to determine the level of risk that a person may have for a heart attack with an average accuracy of 92.57%. Neural networks have been used extensively for training the developed model for analysis purposes. The data has been taken from (Massachusetts Institute of Technology – Boston's Beth Israel Hospital) MIT-BIH Long-term ECG Database and MIT-BIH Arrhythmia Database using Rapid Miner as the platform. The database is divided into equitable data from long-term ECG as well as from arrhythmia patients and labels is assigned to them so as to maintain the legitimacy of the whole dataset.

Biographical notes: Kartik Bhanot is a third year student of BTech in Computer and Communication Technology in MIT, Manipal. He has keen interest in the field of e-healthcare and smart sensors and has also published papers in this field.

Reshub Kr. Nigam is a third year student of BTech in Computer and Communication Technology in MIT, Manipal. He has keen interest in the field of data analytics.

1   Introduction

The global healthcare scenario is fast changing. To cater to the demands of a matured and ageing population, the existing healthcare systems should be modified to adapt to the recent environment. Diagnosis of cardiovascular disorders is very relevant for their impact on patient's life and on overall social paradigm. There has been a gradual transition by the patients and well as the practitioners to move towards e-healthcare as it provides flexibility as well as alternative opinion on their health condition. It serves as a healthcare delivery platform which is supported by the data taken from the patient and further analysis of this data is done by data mining tools like RapidMiner. The electrocardiogram (ECG) is an easy and inexpensive diagnostic test that can be used for routine assessment of patients with heart ailments as it non-invasive, simple and low-cost.

Early assessment of heart ailments risk through ECG analysis is of utmost importance as cardiovascular disease is the leading cause of death across the world. Pre-monitoring and pre-diagnostic of these ECG signals may reduce the rate of heart attacks and strokes significantly. In particular, early detection of risk of arrhythmia will make it possible for the patient to choose appropriate anti-arrhythmic drugs. The features which are extracted from ECG include the amplitudes variation of the waves and peak to peak intervals between the different QRS complex as suggested by Paolettiet al. (2006) and Eftehkhar et al. (2005).

In this paper, a deterministic risk analysis model to evaluate the risk of heart ailments in any ECG data by using neural network technique has been developed by the authors to train the model. The main goal of this paper is to efficiently determine the risk of heart ailment by feeding a patients ECG data to the developed model and categorically state weather that patient in low risk or high risk zone for heart ailments in the future. The problem statement has been formulated keeping in mind the intricacies of the ECG wave taking 'ECG1' and 'ECG2' as parameters and 'riskanalysis' as the label. This model is subjected to training by neural network by filtering out the unlabelled data from the training dataset referred from Rabi et al. (2015).

The rest of the paper is structured as follows. Section 2 shows the literature review about other related work in ECG analysis with neural networks in healthcare domain. Section 3 describes the case study of ECG waves. Section 4 of the paper describes the development of smart health risk analysis model. Section 5 defines the experimental setup. Finally, Section 6 comprises of the conclusion and an insight into the future work.

2   Literature review

In this section, the authors have tried to showcase the recent work done in the area of ECG analysis in combination with soft computing techniques like neural network, honey bee algorithm and some others. Brief research review which comprises of the titles, approaches and the algorithm used, different ECG analysis, platform used and results derived from them given in Table 1.

Table 1 Existing approaches to ECG analysis using soft computing

Title of the paper	Algorithm used	Methodology	Outcome	Reference
Classification of ECG arrhythmia using multi-resolution analysis and neural networks	Neural networks	Wavelet analysis	Basic classification of ECG data	Prasad et al. (2003)
ECG analysis using neural network (2012)	Neural networks	Signal processing	Feature extraction	Karande-Raut et al. (2012)
Discovering dangerous patterns in long-term ambulatory ECG recordings using a fast QRS detection algorithm and explorative data analysis	QRS detection algorithm	Explorative data analytics	Dangerous ECG patterns recognition	Paolettiet et al. (2006)
Evolving a Bayesian classifier for ECG-based age classification in medical applications	Greedy hill-climb search and genetic algorithm	Bayesian network classifier	Comparative study with greedy algorithm and naïve Bayesian algorithm	Wiggins et al. (2008)
Fusion beats extraction from ECG using neural network-based soft computing techniques	Perceptron and back propagation	MATLAB based neural network	Fusion beats accuracy estimation	Ayub et al. (2010)
A model to find optimum percentage of training and testing data for efficient ECG analysis using neural network	Neural networks	R-based data division technique	Optimal division of dataset using R for neural networks	Bhanot (2015)

