International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Classification of Diagnostic Codes of Chronic Condition and Performance Evaluation of Various Approaches

Mohan Kumar K N, S.Sampath, Mohammed Imran

Abstract: ***Health in simple words is normal functioning of human body and disease is abnormal condition that affects normal functioning of human body without any external injury. Health care is all about the prevention of diseases by diagnosis and treatment. The majority of the population across world suffers from chronic disease which is a long term disease leading to multiple ailments if not taken care and chronic diseases cannot be prevented by vaccines or cured by medication, nor do they just disappear. Efforts are needed to build an efficient system which can predict, classify diseases and detect anomalies from health records. Electronic medical records are not better than the old manual records. This paper focuses on Medi-Claim data as it standout uniquely due to its authenticity, volume and demography attributes. Importantly HCC and ICD based coding are compatible with claim data set. This nature of ICD and HCC coding encouraged us to work with Medi-claim data set and HCC coding to build a Machine Learning model for preventive care of chronic diseases. The correlation between diabetes chronic disease and other chronic diseases is established through HCC codes using Machine Learning approaches. Effective inferences are drawn from the perspective of clinical relevance.***

Index Terms: HCC, Chronic Condition, diabetes, Ensemble.

I. INTRODUCTION

'Health is wealth' is an old adage gaining significance than ever before in today's world. Though many epidemics have been eliminated from the society with medical inventions chronic ailments like Diabetes are threatening the well being of human race. Preventive measures for patients suffering from ailments using existing clinical tests and through available limited data records with practitioners have proven to be less efficient [1].

Different methods have been used in healthcare to classify and identify diseases. Risk Adjustment Factor- RAF is a parameter used in healthcare to assess the probable cost incurred by an individual towards the treatment of a disease. Individuals with family history of a particular ailment and old people are given High value of RAF and youngsters and individuals with no family history are assigned a smaller value. Similarly another parameter employed is interaction score. It is a score that predicts a person with X ailment has the risk of getting affected by Y disease. Original Reason for Entitlement Code (OREC) score has limited scope as it tries to identify health issues occurring among only old age people. Co morbidity is a condition where two pathological conditions occur simultaneously in a patient. Moving further World Health Organization has adopted a coding system called International Classification of Diseases (ICD) [2][3] and Hierarchical Condition Categories (HCC). The code assigned in ICD is a 10 Digit number and each code is assigned to 70000 different diseases. The ICD code provides information pertaining to symptoms, patient complaints, reasons for injury and mental disorders. Also the ICD code can be entered into electronic health record of patient for further diagnosing, billing and reporting purposes. Further HCC helps in codifying 79 chronic diseases. HCC coding is mapped with ICD to fetch the ICD's associated with hierarchical chronic diseases. HCC can be used to correlate the chronic diseases which are suspected to be consecutive in nature and thus helps in preventive care of diseases. HCC also predicts the costs incurred in treating different category of patients like inpatient, outpatient and patients in office settings [2][3][4]. Usually the health professionals rely upon different kinds of data sets to detect ailments in a patient. Prominent among them are Hospital Records, Diabetic History [4][5], Electronic Medical Record, Medi-claim data, Heart health record, Hospital Service appointments, National Health Nutrition records etc. Many times these electronic medical records act as mere replacements of old paper charts and are less reliable. Amongst afore mentioned medical records that we analyzed Medi-Claim data standout uniquely due to its authenticity, volume and demography attributes. Importantly HCC and ICD based coding are compatible with claim data set. This nature of ICD and HCC coding encouraged us to work with Medi-claim data set and HCC coding to build a Machine Learning model for preventive care of chronic diseases [6]. In healthcare the best machine learning (ML) [6][7] tool is the Doctor's Brain. But here the main limitation is the incapability of a single human brain to learn, store and analyze data of demographically dispersed and huge volume of data. So, Machine learning models with the ability to determine a set of rules using vast amounts of compute power which a human brain is incapable of processing.

Revised Manuscript Received on 22 May 2019.
***** Correspondence Author
Mohan Kumar K N*, Dept. of CSE., SJB Institute of Technology, Bangalore, India.
S.Sampath, Dept. of ISE., Adichunchangiri Institute of Technology, City Chikkamagaluru, India.
Mohammed Imran, Research and Development, Ejyle Technology, Bangalore, India.

Hence there is need to analyze and design an efficient system which can classify diseases and detect anomalies from electronic health records.

