Journal of Contemporary Issues in Business and Government Vol. 27, No. 3,2021
https://cibg.org.au/

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.227

AN EFFICIENT PREDICTIVE SYSTEM FOR HEART DISEASE USING A MACHINE LEARNING TRAINED MODEL

[1]Prabhavathi K,[2]Dr. V Mareeswari
[1]Department of Computer Science and Engineering,Amruta Institute of Engineering and Management Sciences, Bangalore, Karnataka, India-603 203
[2]Department of Computer Science and Engineering, ACS College of Engineering, Bangalore, Karnataka, India- 603 203
Visvesvaraya Technological University, Belagavi
[1]prabhavathi138@gmail.com,[2]maresh.prasanna@gmail.com

ABSTRACT : Heart is the most important organ of a human body. Through blood, it transfers oxygen and vital nutrients to the various parts of the body and helps in the metabolic activities and it removes wastes of metabolic. Thus, even minor problems in heart can affect the whole organism. Researchers are diverting a lot of data analysis work for assisting the doctors to predict the heart problem. So, an analysis of the data related to different health problems and its functioning can help in predicting with a certain probability for the wellness of this organ. In order to assist the physicians, identification of heart disease by machine learning and data mining technique has been implemented. Heart as one of the essential organ of the human body and with its related disease such as cardiovascular diseases accounts for the death of many in our society over the last decades, and also regarded as one of the most life-threatening diseases in the world. Today healthcare industry is rich in data however poor in knowledge. There are different data mining and tools and algorithms of ML are available for extraction of knowledge from data storeand to use this knowledge for more accurate diagnosis and decision making. The main contribution of this review is tosummarize the recent research with comparative results that has been done on heart disease prediction and also make analytical conclusions. From the study, it is observed Naive Bayes with Genetic algorithm; Decision Trees and ANN techniques enhance the accuracy in predicting heart disease in different scenarios.

Keywords:-Algorithm, Heart disease, Health care, Machine learning, Prediction.

1. Introduction

Sleepiness Heart is an important organ also termed as the centrality forhuman body generates blood to the other part of the organs in the body and where it fails in its responsibility, which causes the immediate death of a person. Hence it is observed from research that works overload, mental stress, depression, change of lifestyle and bad food or eating habits which occurs

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.227

basically among mostly in adults are responsible for the rate of several heart-related diseases in our modern world. To diagnose heart-related diseases or cardiovascular is so complicated and at times difficult to accurately and efficiently made thereby leading to the wrong diagnosis from the healthcare provider which makes the costs of medical care to the patients to be very expensive in most cases. Therefore a predictive system is developed using a supervised Machine learning trained model in MATLAB2018 for prediction of the accuracy and evaluation of the efficiency of the heart-related disease in a person. This will in turn help in diagnosing the disease and lessen the patient medical costs and treatments though this is based on the experience of the doctor and the patients' current result test Also, Receive Operation Characteristic Curve (ROC) which is a pictorial representation that represents the diagnostic ability of a binary classifier system is used with three classified algorithms in order to determine and evaluate the efficiency with respect to real-time environment which is used in by the medical doctor to monitor the patients who are at increased risk of dangerous heart conditions and thereby removal of unskilled clinician diagnostically approach.

Table 1 shows the various categories of heart diseases.

SL No	Type of Heart disease	Description
1	Arrhythmia	An arrhythmia is also called as dysrhythmia. It is an irregular or abnormal heart beat.
2	Cardiac arrest	A cardiac arrest happens without warning. If someone is in cardiac arrest, they collapse suddenly and it will be unconscious and unresponsive
3	Congestive heart failure	CHF is a chronic progressive condition which causes the pumping power of heart muscles. It is also called as simply "heart failure," CHF specifically refers to the stage in which fluid builds up around the heart and causes it to pump inefficiently.
4	Congenital heart disease	Congenital heart disease or defect is a heart abnormality occurs at birth. This causes the heart walls, valves and the blood vessels
5	Coronary artery disease	CAD is the blockage of the coronary arteries, usually caused by atherosclerosis. Atherosclerosis is the construction of cholesterol and fatty deposits on the inner walls of the arteries..

