Heart Disease Diagnosis Using Reconstructive Radial Basis Function Networks with Overlapping Prevention Method

Mashail Alsalamah, Saad Amin, John Halloran
Faculty of Engineering and Computing, Coventry University, Priory St, Coventry CV1 5FB, United Kingdom

Abstract – The term "Heart disease" applies to any abnormal heart condition affecting the heart itself or the blood vessels. It is prevalent today as it is the leading cause of deaths in the U.S. Because the heart is the engine of blood circulation, any issue it suffers directly affects the whole body.

In this paper, the design and implementation of a Radial Basis Function Network is presented to interpret, via data classification, the diagnoses of heart disease patients.

In the designed classifier, the Gaussian distribution function is used as a kernel of RBFs to build up the network, and peak RBF values are determined between -100% and +100% according to whether a patient has a certain disease or not. This creates smooth gradients between different RBFs, allowing the network to act as a fuzzy system.

The designed classifier has special training and optimisation algorithms, with those it aims to use the classification space at its maximum potential. This classifier is implemented as a standalone computer software.

An experiment using the designed system on two datasets collected from Prince Sultan Cardiac Center, Saudi Arabia, and UCI Machine Learning Repository, achieved good results.

Index Terms – Artificial intelligence, data classification, heart diseases, Radial Basis Function Networks.

1. INTRODUCTION

The human body is an interconnected complex machine. Each part more or less reacts to the changes of other parts, and the whole body tries to keep itself as stable as possible. When stability is not possible, such as cases of unstoppable bleeding, death is likely to occur. If it were possible to write a mathematical formula that accounted for every variable and constant of a human body, steps could then be taken to stabilise the body under different conditions, as is done in physics [1].

Today, this is still far from reality. However, by lowering the goal of stabilisation to a specific disease, collectable information of body as its current state (e.g. age, blood pressure, sex), inputs (e.g. diet, air quality), and outputs (e.g. saliva, blood content) becomes valuable clues about the cause of instability i.e. diseases of body, and thereby allows physicians to give advise and/or medicine to reverse the situation.

Coronary Heart Disease (CHD) refers to the state when a waxy plaque occurs inside the coronary arteries. This is illustrated in figure 1.

Arteries carry oxygen-rich blood to the heart. The built-up plaque restricts blood flow, thereby starving the heart of oxygen. This situation generally results with angina (discomfort of the chest) or heart attack. So, by taking this situation of a patient as a clue which is a collectable information, a percentage probability could be given to CHD.

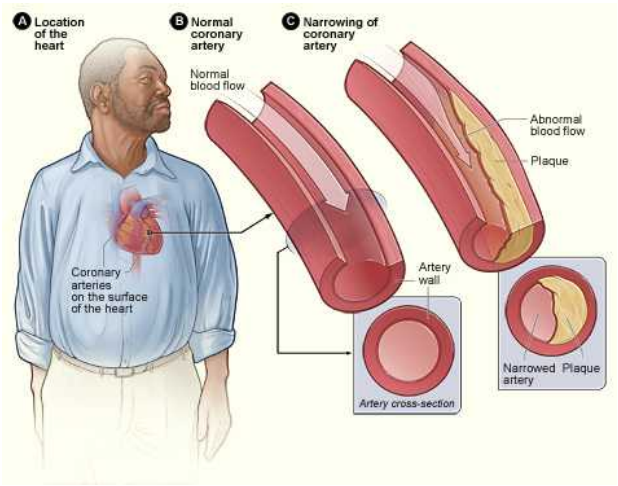

Figure 1: Showing how waxy plaque builds up in arteries [1]

In World Health Federation's online document [2], 26 different heart diseases are defined. In addition, a heart disease patient record database that dates back to 1988 identifies 13 different collectable information [3]. It is possible that today, more data can be collected from patients, and in the future, the number of heart disease types could increase. To be able to process this much collectable information and determine a probability for every different heart disease, a classifier based on the structure of Radial Basis Function Networks (RBFN) is presented in this paper. Generating probability values is required by a classification system to process these data to predict heart disease types.

