Early Diagnosis of Coronary Artery Disease using UCI Data set

SP. Chokkalingam, N. Deepa

Abstract: ***The focus of this project is based on both processing potential clinical features and implementing the classification architecture for detection of cardiac abnormality. The milestone of this first year involves analysis and investigation of different feature selection and transformation methods and theoretical modeling of single and hybrid systems by optimizing associated systematic parameters for better precision and recall. The importance of this paper is due to its clear objectives where an optimized and advanced system is designed and implemented for the cardiac disease utilizing computer aided diagnosis techniques for data and signal processing. The methodology is clear and trait forward using the hybrid approach of data mining techniques integrated to deliver enhanced performance on desired data set. In this paper comparative classification approaches are integrated to enhance system detection rate and decrease false alarms. The study focuses on feature pre-processing to select suitable feature subsets for classification algorithms like clustering (unsupervised learning) and SVM (supervised learning) which helps in generalizing the diagnosis system to detect unseen abnormality. For this study, will first apply statistical measures such as scoring ranking for clinical datasets consisting the electrocardiogram (ECG) features to reduce its dimension by eliminating irrelevant features. In the second part, will apply parametric tuned classification algorithms for selected feature subsets. The third part is to quantify the severity of CAD. At the last performance of the proposed system will be compared with other applied classification techniques in terms of accuracy, sensitivity and specificity.***

I. INTRODUCTION

In the present situation many fields such as hospitals, satellite communication, medical industry, banking sector, military and other public services that depend on the data storage. The most challenge for engineers, business people and scientists is to extract rapidly precious information drawn from raw images. The primary purpose of image processing is to convert images into information [1]. Today's medical imaging systems produce massive amounts of high-resolution diagnostic images. The production of visual representations of body parts, tissues, or organs, used in clinical diagnosis, allows doctors to uncover diseases earlier and improve patient outcomes.

White blood cells (WBCs) and red blood cells (RBCs) collected from blood samples based on microscopic images. The process of life is sustained by blood, a specialized body fluid consisting of plasma and blood cells.

Revised Manuscript Received on July 13, 2019.

SP.Chokkalingam, Professor, Saveetha School of Engineering, SIMATS, Chennai

N.Deepa, Assistant Professor, Saveetha School of Engineering, SIMATS, Chennai

To arrive at a proper diagnosis of a disease, blood cells must be identified and their relative quantity in the blood samples known. Due to the developments in technology, this traditional blood examination is digitized. By connecting a high-resolution digital camera to a microscope, blood cell images are captured after adjusting the microscope's magnification to obtain a fairly good resolution image [10].

WBC is classified into basophils, neutrophils, eosinophils, lymphocytes and monocytes. In white blood cells, lymphocytes are around 15% to 40% ranges and depends on the machine mentioned ranges may differ. When the body's general defence systems have been penetrated by dangerous, conquering microorganisms, lymphocytes help provide a specific response to attack these predators.

Image processing is used to enhance the image and to extract useful information from the image. An image is an array or a matrix of square pixels (picture elements) arranged in columns and rows. Image acquisition is the initial process required to collect images from resources like laboratories, online databases, and hospitals. Pre-processing is required during image acquisition because of the propinquity of noise in the image. Image enhancement is the process of improving user perception, in addition to providing an image of good quality. This enhanced image, in turn, undergoes further processing to extract valuable information. Segmentation is a method used to split the image into different pixel regions with similar attributes. Thresholding is the process used to segment the image [7]. Extracting features from the region of interest (ROI) is a major step in image processing. These features are used to identify the specific characteristic and parameters of an image, which helps classifications, are explained.

II. LITERATURE SURVEY

Image acquisition is the initial process required to collect images from sources like doctors, laboratories and hospitals. Capturing images from a patient using a highly-configured microscope and a Sony camera with a 300x magnification is the essential process of preprocessing. A single cell formed with manual cropping of white blood images, and its nucleus, cytoplasm and background are all cropped from the same image. It was observed that segmentation and cropping, manually done, involved much time. The importance of white blood cells, and how they help doctors diagnose patients' symptoms, was clearly demonstrated. The automatic counting of WBCs helps diagnose disease rapidly. The WBC is considered the input and features are extracted from the input image using the eigenvector method, with the eigen value and eigenvector being vital features for the

classification of WBCs [6]. This system for the segregation of white blood cells, tested with 50 samples, produced accuracy close to 92%.