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.227

6	High Blood Pressure	It is also called as Hypertension, occurs when your BP increases to unhealthy levels. It depends on how much blood is passing through your blood vessels and the amount of resistance the blood meets when the heart is pumping.
7	Peripheral artery disease	This leads to a poor quality of life and a high rate of depression.
8	Stroke	A stroke generates when the blood supply to part of your brain is reduced, it prevents brain tissue from getting oxygen and nutrients.

Table 1: various types of heart disease.

Figure shows the various parts of human heart.

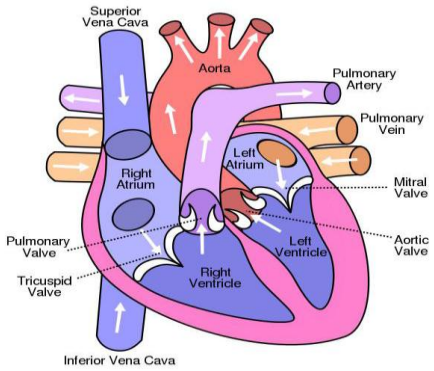

Figure 1: Human Heart

This paper is described as follows. Part 2 explains background knowledge. Part 3 explains an overall related work. In Section 4 and 5 ,observation and findings and conclusion and are described.

1. Prior Knowledge

In every field of education we need prior knowledge to understand and analyze that field very well, prior knowledge become base for successful understanding and analyses of any study. So before we start to study the actual content of this paper we have to study and understand the basic concepts related to the paper that will help us to understand and comprehend the paper very well.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.227

2.1 Classification: Classification algorithms are used to classify a record. It is used for questions which can have only a limited number of answers. When you have only two choices, it's called binary classification, if you have more than two choices it's called multi class classification.

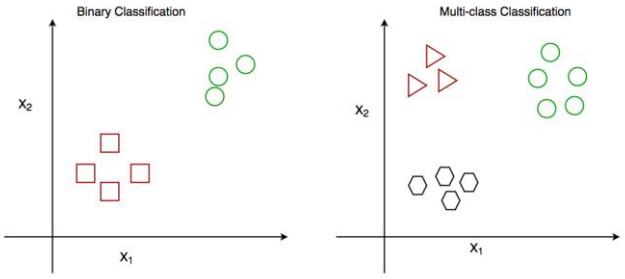

Figure 2: Representation of classification

2.2Clustering: It helps you to understand the structure of dataset. These algorithms separate the data into groups or clusters, to ease out the interpretation of the data.

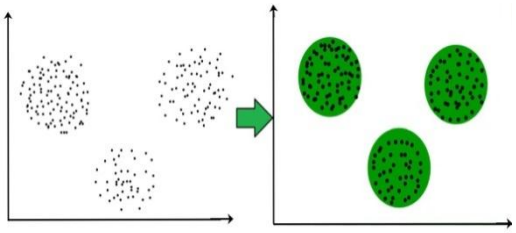

Figure 3: Representation of clustering

2.3 Decision trees: DT is one the predictive modelling technique used in ML, statistics. This method divides the data set based on some metrics called gain and entropy. It is one of commonly used supervised and non-parametric learning technique used for classification and regression.

2.4 Naive Bayes:A Naive Bayes classifier follows the principles of Bayes theorem. It is probabilistic ML classifier model used for discriminate various objects depending on some features.

2.5 Artificial Neural Networks:A collection of nodes which are interconnected and inspired by biological neural network. In this representation, an arrow represents the connection from one

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.227

neuron to another. Gradient descent and back propagation algorithms are the most widely used algorithm based on ANN.

2.6 Genetic Algorithm: Optimization problems are solved by GA. The GAs generates successor hypotheses by mutation and crossover of the best currently known hypotheses. So in each step a part of the current population is replaced by offspring of the fit hypotheses. In order to identify the best hypothesis, a space of candidate hypotheses is searched, so-called hypothesis fitness.