In Noor A. Setiawan's paper [4], a fuzzy decision support system is presented for the diagnosis of coronary artery disease. This system takes the collected information as evidence, and creates fuzzy rules based on them.

In another recent paper [5], the Support Vector Machine (SVM) classifier is used, along with a genetic algorithm to implement a decision support system for heart disease classification. In this system, a genetic algorithm is used to get rid of a part of the collected data that is seen as less or non effective in classification results. It is claimed that this process increases the accuracy of SVM.

It is difficult to separate diseases from each other. Before the effects of one disease type ends, another has already started. Thus, using a separation-based classifier like SVM, or generating only true-false results does not suit heart disease

MOBIHEALTH 2015, October 14-16, London, Great Britain

DOI 10.4108/eai.14-10-2015.2261993

classification. An RBF, with its ability to reach infinity, can cover a large portion of the classification space for a single heart disease type, and is able produce results at varying degrees as would a fuzzy system. These qualities lend an RBFN heart disease classification very well.

The main purpose of this paper is to use the basic RBFN structure to develop some special training and optimisation algorithms to analyse heart disease patient records. Unlike other studies those use very well known machine learning tools by changing their parameters, and connecting them each other, this study will use specially designed and developed algorithms, and a computer software to implement the proposed classification system. Two datasets from Prince Sultan Cardiac Center, Saudi Arabia, and UCI Machine Learning Repository are used in this paper [3].

This paper starts with the theory of the algorithms, and ideas behind them. Later, patient record datasets, the implemented system, and the dataset experiment are described. Lastly, conclusions regarding the system are presented.

This paper builds on one that was previously published [6] in a way that it simplifies the whole system, provides extensions, and fixes some drawbacks of training process. It means that the work presented in [6] used a method that was more complex in implementation.

2. PROPOSED RBFN SYSTEM

2.1. Classification Procedure

A given record is the combination of features values, and belonging to all defined labels of classifier. A Radial Basis Function (RBF) sees each feature of a record as a unique dimension of a classification space, and uses feature values to determine a point in that space. An example of this is seen in figure 2.

In the designed classification system, a record can belong to all defined labels at various degrees. The degree of belonging can be any value between +100% and -100%, as below:

- +100%: Record belongs to given label completely.
- -100%: Record certainly doesn't belong to given label.
- 0%: It is unknown whether the record belongs to a given label or not.

Having percentage based belonging information helps human users to determine how much serious to take a certain disease on a patient.

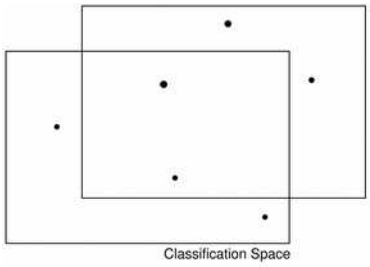

Figure 2: Illustrates RBF points in three dimensional classification space

An RBF's value is dependant on the distance of a point to its centre; the direction between two makes no difference. Because each point (i.e. a record), has a separate belonging level for each defined label, any point in the classification space can be classified based on its closeness to known points. In figure 3, records with labels are represented by circles, and those yet to be classified by a square.

The designed classifier averages the value of each unique label to determine the value of labels of an unknown point.

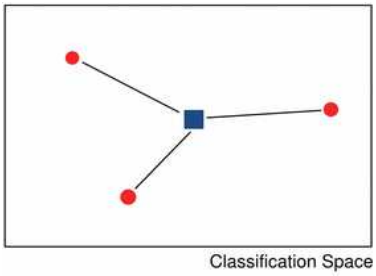

Figure 3: Shows how classification is done where circles are known records, square is unknown.

Let us assume that label *L1* is defined in the classifier. At the centre of known points (RBFs), peak percentage values are given for label L1 as +80%, -34%, and +5%. Based on used kernel function of RBF (e.g. Gaussian distribution function), the percentage value of a label will change (typically decreasing) as it distances from the centre point. As an example, at distance of square point to centre of each RBF, these percentage will be +40%, -12%, +0.7%, their average value is calculated at +9.56%. The unknown square point is then labelled as L1 at +9.56%.