Segmentation of white blood cells using the active contour, balloon model and gradient vector. This method was found unfeasible for classifying monocytes, since their shape is larger than that of others. The neighborhood of the initial shape of the nucleus was found from the blood smear image and cropped by means of the active contour algorithm. The active contour is applied to the gradient-flow-vectored image for cropping the region of interest from the blood smear image. The major and minor axes are used to determine the centroid, and the nucleus found from the centroid. The adaptive contour method is applied and cytoplasm segmented from the white blood cell. This is a time-consuming process, suitable only for regular images, since the centroid depends on the axis [9].

The features of the nucleus alone have so far been considered sufficient to classify white blood cells. This work investigates whether the premise stated is true. The aim is to segment the nucleus from the cell, but due to problems with overlapping, the entire cell is segmented and, thereafter, the nucleus alone needs to be separated. The classification rate is 77% and the disadvantage of the system is that misclassification occurs in adjacent classes since this method cannot differentiate between the classes [11].

The nucleus is segmented by the poisson equation-based approach and, in this straightforward method; the shape of the nucleus is represented with the area between inner distances are measured. Experimental results have shown that the poisson equation method is better than the others. However, the notion of random walks and other criteria are required for the poisson equation method.

A rough set-based clustering approach to detect leukemia automatically from a blood smear image is presented, and the same approach is followed for color-based segmentation of WBCs. The irregular boundaries of the nucleus are measured using the hausdorff dimension and contour signature. Along with this measure, leukocyte features like shape, color and texture are also considered to obtain greater accuracy. The fact that this approach is unsuitable for images dependent on stain and illumination is a drawback. The Gram-Schmidt method was used to segment the nucleus and cytoplasm with the help of an orthogonal basis of eigenvectors, orthogonal approaches proving better than theoretical algorithms. After segmenting the WBCs, the remaining RBCs and chromatin dots are separated. The confusion matrix defines sensitivity at 83% and specificity at 98% for the 55 images annotated, with the segmentation of RBCs and WBCs not being perfectly achieved [12].

The double histogram threshold technique is implemented, initially, to separate white blood cells and red ones from the blood smear image. Gray-level co-occurrence matrix (GLCM), one of the most popular image-analysis techniques used to extract texture features, perform a classification and morphological operations, and proposed a method of automatic segmentation, size determination, counting and classification of white blood cells. Peripheral blood stem cells or bone marrow as the graft source for the allogeneic hematopoietic cell transplantation method is proposed [14,15].

Biomedical data processing and classification are vivid scientific topics; this is due to the huge number publications per year and the need of optimized real-time systems for disease detection. Today in Kingdom, a biomedical signal (ECG) investigation is a basic step for diagnosing heart diseases. There are around 20,877 deaths in Saudi Arabia due to Coronary Heart Disease (CHD) which means 23.98% of total deaths in the Kingdom. The Coronary Heart Disease (CHD) age adjusted death rate is around 180.6 per 100,000 in Saudi Arabia which is ranked 32 in the world. Heart disease is one of the leading causes of deaths around the globe. Coronary Artery Disease (CAD) in which the coronary arteries contracted and slowly solidify which leads to chronic type of heart attacks. In one of the American heart association (AHA) survey, CAD is considered main killer of US citizens [5]. Among different CAD related factors like sex, age cannot be changed but other factors regarding lifestyle can be adjusted such as lowering blood pressure and cholesterol, having physical activities etc. can decrease the CAD risks. Diabetic patients can also prevent from severity by taking proper diet and physical routines [6, 7]. Doctors observe earlier CAD cases as well to assist on-going patient's tests. Factors can also analyze to achieve effective procedure for patients, but it becomes complicated when considering various patients and their related factors which vary from one patient to another. Few CAD patients have symptoms like high blood pressure or cheat pain while some don't have any symptoms unless heart attack takes places because of blocked coronary arteries [8].

In recent years, computerized diagnostic systems are being designed to enhance the abilities of cardiologists in detecting CAD patients with higher accuracy rate [9]. Numerous researches related to the diagnosis of heart disease have motivated this research proposed for performance enhancement of CAD detection system. Different feature extraction and selection algorithms have been integrated with classifiers like Artificial Neural Networks, Support Vector Machines, Decision Tress and Clustering approaches, etc. This research is focused entirely on optimizing the classification architecture by reducing complexity and processing overhead in the detection of pattern behavior for disease identification. Our research group is using this proposed methodology on standard UCI Heart datasets for extracting their sensitive features to enhance the detection rate of CAD and normal cases.

