EFFICIENT CLASSIFICATION TECHNIQUES FOR HIGH DIMENSIONAL TRADITIONAL CHINESE MULTI-LABEL MEDICAL DATA

A PROJECT REPORT

Submitted by

PUVIYARASI K (810015104071)

LAKSHMI PRIYA S (810015104723)

In partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

UNIVERSITY COLLEGE OF ENGINEERING – BIT CAMPUS

TIRUCHIRAPPALLI - 620 024

ANNA UNIVERSITY : CHENNAI 600 025

APRIL 2019

UNIVERSITY COLLEGE OF ENGINEERING, BIT CAMPUS
TIRUCHIRAPPALLI-620 024

BONAFIDE CERTIFICATE

Certified that this project report **"EFFICIENT CLASSIFICATION TECHNIQUES FOR HIGH DIMENSIONAL TRADITIONAL CHINESE MULTI-LABEL MEDICAL DATA"** is the bonafide work of **K.PUVIYARASI (810015104071)** and **S.LAKSHMI PRIYA (810015104723)** who carried out the project work under my supervision.

SIGNATURE
Dr. D. Venkatesan
HEAD OF THE DEPARTMENT
Assistant Professor
Computer Science and Engineering
University College of Engineering
BIT Campus
Triuchirappalli – 620 024

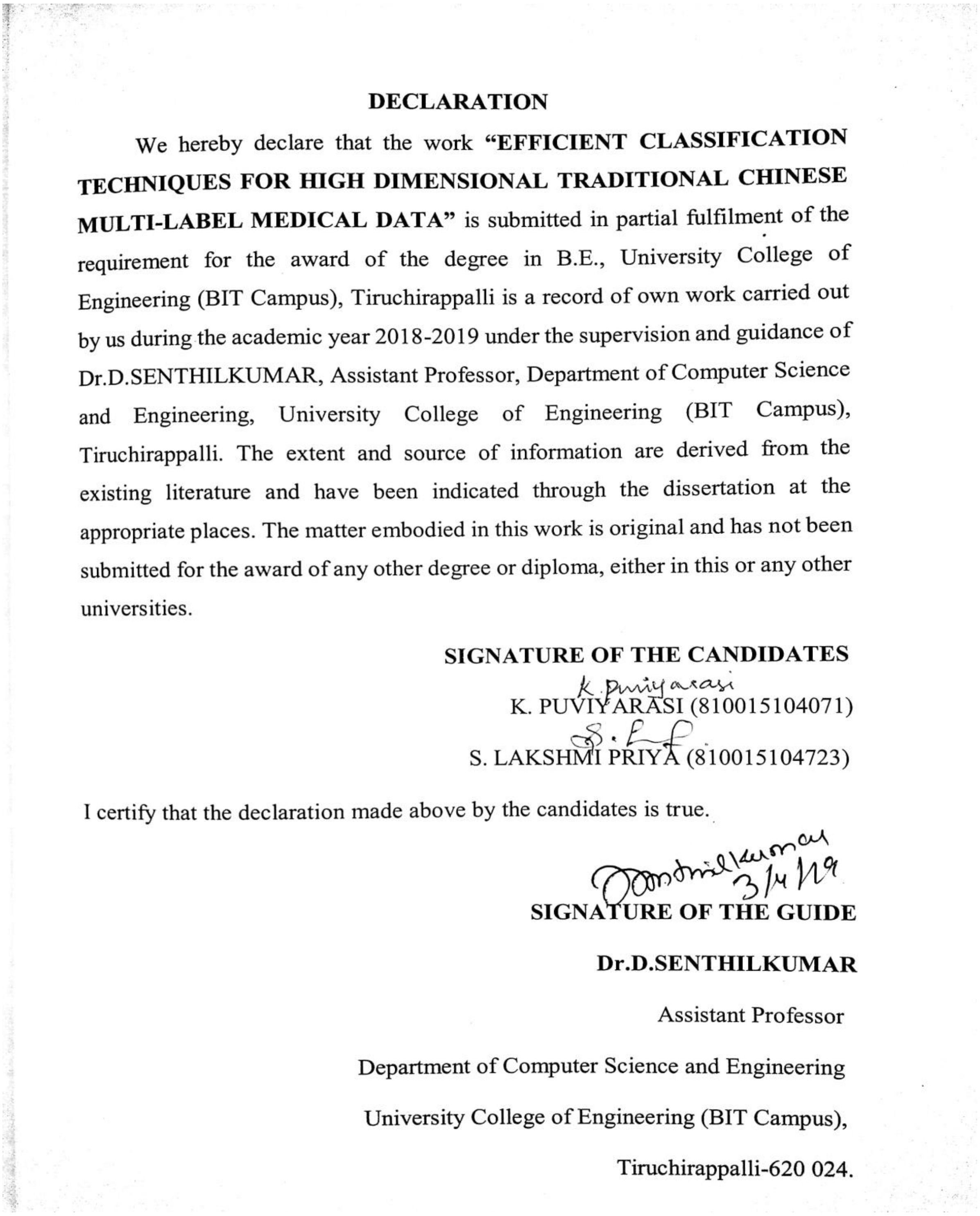
Senthilkumar
3/4/19

SIGNATURE
Dr. D. Senthilkumar
PROJECT GUIDE
Assistant Professor
Computer Science and Engineering
University College of Engineering
BIT Campus
Triuchirappalli – 620 024

Certified that **K.PUVIYARASI** and **S.LAKSHMI PRIYA** was examined in the Project Viva Voce examination held on ______________________

Internal Examiner

External Examiner



DECLARATION

We hereby declare that the work **"EFFICIENT CLASSIFICATION TECHNIQUES FOR HIGH DIMENSIONAL TRADITIONAL CHINESE MULTI-LABEL MEDICAL DATA"** is submitted in partial fulfilment of the requirement for the award of the degree in B.E., University College of Engineering (BIT Campus), Tiruchirappalli is a record of own work carried out by us during the academic year 2018-2019 under the supervision and guidance of Dr.D.SENTHILKUMAR, Assistant Professor, Department of Computer Science and Engineering, University College of Engineering (BIT Campus), Tiruchirappalli. The extent and source of information are derived from the existing literature and have been indicated through the dissertation at the appropriate places. The matter embodied in this work is original and has not been submitted for the award of any other degree or diploma, either in this or any other universities.

SIGNATURE OF THE CANDIDATES

k. puviyarasi

K. PUVIYARASI (810015104071)

S. L

S. LAKSHMI PRIYA (810015104723)

I certify that the declaration made above by the candidates is true.

Senthilkumar 3/4/19

SIGNATURE OF THE GUIDE

Dr.D.SENTHILKUMAR

Assistant Professor

Department of Computer Science and Engineering

University College of Engineering (BIT Campus),

Tiruchirappalli-620 024.

ACKNOWLEDGMENT

We would like to thank the Almighty for all the blessings he bestowed on us, which drove us to the successful completion of this project.

We would like to extent our heartfelt gratitude to our respected Dean of Anna University-BIT Campus, Tiruchirappalli **Prof. Dr. T. SENTHILKUMAR,** who is the guiding light for all the activities in our college.

We would like to express our special thanks to our beloved Head of the Department **Dr. D. VENKATASAN,** Head of Department/CSE for his kind guidance towards the success of this project.

We would like to thank and express our deep sense of gratitude to our project Guide **Dr. D. SENTHILKUMAR,** Assistant Professor, Department of Computer Science and Engineering, for his valuable guidance, encouragement and constant support throughout our work.

We also thank all the teaching and non-teaching staffs of the Department of CSE, our beloved parents and friends, for their help and support to complete our project successfully.