3 Experimental setup based on ECG case study

3.1 Electrocardiogram

ECG signals are used to access the muscular and electrical functions of the heart by detecting and amplifying the received electrical signals during the de-polarisation of the heart muscle during every heartbeat. A charge is developed across the outer wall of the heart muscle during the state of rest. During de-polarisation, the charge that was produced is reduced to zero thus activating the mechanism of contraction in heart. Heartbeat can be defined as the progression of a wave in an ordered form which can be measured by fluctuations in voltage between the two electrodes placed on both sides of the heart. Figure 1(a) shows smart ECG sensors having three leads (+ ve, – ve and neutral) to be placed at their specified positions shown in 1(b).

Figure 1   (a) Smart ECG sensors having three leads (b) Specified positions of ECG sensors
(see online version for colours)

|	
(a)	(b)

3.2   Obtaining ECG data

The ECG values in the form of analogue values are returned to the user. These values are measured in voltage which varies in between –5 to +5 volts for every cycle of a heartbeat as suggested by Park and Kang (2014).

3.3   QRS complex

The depolarisation of the left and right ventricles of the human heart contains a combination of three graphical deflections called the QRS complex whose length varies between 0.06–0.10 s in case of adults. For children and during physical activity, this duration may be shorter than the normal duration. A typical ECG wave comprising of the QRS complex and the different intervals of the ECG wave are shown in Figure 2.

Figure 2   A typical ECG waveform
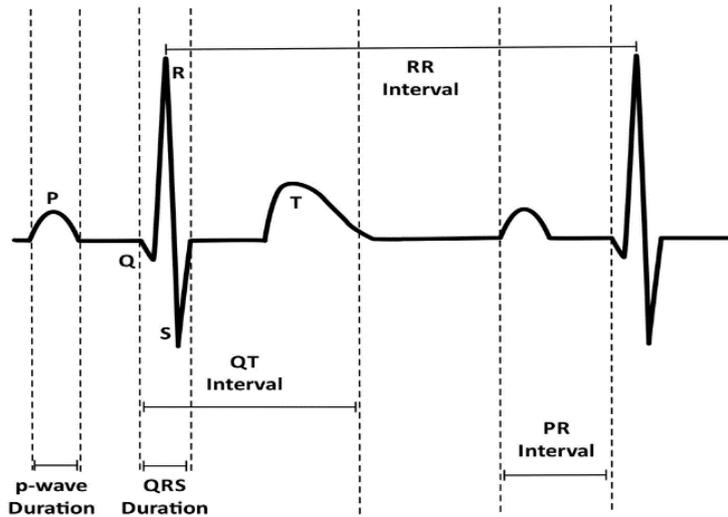
An ECG wave comprises of five deflections from 'P' to 'T'. The Q, R, and S part is usually considered together as the waves occur in quick succession and does not appear in all the leads. After the P wave is completed, a downward deflection of Q wave begins. After the P wave, R wave follows as an upward deflection and the S wave is the downward deflection after the R wave. The S wave is followed by the T wave and sometimes, an additional U wave is also present in the cardiac cycle which follows the T wave. These all deflections of the ECG wave are described in Figure 2.

3.4 *Pre-processing and feature extraction of ECG data*

Pre-processing of the signal is aimed at improving the quality of the signal for precise analysis of ECG data. Techniques adopted for de-noising and pre-processing of the ECG signals are low frequency baseline wander which is caused by respiration and body movements and high frequency random noises caused by main interference (50 or 60 Hz) as suggested by Bellos et al. (2010). Linear filtering and polynomial fitting techniques are employed for baseline wander removal. Inter beat (RR) interval and heart rate variability (HRV) are the features that can be extracted from ECG signals using the Pan-Tompkins technique.

4 Development of smart health risk analysis model

A model has been developed by the authors to predict the risk a patient may have of heart ailments by applying Neural Networks techniques on the ECG data obtained from the person in concern. MIT-BIH Long Term ECG database (LTDB) and MIT-BIT Arrhythmia Database (MTDB) data has been taken in the current paper for experimental and analysis purposes. The above data sets were then used to train the model and also to test the model and get predictions as results. The model comprises of neural network using RapidMiner as the software platform, which is further described below. The authors also prepared a model using the decision tree algorithm, same data set and platform to predict the results and performed a comparative analysis.

