A Hybrid Ensemble Classification Approach to Determine the Impact of Asthma in Association with Gastro Esophageal Reflux Symptoms

Article *in* Indian Journal of Public Health Research and Development · February 2019

DOI: 10.5958/0976-5506.2019.00284.5

2 authors, including:

Kasturi Karuppiah
Vels University
29 PUBLICATIONS **44** CITATIONS

SEE PROFILE

DOI Number: 10.5958/0976-5506.2019.00284.5

A Hybrid Ensemble Classification Approach to Determine the Impact of Asthma in Association with Gastro Esophageal Reflux Symptoms

K.Kasturi [1], S.Prasanna[2]

[1]Research Scholar, Department of IT, [2]Associate Professor, Department of Computer Applications, VISTAS

ABSTRACT

Objectives: This implementation work focuses on the predicting severity of respiratory problems of asthmatic patients from the dataset of the PFT report with the significant parameters of Gastro Esophageal Reflux symptoms (GER).

Methodology: The pulmonary functionality test (PFT) report of the asthmatic patients is associated with the significant parameters of GER symptoms to determine the impact of GER symptoms on asthma using a proposed hybrid ensemble classification.

Methods/Statistical Analysis: Using R statistical tool a model has been developed for ensemble classification by stacking the SVM and Random Forest algorithms and boosting with the improved Gradient Boosting algorithm.

Findings: It has been identified that the asthmatic patients who have been reported as 'normal' or 'mild' in the PFT report also have the respiratory problems often and urge for frequent check – ups. This can be due to the implications of significant symptom parameters of GER.

Applications: The outcome of the developed model HMMC describes about the classification accuracy of the applied dataset of the asthmatic patients with GER symptoms and predicts the severity of asthma in asthmatic patients more accurately rather than the outcome of the existing classification techniques.

Keywords: *boosting, asthmatic Patients, PFT, GER, ensemble classification, HMMC.*

INTRODUCTION

Prediction and assortment of medical datasets serve in reducing the count of diagnostic issues to recognize the diseases there by affording economical solutions for healthcare systems and medical diagnosis software system. Data pre-processing operates an essential role in prime powerful and capable data for data mining. Feature selection aid in contributing essential characteristics for construction of extensive predictive models. The processing of medical data mining has more potential for exploring the hidden patterns in the knowledge sets of the medical domain. These patterns will be used for clinical diagnosing. However, the obtainable raw medical knowledge is heterogeneous and voluminous in nature. This collected information is often integrated to give a user orienting approach to novel and hidden patterns within the information. In my previous work the significant parameters of the GER symptoms have been identified [1] . And this work deals with the implications of the significant GER parameters of the asthmatic patients with their PFT report value. This leads to the way to find the accuracy of the classification prediction.

EXISTING METHODOLOGY

Chun-Rong Huang et al [2] proposed that "GER can be diagnosed by typical symptoms of acid regurgitation or heartburn sensation at the epigastric or mid chest regions".E.A.Boiler et al [3] proposed that "In western countries, 10–30% of the population experiences symptoms of gastro esophageal reflux (GER)". Sudha Pandit et al [4] propose that "The prevalence of GER

is primarily based on the acid reflux symptoms and these symptoms are not always present in patients with endoscopic evidence of esophagitis". D. Vincent et al [5] proposes that "Gastro-esophageal reflux (GER) has been suspected as a causal factor, but the relationship between GER and asthma remains controversial". The existing work focuses only upon the environmental factors such as living, working environment and parental history as the main causes of respiratory diseases [6] .But the fact is GER may cause, trigger or exacerbate pulmonary disease like bronchial asthma and many other diseases.

PROPOSED METHODOLOGY

A questionnaire can be scientifically accepted by following the standard criteria such as valid, responsive and reliable and it is well determined and established by psychometric methods elsewhere. Such a questionnaire is often administered directly by the patient, filled in, and is straightforward to know as these attributes guarantee prime quality information assessment. The GER symptoms are not always known in the endoscopy test and at the same time the asthmatic patient reported as 'normal' is not always having the normal respiration. This is due to the fact that GER has implications over asthma. This leads to conduct a prospective assessment of GER prevalence in a population of asthmatic patients and to find the prediction accuracy of severity of asthma using the proposed classification model called Hybrid Meta Model Classification(HMMC).

SELECTION OF ALGORITHMS

Random Forest is the flexible machine learning algorithm and produces accurate result with hyper parameter tuning. This supervised learning algorithm builds multiple decision trees, merges them together to get a more accurate and stable prediction [7]. It can also handle missing values and large dataset with higher dimensionality effectively [8].Support Vector Machine (SVM) is used in binary classification and also applied for pattern recognition and it is a promising classification approach for prediction purpose [9].