PUVIYARASI K
LAKSHMI PRIYA S

ABSTRACT

TCM (Traditional Chinese medicine) is a old style of medicine depend on more than 3000 years of Chinese medical field practice that contain different forms like massage, dietary therapy, exercise and acupuncture. One of the elementary comments of TCM is that the body's important energy circulates via channels known as meridians. In TCM diagnosis, a patient may be affected with more than one syndrome labels and its computer aided diagnosis is a distinctive application in the domain of multi-label learning of high dimensional data. TCM datasets are mostly multi-label. Multi-label classification was mainly motivated by the task of text categorization and medical diagnosis in the part. In classification, a dataset is said to be imbalanced when the number of cases which represents one class is much smaller than the ones from other class. This project proposes to machine learning algorithm, random forest and GBM. To reduce the imbalance for the dataset four different models are generated (Balanced + Not stratified, Balanced + Stratified, Unbalanced + Not Stratified, Unbalanced + Stratified) using random forest. Comparing the four different models the Unbalanced + Not Stratified and Unbalanced + Stratified is the best model with the random forest. The best model of Random Forest is compared with Gradient boosting Machine. From the experimental results Gradient Boosting Machine gives the better performance accuracy in terms of imbalanced error.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

TCM	Traditional Chinese Medicine
MLD	Multi Label Dataset
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
CF	Chronic Fatigue
GBM	Gradient Boosting Machine
RF	Random Forest
MLL	Multi Label Learning
CP	Conformal Predictor
PT	Problem Transformation
NBC	Naive Bayes Classifier
KNN	K Nearest Neighbours
HC	Hypertensive Crisis
CHD	Coronary Heart Disease
RAD	Relative Associated Density
MSE	Mean Square Error
RMSE	Root Mean Square Error

CHAPTER 1

INTRODUCTION

1.1 Overview

Data mining is the process of reduce the large data sets to identify patterns to solve problems through data analysis method. Data mining is also called as data analysis. Data mining contains various techniques are Classification, clustering, Regression, Association rules, Outer detection, Sequential Patterns, and prediction.

Classification techniques are used to recover important and applicable information about data, and Meta data. It helps to classify data in multiple data. This project mainly used to Multi-label classification techniques in Machine learning. In these techniques used to make efficient multi label classification.

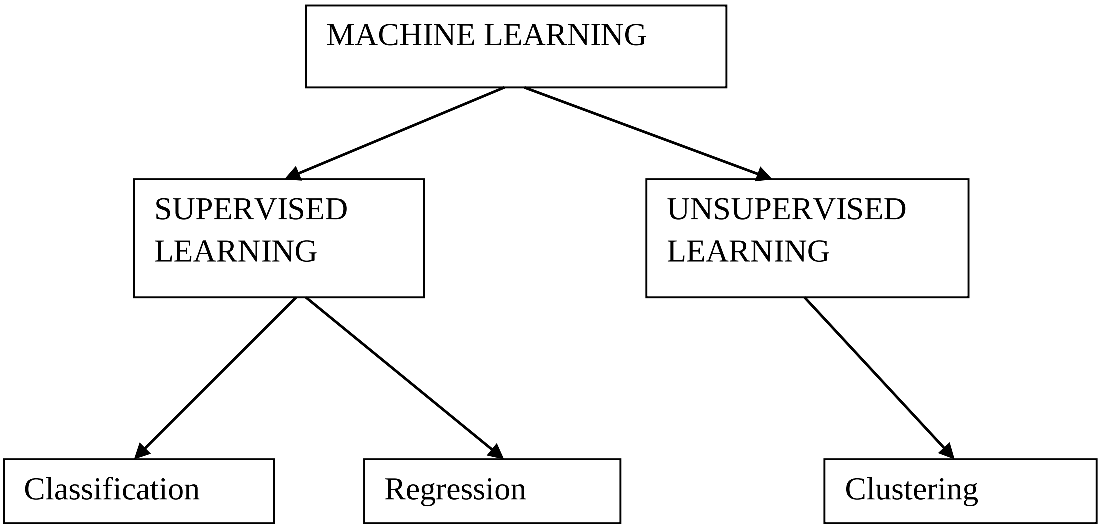


Figure 1.1 Data mining types

1.2 Traditional Chinese Medicine (TCM)

Traditional Chinese medicine (TCM) is a style of traditional medicine based on more than 2500 years of Chinese medical practice that includes various forms such as herbal medicine, acupuncture, massage, exercise and dietary therapy. It is ancient treatment of the world but recently introduced modern western medicine. Traditional Chinese medicine is One of the basic tenets of TCM is that the body's vital energy circulates through channels called meridians. Meridians have branches connected to bodily organs and functions. "Traditional Chinese" is used to contrast traditional characters with Simplified Chinese characters, a standardized character set introduced by the government of the People's Republic of China. According to TCM, illness arises as a result of specific disease *imbalances* of the *Functional Entities*. If there is an imbalance within the any of the functional entities, it will not be able to perform their cardinal functions, and as result, illness may arise.

1.3 Multi-Label Dataset (MLD)

Multi-label dataset means that each instance can be labelled with anywhere from a single to labels, where is the total number of different labels in the dataset. This information represented by binary matrix, where $M\,ij$=1 if instance has a label, and otherwise $Mij = 0$.the problem of classifying instances into one of three or more classes in multiclass classification. Mostly medical data are multi data.

1.4 Multi-label Classification

Multi-label Classification and the strongly related problem of multi-output classification are variants of the classification problem where multiple labels may be assigned to each instance. Multi-label classification is a generalization of multi-purpose classification. Multi-class classification is classifying instances into one of three or more classes.

Multi class classification problem by dividing the output in a tree. Each parent node is divided into multiple child nodes and each child node represents only one class. Multi label classifications contains more data set are roughly equal classes, so one class exceeds another class so, class imbalance occur in the Multi label data set.

1.5 Class Imbalance

Class Imbalance is defined as total number of class of data is less than the total number of another class or total number of positive data less than the total number of negative data. Many classification learning algorithms have low predictive accuracy for the infrequent class. Data are said to suffer the Class Imbalance Problem when the class distributions are highly imbalanced.so, introduced the concept of True Positive, True Negative, False Positive and False Negative.

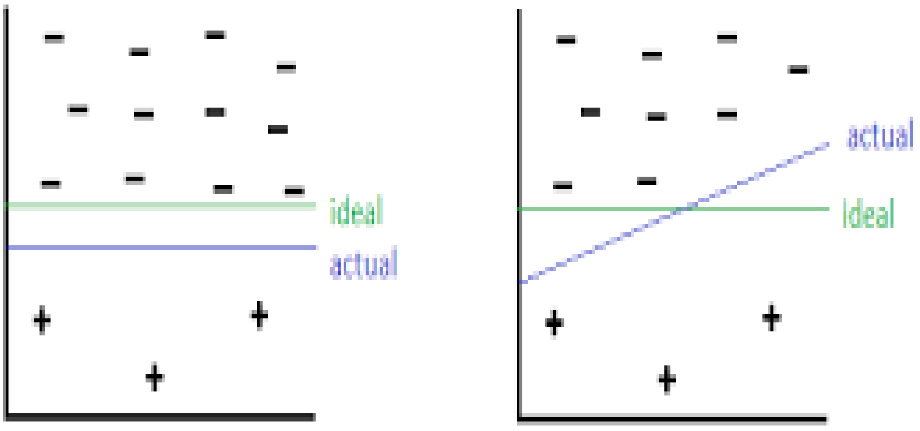


Figure 1.2 Class Imbalance

1.5.1TRUE POSITIVE (TP)

True positive example is positive and is classified correctly as **positive.**

TRUEPOSITIVE (TP) RATE=TP/(TP+FP)

The closer to 1, the better.TP Rate=1 when FP=0.no false positives.

1.5.2 TRUE NEGATIVE (TN)

True negative example is negative and is classified correctly as **negative.**

TRUE NEGAVTIVE (TN) Rate=TN/(TN+FN)

The closer to 1, the better.TN Rate =1 when FN=0.no false negatives.

1.5.3 FALSE POSITIVE (FP)

False negative example is **negative** but is classified wrongly as **positive.**

FALSE POSITIVE (FP) Rate=FP/(FP+TN)

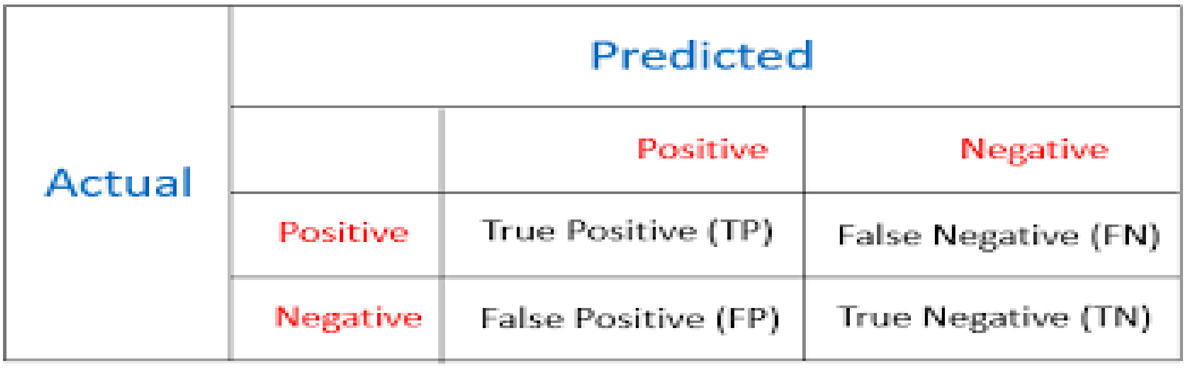
The closer to 0, the better. FP Rate = 0 when FP=0.no false positives.