Coronary Artery Disease (CAD) is a complicated clinical syndrome which results in functional disability of heart to eject or fill with blood. CAD causes reduction of blood flow in coronary arteries for pumping blood to heart muscles and carrying depleted blood out. Clinical diagnosis of heart anomalies in the early stage of CAD with the help of biomedical technologies is essential yet difficult to achieve precisely and efficiently. It requires detailed information from patients when there are no obvious symptoms like chest pain, cholesterol, obesity, hypertension, etc. The goal of this project is to develop an optimized CAD detection system with an ensemble approach for features extraction and classifiers to extract sensitive features which can be

classified with maximum detection rate. For this purpose, we are using UCI heart datasets to develop CAD detection system by proper training and cross-validation of sensitive features to distinguish abnormal cases (CAD patients) from normal ones represented in Figure [1].

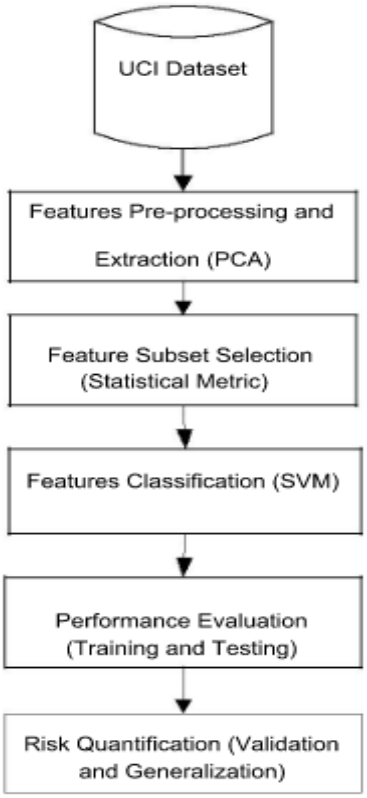

Fig. 1 Phases for CAD Detection System

In this report, the activities of the first year are explained, which are the detailed study of the computer aided diagnosis systems, their requirement, earlier work done in this field and the comparative analysis of data mining techniques. After this, the formation of hybrid system with integration of different techniques is designed to proceed for implementation by optimizing associated parameters and functions. The features of the selected dataset are processed to transform them in acceptable form for further extraction. The selection of the features is done on the basis of their importance to enhance the overall system performance. In later phases of second year, the system will be cross validated, trained and tested with desired subsets from the dataset. At the end, different data sets (subsets) will be used for the CAD detection including ECG signals. To identify CAD presence, we are using a UCI heart dataset with thirteen attributes like age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG result, maximum heart rate achieved, the occurrence of exercise induced angina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, a number of major vessels colored by fluoroscopy, and thal. These attributes are enough to assist us in detecting CAD patient by taking help of angiography result like zero means patient is normal while one for single-vessel, two for double-vessel disease, three for triple-vessel disease, and four for left main Coronary Artery Disease [2-4].

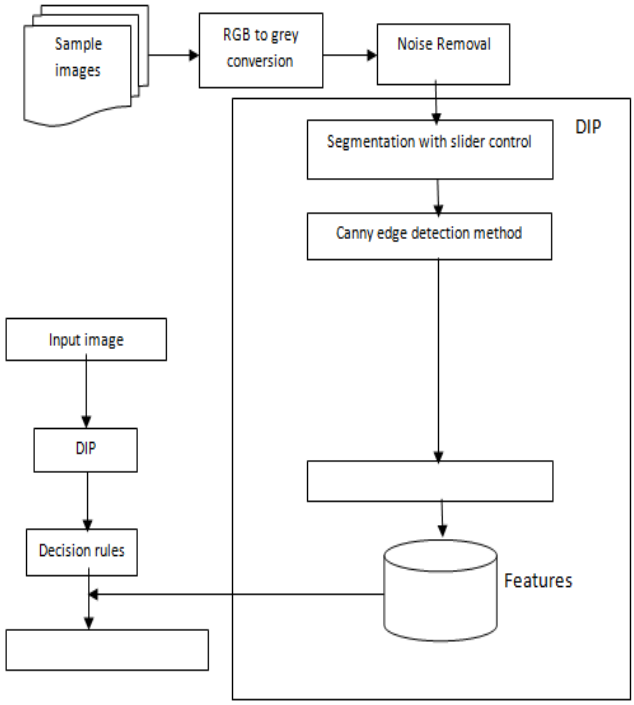

Fig. 2 CAD Diagnosis Process from Images

Sample images are collected from various laboratories. Noise may accumulate in images during image acquisition due to excessive staining. Wiener filter is found as most suitable for noise removal. Segmentation method with slider control is proposed for segmenting the ROI. The canny edge and bounding box algorithm are used to detect the ROI edges. Features are extracted from ROI and stored in the database. Decisions rules are framed to ascertain whether or not the image is affected are explained in Figure 2.