4.1 *Neural networks*

A computational model or a mathematical model which is inspired by the structure of biological neural networks is called an artificial neural network (ANN). A connectionist approach is used by ANN which consists of an interconnected group of artificial neurons for computation of the data received. Neural network is mostly an adaptive system whose structure changes based on internal or external data which flows through the network during the learning phase. Nowadays, neural networks are generally used to model a structure that trains itself with the relation between the input and output to propose a pattern, which can predict further results as referred from Dreiseitl and Ohno-Machado (2002).

A feed-forward neural network may be defined as an ANN where no cycle is formed between the connection units. In this type of network, there is unidirectional flow of data i.e. moving forward from the input nodes, through the hidden nodes (if any) and finally to the output nodes. The network does not contain any loops or cycles as referred from Santos (2014).

There are two phases in back propagation algorithm: propagation and weight update. It is a supervised learning method. These two phases are repeated till the performance of the network isn't good. In this algorithm, the values of output are compared with some predefined correct answer to calculate the value of error-functions. By some technique, this error function is fed back to the network. Then the value of these functions is used by the algorithm to adjust the weight of each connection in order to reduce the value of error function in each cycle. Then this process is repeated for a numerous number of times until the value of the error function is small. Then the network is said to have learned a certain target function.

4.2   RapidMiner

RapidMiner Studio is an environment for designing of advanced analytic processes with the concepts of data mining, machine learning, business analytics, predictive analysis and text mining. It is a coding-free environment, which makes it more users friendly. It is enriched with a numerous number of operators, which just have to be dragged and dropped onto the screen to create the desired process and obtain the results in a very systematic manner. Some of the commonly used operators being retrieve excel, retrieve csv, filter example, filter range, performance analysis etc. RapidMiner Studio also supports advanced operators like decision tree, neural net etc., to prepare models and analytic workflows. The main operators being used to prepare the deterministic risk analysis model have been described below.

4.2.1   Read csv/excel operator

This operator can be used to read an examples set from excel worksheets or csv (comma separated values) files. This operator is able to read data from different versions of excel. The user while loading the excel workbook, should specify which of the sheets should be included in the data table. The tables must be in a format that each row should have example set and columns denote the attributes. Here, the first row of the excel sheet might be used for attribute names which can be indicated by a parameter.

4.2.2   Decision tree operator

Decision trees are graph-like or tree-type model. This representation of data is meaningful and easier to understand. Hence, it has an advantage over other methods of representation. This operator in rapid miner studio generates the decision tree for the classification of numerical and nominal data. Its goal is to predict the value of a target attribute, often known as a class or label based on several input attributes and finally creates a classification model. Here each of the input attributes is represented with an interior node and the number of edges of a nominal interior node is equal to the number of values possible for the corresponding input attribute.

4.2.3   Apply model operator

A model is first trained on an example set and it learns all the basic information related to that example set. Then that model can be applied to another example set of same type, usually for prediction purposes. So, this operator in rapid miner studio helps one apply a

trained model to an example set for prediction or performance testing functions. All the needed parameters are stored within the model. So, it is required that the input example set on which model has to be applied, has attributes of same type and in the same order, else it could lead to serious errors. Several models pertaining to data analysis can also be grouped together using the group model operator, if multiple models have to be applied to an example set.

4.2.4 Filter examples operator

This operator in rapid miner is used to create a new example set for input from a given input data set, based on the conditions set as parameter for the operator. Various pre-defined conditions are available, which a user can select or can also create their own condition. This operator may reduce the data set given as input, but does not affect the attributes. Some of the available filtering conditions are 'is missing', 'is not missing', 'contains', 'starts with' etc.

4.2.5 X-validation operator

In order to estimate the performance of a learning operator which usually operates on unseen data sets, cross-validation operation is performed using the X-validation operator (read as cross-validation) in RapidMiner Studio. It mainly estimates in practice how accurately a model would perform after learning from any given model.

5 Experimental setup

In this paper, the data have been taken from the LTDB of patients with record numbers 14046, 14134, 14149, 14157, 14172, 14184 and 15814 respectively and from MIT-BIH Arrhythmia Database (MITDB) of patients with record numbers 100, 101 and 102. The data considered for all the above patients are of ten seconds duration and in standard format.

The ratio in which the testing data was feed to train the model was 7:2 to create equitable number of samples, as ECG signals of an arrhythmia patient were sampled at every 3 seconds as compared to 8 seconds for a normal patient. Then the steps followed are as follows:

1 We first used the read excel operator to load the example data set, that we prepared according to the ratio mentioned earlier.