2.7 Cross Validation: Cross validation technique is used to evaluate the estimator's performance. It is a method used for evaluation of predictive models. In this technique the dataset is divided into two parts i.e. Training set used to train the model and test set for evaluation. The K--fold cross-validation technique divides the original sample into k subsets of equal size.

3.Literature Survey

Till date different studies have been done for predicting heart disease. The different practical results have been achieved for different proposed ML and data mining techniques implemented on clinical dataset. But, still today we are facing a lot of problem faced by the heart disease. Some of the recent research papers are as follows

In [1] different prescribed data of 1094 patients from different parts of India have been analyzed. In order to predict whether a new out-of-sample data has a probability of having any heart attack or not, model is constructed which gets trained by the input data. This model is useful to make decision with the doctor so that doctor can treat the patient well and creates transparency between the doctor and the patient. We hence stated one new metric called Selection Value which takes care of these scenarios and selects that algorithm which gives maximum S.V. We do not want to bias the doctor with the results of the classification rather as discussed in the proposed scenario section; we try to give the doctor with the better option with the history similar data results. Using these data, the doctor can have a transparency with the patient and the patient won't feel cheated at the end.

In [2] dataset is analyzedby Naive Bayes, algorithm based on risk factors. Heart disease is predicted by DT and combination of algorithmsalong with same attributes. Classification technique is performed on heart dataset downloaded from UCI library site. Feature selection techniques and decision tree based techniques are applied to produce high accuracy with four features. Experimental outputs are generated by using WEKA tool which proved that CART is more efficient than other methods even after applying FST as 84 % accuracy with 4 features only.

[3] Summarized the current researches related to heart disease prediction by ML techniques. In order to classify the patients whether they are having heart diseases or not depending on the information obtained in the health care data, logistic regression method has been applied as classifier algorithm. Prediction is also performed on data set. The major contribution for this

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.227

study is to apply logistic regression algorithm for prediction along with classification. Propos model reduce the time complexity and can be used by any employee of non-medical to predict heart disease.

[4] For classification of healthy subjects and heart disease, an Identification system has been proposed by applying Machine learning models. For selection of more appropriate attributes for improving the classification accuracy and to decrease the predictive system's computational time, Sequential backward selection of feature algorithm has been applied in this work.70% of Cleveland heart disease input is applied for training the model and 30% of input is used for validation. Evaluation metrics have been used for the measurement of performance of developed system.

In [5] Random Forest algorithm is applied for implement the system for finding possible diseases of heart. Detection of heart disease based on medical data collected from different patients is the main aim of this study. Diagnosis has been performed based on the information of clinical data and test results.Random Forest algorithm produced an overall accuracy of 84.448%.

[6] In order to keep track the current health status of patient , heart disease prediction system has been implemented in a real time. The contribution of this study is to optimal ML algorithm that produces the high accuracy. For selecting important attributes from dataset, univariate feature selection and Relief feature selection algorithms are applied. Compared algorithms in this study are DT, SVM, RF,LR.Cross-validation has been applied to improve the accuracy and hyper parameter tuning. The major strength of the developed system to manage data collected from twitter which includes patient's data. This is achieved by deploying Apache Kafka with Apache Spark as the base architecture of the model. Highest accuracy at 94.9% is obtained by Random forest classifier.

In [7] the main aim of the proposed novel method is to detect the significant attributes by employing ML approaches and also to improve the CAD accuracy. For model prediction various combinations of attributes and different classification techniques have been applied. And also improved the level of performance with an accuracy of 88% by predicting model by HRFLM algorithm.

In [8] , An automated decision support system has been implemented to detect the heart disease based on artificial neural network (ANN). Refinement of features and elimination of the problems posed by the predictive model, i.e., the problems of under fitting and over fitting are also focused here.F2 statistical

model has been proposed to eliminate irrelevant features and DNN model been proposed for classification By comparing its performance with conventional ANN and DNN models, f2 DNN has been evaluated. And finally 93.33% of accuracy is obtained by the proposed model.