III. RESEARCH METHODOLOGY

Based on the necessary research in this field, our main focus will be on the optimal performance of the proposed architecture for higher detection rate and lower false alarms. The system methodology as shown in the Figure 1 presents the hybrid integration of data mining approaches for features pre-processing and classification to detect clinical abnormality in cardiac patients. In this project we have chosen the hybrid combination because we need to achieve better and enhanced results. The methodology is clear and well-defined, this system will be developed using Matlab, Neuro Solutions to testify the performance of experimental data for required sensitivity and specificity. The steps undertaken are towards the prototyping of the system by analyzing the data such as: clinical and laboratory findings, ECG signals, etc. of the patients along with the label indicating absence or presence of the CAD disease. Overall, the methodology is improved and modified as compared to the classical one. It includes theoretical procedures, data processing and extraction, experimental setup (classification architecture), fixing the parameters (cluster sizes and tuning classifier cost and gamma values), conquering the obstacles (time complexity and overhead) with simple and clear explanation.

The results found in the first year are submitted and accepted for publications as mentioned in this section. The key findings, in this year (Phase 1 and Phase 2) are the theoretical review of earlier approaches in CAD detection and then analyzing the selected dataset for processing and extraction. In this research we are using datasets taken from UC Irvine Machine Learning Repository [10]. Our finding from these datasets will assist us in comparing performance of our CAD system with previous work. Different algorithms based on data mining techniques are being evaluated to figure out the most relevant and effective method for feature selection and transformation. Linear and non-linear transformation, supervised and unsupervised classification approaches, all are reviewed and analyzed to proceed further for hybrid system development. Apart from selecting appropriate classifier and feature processing methods, parameters associated with classifier architecture also need to be tuned well to get optimized accuracy and detection rate. Another obstacle is to develop a system, with the generalization ability to detect unseen records with maximum sensitivity and specificity.

IV. SIGNIFICANT LEVELS AND RANGES

The clinical data set contains five critical parameters such as area, perimeter, circularity, solidity, and roundness, related to CAD. The data set is classified into two types, namely, Non-Inflamed and Inflamed. The non-inflamed is denoted as LEVEL-0, and the inflamed are further classified as Low, Moderate and High.

The following are the guidelines for the range of values of the critical parameters to represent the image in a specific group.

LEVEL – 0: Non - Inflamed

The following are the lower and upper ranges of values for the moderate inflamed type

Area (P1) : 9036-12989
Perimeter (P2) : 471-630
Circularity (P3) : 0.4084- 0.5116
Solidity (P4) : 0.8356-0.8866
Roundness (P5) : 0.79-0.1123

If the area values lie in the range of 9036-12989, then the particular image sample can be considered as LEVEL-0 type. For Level 0 type, the perimeter value lie in the range of 471-630, circularity can fall in the range of 0.4084-0.5116, solidity stays in the range of 0.8356-0.8858, and the roundness value lie in the range of values between 0.79 and 0.1123 (e.g:0.81). The medical values retrieved from the microscopic data of the inflamed type can be further classified into three broad levels, namely, low, moderate, and high. The following are the upper and lower limit values found in the image set corresponding to the low inflamed type.

Inflamed: Low Level

Area (P1) : 8639-8970 & 13010 - 13757
Perimeter (P2) : 453-469 & 631-652
Circularity (P3) : 0.4065-0.4133& 0.5122-0.5290
Solidity (P4) : 0.8259-0.8352 & 0.8865 – 0.8906
Roundness (P5) : 0.0748-0.0778 & 0.1084 – 0.1257

Inflamed: Moderate Level

Area (P1) :7583 – 8599 & 13861 - 14957
Perimeter (P2) : 398-451 & 652 -676
Circularity (P3) : 0.4054-0.4148 & 0.5301- 0.6013
Solidity (P4) : 0.8107 – 0.8245 & 0.8912 – 0.9040
Roundness (P5) : 0.0670 – 0788 & 0.1148 - 0.1420

Inflamed: High Level

Area (P1) : 6581- 7491 & 15106 - 26090
Perimeter (P2) : 359-395 & 677- 894
Circularity (P3) : 0.3955-0.4165 & 0.6030 – 0.6413
Solidity (P4) : 0.6535 - 0.8088 & 0.9052-0.9167
Roundness (P5) : 0.0585 - 0.0654 & 0.1269 – 0.2012

V. ANALYSIS USING T-TEST

Table 1 includes the results obtained from t-test tool for inflamed versus non-inflamed categories with respect to area. The normality and equal variance tests failed for this t-test analysis. The difference in the mean values of the two groups is greater than would be expected by chance; there is a statistically significant difference between the input groups (P = <0.001).