2 In the load file configuration wizard, make 'ECG1' and 'ECG2' signal values as input attributes and the result attribute having values either 'low risk' or 'high risk' as the label.

3 Next we fed this example data set to the neural network operator to train itself with the data and prepare a model.

4 Now, this model could be applied to similar data sets to predict the results of the label attribute. So we used the apply model operator, with prepared model as one input and unlabeled data set as another.

5   The unlabeled data set fed to the apply model operator firstly was of pure normal patients to predict accurate results.

6   The whole process was then run, and the prediction results and statistics were recorded (Figure 6) along with the neural network prepared to obtain the results (Figure 5).

7   Now, authors aim is to repeat the whole process, but this time to train the model with the help of decision tree. So, above steps are repeated, with the decision tree operator instead of neural net operator.

8   The decision tree prepared was recorded (Figure 3) along with the statistics of the predicted data with the help of obtained decision tree (Figure 4).

According to the prediction of results, by both the models for the same data set, accuracy percentage of neural network model was obtained to be 92.57% as compared to 82.42% for the decision tree model.

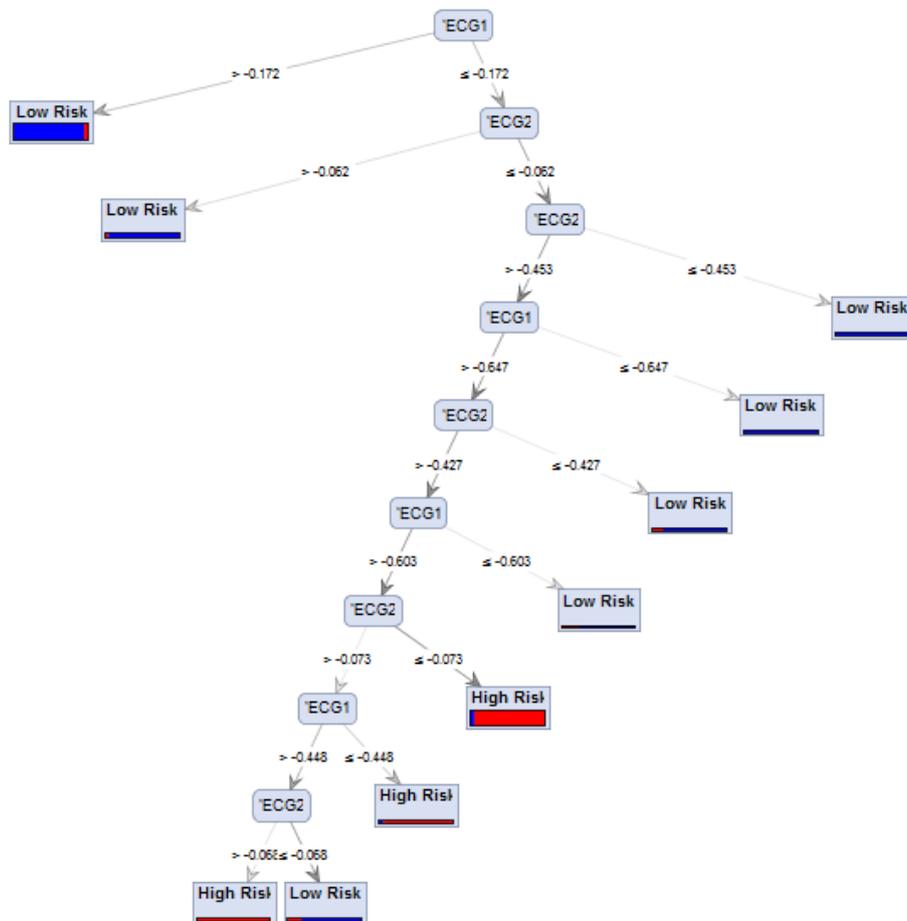**Figure 3**   Decision tree prepared from the experimental model (see online version for colours)