1.5.4 FALSE NEGATIVE (FN)

False Negative example is positive but is classified wrongly as negative**.**

FALSE NEGATIVE (FN) Rate=FN/(FN+TP)

The closer to 0, the better. FN Rate =0 when FN=0.no false negatives.



		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 1.1 Evaluation metrics

1.6 Motivation of this project

Data and Information or Knowledge has a significant role on human activities. Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspectives and summarizing it into useful information. Due to the importance of extracting knowledge/information from the large data repositories, data mining has become an essential component in various fields of human life including business, education, medical and scientific. Data mining contains various techniques are classification, clustering, regression, association rules, outer detection, sequential pattern and prediction Classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation

1.7 Objective of this project

Chronic Fatigue (CF) still remains unclear about its diagnostic criteria in the medical community. Traditional Chinese medicine (TCM) adopts a unique diagnostic method, namely syndrome differentiation, to diagnose the CF with a set of syndrome factors, which can be regarded as the Class imbalance problem. To obtain an effective and reliable diagnostic tool, we use Gradient Boosting Machine (GBM) and Random Forest (RF) for the syndrome differentiation of CF.

1.8 Contribution of this project

In this project, the 95 input variables is classify the output variables 4. This project rectify the class imbalance problem using multi label classification method. Finally, the proposed approach is compared with some popular methods, such as Gradient Boosting Machine (GBM) and Random forest. The experimental results shows that the proposed multi label model has the best classification ability.

1.9 Structure of this project

- Chapter 1 starts with a brief introduction about the TCM, multi label classification, class imbalance and classification
- Chapter 2 briefly discuss the literature review related to this project.
- Chapter 3 shows the proposed technique in this project. The proposed technique describes about the modules used here.
- Chapter 4 presents the software and hardware require in this project.
- Chapter 5 discuss the result obtained from the experiment. Then it compares the various existing algorithms proposed technique.
- Finally this project end with a comprehensive summary, conclusion and result of this project.

CHAPTER 2

LITERATURE SURVEY

This chapter discuss the literature review related to current work. Initially the techniques already used for the TCM data is discussed. Then the review about multi label approach is discussed.

Guo et al [1] presented "Patient classification of hypertension in Traditional Chinese Medicine using multi-label learning techniques". Hypertension is the important risk aspects for cardiovascular diseases. In this work, the BrSmoteSvm had a better performance to compare various multi-label classifiers in the estimation measures of Average precision, Coverage, One-error, Ranking loss. BrSmoteSvm can model the hypertension's syndromes differentiation better performance of imbalanced problem. Advantage of BrSmoteSvm it helps to run time and storage problems by decreasing the number of training data samples when training dataset is large. Disadvantage of BrSmoteSvm it can remove potentially useful information.

Wan et al [2] presented "Reliable Multi-Label Learning via Conformal Predictor and Random Forest for Syndrome Differentiation of Chronic Fatigue in Traditional Chinese Medicine". In this work, the Chronic Fatigue (CF) still remains unclear about its diagnostic measures in the medical communal. To achieve an efficient diagnostic tool, use Conformal Predictor (CP), Random Forest (RF) and Problem Transformation method (PT) for the syndrome differentiation of Chronic fatigue. In this work, using PT method, CP-RF is prolonged to manage MLL problem. CP-RF relates RF to measure the level (p-value) of every label being the true label, and picks multiple labels whose p-values are greater than the pre-defined significance level as the region prediction. In this study, CP-RF are compare with CP-NBC(Naive Bayes Classifier), CP-KNN(K-Nearest Neighbours) and ML-KNN on CF dataset. CP-RF demonstrates an outstanding performance beyond CP-NBC, CP-KNN and ML-KNN under the general metrics of subset accuracy, hamming loss, one-error, coverage, ranking loss and average precision. Furthermore, the performance of CP-RF remains steady at the large scale of confidence levels from 80% to 100%, which indicates its robustness to the threshold determination. It yields a efficient accurate classifier and learning is fast. No understandable.

Fan et al [3] presented “Using Random Forest for reliable classification and cost-sensitive learning for medical diagnosis”. Most machine-learning classifiers output label estimates for new examples without demonstrating how reliable the predictions. The applicability of these classifiers is limited in critical domains where incorrect predictions have serious significances, like medical diagnosis. Further, the default hypothesis of equal misclassification costs is most likely violated in medical diagnosis. The process of using RF outlier amount to design a unconventionality measure benefits the subsequent predictor. The output of minimizing the risk of misclassification is attained by agreeing the different confidence level for different class. Dominant and accurate. Over fitting can simply occur.

Martin et al [4] presented “Hypertensive crisis: clinical epidemiological profile”. Hypertensive crisis is a kind of blood pressure (BP) and it can apparent as hypertensive emergency usually attended by levels of diastolic BP≥120 mmHg. There was to illustrate the clinical-epidemiological shape of HC over the way of 2 year in a university condition hospital and achieve a analysis of the works. HC is a clinical object related with high illness in the spare room. Entities with HE is mature and inactive and have lower rates of antihypertensive treatment. Suitable control of BP should be followed as a way to avoid this severe problem of hypertension.

Liu et al [5] Intimated “Modelling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi label learning”. Coronary heart disease (CHD) is a common cardiovascular disease that is extremely harmful to humans. In Traditional Chinese Medicine (TCM), the diagnosis and treatment of CHD have a long history and ample experience. Standardization scale on inquiry diagnosis for CHD in TCM is designed, and the inquiry diagnostic model is constructed based on collected data by the MLL techniques. A total of 555 cases are collected for the modelling of inquiry diagnosis of CHD. The patients are diagnosed clinically by fusing inspection, pulse feeling, palpation and the standardized inquiry information. Models of six syndromes are constructed by ML-kNN, RankSVM, BPMLL and kNN, whose mean results of accuracy of diagnosis reach 77%, 71%, 75% and 74% respectively. After removing symptoms of low frequencies, the mean accuracy results of modelling by ML-kNN, RankSVM, BPMLL and kNN reach 78%, 73%, 75% and 76% when 52 symptoms are remained. In this project MLL techniques facilitate building standardized inquiry models in CHD diagnosis and show a practical approach to solve the problem of labelling multi-syndromes simultaneously.

Li et al [6] presented “Inquiry diagnosis of coronary heart disease in Chinese medicine based on symptom-syndrome interactions”. It labelled a long history of coronary heart disease (CHD) diagnosis and treatment in

Chinese medicine (CM), but a formalized description of CM knowledge is still unavailable. This work analysed a set of CM clinical data, which is important and urgent. Relative associated density (RAD) was used to analyse the one-way links between the symptoms or syndromes or both. RAD results were further used in symptom selection. Using RAD to select symptoms based on different classifiers improved the accuracy of syndrome prediction. Compared with other traditional symptom selection methods, RAD provided a higher interpretability of the CM data. The RAD method is effective for CM clinical data analysis, particular for analysis of relationships between symptoms in diagnosis and generation of compact and comprehensible symptom feature subsets.

Poon et al [7] presented “A novel approach in discovering significant interactions from TCM patient prescription data”. The efficiancy of a traditional Chinese medicine medication derives from the complex interactions of herbs or Chinese Physical Medical in a formula. It offers a new approach to systematically generate combinations of interacting herbs that might lead to good outcome. This was tested on a data set of prescriptions for diabetic patients to verify the effectiveness of detected combinations of herbs. It presents an exploratory analysis of clinical records using a pattern mining approach called Interaction Rules Mining.

Wang et al [8] presented “ Multi label learning via random label selection for protein subcellular multi location prediction”. It described Prediction of protein subcellular localization is an important but challenging problem, particularly when proteins may simultaneously exist at, or move between, two or more different subcellular location sites. Most of the existing protein subcellular localization methods are only used to deal with the single-location proteins. In the past few years, only a few methods have been proposed to tackle proteins with multiple locations. This method named as a random label selection (RALS) (multi label learning via RALS), which extends the simple binary relevance (BR) method, is proposed to learn from multi location proteins in an effective and efficient way. Experimental results on two benchmark data sets also show this methods achieve significantly higher performance than some other state-of-the-art methods in predicting subcellular multi location of proteins.