This work uses Medi-claim data to achieve aforementioned objective of classifying chronic diseases for preventive care. We have considered claim data as it is a proven data compared to other sources which would be incomplete, inconsistent and small in size leading to wrong inferences.

The flow of this paper is as follows: The related work is detailed in section II which highlights data extraction, pre-processing and classification of health records. Section III describes the methods and models followed in the experiment. Section IV provides the experimental setup. The inferences drawn from the results is discussed in Section V. Section VI contains the summary and directions for future enhancement.

II. RELATED WORK

Exploration of existing techniques would give better visibility in the context discussed above; this section provides a detailed overview of different approaches used for disease classification. Divya et. al. [8] had come up with a comprehensive survey on the usefulness of classifier systems in chronic disease prediction. The primary focus of authors is on usage of classification system which are Parallel and Adaptive in nature for the diagnosis of chronic diseases. Interesting findings of the work is Adaptive and Parallel classification system reduces the decision making time and can detect chronic diseases accurately in comparison to other systems. The authors emphasize developing hybrid classification methods so that efficiency of computing and classifier accuracy is improved. Ashok et. al. [9] had carried out a comparative study of five classification algorithms (viz. Support Vector Machine, Regression, BayesNet, NaiveBayes and Decision Table) and have come up with results depicting the performance of these algorithms in classifying diabetic patients. The researchers have implemented a hybrid model for comparative study of aforementioned machine learning algorithms. The study is based on UCI repository data set. The findings are without feature selection. The BayesNet Algorithm outperformed other methods. While Decision Table algorithm performs better with feature selection. The authors emphasise that Hybrid classification models can be constructed in future to predict other diseases and ask the research community to find out ways to improve the performance of their results by applying other algorithms. Araujo et. al. [10] focuses on two aspects. Firstly to improve the quality of data used in health insurance sector. Secondly a comparative study is carried out using Decision Tree techniques to understand the medical reviewer behaviour. They have employed data set from non profitable health insurance companies in the experiments. Shen et. al. [11] in their work have applied Association Rule Mining (ARM) methods within a narrow context so that overwhelming rule sets generated by ARM methods is filtered down to a set of interesting rules applicable for knowledge discovery. They have proposed a hybrid clustering-ARM approach to identify risk of chronic heart ailment Myocardial Infarction (MI) or heart attack. Their experiments were based on the Framingham Heart Study dataset and have come up with a Clustered CARs framework that enables identifying and characterizing patients with risk of MI. They have managed to obtain CAR clusters with the best balance of precision, recall and low overlap on dataset used. Their framework substantially minimizes a large set of association rules to a manageable set of clusters that can identify distinctive characteristics of at-risk patients. Further authors propose their research work can be extended by using datasets from New Zealand VIEW and Framingham Heart Study cohort. They have also identified that a certain challenges that the researchers can find in these data sets are the availability of variables of greater range and temporal patterns. This enables researchers to obtain features based on treatment patterns and physiological changes. Varun Chandola et. al. [12] in their work have proposed how effectively the developments in arena of Big Data can be collaborated with data from healthcare sector to enable the healthcare professionals to leverage the developments in Data analytics arena. Authors have used analysis problems from popular data mining techniques, text mining, social network analysis and temporal analysis and higher order feature construction on healthcare domain. They have given exclusive case studies with the primary focus on improving the ratio of cost-care for healthcare domain and minimizing the cases of fraud and abuse. Chih et. al. [13] ha applied classification techniques to identify people suffering from five chronic disorders such as hypertension, diabetes, cardiovascular, liver and renal ailments. It proposes an early warning system based on critical values of risk factors so that a preventive healthcare measure is taken against chronic diseases. The authors have employed KNN, Sequential Forward Selection (SFS) and Linear Discriminant Analysis (LDA) algorithms to classify and detect risk factors associated with the five types of chronic ailments. In the research authors have employed LDA and SFS to identify key risk factors leading to diseases. The K-NN is used to predict and warn the member patient to take preventive measure. The limitation identified by authors is that the data set used is based on clinical tests and always there is scope to expand the data set thus improving the performance of the overall system. Asha et. al. [14] had proposed a cascaded model for classifying Pima Indian Diabetic Database (PIDD). They have employed K-means clustering and Decision tree C4.5 in their model. K-means clustering eliminates incorrect classified instances whereas the Decision Tree fine tunes the classification. The experimental results prove that cascaded K-means and Decision Tree C4.5 enhances classification accuracy. Further authors find that rules generated by Cascaded C4.5 model is easier to interpret compared to standalone C4.5 with continuous data. Authors conclude that the cascaded classifier of K-means and Decision Tree provide accurate results against other Decision Tree methods reported in literature. Duygu et. al. [15] had developed an automated diabetes diagnosis system based on Morlet Wavelet Support Vector Machine Classifier (MWSVM) and Linear Discriminant Analysis (LDA). The combo of LDA-MWSWM works in three different stages. In first stage LDA is employed for feature extraction and reduction. MWSVM classifier classifies the data in second stage. Further in next stage accuracy of diagnosis is calculated through sensitivity, specificity analysis and classification accuracy and confusion matrix.