Table. 1 t-test results for Inflamed versus non-inflamed

Group Name	N	Missing	Mean	Std.Dev.	SEM
Inflamed	227	0	14529.52	4650.75	308.681
Non Inflamed	163	0	10863.54	1054.07	82.562

Similar to t-test analysis for area parameter, all other parameters of PARCS is also analyzed. Both normality and equal variance tests shows that the sample groups statistically differ between the inflamed and non-inflamed categories for all the other parameters. The chi-square statistical testing with all the parameters of multiple categories or levels, yields positive statistical results. The proportions of observations in different columns of the contingency table for all the levels vary from row to row. The two characteristics that define the contingency table are significantly related. (P = <0.001).

VI. CONCLUSIONS

The main focus of the project is to achieve optimized CAD detection with hybrid integration of enhanced classification and feature processing techniques. This year was more focused on the theoretical review of earlier work including their limitations along with the study of the selected dataset for transformation and extraction to lead better system formation. Further, analysis of data mining methods is needed to derive best possible diagnostic system for cardiac patients. The objective of this CAD diagnosis system is to identify possible cardiac abnormalities from unseen clinical records once the system is trained enough with standard dataset. Once an optimized system is developed, it is ready to assist medical staff as per their requirements.

These systems are trained, cross validated and tested well before they can be used in real-time environment. Comparative classification approaches are used to enhance system performance as much possible and decrease false alarms. The expected results of this research can also be applicable in detecting different types of heart diseases using ECG signals and other symptoms. These approaches can be further extended to the other diseases within the medical domain like cancers and chronic health problems etc.

REFERENCES

1. N. A. Al-Baghli, et al., "Awareness of cardiovascular disease in eastern Saudi Arabia," J Family Community Med, vol. 17, pp. 15-21, 2010.
2. L. Wei and R. B. Altman, "An Automated System for Generating Comparative Disease Profiles and Making Diagnoses," IEEE Transactions on Neural Networks, vol. 15, p. 597, 2004.
3. R. Detrano, et al., "Algorithm to predict triple-vessel/left main coronary artery disease in patients without myocardial infarction. An international cross validation," Circulation, vol ,83 .pp. III89-96, May 1991.
4. (2013, 27 July), UCI Respository for Cleveland and Statlog Heart Dataset. Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease
5. http://www.americanheart.org. (2013). Available: http://www.americanheart.org
6. R. Alizadehsani, et al., "Diagnosis of Coronary Artery Disease Using Data mining based on Lab Data and Echo Features " Journal of Medical and Bioengineering, vol. 1, 2012.
7. M. M. Al-Nozhaa, et al., "Smoking in Saudi Arabia and its relation to coronary artery disease," Journal of the Saudi Heart Association, vol. 21, pp. 169-176, 2009.
8. N.Deepika, et al., "Association rule for classification of Heart-attack patients," International Journal of Advanced Engineering Sciences and Technologies, vol. 11, pp. 253 - 257, 2011.
9. R. Das, et al., "Effective diagnosis of heart disease through neural networks ensembles," Expert Systems with Applications, vol. 36, pp. 7675–7680, 2009.
10. U. I. M. L. R. http://archive.ics.uci.edu/ml/. (2013). Available: http://archive.ics.uci.edu/ml/
11. SP.Chokkalingam, Komathy. K, Mohan Kumar. P, DuraiMurugan. S Apr 2015, 'CAD Dataset Analysis and Case Studies', International Journal of Applied Engineering Research, vol. 10, no. 4, pp. 2948-2952, ISSN: 9734562.
12. PawanPatidar, Manoj Gupta, SumitSrivastava, Ashok Kumar Nagawat, "Image de-noising by various filters for different types of noises", Published in International Journal of computer applications, Vol-9, November 2010.
13. G.Padmavathi, Dr.P.Subashini, M.Muthu Kumar and Suresh Kumar Thakur, "Comparison of the filters used for underwater Image-Preprocessing", IJCSNS International Journal of Computer Science and the Network Security, VOL.10 No.1, January 2010.
14. Y.Murali Mohan Babu, Dr.M.V. subramanyam, Dr.M.N.Giri Prasad, "PCA based image de-noising", Published in SIPIJ,Vol 2, April 2012.
15. WL. Clapp."The Renal Anatomy". In: XJ. Zhou, Z.Laszik ,T.Nadasdy , VD. D'Agat i, FG.Silva, eds. Silva's Diagnostic Renal Pathology. New York: Cambridge University Press; 2009.