[9] This research work aims to develop an improved fuzzy logic based artificial neural network classifier for Real time data are obtained and Several decision support systems are built for

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.227

diagnosing diseases among patients. In this research work the aim of the proposed IFANN classifier is to attain maximum prediction accuracy for CAHD among diabetic patients. Both male and female diabetic patient records are obtained from the reputed medical centers along with the class label of CAHD occurrence. The results are promising and it is inferred that 86.32% accuracy is obtained for male diabetic patients and 85.29% accuracy is obtained for female diabetic patients. Yet there is more scope for further improving the prediction accuracy and in the near future some optimization techniques are aimed to be built for attribute selection.

In [10 classification technique is performed on heart dataset downloaded from UCI library site. Feature selection techniques and K means and apriori algorithms are applied to produce high accuracy with four features for CAD analysis. Experimental outputs are generated by using MATLAB tool which proved that CART is more efficient than other methods even after applying FST as 84 % accuracy with 4 features only.

[11] analyzed the status of patient's nutrition depending on their intake of foods. A conceptual framework has been implemented with the use of RStudio with cross validation in order to compare and analyze the results. Predictive analysis on medical data is carried out in effective manner. For the collected data the proposed system first analyze the contents based on predefined features, then classify the data based on similarities and decision tree is constructed to provide efficient results. Finally predictive and performance analysis is carried out to produce the statistical report.

In [12] presented a classifier model for predicting heart disease on Cleveland dataset to assess the accuracy of model. BPNN and LR classification models are applied for the study. Un- biased estimate of this classification model is measured by 10-fold cross validation techniques. In order to determine the best classifier for the prediction of existence of heart disease, experimental results are conducted. The dataset contains 13 fields and 270 records which have been collected from Cleveland. BNN produced 85% of accuracy and LR produced 92% of accuracy.

In [13] presented a ML based system to predict to heart disease. Heart Disease Dataset has been collected from UCI repository. And finally combined and trained Cleveland, Hungarian and Switzerland datasets with the help of SVM. Proposed model achieved higher and better accuracy than other classification models that uses SVM or Naïve Bayes, with just one of the three UCI databases and without imputing missing values. Imputing missing values and using larger data than other models helped to obtain quite promising results in classifying the possible heart disease patient with an accuracy of 87% for SVM and 86% for Naïve Bayes. The purpose of our proposed technique is to achieve more accurate prediction of heart diseases using SVM and Naïve Bayes. The presented approach consists of combining multiple datasets to have more input data to train the model, then using imputation techniques with KNN algorithm to fill missing values. As a perspective, we intend to implement an optimized approach to increase the accuracy while training the model on a big dataset.

[14] Presented a ML based system to predict to heart disease. Heart Disease Dataset has been collected from UCI repository. And finally combined and trained Cleveland, Hungarian and

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.227

Switzerland datasets with the help of SVM, Naïve Bayes algorithms. Proposed model achieved higher and better accuracy than

other classification models that uses SVM or Naïve Bayes, with just one of the three UCI databases and without imputing missing values. Imputing missing values and using larger data than other models helped to obtain quite promising results in classifying the possible heart disease patient with an accuracy of 87% for SVM and 86% for Naïve Bayes. The purpose of our proposed technique is to achieve more accurate prediction of heart diseases using SVM and Naïve Bayes. The presented approach consists of combining multiple datasets to have more input data to train the model, then using imputation techniques with KNN algorithm to fill missing values. As a perspective, we intend to implement an optimized approach to increase the accuracy while training the model on a big dataset.

[15] To develop the system, 299 heart sounds from patients were obtained and labeled as normal and abnormal heart sound. Features were extracted and labeled as dataset; K Nearest Neighbor (KNN), Support Vector Machine (SVM) and Decision Tree (DT) algorithm were used as the training platform. From the classification analysis developed using the supervised ML trained model in MATLAB2018 in conjunction with system software features for the prediction of the heartbeat for both current and predefined of a heart condition algorithms used in training the dataset for the prediction when principle component analysis was enabled, the result shows that KNN algorithm has the highest and best accuracy of 94.4%, followed by the SVM with 84.4% and DT had 81.1%. While from the evaluation analysis, KNN on Receive Operation Characteristic Curve (ROC) with 90% variance and training time of 12.88 seconds on positive class of abnormal over false classes of normal heart sound has AUC as 0.94 and on ROC curve with PCA 90% variance. Hence the analysis from the result shows that out of the three classified algorithms used, KNN predicts and have the highest accuracy and is more efficient with respect to real-time environment.