This classifier supports multiple labels and RBFs for same points. To calculate the value of each label for an unknown point, separate calculations must be done.

2.2. Training Procedure

The designed training procedure takes multiple records, and builds up RBFN by filling with RBFs. Some items to be noted before explaining the procedure are:

1. Whenever the network is to be trained, all previously created RBFs must be removed. This facilitates better decisions and optimisation processes.
2. At the total number of defined labels, for each record, a new RBF is created; however, the total number of created RBFs may decrease during the optimisation process.
3. If the value of a label is 0% for a record, it will not add any new information to network, thereby will be ignored.
4. A label is positive if its value is greater than 0%, and negative if its value is less than 0%.
5. Training can be done for a specific label if there is more than one record; of these, there must be at least one positive and one negative.
6. For each label, the training procedure is repeated.

The training procedure starts by creating a new RBF for each record. To demonstrate, it will be assumed that this training procedure is being done for label *L1*. The RBF's centre position in the classification space is determined based record's feature values, and its label is set to L1 with L1's percentage value. At this point, the classification space has two or more RBFs with positive or negative labels. In figure 4, circles represent positive label RBF points, and squares negative label RBF points in a two dimensional classification space.

Each RBF has a kernel function that's value changes

depending on the distance value given to it. This training procedure assumes that the value of kernel function reaches its minimum after its radius.

At first, the radius value of each RBF is set to the distance between the two closest RBFs of different signatures (positive or negative). To demonstrate, if an RBF label has a positive percentage value at its centre, the distance between it and the closest negative RBF should be calculated to determine the positive RBF's radius. Figure 5 illustrates this step.

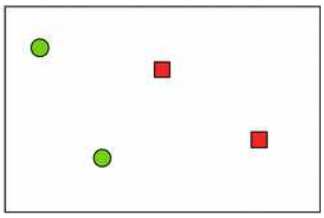

Figure 4: Circles are positive, squares are negative RBFs

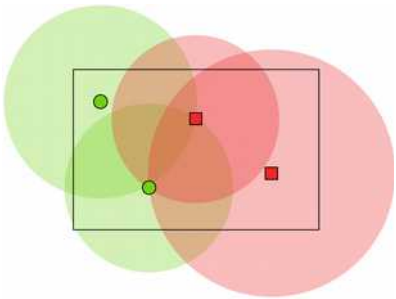

Figure 5: Illustration of how radius is calculated for each RBF

At this point, total number of in the classification space will be multiplication of total number of labels and total number of training records.. Thus an optimisation procedure is developed to decrease this number. If a positive RBF's coverage area includes another positive RBF, the two should be removed and a new RBF is created in between to represent them both. The same operation applies to negative RBFs. After this optimisation process is completed, radius of all RBFs are calculated again in the same manner.

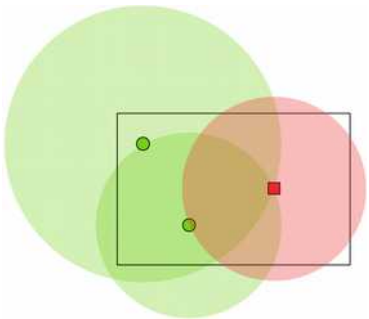

Figure 6: Illustration of post-optimization and radius values

Figure 6 illustrates the post-optimisation classification space of the RBFs seen in figure 5. With the latest radius values, parameters of all RBFs are updated. Finally, the classifier has its own RBFs, with label values, at specific points. Classification becomes possible when the training process has been completed for all labels.

2.3. Applying Classification and Training Procedures to Heart Disease Patient Records

Information may be collected from patients in three different ways:

1. **Patient input**: What the patient eats, how active the patient is in the day, smoking habits, alcohol usage, etc.
2. **Direct**: Sex, age, etc.
3. **Measurement**: Blood pressure, heartbeat rate, cholesterol, etc.

This information determines the centre point of each RBF in the classification space.