Authors feel LDA and MWSVM classifier will be useful for medical practitioners in accurately diagnosing diabetes. As future work authors suggest to apply advanced classifier methods to improve the accuracy.

The overview of the research work discussed in literature on disease classification has encouraged us to focus on using of ML approaches on CMS claim data along with HCC codes. This helps to build an efficient system to establish the correlation among the chronic diseases and classify them accordingly with respect to HCC codes. Mapping of HCC codes over ICD code would eliminate the redundancy of diagnostic code which makes representation of chronic diseases simple. This work focuses on classification of diagnostic codes, which helps to study the hierarchy associated with chronic diseases. The outcome of this investigation would help the member patient to prevent future consequences of the health based on the current health status and avoid monumental cost of treatment. The experiments in this work uses CMS claim data. The consecutive sections will focuses on the data used.

III. METHODS AND MODELS

Over the past decades, many supervised techniques have been developed for aforementioned problem. Few important algorithms for classification are Support Vector Machine (SVM) [16], Artificial Neural Network (ANN), Decision Tree, Random Forest Logistical Regression Classifier and Ensemble methods. In Supervised learning technique, the dependent feature a_j €A exists is examined. The term learning means, the inference drawn from the dependent feature on the training set features A_i and verified from the set of test samples. The emphasis of the classification model is not the estimation of parameters but the flexibility of the model or capacity of the model which directly relates to complexity of the model.

SVM is based on statistical theory for risk minimization on structures. It is used to find the hyper plane which is optimal enough to separate samples of two classes in a feature space S. The hyper plane margin should be maximised between the two classes [16].

ANN model is mathematical representation of biological neurons in brain. ANNs are structured in layers and each neuron is an activation function and gets stimulus from other neurons as inputs and are associated with weights. The main aim of ANN is to decrease the empirical learning risk. The non linear optimization technique is employed to determine the problem of minimization. Multilayer ANN is the known method to find the set of weights w^1 and w^2 and here the convergence of minimum error is not assured. ANN is mainly used for classification [17].

Decision tree is one among the extensively used classification approaches. Each node in a decision tree is a set of rules based on split criteria that represents the outcomes.

Random forest is an ensemble (bagging) version of decision tree in which multiple decision trees are constructed with root node of each tree based on different split criteria. The outcome of all decision trees with different split criteria is consolidated to represent the best outcome.

Logistical regression is one of the fundamental classifier similar to linear regression. This model uses a sigmoid function as a logistic function. It considers any real values between zero and one. The decision boundary distinguishes probabilities into two classes namely positive and negative.

The above detailed models would produce results individually which needs to be improved. Ensemble model enable us to improve the accuracy by polling. In this technique we combine different individual models along with carefully selected parameters to enhance the accuracy [18].

The different ensemble methods are Bagging, Boosting and Stacking[19].

The interesting part of bagging approach is that out of given X objects, the model produces Y objects (Y<X) using bootstrap re-sampling with replacement. Each X_i is used to train different models. Outcomes of individual models are averaged or majority voting concept will be used.

Boosting [15] model uses different sampling methods than Bagging. Boosting evaluates the individual models using constant probability for the set of samples on every instance. The probability is changed from time to time to improve the model accuracy.

Stacking is an ensemble method in which a model is trained using the outcomes of two or more individual models. Stacking yields accuracy better than any single trained models. Stacking is also used to estimate Bagging's model error rate.

IV. EXPERIMENTAL SETUP

A. Data Acquisition

The experiment got initiated with Data acquisition (data collection). The med-claim data (MCD) was gathered from Centre for Medicare and Medicaid services (https://www.cms.gov/) [21]. MCD contains separate files for beneficiary information, inpatient, outpatient, carrier claims and prescription codes. The beneficiary file contains demographic information of the member patient. Inpatient, outpatient, carrier claims files contains the claims details of the beneficiary in terms ICD. Prescription code is excluded in our work. The med-claim data provides 20 samples of data set, of which sample1 is used in our work. The claim files of 2008 were extracted. The data are stored and handled in Hadoop distributed file system (HDFS). The platform used for development is python anaconda and pycharm IDE.