Galar et al [9] presented "A review on ensembles for the class imbalance problem bagging, boosting, and hybrid based approaches". Classifier learning with datasets that suffer from imbalanced class distributions is a challenging problem in data mining community. This issue occurs when the number of examples that represent one class is much lower than the ones of the other classes. Its presence in many real-world applications has brought along a growth of attention from researchers. In machine learning, the ensemble of classifiers are known to increase the accuracy of single classifiers by combining several of them, but neither of these learning techniques alone solve the class imbalance problem, to deal with this issue the ensemble learning algorithms have to be designed specifically. In addition, they develop a thorough empirical comparison by the consideration of the most significant published approaches, within the families of the taxonomy proposed, to show whether any of them makes a difference. In this project, the results show empirically that ensemble-based algorithms are worthwhile since it outperform the use of pre-processing techniques before learning the classifier, therefore justifying the increase of complexity by means of a significant enhancement of the results.

Chawla et al [10] presented “Special issue on learning from imbalanced data sets”. There have been many skewed cancer gene expression datasets in the post-genomic era. Extraction of differential expression genes or construction of decision rules using these skewed datasets by traditional algorithms will seriously underestimate the performance of the minority class, leading to inaccurate diagnosis in clinical trails. This presented a skewed gene selection algorithm that introduces a weighted metric into the gene selection procedure. The extracted genes are paired as decision rules to distinguish both classes, with these decision rules then integrated into an ensemble learning framework by majority voting to recognize test examples; thus avoiding tedious data normalization and classifier construction. The mining and integrating of a few reliable decision rules gave higher or at least comparable classification performance than many traditional class imbalance learning algorithms on four benchmark imbalanced cancer gene expression datasets.

Khalida et al [11] presented “An analysis of ambulatory blood pressure monitoring using multi-label classification”. Ambulatory blood pressure monitoring (ABPM) involves measuring blood pressure by means of a tensiometer carried by the patient for a duration of 24 h, it currently occupies a central place in the diagnosis and follow-up of hypertensive patients, it provides crucial information which allows to make a specific diagnosis and adapt therapeutic attitude accordingly. The traditional analysis process suffers from different problems. This project attempted to improve the analysis of ABPM data using multi-label classification methods, where a record is associated with more than one label (class) at the same time. Seven algorithms are experimentally compared on a new multi-label ABPM-dataset. Experiments are conducted on 270 hypertensive patient records characterized by 40 attributes and associated with six labels. Results show that the multi-label modelling of ABPM data helps to investigate label dependencies and provide interesting insights, which can be integrated into the ABPM devices to dispense automatically detailed reports with possible future complications.

Miao et al [12] presented “Syndrome differentiation in modern research of traditional Chinese medicine” Syndrome differentiation (Bian Zheng) in traditional Chinese medicine (TCM) is the comprehensive analysis of clinical information gained by the four main diagnostic TCM procedures: observation, listening, questioning, and pulse analysis, and it is used to guide the choice of treatment either by acupuncture and/or TCM herbal formulae, that is, Fufang. TCM syndrome differentiation can be used for further stratification of the patients’ conditions with certain disease, identified by orthodox medical diagnosis, which could help the improvement of efficacy of the selected intervention. In modern TCM research it is possible to integrate syndrome differentiation with orthodox medical diagnosis leading to new scientific findings in overall medical diagnosis and treatment. In this review, the focus is to screen published evidence on the role of syndrome differentiation in modern TCM research with particular emphasis on basic and clinical research as well as, pharmacological evaluation of TCM herbal formulary for drug discovery.

CHRIS et al [13] presented “Minimum Redundancy Feature Selection from microarray gene expression data” How to selecting a small subset out of the thousands of genes in microarray data is important for accurate classification of phenotypes. Widely used methods typically rank genes according to their differential expressions among phenotypes and pick the top-ranked genes. We observe that feature sets so obtained have certain redundancy and study methods to minimize it. In this project, a minimum redundancy maximum relevance (MRMR) feature selection framework. Genes selected via MRMR provide a more balanced coverage of the space and capture broader characteristics of phenotypes. It lead to significantly improved class predictions in extensive experiments on 6 gene expression data sets: NCI, Lymphoma, Lung, Child Leukemia, Leukemia, and Colon. Improvements are observed consistently among 4 classification methods: Naive Bayes, Linear discriminant analysis, Logistic regression, and Support vector machines.

Jiang et al [14] presented "An improved *K*-nearest-neighbour algorithm for text categorization Ext categorization is a significant tool to manage and organize the surging text data". Many text categorization algorithms have been explored in previous literatures, such as KNN, Naive Bayes and Support Vector Machine. KNN text categorization is an effective but less efficient classification method. In this paper, an improved KNN algorithm for text categorization, which builds the classification model by combining constrained one pass clustering algorithm and KNN text categorization. Empirical results on three benchmark corpora show that the algorithm can reduce the text similarity computation substantially and outperform the-state-of-the-art KNN, Naive Bayes and Support Vector Machine classifiers. In addition, the classification model constructed by this algorithm can be updated incrementally, and it has great scalability in many real-word applications.

CHAPTER 3

PROPOSED SYSTEM

Chronic Fatigue syndrome (CFS) is a complicated disorder characterized by extreme fatigue that can't be explained by any underlying medical condition. The patient classification of Chronic fatigue syndrome has become an important topic because Traditional Chinese Medicine lies primarily in “treatment based on syndromes differentiation of the patients”. Syndromes differentiation was modeled as a patient classification problem in the field of data mining. This scheme proposes efficient classification techniques for high – dimensional traditional Chinese multi-label medical data. The GBM and Random forest classifier algorithm used here will improve the performance and accuracy of existing evaluation criteria.

3.1 SYSTEM ARCHITECTURE

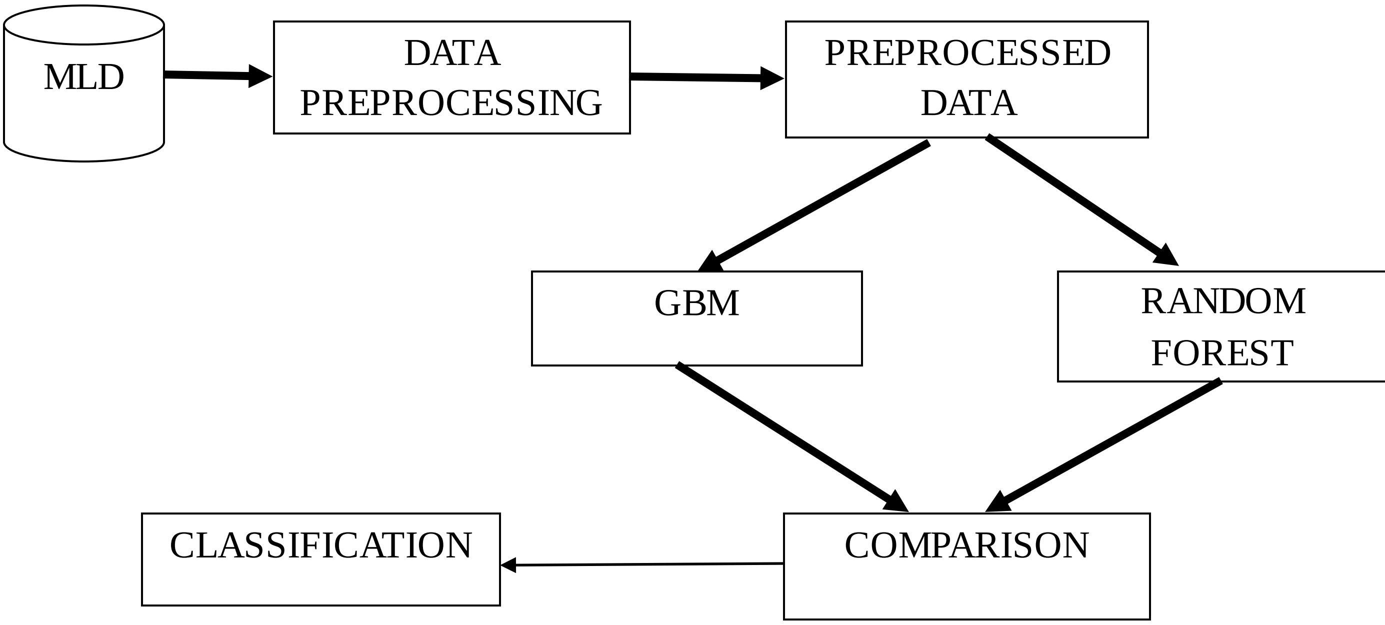


Figure 3.1 Architecture of the proposed system

3.2 MODULE 1 Dataset Collection

Dataset collection is the process which plays an important role in projects and researches. Data is a piece of information that should be collected carefully so that it is useful. Data collection is a process of gathering of information used for various processes in the project such as preprocessing, predictions. In our proposed scheme the dataset to be collected is a medical dataset. Collecting data from and on behalf of medical patients is a critical component of healthcare, particularly when that data needs to be analysed to provide the best and most proper care. The collected dataset consist of patient details suffered from disease named Chronic Fatigue (CF).This dataset is used for the next step that is data preprocessing.