Figure 4 Prediction statistics obtained of decision tree model (see online version for colours)
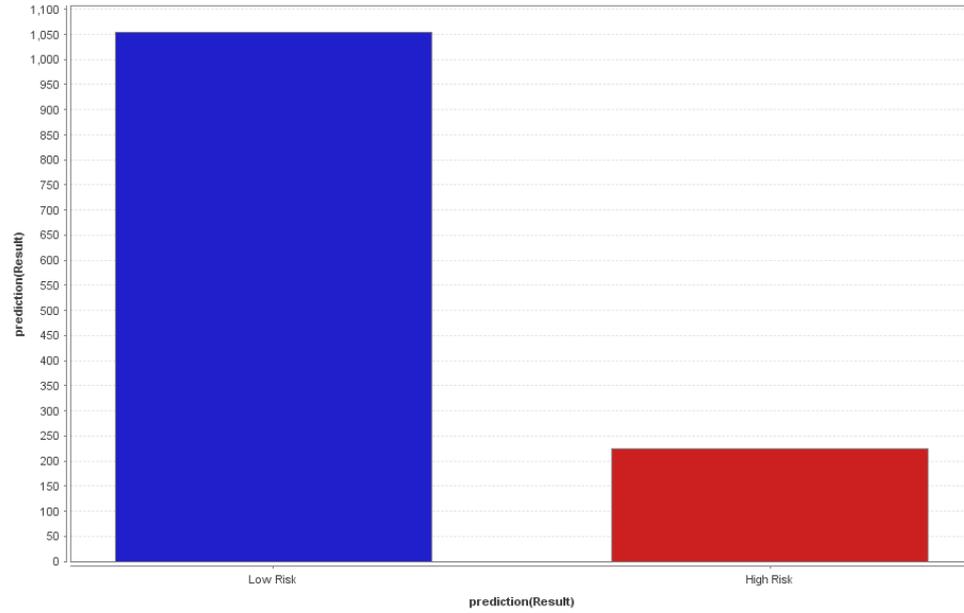

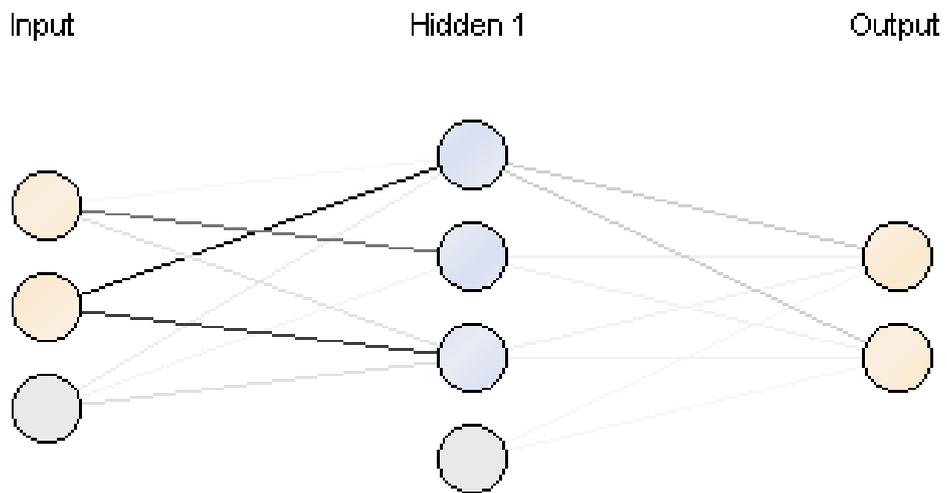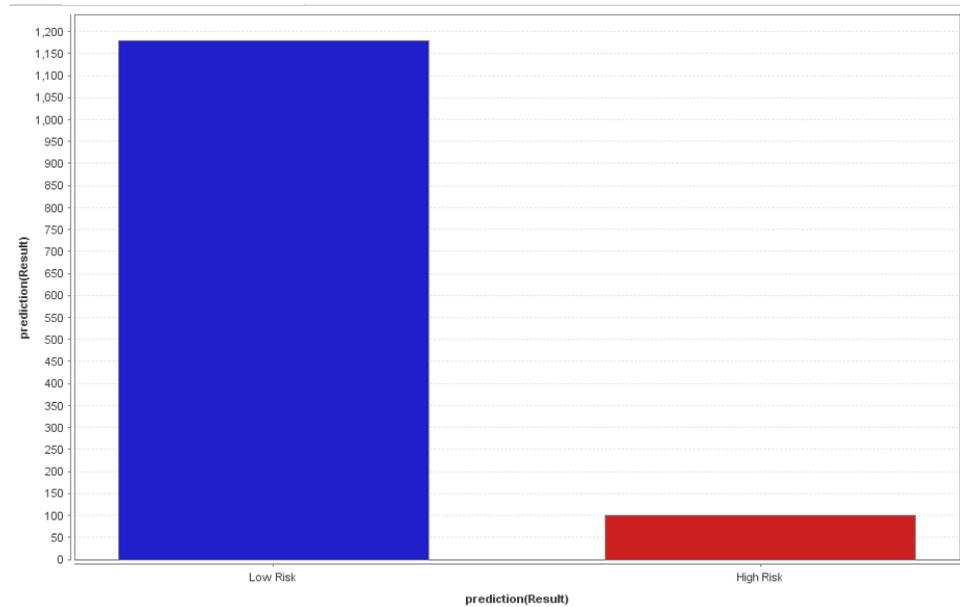**Figure 5** Neural network node diagram (see online version for colours)