From the classification analysis result that is developed using the supervised machine learning trained model in MATLAB2018 in conjunction with the system software for the prediction of the heartbeat for both current and predefined of a heart condition obtained from the different algorithms used in training the dataset for the prediction when principle component analysis (PCA) was enabled, that K Nearest Neighbour algorithm has the highest and best accuracy of 94.4% from 5 features out of 26, followed by the Support Vector Machine with 84.4% accuracy while Decision Tree with 81.1%.

Also from the evaluation analysis, it shows that KNN on ROC curve with 90% variance and training time of 12.88 seconds plotting positive class of abnormal over the false classes of normal heart sound has the AUC to be 0.94 while on ROC curve with PCA 90% variance and training time of 1.7119 seconds with 5 out of 26 features plotting positive class of normal over negative classes of abnormal heart sound has the AUC to be 0.89 efficiency.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.227

Table 2 shows comparative analysis of different techniques used in the existing work.

ML techniques	Author	Year	Data set	Tool	Accuracy
Hybrid (GA + SVM)	Tan et al .	2009	UCI	LIBSVM and WEKA	84.07%
SVM	Parthiban and Srivatsa	2012	Research Institute in Chennai	WEKA	94%
J48	Chaurasia, Pal	2013	UCI	WEKA	84.3%
Naïve Bayes	Vembandasamy et al.	2015	Diabetic Research Institute in Chennai	WEKA	86%
SVM	Otoom et al.	2015	UCI	WEKA	85%
SVM	Nimai Chand Das Adhikari	2017	unspecified	LIBSVM	80%
ANN, LR	Shrinivas D. Desai	2018	Cleveland HD dataset	WEKA	85%, 92%
ANN+DNN(x^2 *DNN)*	LIAQAT ALI	2019	UCI	WEKA	93%
HRFLM.	SENTHILKUMAR MOHAN	2019	UCI	WEKA	88.7%
Random Forest	Hager Ahmed	2019	Cleveland HD dataset	Apache Kafka	94.9%
KNN	Amin UlHaq	2019	Cleveland HD dataset	WEKA	90%
N2Genetic-nuSVM	MoloudAbdar	2019	UCI	WEKA	93%

Table 2: Analysis of various ML algorithms used in the existing work for heart disease prediction.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.227

4. Observations and Findings

From this study we come up with following observations that should be taken in consideration in future research work for high accuracy and more accurate diagnosis of heart disease by using intelligent prediction systems.

 In most experiments Small and same dataset has been used to train prediction models. So, we have to take real data in a large quantity of heart disease patients from reputed medical institutes of our country and use that data to train and test our prediction models. Then we have to examine the accuracy of our prediction models on large datasets.

 We have to consult highly experienced experts of cardiology to prioritize the attributes according to their effect on patient's health and also if necessary add more essential attributes of heart disease for more accurate diagnosis and high accuracy.

 There is need to develop more complex hybrid models for accurate prediction by integrating different techniques of data mining and machine learning and also include text mining of unstructured medical data available in large quantities in medical institutes. Also use of Genetic algorithm for optimization and feature selection make intelligent prediction models much better in overall performance.

 In this study we find more focus was given on classification techniques as compared to regression and association rule. So, for better comparative results in future research we have to take these things in our consideration.
 Accuracy of research is directly proportional to the selection of research tools and procedures. So, Choice of appropriate experimental tool (WEKA, METLAB etc.) for implementation of techniques is also an important parameter.