Multi-label dataset means that each instance can be labelled with anywhere from a single to labels, where is the total number of different labels in the dataset. This information represented by binary matrix, where $M ij$=1 if instance has a label, and otherwise $Mij = 0$.the problem of classifying instances into one of three or more classes in multiclass classification. Mostly medical data are multi data.

CF dataset, which consists of 736 cases. Specifically, 95 symptoms are used to identify CF, and four syndrome factors are employed in the syndrome differentiation, including spleen deficiency, heart deficiency, liver stagnation and qi deficiency.

3.3 MODULE 2 Data Pre processing

The size of medical datasets is usually very large, which directly affects the computational cost of the data mining process. Feature selection is a data preprocessing step in the knowledge discovery process, which can be employed to reduce storage requirements. The objective data preprocessing is to filter out outliers (or noisy data) from a given (training) dataset. However, when the dataset is very large in size, more time is required to accomplish the feature selection process. This scheme uses a tool called WEKA which converts the dataset of .arff format to .csv file so as to reduce the possibilities of data loss. Further data preprocessing is automatically done by the algorithms used in our scheme which are explained in the following section.

3.4 MODULE 3 Ensembles

While predicting the target variable using any machine learning technique, the main causes of difference in actual and predicted values are **noise, variance, and bias**. An ensemble is just a collection of predictors which come together to give a final prediction. The reason for using ensembles is that many different predictors trying to predict same target variable will perform a better job than any single predictor alone. Ensembling techniques are further classified into Bagging and Boosting. **Bagging** is a simple ensembling technique that builds many *independent* predictors or models and combines them using some model averaging techniques, example Random Forest. Boosting is an ensemble technique in which the predictors are not made independently, but sequentially, example Gradient Boosting Machine. The following Figure 3.2 shows the classification of Ensemble technique:

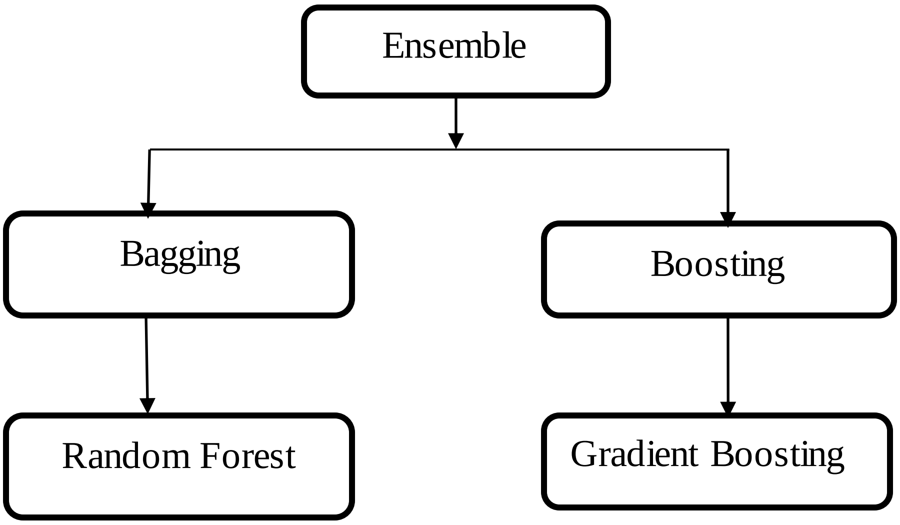


Figure 3.2 Classification of Ensemble

3.5 MODULE 4 Gradient Boosting Machine

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Gradient boosting involves three elements: a loss function to be optimized, a weak learner to make predictions, an additive model to add weak learners to minimize the loss function. Most common form of GBM that optimizes the mean squared error (MSE), also called the L_2 loss or cost. The mean squared error is the average of the square of the difference between the true targets and the predicted values from a set of observations, such as training or validation set. Many algorithms, focus on minimizing the residuals and, therefore, emphasize the MSE loss function. The algorithm discussed in the previous section outlines the approach of sequentially fitting regression trees to minimize the errors. This specific approach is how gradient boosting minimizes the mean squared error (MSE) loss function. However, often we wish to focus on other loss functions such as mean absolute error (MAE) or to be able to apply the method to a classification problem with a loss function such as deviance. The name gradient boosting machines come from the fact that this procedure can be generalized to loss functions other than MSE.

Gradient boosting is considered a gradient descent algorithm. Gradient descent is a very generic optimization algorithm capable of finding optimal solutions to a wide range of problems. The general idea of gradient descent is to tweak parameters iteratively in order to minimize a cost function. Suppose you are a downhill skier racing your friend. A good strategy to beat your friend to the bottom is to take the path with the steepest slope. This is exactly what gradient descent does - it measures the local gradient of the loss (cost) function for a given set of parameters and takes steps in the direction of the descending gradient. Once the gradient is zero, we have reached the minimum.

Fit a model to the data, $F_1(x) = y$

Fit a model to the residuals, $h_1(x) = y - F_1(x)$

Create a new model, $F_2(x) = F_1(x) + h_1(x)$

It's not hard to see how we can generalize this idea by inserting more models that correct the errors of the previous model. Specifically,

$$F(x) = F_1(x) \mapsto F_2(x) = F_1(x) + h_1(x) \ldots \mapsto F_M(x) = F_{M-1}(x) + h_{M-1}(x)$$

Where $F_1(x)$ is an initial model fit to y

Since we initialize the procedure by fitting $F_1(x)$ our task at each step is to find

$h_m(x) = y - F_m(x).$

3.5 MODULE 5 Random Forest (RF)

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees. In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results. Random forest undergo cross validation is a powerful tool that is used for estimating the predictive power of the model, and it performs better than the conventional training and test set. Using cross validation, it is possible to create multiple training and test sets and average the scores to give us a less biased metric. In our scheme we use four different models for each labels assigned in the dataset and the models are as follows:

M1= Balanced + Not stratified

M2= Balanced + Stratified

M3= Unbalanced + Not stratified

M4= Unbalanced + Stratified

Here the word Stratified specifies an attempt to evenly distribute observation from the different classes to all sets when splitting dataset into training and validation.

3.6 MODULE 6 Classification

Classification techniques in data mining are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data. The term could cover any context in which some decision or forecast is made on the basis of presently available information. Classification procedure is recognized method for repeatedly making such decisions in new situations. Classification techniques predict discrete responses for example, whether an email is genuine or spam, or whether a tumor is cancerous or benign. Classification models classify input data into categories. Typical applications include medical imaging, speech recognition, and credit scoring. In our proposed scheme we use boosting and bagging algorithm for classification as mentioned previously.

CHAPTER 4

REQUIREMENT SPECIFICATIONS

4.1HARDWARE AND SOFTWARE SPECIFICATION

4.1.1 HARDWARE REQUIREMENTS

RAM	: 2 GB
Hard Disk	: 500 GB
Processor	: Intel processor
Monitor	: 15’ LCD Monitor

4.1.2 SOFTWARE REQUIREMENTS

Front End	: Java
Tools	: R, R studio
Back End	: MS-Excel
Operating System	: Windows 10 pro

CHAPTER 5

EXPREMENTAL RESULTS

The data used in the proposed scheme is obtained from TCM type of medical dataset. The work progress through the following stages; the preprocessor, WEKA, GBM and Random Forest classifiers. The preprocessing stage involves rectification of training dataset to avoid noises in the dataset. The WEKA tool is used for converting ARFF to CSV format so that it can reduce the possibilities of data loss. The following Table 5.1 shows the common measures of GBM in which the MSE, RMSE and Log loss.