B. Data Modeling and Pre-Processing.

The primary challenge was the records in the acquired data contained a large number of attributes, and many of them are irrelevant in our experimental context. Data modelling involves careful selection of attributes from the extracted data to fit the context. Pre-processing of data involves the process of transforming raw data into the form suitable for ML algorithms. In the nutshell we have used big data technique (Hadoop) to store and process the data. HDFS is used for storage of data and Hive is used for data processing. With HDFS and Hive that data can be parallel processed and performance can be tuned by configuring the number of mappers and reducers whereas SQL doesn't provide this advantage. Fig. 1 details model of the proposed approach.

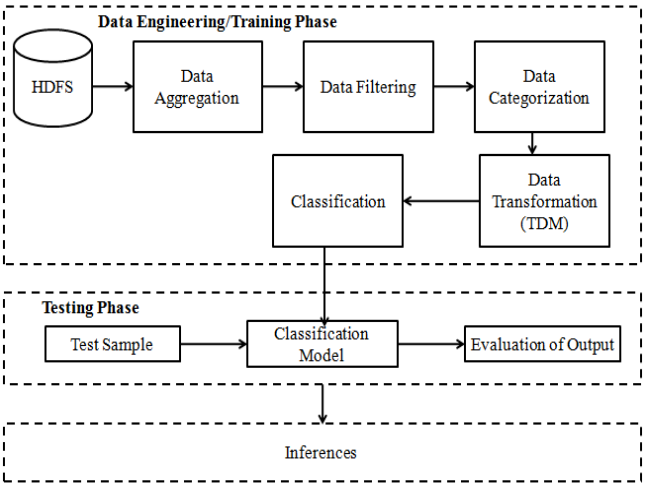

Fig 1. Abstract Model of the proposed approach

The records of inpatient, outpatient, and carrier claims files are moved into HDFS system. Hive tables are created for each type of claim and beneficiary file[20][21]. One more table called super data (SD) is created to contain the consolidated data of all the three claim files mentioned. The claim tables are populated with claim file data. The primary key for all the tables is beneficiary id. The SD table is populated with ICD attributes from all the claim tables. Our focus is only to deal with chronicle diseases hence ICD of non chronic diseases should be removed [22]. We have mapped HCC code with ICD code to remove the codes of non chronic diseases. For this ICD code in SD table is mapped with HCC code and resultant HCC codes are stored in a table called HCC data (HD) table. The records in HD table are again divided into diabetic and pre diabetic records by the first occurrence of diabetes HCC code for every member patient and stored separate tables called diabetic and non diabetic table. Finally Training Data is extracted from diabetic and non diabetic table and stored in table called final data (FD) table with additional attribute called class. The class value is 0 for non diabetic record and 1 for diabetic record. The records of the train final data table are exported in to CSV file. The CSV is finally concatenated with race, age and gender of beneficiary *i.e*. member patient.

Text data cannot be the input to any ML algorithm; hence the final data should be transformed to meet the needs of ML algorithm. Final data CSV is transformed into document term matrix (TDM) which is sparse matrix in which column attributes are class, HCC code race, age and gender and Row represents records of the member patient. The occurrence of HCC code is represented by 1, indicating the chronic ailment that the respective member patient is suffering from and 0 for absence of that ailment. The chronic ailment that we are dealing with is diabetes.

C. Training and Testing.

After pre-processing and consolidating records of the claim files, the total number of records is around 19 lakhs and these records are used for training and testing. We have made sure that there is uniform distribution of data among the classes. After data engineering and generating TDM file, the various classifiers of ML models are trained and tested for accuracy. TDM file data is partitioned into two subsets called train and test data in the ratio of 8:2. This partitioned subset data is trained and tested using a classification technique. Classification is a prediction technique of ML method used to classify a target class. Every Machine Learning model has training and testing phase in which the learning algorithm correlates the features and classify the target class. The few classification techniques are artificial neural network multi layer perceptron (ANN MLP) [17], Decision tree, Logistical Regression Random Forest, Support Vector Machine (SVM) [16]. Confusion matrix and classification report is used for the evaluation of classifiers.