When a physician gives a decision about what type of heart diseases a patient has, this information combined with collected information to become a training record for the classifier. The more accurate the physician's decision is, the better the classifier is trained. Figure 7 illustrates how the radius determining process (as shown in figure 5) can be applied to heart disease records.

When records that have not been determined by a physician are applied to the system, they are classified based on how the training procedure used the records to build up the network.

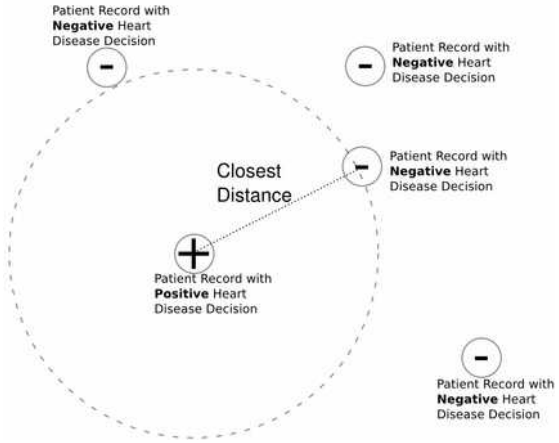

Figure 7: Determining the radius of RBF for positive heart disease decision

2.4. Gaussian Distribution Function

This function has following form:

$$g()=a e^{-\frac{b^2}{2c^2}}$$

where the parameter “a” is the peak value of function, “b” is distance, and “c” is how wide the bell curve of function is.

3. DATASETS

Two separate datasets were used for classification.

The first dataset of heart disease patient records was provided by the Prince Sultan Cardiac Center, in Saudi Arabia. They provided 89 records. For testing purposes, 60 of these records were used for training and the remaining 29 records for testing. These records had seven inputs as patient features and nine different heart disease types as output. Unfortunately, this dataset was limited in its number of records and input attributes, and diagnostic results were not verified. The structure of this dataset is represented in the table below.

Attribute	Explanation
Input Data	
Age	
Sex	“Male”, “Female”
Bp upper	Upper blood pressure
Bp lower	Lower blood pressure
Pulse	Heartbeat
Spo2	Oxygen saturation level
Hb	Hemoglobin
Output Data	
Arrhythmia	“Positive”, “Negative”
Congestive heart failure and cardiomyopathies	“Positive”, “Negative”
Ischemic acute myocardial infarction	“Positive”, “Negative”
Ischemic stable angina	“Positive”, “Negative”
Ischemic unstable angina	“Positive”, “Negative”
Valvular aortic regurgitation	“Positive”, “Negative”
Valvular aortic stenosis	“Positive”, “Negative”
Valvular mitral regurgitation	“Positive”, “Negative”
Valvular mitral stenosis	“Positive”, “Negative”

Table 1: Structure of first dataset

The second dataset was collected from the University of California, Irvine (UCI)'s machine learning repository [3]. This dataset has 303 patient records from the Cleveland Clinic Foundation database, each with 14 attributes. Six of these records had missing features, and were thus ignored. Thirteen attributes were input, one being the indicator for whether the patient has heart disease or not. From the remaining 297 patient records, 138 of them were chosen randomly for training purposes, and remaining 159 for classification. In this dataset, output was given as a number between 0 and 4, in which 0 indicates a patient without heart disease, and those with heart disease are represented by number 1 to 4. For this reason, number 0 was seen as -100% and other values as +100% for RBFs. The structure of this dataset is explained in the table below.