Mean Squared Error (MSE), which measures the average error performed by the model in predicting the outcome for an observation. Mathematically, MSE is the average squared difference between the observed actual outcome values and the values predicted by the model. So, MSE = mean ((observations predictions) ^ 2). The lower value of MSE is better model.

Root Mean Squared Error (RMSE), which measures the average error performed by the model in predicting the outcome for an observation. Mathematically, the RMSE is the square root of the mean squared error (MSE), RMSE= sqrt (MSE).The lower value of RMSE is better model.

Log loss (Logarithmic loss) measures the performance of a classification model where the prediction input is a probability value between 0 and 1. Log loss increases as the predicted probability diverges from the actual label

GBM

	T1	T2	T3	T4
MSE	0.000217525	0.000530746	0.01488992	1.017642
RMSE	0.01474872	0.2303793	0.1220243	0.00319005
Log loss	0.006491229	0.006329349	0.04517508	0.003036938

Table 5.1 GBM for four Targets

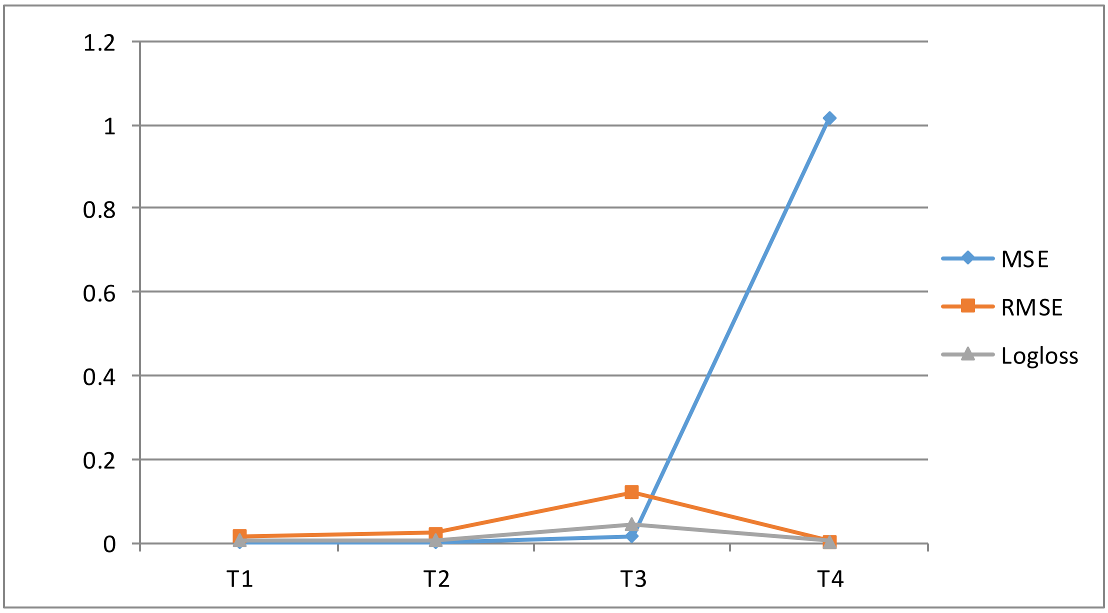


Figure 5.1 GBM for four Targets

The following Table 5.2 shows the common measures achieved by the four different model for one label (T1). Similarly, the Table 5.3, 5.4, 5.5 shows the common measures achieved for other three labels T2, T3, T4 respectively. Results showed that model 3 and 4 i.e., M3, M4 performs well when compared with M1 and M2.

RANDOM FOREST

TARGET 1 (T1)

	M1	M2	M3	M4
MSE	0.003974711	0.003073636	0.002785161	0.002785161
RMSE	0.06304531	0.5544038	0.05277462	0.05277462
Log loss	0.03480119	0.03028241	0.02907588	0.02907588

Table 5.2 Random Forest for Target 1 (T1)

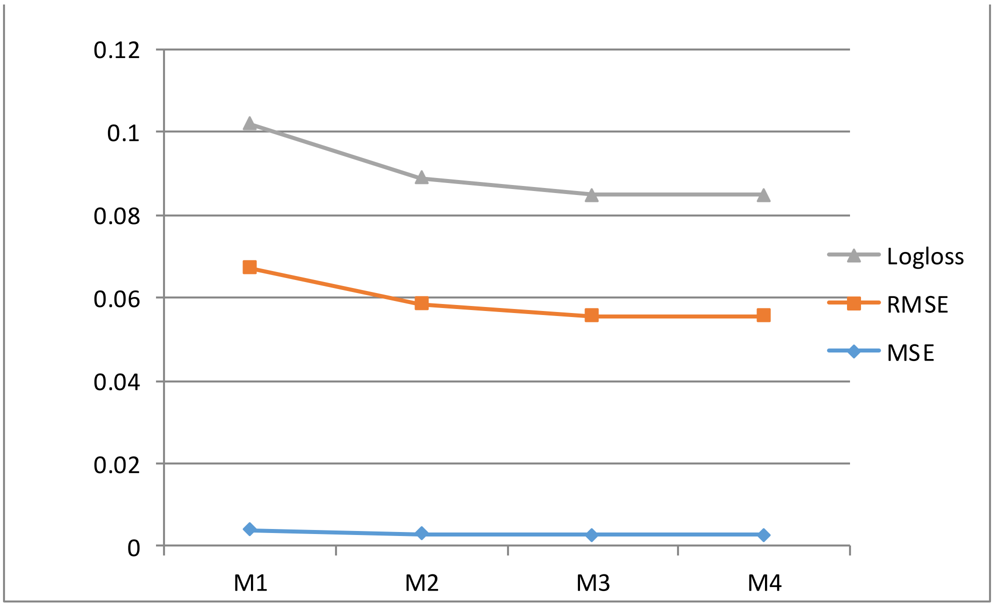


Figure 5.2 Random Forest for Target T1

TARGET (T2)

	M1	M2	M3	M4
MSE	0.01184767	0.01184767	0.01170807	0.01170807
RMSE	0.108847	0.108847	0.1082038	0.1082038
Log loss	0.08296549	0.08296549	0.08681097	0.08681097

Table 5.3 Random Forest for Target T2

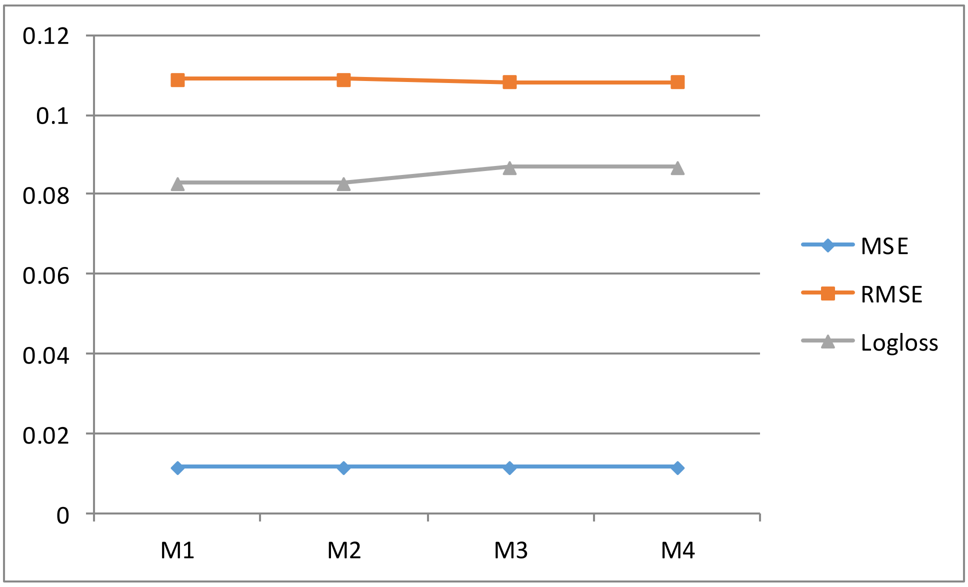


Figure 5.3 Random Forest for Target T2

TARGET 3 (T3)

	M1	**M2**	**M3**	**M4**
MSE	0.006190471	0.006190471	0.006470311	0.006470311
RMSE	0.07867955	0.07867955	0.08043825	0.08043825
Log loss	0.03990048	0.03990048	0.03846515	0.03846515