DATASET DESCRIPTION

- The PFT results for the asthmatic patients are got as raw data as .xps files from the Asthma and Allergy Resource Centre, Chennai, Tamil Nadu, India. Around 2000 raw data has been collected and entered in a spread sheet and stored as .csv file.
- The GER symptom parameters have been assessed to the above asthmatic patients through the questionnaire form and entered in a spread sheet and stored as .csv file.

The resultant significant GER symptom parameters from my previous research work and the PFT report data are taken under consideration for the purpose of classification.

The dataset comprises of

- PID(Patient ID)
- Age
- Gender
- Heartburn (GER symptom)
- Dysphagia (GER symptom)
- Nausea (GER symptom)
- PFT Report
- FreqCheckup

Table 1: Possible values of attributes

Heartburn	Dysphagia	Nausea	Report	Freq Checkup
voften often noften notAtAll	voften often noften notAtAll	voften often noften notAtAll	normal mild severe	yes no

EXPERIMENTAL RESULTS

CLASSIFICATION TREE

The Classification tree has the structure of a flowchart in which each node represents a condition test on an attribute and that each branch indicates the outcome of the condition test. Each terminal or leaf node indicates the class labels**.** Fig 1 shows the distribution of class labels of the PFT report with GER key attributes.

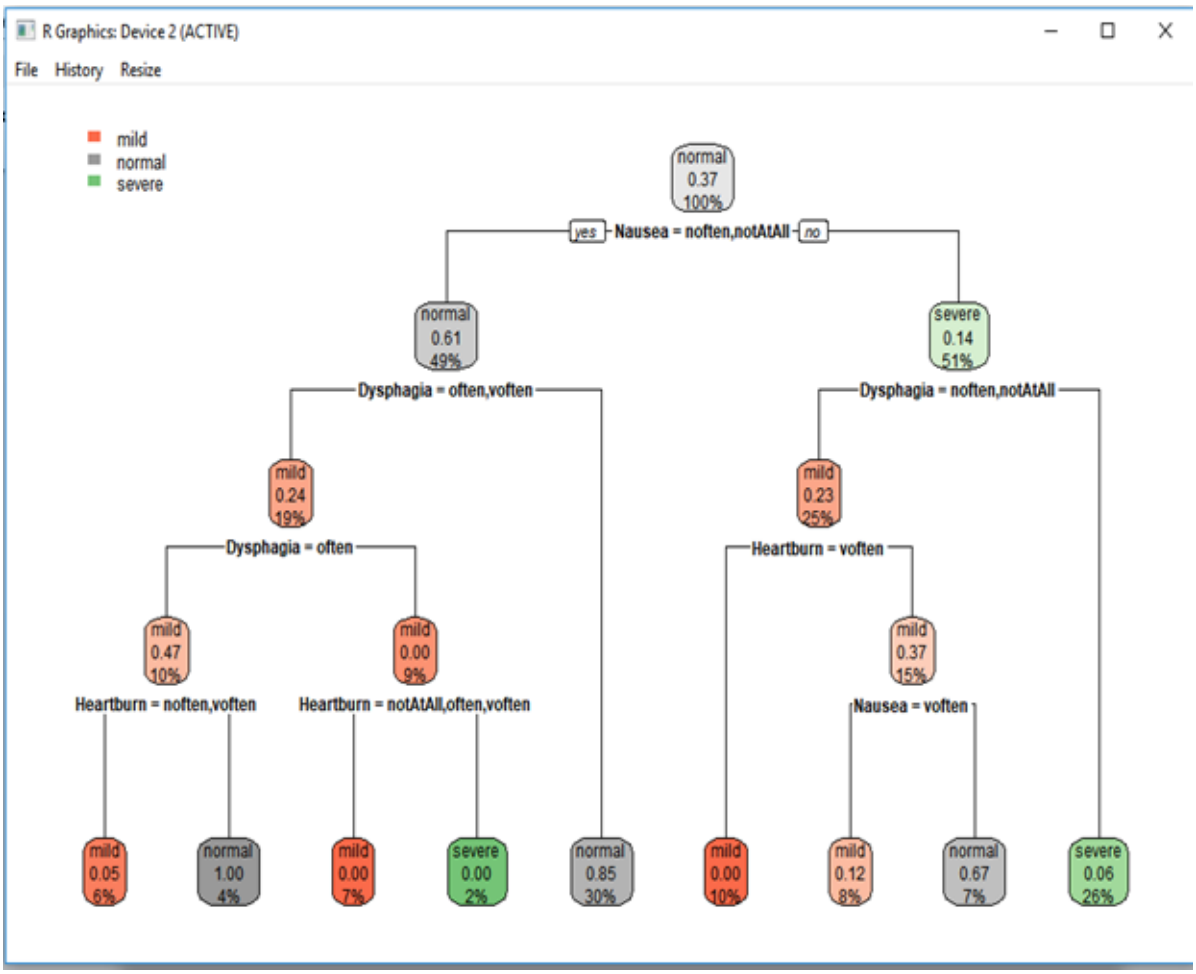

Fig 1: Classification Tree

CLASSIFICATION ALGORITHMS

In random forest algorithm the importance of a variable can be estimated by looking the rate of increase in prediction error when data for that variable is permuted while all the others are left unchanged. The calculations are carried out tree by tree as the random forest is developed [10].

"SVM converges fast and leads to high accuracy. When scores of multiple parameter datasets are combined, majority voting reduces noise and increases recognition accuracy" [11]. It also includes avoidance of over fitting effectively [12] .The built SVM model has the accuracy of 94.85% [Fig 2] and the accuracy of the Random Forest model has the greater accuracy of 95.15% [Fig 3].

BOOSTING ALGORITHMS

C5.0 is the extension of C4.5 algorithm. It gives a binary tree or multi branch tree and uses the gained information for splitting criteria. For estimating missing values it uses other attributes .It represents how the rules are generated with high accuracy and low memory usage [13] .Stochastic Gradient Boosting

is one of the powerful classification algorithms that attains maximum accuracy 96.97% whereas the accuracy of C5.0 is 93.64% which is lesser than SGB. SGB is powerful in terms of classifier and prediction at less execution time.