V. RESULTS AND DISCUSSIONS

The experiment is tested on various classification techniques [22] to find out optimal solution. Results of different single classifiers and ensemble classifiers are tabulated in Table 1. The important intuitive measure of performance is accuracy. Accuracy is the ratio of correct predictions to the total observations. It is evident from Table1. that the ANN multilayer perceptron, logistical regression and SVM yield an accuracy 48% to 70% which is not acceptable. The accuracy of any health care system should be in the acceptable range; otherwise it will lead to loss of life.

Table I: Results of Classification Techniques

Sl. No.	Classification Techniques	Precision	Recall	f1-score	Accuracy in %
1	ANN MLP	0.23	0.48	0.31	48
2	Decision Tree	0.83	0.82	0.82	82
3	Logistical Regression	0.61	0.62	0.61	62
4	Random Forest	0.84	0.84	0.85	84
5	SVM	0.71	0.70	0.68	70
6	Ensemble Bagging	0.78	0.76	0.75	83
7	AdaBoost	0.79	0.79	0.79	84
8	Gradient Boosting	0.81	0.81	0.80	86
9	XGB Boosting	0.80	0.80	0.79	86
10	Ensemble Stacking	0.90	0.90	0.90	84

Decision tree classifier provides a result of 82% accuracy but we cannot consider this accuracy as final because decision tree consider only one of the features as split criteria of tree node to derive the rule to classify the input. For this reason we opted for Random forest classifier which produces an accuracy of 84% by considering all features as split criteria tree node before deriving a rule to classify. The above results are obtained with single classifiers. Furthermore to improve the accuracy of the learning model, we decided to use ensemble of classifiers [9][23]. The reason for using ensemble of classifier is statistical, computational and representation. Through ensemble approach estimating the performance and generalization accuracy of the model on unseen data can be achieved.

The performance can be increased by calibrating the learning model and with help of a hypothesis space a best-performing model is chosen. With this it is possible to identify the machine learning model that fits the context, we are dealing with. Ensemble Bagging produces an accuracy of 83% and booting algorithm such as Adaboost, Gradient boosting and XGB boosting produces an accuracy of 84%, 86% and 86 % [18] [19] [20].

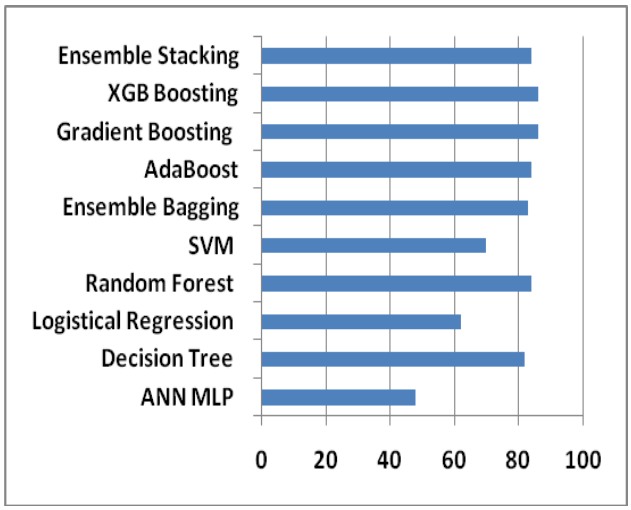

Fig. 2: Accuracy of Classification Techniques

The main variation between different boosting algorithms is their method of weighting training data points and hypotheses. Stacking technique produces an accuracy of 84 %. The accuracy of the various classifiers is depicted in Fig. 2.

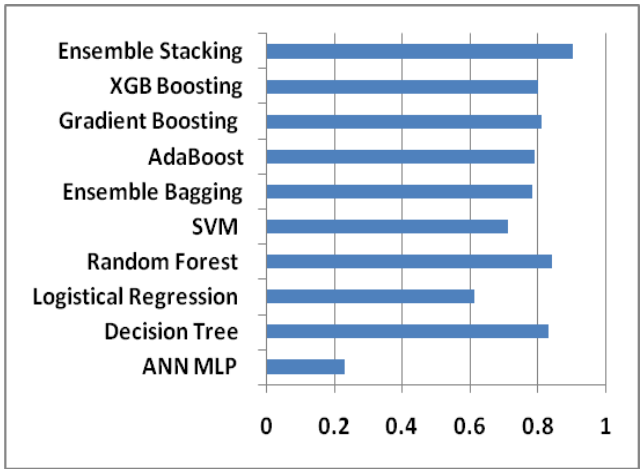

Fig. 3: Average Precision of Classification Techniques

Precision is given by the ratio of positive observations correctly predicted to the total positive observations predicted. Fig. 3 depicts the average precision of various classifiers.