Table 5.4 Random Forest for target T3

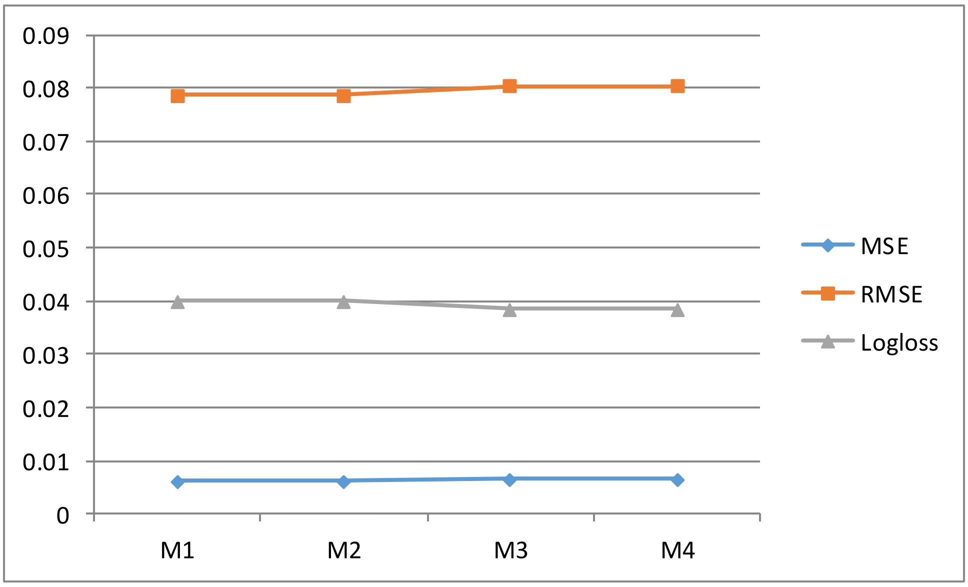


Figure 5.4 Random Forest for Target T3

TARGET 4 (T4)

	M1	M2	M3	M4
MSE	0.000753418	0.000753418	0.00081447	0.00081447
RMSE	0.02744847	0.02744847	0.02853891	0.02853891
Log loss	0.01573072	0.01573072	0.01755831	0.01755831

Table 5.5 Random Forest for Target T4

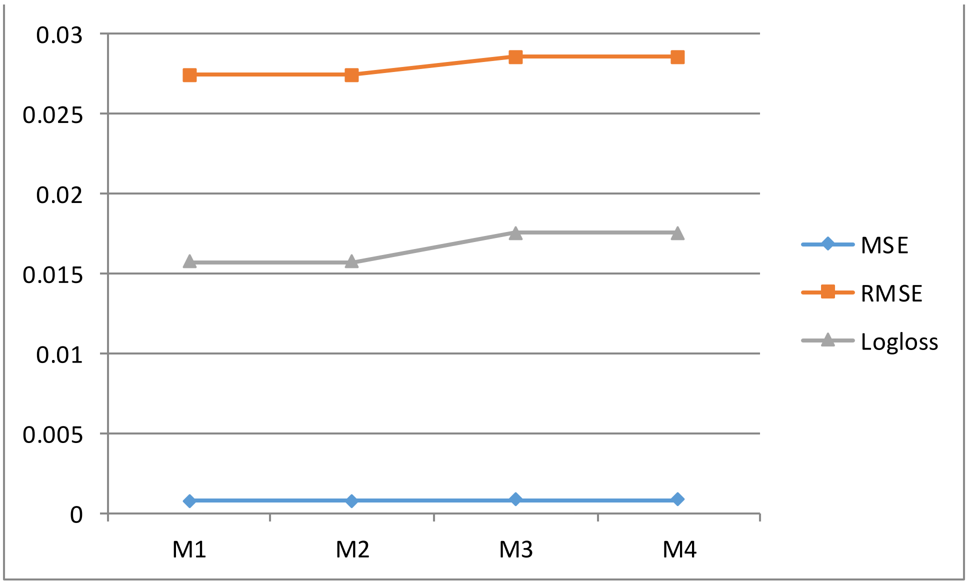


Figure 5.5 Random Forest for Target T4

The Random Forest algorithm uses four models for each label in the dataset. It has been observed that Random forest shows a better performance when there is a growth in the size of training dataset.

MSE	GBM	RF
T1	0.000217525	0.002785161
T2	0.000530746	0.01170807
T3	0.01488992	0.006470311
T4	1.017642	0.00081447

Table 5.6 Compare GBM and RF for MSE

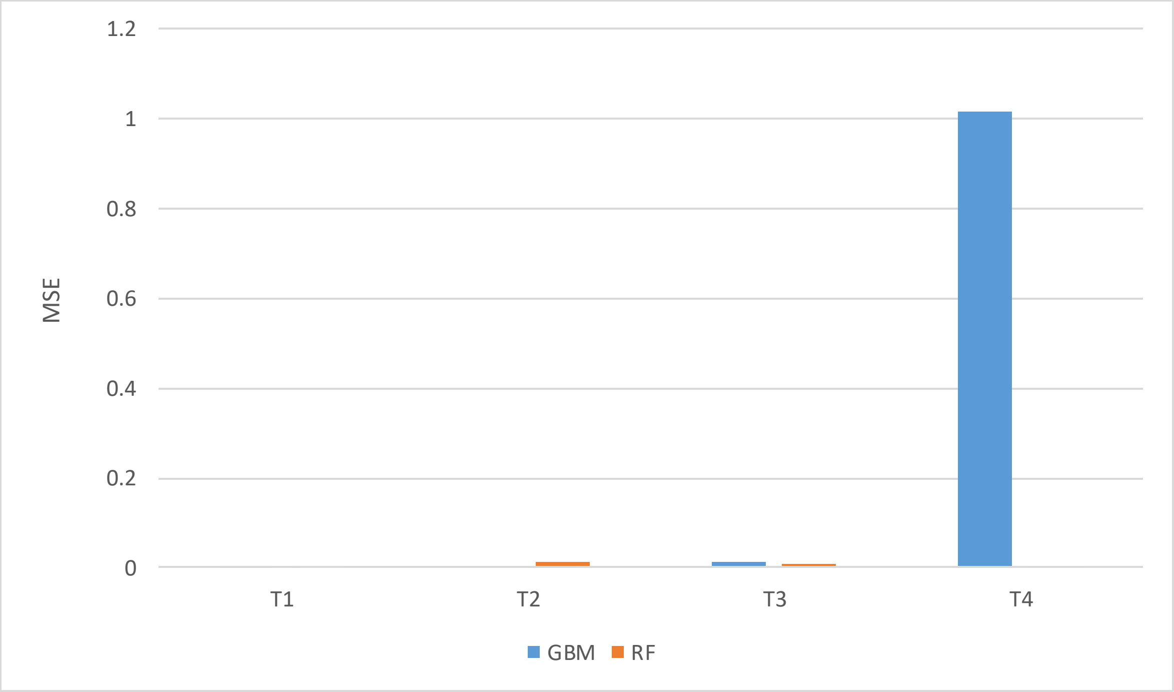


Figure 5.6 Compare GBM and RF for MSE

RMSE	GBM	RF
T1	0.01474872	0.05277462
T2	0.2303793	0.1082038
T3	0.1220243	0.08043825
T4	0.00319005	0.02853891

Table 5.7 Compare GBM and RF for RMSE

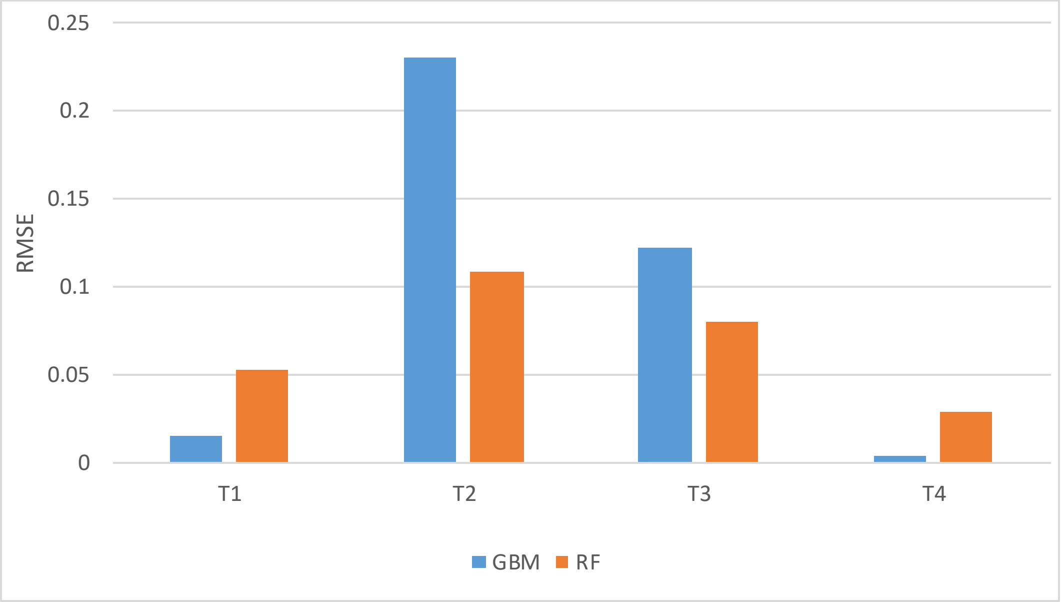


Figure 5.7 RMSE compare GBM and RF

Log loss	GBM	RF
T1	0.006491229	0.02907588
T2	0.006329349	0.08681097
T3	0.04517508	0.03846515
T4	0.003036938	0.01755831