Attribute	Explanation
Age	
Sex	1 = male; 0 = female
Chest pain type	**Value 1**: typical angina **Value 2**: atypical angina **Value 3**: non-anginal pain **Value 4**: asymptomatic
Resting blood pressure	In mm Hg on admission to the hospital
Serum cholesterol in mg/dl	
Fasting blood sugar > 120 mg/dl	1 = true; 0 = false
Resting electrocardiographic results	**Value 0**: normal **Value 1**: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) **Value 2**: showing probable or definite left ventricular hypertrophy by Estes’ criteria
Maximum heart rate achieved	
Exercise induced angina	1 = yes; 0 = no
ST depression induced by exercise relative to rest	
The slope of the peak exercise ST segment	**Value 1**: upsloping **Value 2**: flat **Value 3**: downsloping
Number of major vessels	(0-3) coloured by flourosopy
Thal	3 = normal; 6 = fixed defect; 7 = reversible defect
Diagnosis of heart disease (output)	0 = not present; 1,2,3,4 = present

Table 2: Structure of second dataset

4. IMPLEMENTED SYSTEM

The developed system is implemented using C# programming language in Microsoft Visual Studio 2013. The system allows for the creation of new classifiers with desired features, labels, and the ability to store records, train the classifier with these records, and conduct classification. In figure 8, the definition and creation processes of new classifiers are illustrated.

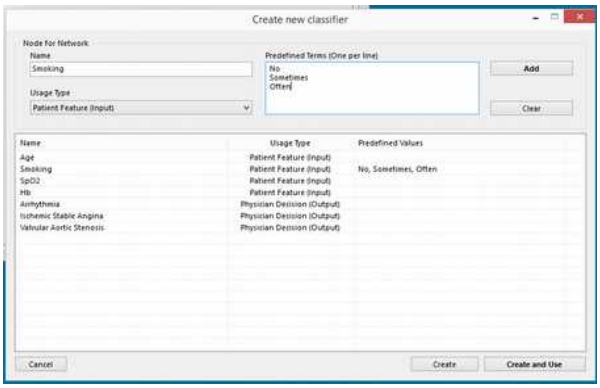

Figure 8: Define and create a new classifier

Based on how a classifier is defined, the system provides a graphical user interface to enter new records. In consideration of comma-separated values (CSV) Excel files, a copy/paste process is allowed to ease this process. In figure 9, an example of this screen can be seen.

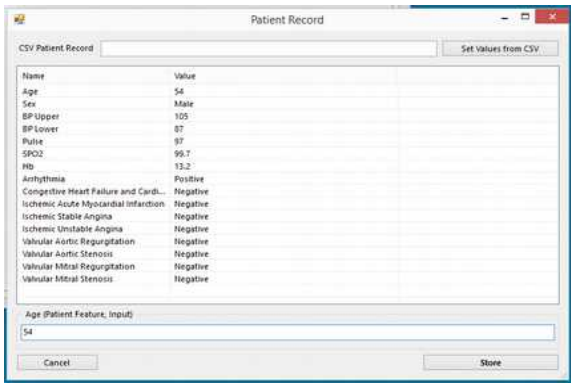

Figure 9: Adding new records for classifier

In the same way, the system provides another interface to enter features from a record and presents classification results immediately. CSV support is found on this interface as well. It can be seen in figure 10.

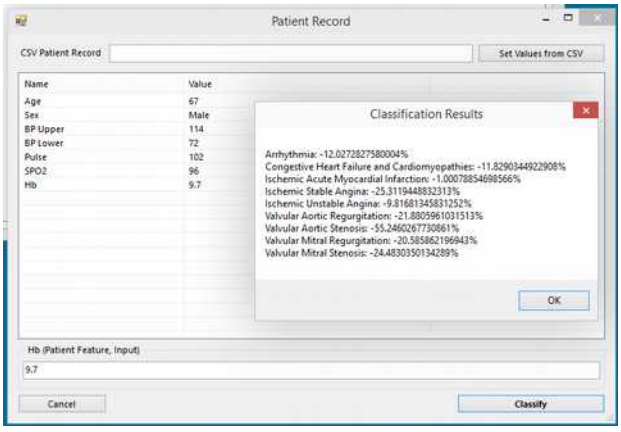

Figure 10: Classification process on developed system

The process of classifier training contains many steps as listed in the second part of this paper. This progress is shown in the system as well. Presented technical information is helpful in learning the details of the training algorithm, as demonstrated in figure 11. In it, the number of RBF nodes in the classifier, before and after optimisation, are also indicated.