5. Conclusion

From the study of various recent research papers written on heart disease prediction .We find that different techniques of DM and ML are applied for heart disease predictionwith the help of different experimental tools such as WEKA, MATLAB etc. Different datasets of heart disease patients are used in different experiments.

References

[1].Nimai Chand Das Adhikari ,"hpps: heart problem prediction system using machine learning",doi : 10.5121/csit.2017.71803, cs& it-cscp 2017

[2] RajshN,Manesha" Prediction of Heart Disease Using Machine Learning Algorithms" *International Journal of Engineering & Technology, 7 (2.32) (2018) 363-366*

[3] Reddy Prasad, Pidaparthi Anjali "Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning", IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-3S, February 2019

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.227

[4] ” Heart Disease Prediction System Using Model Of Machine Learning and Sequential Backward Selection Algorithm for Features Selection”, 5th International Conference for Convergence in Technology (I2CT),Pune, India. Mar 29-31, 2019.

[5]. Ricardo Buettner ,” Efficient machine learning based detection of heart disease” Springer volume 3 ,issue2 2019.

[6] Hager Ahmed “Heart disease identification from patients’ social posts, machine learning solution on Spark” Elsevier 2019.

[7] SENTHILKUMAR MOHAN 1, CHANDRASEGAR THIRUMALAL ,” Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”,IEEE, VOLUME 7, 2019

[8]LIAQAT ALI , ATIQUR RAHMAN, “An Automated Diagnostic System for Heart Disease Prediction Based on _2 Statistical Model and Optimally Configured Deep Neural Network”,IEEE, VOLUME 7, 2019.

[9] B. Narasimhan.” IMPROVED FUZZY ARTIFICIAL NEURAL NETWORK (IFANN) CLASSIFIER FOR CORONARY ARTERY HEART DISEASE PREDICTION IN DIABETES PATIENTS”, Volume-9 | Issue-4 | April-,IEEE journal 2019

[10] Hadia Amin, Abita Devi, NidaUl Amin ,“PREDICTIVE ANALYSIS OF HEART DISEASE USING K-MEANS AND APRIORI ALGORITHMS”, IEEE journal ,Volume VI, Issue VI, JUNE 2019.

[11] RoohallahAlizadehsani, MoloudAbdar,” Machine Learning based coronary Artery Disease Diagnosis: A Review”, https:// doi.org/10.1016/j.compbiomed.., Elsevier june 2019

[12] Shrinivas D. Desai, ShantalaGiraddi, PrashantNarayankar” Back-Propagation Neural Network Versus Logistic Regression in Heart Disease classification” https://doi.org/10.1007/978-981-13-0680-8_13,sprinzer,july 2019

[13] FadouaKhennou, CharifFahim **,“**A Machine Learning Approach: Using Predictive Analytics to Identify and Analyze High Risks Patients with Heart Disease” *IEEE Journal of Machine Learning and Computing, Vol. 9, No. 6, December 2019.*

[14]MoloudAbdar ,WojciechKsi , ˙zek**”** A new machine learning technique for an accurate diagnosis of coronary artery disease”, https://doi.org/10.1016/j.cmpb.2019.104992, Elsevier 2019

[15] Ozichi N. Emuoyibofarhe, Segun Adebayo” Predictive System for Heart Disease Using a Machine Learning Trained Model”, *IJC 2019,Volume 34, No 1.*

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.227

Author's Profile

Prabhavathi K, received her B.E, M.Tech degree specialized in Computer Science and Engineering. Currently she is pursuing her Ph.D degree from Visvesvaraya Technological University, Belgaum. Currently she is working as Assistant Professor in Department of Computer Science and Engineering, Amruta Institute of Engineering and Management Sciences Ramanagara, Karnataka. And having 6 years of experience in teaching. Her research interests include artificial intelligence, machine learning and big data.

Dr.VMareeswarireceived his B.E, M.E and Ph.D degree specialized in Computer Science and Engineering. She is a member of various professional associations and has 16 years of experience. Currently he is working as Associate Professor and Head in Department of Computer Science and Engineering, ACS College of Engineering, India. Her research interests include Information Retrieval, Data mining