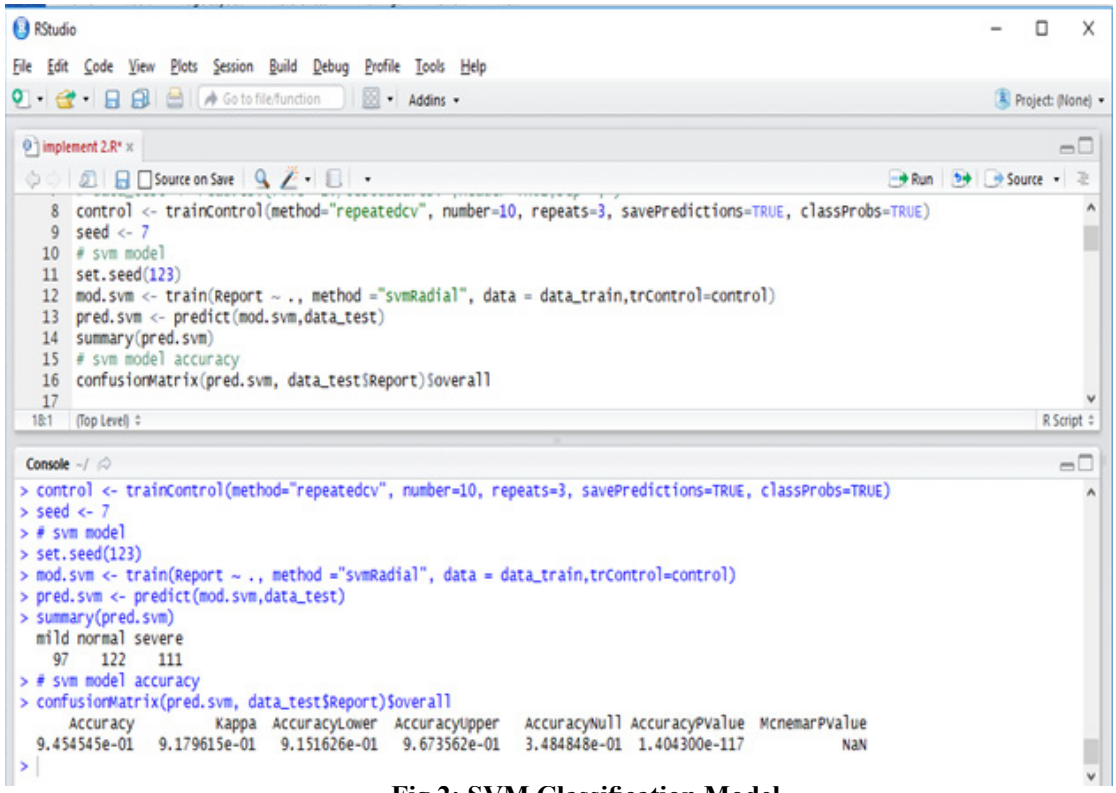

Fig 2: SVM Classification Model

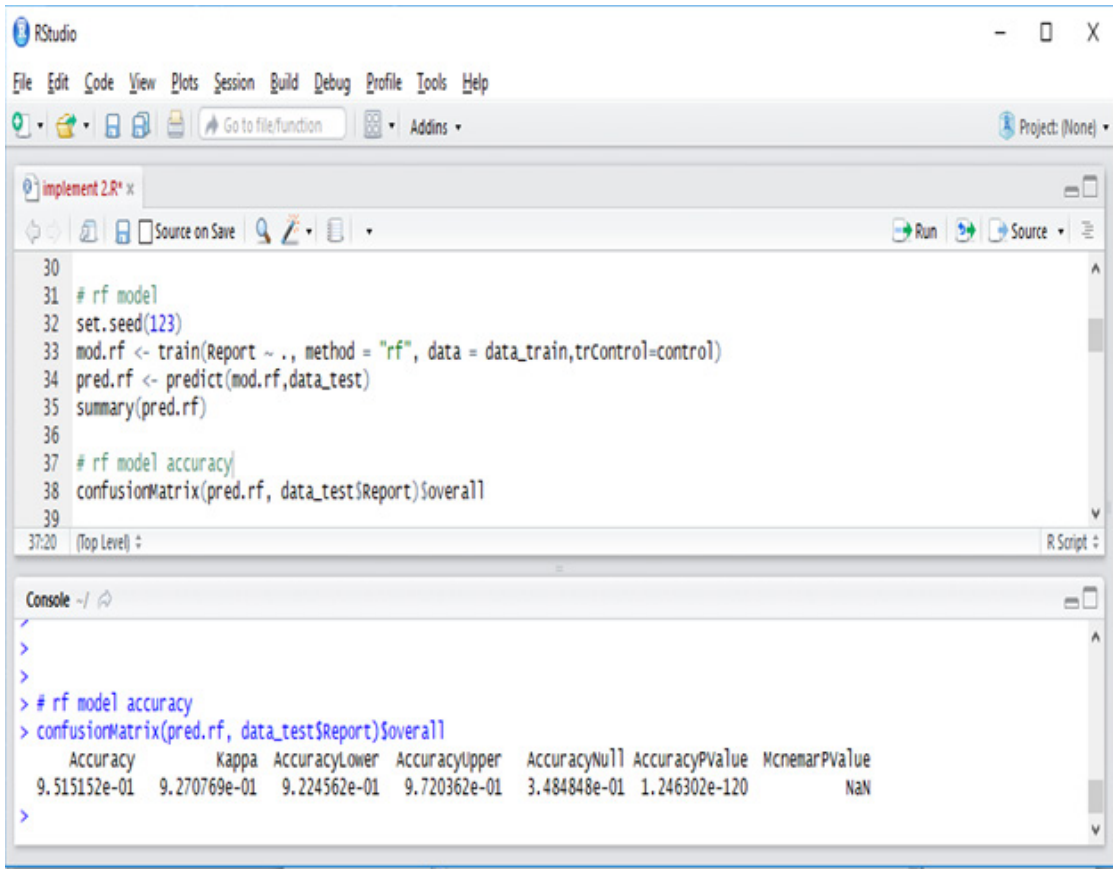

Fig 3: Random Forest Classification Model

PROPOSED MODEL

The hybrid function model can be generated for the predictive purpose with less error and high accuracy without affecting the data consistency [14]. Hybrid model reduces the diagnostic procedure to identify the diseases and to mine the hidden pattern of knowledge [15]. A Hybrid model is needed to diagnose the complexity of disorders and requires more empirical evidences before incorporated into clinical practise [16].

HYBRID META MODEL CLASSIFICATION (HMMC)

Step 1: Splitting the dataset into training and testing data.

Step 2: The developed Random Forest Classification Model and the SVM Classification Model is Combined by the ensemble classification technique called Stacking.

Step 3: Then the stacked outcome is fed into another ensemble classification technique called **Boosting.**

Step 4: The Boosting concept is implemented by the proposed **Improved Gradient Boosting algorithm** to get the maximum accuracy in less time.