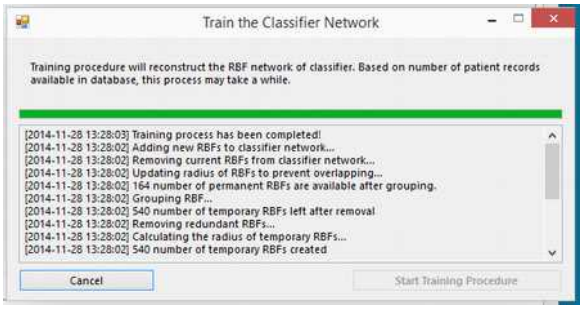

Figure 11: Training Process

5. Experiment Results

The first dataset has 7 input fields, and 9 different disease types as output. “Positive” and “Negative” values indicate expected outputs, and percentages indicate result of classifier.

	Rec 1	Rec 2	Rec 3
Age	49	53	25
Sex	Male	Female	Male
BP Upper	150	128	130
BP Lower	80	68	80
Pulse	102	77	75
SPO2	98	97	99
Hb	17.2	11.4	18
Arrhythmia	Positive 2.78%	Positive -12.84%	Positive .64%
Congestive Heart Failure and Cardiomyopathies	Negative -7.73%	Negative -16.69%	Negative -8.15%
Ischemic Acute Myocardial Infarction	Negative -7.37%	Negative -9.06%	Negative -2.59%
Ischemic Stable Angina	Negative -22.99%	Negative -32.34%	Negative -14.70%
Ischemic Unstable Angina	Negative -5.53%	Negative -13.21%	Negative -4.19%
Valvular Aortic Regurgitation	Negative -23.67%	Negative -40.92%	Negative -14.84%
Valvular Aortic Stenosis	Negative -59.97%	Negative -81.09%	Negative -53.56%
Valvular Mitral Regurgitation	Negative -33.03%	Negative -50.86%	Negative -23.86%
Valvular Mitral Stenosis	Negative -24.80%	Negative -34.90%	Negative -14.84%

Table 3: Results of classification of first dataset

The second dataset has output values as either true or false. However, the developed classifier's outputs are based on percentages. To be able to measure accuracy, positive percentages are accepted as true, and negative percentages as false; and they are then compared to dataset values. From 159 testing records, 110 of those records were classified correctly, an accuracy rate of 69.2%. Some of these testing results and expected results are in the below table.

	Rec 1	Rec 2	Rec 3	Rec 4	Rec 5
Age	63	67	62	57	63
Sex	1	1	0	0	1
Chest paint type	1	4	4	4	4
Resting blood pressure	145	160	140	120	130
Serum cholesterol in mg/dl	233	286	268	354	254
Fasting blood sugar > 120 mg/dl	1	0	0	0	0
Resting electrocardiographic results	2	2	2	0	2
Maximum heart rate achieved	150	108	160	163	147
Exercise induced angina	0	1	0	1	0
ST depression induced by exercise relative to rest	2.3	1.5	3.6	0.6	1.4
The slope of the peak exercise ST segment	3	2	3	1	2
Number of major vessels	0	3	2	0	1
Thal	6	3	3	3	7
Expected Output	Neg	Pos	Pos	Neg	Pos
Result(%)	-0.71	8.55	1.96	-0.07	2.80

Table 4: Results of classification of second dataset

6. CONCLUSIONS

The designed and implemented RBFN-based classifier system includes a special training algorithm aimed at heart diseases. This allows patients and diseases to be seen from the perspective of physicians by the way how data is collected, and results presented.

The training algorithm's optimisation procedure also aims to reduce the number of RBFs with minimal change to classification space coverage.

Classification accuracy was considerably good in this stage, but due to the lack of records available for training, it is not yet clear how successful the training algorithm is. Based on the ideas backing the system's design, and increased number of training records should greatly improve the accuracy of the classifier.

The designed classifier has the advantage of accepting and generating a percentage-based classification value that stretches from -100% to +100%. The addition of negative percentages makes it possible to see to what degree diseases are unlikely in patients. With its thoughtfully designed features, the system has become a suitable base for heart disease classification.