Figure 6 Prediction statistics of neural network model (see online version for colours)

6   Conclusions and future work

In this paper, authors tried to develop a deterministic model for predicting the risk of heart ailments using two different models and concluded the possibility of heart disease risk for the person in focus. Additionally, we have compared the performance of the decision tree model with neural network model and calculated percentage accuracy for both the models. The major advantages of using this proposed model is that the model can learn on its own based on the sample space and can make new decisions, forecasts and classifications. The major disadvantage of the proposed model is the complexity involved and the decline in performance when dealing with noise or outliers.

In future, authors would like to expand the spectrum of diseases whose risk can be analysed by this model and the data which can be fed to it so as to provide a wholesome analysis of the risk associated with the every disease to the patient. Also, we would like to integrate cloud services to our model so as to manage large databases efficiently and to decrease the execution time of our model.

References

Ayub, S. et al. (2010) 'Fusion beats extraction from ECG using neural network based soft computing techniques', *Pelagia Research Library Advances in Applied Science Research*, Vol. 1, No. 2, pp.76–83.

Bellos, C.C., Papadopoulos, A., Rosso, R. and Fotiadis, D.I. (2010) 'Extraction and analysis of features acquired by wearable sensors network', *Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine*.

Bhanot, K. (2015) 'A model to find optimum percentage of training and testing data for efficient ECG analysis using neural network', *International Journal of System Assurance Engineering and Management*, Vol. 7, pp.1–6, DOI 10.1007/s 13198-015-0398-7.

Dreiseitl, S. and Ohno-Machado, L. (2002) 'Logistic regression and artificial neural network classification models: a methodology review', *Journal of Biomedical Informatics*, October 2002, Vol. 35, Nos. 5–6, pp.352–359.

Eftekhar, B., Mohammad, K., Ardebili, H.E., Ghodsi, M. and Ketabchi, E. (2005) 'Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data', *BMC Med Inform Decis Mak*, 15 February 2005, Vol. 5, p.3, DOI: 10.1186/1472-6947-5-3.

Karande-Raut, D.S. et al. (2012) 'ECG analysis using neural network', *International Journal of Systems, Algorithms and Applications*, December 2012, Vol. 2, No. 12, pp.5–8, ISSN online: 2277-2677.

Paolettiet, M. et al. (2006) 'Discovering dangerous patterns in long-term ambulatory ECG recordings using a fast QRS detection algorithm and explorative data analysis', *Computer Methods and Programs in Biomedicine*, Vol. 8, No. 2, pp.20–30, Elsevier Journal.

Park, J. and Kang, K. (2014) 'PcHD: personalized classification of heartbeat types using a decision tree', *Computers in Biology and Medicine*, November, Vol. 54, pp.79–88, doi: 10.1016/j. compbiomed.2014.08.013; Epub 2014 August 19.

Prasad, G.K. et al. (2003) 'Classification of ECG arrhythmia using multi-resolution analysis and neural networks', *TENCON 2003*, Bangalore, India, 0-7803-765I – x/03.

Rabi, A., Hadzima-Nyarko, M. and Šperac, M. (2015) 'Modelling river temperature from air temperature: case of the River Drava (Croatia)', *Hydrological Sciences Journal*, Vol. 60, No. 9, pp.1490–1507, DOI: 10.1080/02626667.2014.914215.

Santos, J. (2014) *A Trading Agent Framework Using Plain Strategies Machine Learning*, Repositório Aberto da Universidade do Porto, FEUP – Dissertação [online] http://hdl.handle. net/10216/76151.

Wiggins, J.M. et al. (2008) 'Evolving a Bayesian classifier for ECG-based age classification in medical applications', *Applied Soft Computing*, Vol. 8, pp.599–608, PMID: 22010038, PMCID: PMC3193938, DOI: 10.1016/j.asoc.2007.03.009.