PERFORMANCE ANALYSIS

Table 2: Classification Accuracy

Model Built Using	Accuracy
SVM	94.84.%
Random Forest	95.15%
HMMC	99.4%

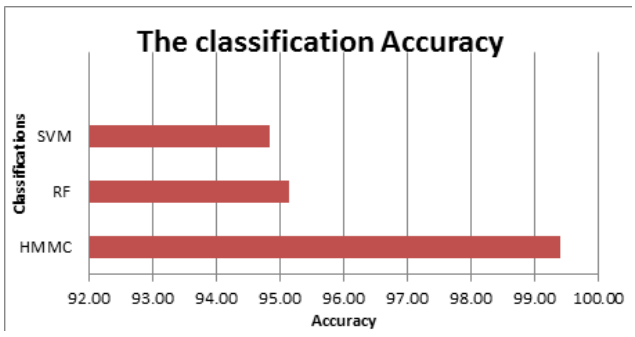

Fig 4: Classification Accuracy Graph

From the Fig 4 the classification accuracy depicts that the proposed model HMMC has the maximum accuracy rather than the existing classification techniques.

CONCLUSION

From the above implementation it has been interpreted that the results of the proposed ensemble model classification HMMC shows improved overall classification accuracy significantly rather than the existing model classification accuracy. This model confirms that the significant GER symptom parameters have high impact on asthmatic patients. Even though the asthmatic patient's PFT report says the condition of the patients as normal or mild, the existence of the significant GER symptom parameters makes the patient's respiration more worst and their by urging them for frequent check-ups.

Ethical Clearance-Not Required (only getting the test report from the hospital).

Source of Funding- Self.

Conflict of Interest – Nil.

REFERENCES

[1] Kasturi K, Prasanna S. Implications of Asthma over Gastro Esophageal Reflux Symptoms Using Regression Analysis. Journal of Advanced Research in Dynamical and Control Systems. 2018[Special Issue]; 11:1066-1072.

[2] Chun-Rong Huang , Yan-Ting Chen ,Wei-Ying Chen , Hsiu-Chi Cheng, Bor-Syang Sheu. Gastroesophageal Reflux Disease Diagnosis Using Hierarchical Heterogeneous Descriptor Fusion Support Vector Machine. IEEE.2016;63(3)

[3] Bolier, E., Kessing, B., Smout, A. and Bredenoord, A. (2013). Systematic review: questionnaires for assessment of gastroesophageal reflux disease. Diseases of the Esophagus, 28(2), pp.105-120.

[4] Pandit S, Boktor M, Alexander JS, Becker F, Morris J. Gastroesophageal reflux disease: A clinical overview for primary care physicians. Pathophysiology. 2018 Mar 1;25(1):1-1.

[5] Vincent D, Cohen-Jonathan AM, Leport J, Merrouche M, Geronimi A, Pradalier A, Soule JC. Gastro-oesophageal reflux prevalence and relationship with bronchial reactivity in asthma. European Respiratory Journal. 1997 Oct 1;10(10):2255-9..

[6] Rohini K, Suseendran G. Aggregated K Means Clustering and Decision Tree Algorithm for Spirometry Data. Indian Journal of Science and Technology. 2016 Nov 30;9(44)..

[7] Casanova R, Saldana S, Chew EY, Danis RP, Greven CM, Ambrosius WT. Application of random forests methods to diabetic retinopathy classification analyses. PLOS one. 2014 Jun 18;9(6):e98587.

[8] Pantanowitz A, Marwala T. Missing data imputation through the use of the Random Forest Algorithm. InAdvances in Computational Intelligence 2009 (pp. 53-62). Springer, Berlin, Heidelberg.

[9] "Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach1. Journal of molecular biology. 2001 Apr 27;308(2):397-407.

[10] Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002 Dec 3;2(3):18-22.

[11] Ding CH, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics. 2001 Apr 1;17(4):349-58.

[12] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. InProceedings of the fifth annual workshop on Computational learning theory 1992 Jul 1 (pp. 144-152). ACM.

[13] Pandya R, Pandya J. C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. International Journal of Computer Applications. 2015 May;117(16):18-21.

[14] Banihashemi S, Ding G, Wang J. Developing a hybrid model of prediction and classification algorithms for building energy consumption. Energy Procedia. 2017 Mar 1;110:371-6.

[15] Raghavendra S, Indiramma M. Hybrid data mining model for the classification and prediction of medical datasets. International Journal of Knowledge Engineering and Soft Data Paradigms. 2016;5(3-4):262-84..

[16] Esbec E, Echeburúa E. The hybrid model for the classification of personality disorders in DSM-5: a critical analysis. Actas Esp Psiquiatr. 2015 Sep 1;43(5):177-86.