7. REFERENCES

[1] NASA. “Newton's Laws of Motion”. [Online] Available: <http://www.grc.nasa.gov/WWW/k-12/airplane/newton.html>

[2] World Health Federation. “Different heart diseases”. [Online] Available: <http://www.world-heart-federation.org/cardiovascular-health/heart-disease/different-heart-diseases/>

[3] University of California, Irvine (UCI). Machine Learning Repository. “Heart Disease Data Set”.

[4] Noor Akhmad Setiawan, P.A. Venkatachalam, Ahmad Fadzil M.Hani. “Diagnosis of Coronary Artery Disease Using Artificial Intelligence Based Decision Support System”. Proceedings of the International Conference on Man-Machine Systems (ICcoMMS) 2009.

[5] Sumit Bhatia, Praveen Prakash, G.N. Pillai. “SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features”. Proceedings of the World Congress on Engineering and Computer Science 2008.

[6] Alsalamah, M., Amin, S., Halloarn, J. “Diagnosis of Heart Disease by Using a Radial Basis Function Network Classification on Patients' Medical Records”. IMWS-Bio2014,183-186.IEEE, December 2004.

[7] David E. Newby, et al. “Expert position paper on air pollution and cardiovascular disease”. European Heart Journal (2015) 36, 83–93.

[8] P. A. Tijare, S. N. Sawalkar, and M. B. Wadhonkar, “Classification of heart disease dataset using multilayer feed forward backpropagation algorithm”. International Journal of Application or Innovation in Engineering & Management, vol. 2, no. 4, April 2013.

[9] M. Akhil Jabbar, B.L Deekshatulu, Priti Chandra. EC, Bhongir, India. 2013. Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection. Global Journal of Computer Science and Technology Neural & Artificial Intelligence. Vol 13, Issue 3.

[10] A. J. Howell, and H. Buxton, Face Recognition Using Radial Basis Function Neural Networks, 1996. [Online]. Available: <http://sites.poli.usp.br/d/pmr5406/download/papers/howell96face.pdf>

[11] K. U. Rani, “Analysis of heart diseases dataset using neural network approach,” International Journal of Data Mining & Knowledge Management Process (IJDKP), vol. 1, no. 5, September 2011.

[12] K. Rajeswari, V. Vaithiyanathan, and P. Amirtharaj, “A novel risk level classification of ischemic heart disease using artificial neural network technique – an Indian case study,” International Journal of Machine Learning and Computing, vol. 1, no. 3, August 2011.

[13] A. Afifi, “Laguerre kernels-based SVM for image classification,” International Journal of Advanced Computer Science and Applications, vol. 5, no. 1, 2014.

[14] F. Schwenker, H. A. Kestler, and G. Palm, “Three learning phases for radial-basis-function networks,” Neural Networks, vol. 14, no. 2001, pp. 439–458, December 2000.

[15] Shamsher Bahadur Patel, Pramod Kumar Yadav, Dr. D. P. Shukla. “Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques”. IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS). e-ISSN: 2319-2380, p-ISSN: 2319-2372. Volume 4, Issue 2 (Jul. -

Aug. 2013), pp. 61-64.

[16] Saeed Mehrabi, Mehran Maghsoudloo, Hossein Arabalibeik, Rezvan Noormand, Yones Nozari. Tehran University of Medical Sciences. Iran. “Application of multilayer perceptron and radial basis function neural networks in differentiating between chronic obstructive pulmonary and congestive heart failure diseases”. Expert Systems with Applications (Impact Factor: 1.97). 04/2009; 36:6956-6959.

8. AUTHORS

Mashail Alsalamah

A PhD student in health informatics, Faculty of Engineering and Computing, Coventry University

Dr. Saad A. Amin

PG Programme Manager, Department of Computing, Faculty of Engineering and Computing, Coventry University

Dr. John Halloran

A Senior Lecturer in Human Computer Interaction (HCI) at Coventry University, and a member of the Cogent Computing Applied Research Centre.