Table 5.8 Compare GBM and RF for Log loss

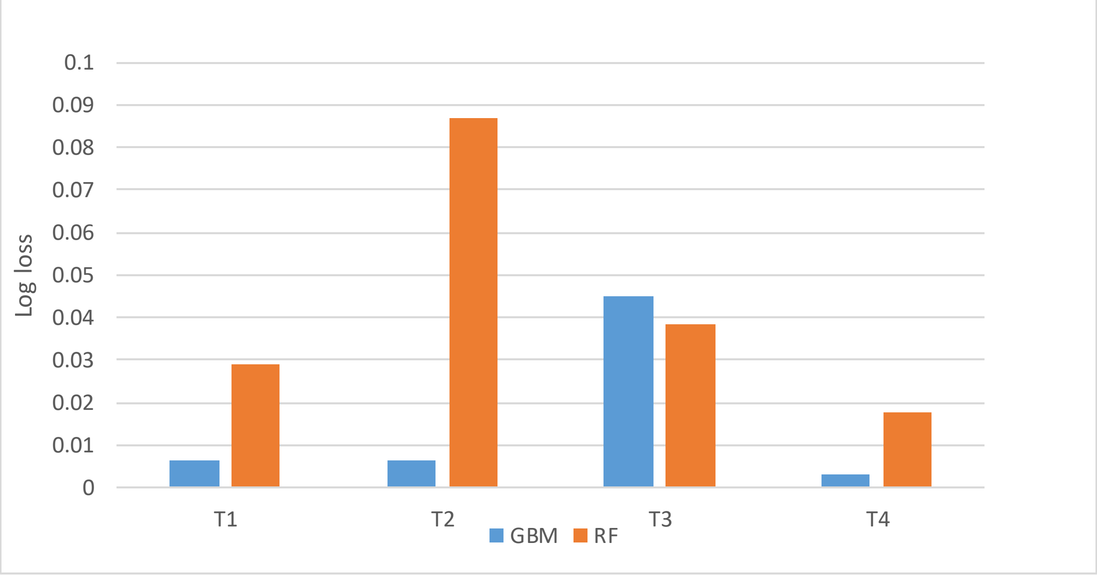


Figure 5.8 compare GBM and RF for Log loss

In our scheme two types of classifiers are used viz GBM and Random Forest that improves the performance of the existing scheme. This is because it attains MSE, RMSE and Log loss rates are seems to be low. Though Random Forest performs better in terms of size of data there exist problems when the dataset needs sampling. The purpose of using another classifier named GBM is to improve the performance of log loss measurement due to which there is a reduction in the existing problem i.e., class imbalance. Table 5.1 shows the common measures of GBM in which the MSE, RMSE and Log loss .are low compared with Random Forest . Hence, it has been proved that GBM outperforms Random Forest regardless of size of the dataset.

CHAPTER 6

CONCLUSION

Classification of Traditional Chinese Medicine has been performed using GBM and Random Forest algorithms. In this system the imbalanced data's are removed which will minimizes the error. In Random Forest algorithm four models were cross validated for classification. The four models are Balanced + Not Stratified (M1), Balanced + Stratified (M2), Unbalanced + Not Stratified (M3), Unbalanced + Stratified (M4). By comparing these four models M3 and M4 gives the reduced error rate than Balanced + Not Stratified (M1), Balanced + Stratified (M2). Hence the generated Unbalanced + Not Stratified (M3) and Unbalanced + Stratified (M4) model for Random forest are compared with the GBM by using the parameters MSE, RMSE and Log loss. While comparing GBM and RF it has been showed that GBM has a better performance accuracy in terms of imbalanced error. GBM minimizes the error which improves the performance of a system.

REFERENCES

[1] Li, G. Z., He, Z., Shao, F. F., Ou, A. H., & Lin, X. Z. (2015). Patient classification of hypertension in Traditional Chinese Medicine using multi-label learning techniques. BMC medical genomics, 8(3), S4.

[2] Wang, H., Liu, X., Lv, B., Yang, F., & Hong, Y. (2014). Reliable multi-label learning via conformal predictor and random forest for syndrome differentiation of chronic fatigue in traditional Chinese medicine. PloS one, 9(6), e99565.

[3] Yang, F., Wang, H. Z., Mi, H., & Cai, W. W. (2009). Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC* bioinformatics, 10(1), S22.

[4] Vilela-Martin, J. F., Vaz-de-Melo, R. O., Kuniyoshi, C. H., Abdo, A. N. R., & Yugar-Toledo, J. C. (2011). Hypertensive crisis: clinical–epidemiological profile. Hypertension Research, 34(3), 367.

[5] Liu, G. P., Li, G. Z., Wang, Y. L., & Wang, Y. Q. (2010). Modelling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi-label learning. BMC complementary and alternative medicine, 10(1), 37.

[6] Li, G. Z., Sun, S., You, M., Wang, Y. L., & Liu, G. P. (2012). Inquiry diagnosis of coronary heart disease in Chinese medicine based on symptom-syndrome interactions. Chinese medicine, 7(1), 9.

[7] Poon, S. K., Poon, J., McGrane, M., Zhou, X., Kwan, P., Zhang, R., ... & Man-yuen Sze, D. (2011). A novel approach in discovering significant interactions from TCM patient prescription data. International journal of data mining and bioinformatics, 5(4), 353-368.

[8] Wang, X., & Li, G. Z. (2013). Multilabel learning via random label selection for protein subcellular multilocations prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10(2), 436-446.

[9] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging, boosting, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4), 463-484.

[10] Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter, 6(1), 1-6.

[11] Douibi, K., Settouti, N., Chikh, M. A., Read, J., & Benabid, M. M. (2018). An analysis of ambulatory blood pressure monitoring using multi-label classification. Australasian physical & engineering sciences in medicine, 1-17.

[12] Jiang, M., Lu, C., Zhang, C., Yang, J., Tan, Y., Lu, A., & Chan, K. (2012). Syndrome differentiation in modern research of traditional Chinese medicine. Journal of ethnopharmacology, 140(3), 634-642.

[13] Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02), 185-205.

[14] Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. Expert Systems with Applications, 39(1), 1503-1509.

[15] Shao, H., Li, G., Liu, G., & Wang, Y. (2013). Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine. Science China Information Sciences, 56(5), 1-13.

[16] Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition, 40(7), 2038-2048.

[17] Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, *3*(02), 185-205.

[18] Zhang, M. L., Peña, J. M., & Robles, V. (2009). Feature selection for multi-label naive Bayes classification. Information Sciences, 179(19), 3218-3229.

[19] Li, F., Zhao, C., Xia, Z., Wang, Y., Zhou, X., & Li, G. Z. (2012). Computer-assisted lip diagnosis on traditional Chinese medicine using multi-class support vector machines. BMC complementary and alternative medicine, 12(1), 127.

[20] Huang, G. B., Ding, X., & Zhou, H. (2010). Optimization method based extreme learning machine for classification. Neurocomputing, 74(1-3), 155-163.

[21] Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.

[22] Li, L., Liu, H., Ma, Z., Mo, Y., Duan, Z., Zhou, J., & Zhao, J. (2014, December). Multi-label feature selection via information gain. In International Conference on Advanced Data Mining and Applications (pp. 345-355). Springer, Cham.