Recall is given by the ratio of positive observations correctly predicted for all observations in actual classes. Fig. 4 depicts the average recall of various classifiers.

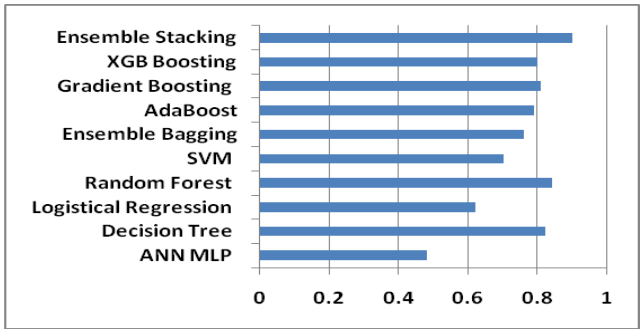

Fig. 4: Average Recall of Classification Techniques

Precision and Recall scores are represented by F1. Precisely F1score is the average weights of Recall and Precision. Both false positive and false negative are taken into consideration during this evaluation. When accuracy is weighed against F1 latter proves to be more useful. Particularly for an uneven class distribution, F1 plays vital role. Accuracy has upper hand when both false positives and false negatives have similar cost. On contrary when the cost of false positives and false negatives are dissimilar Recall and Precision are referred. Various classifiers' average recall is depicted in Fig. 4

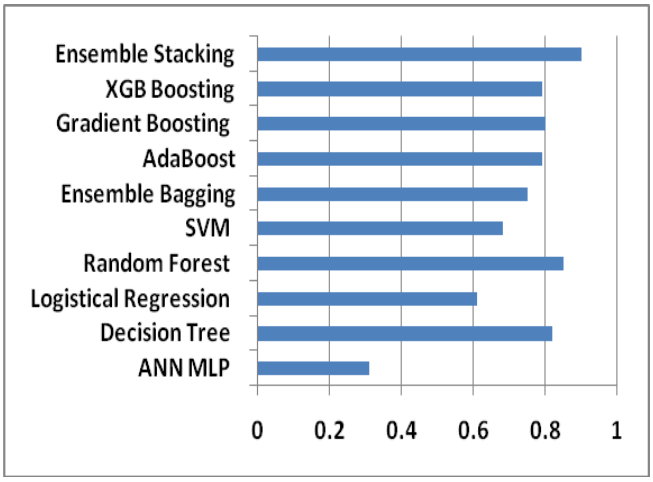

Fig. 5: Average F1-score of Classification Techniques

.

The inference is that gradient boosting and XGB boosting technique produces the highest accuracy of 86 % and it is evident from the Table-1, and is proved in most of the cases. Ensemble models have proven to be stable and thus ensure better performance for majority of the test cases. [23].

VI. CONCLUSION

The focus of our research work is to build a mechanism for preventive care of chronic diseases. In this paper we have explored the Medi-Claim data uniquely due to its authenticity. We have extracted and restructured the data for diabetes disease because of availability of large sample of records, as majority of the population across the globe suffer from diabetes. HCC codes were mapped with ICD and only HCC codes were retained to establish the correlation between hierarchical diseases. ML methods were then applied to establish the correlation between diabetes and other chronicle diseases.

The experiment was conducted using various classification techniques to classify the test case to be diabetic or not along with the indication of hierarchy of occurrences of chronic disease symptoms in case treatment is deferred. The model was evaluated using the confusion matrix. The metrics such as F1 score, precision, recall were considered for evaluation. After thorough investigation it was evident that gradient boosting technique a type of ensemble approach produces the highest accuracy of 86%. Ensemble models exhibit stability and robustness, thus ensuring better performance for majority of test cases. The main overhead of this approach is time. For this reason it is not suitable for real time applications. This work provides the HCC codes and their correlation to diabetic disorder; hence this model helps in prediction of possible occurrence of other chronic diseases.

This work can be extended by including additional features associated with member patients such as prescription codes, habits, profession, location, economical status *etc*.. Including additional features would reduce the bias and variance of data, hence improves the classification accuracy. Also this work provides avenues to predict multiple chronic conditions among beneficiaries. Further attempts can be made to improve the accuracy of classification by adapting advance ML Techniques.

REFERENCES

1. Peter B. Jensen, Lars J. Jensen, Soren Brunak, Mining electronic health records: towards better research applications and clinical care, *Nature Reviews Genetics* 2012, 13:395-405.
2. Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, Lei Hua. ―Data Mining in Healthcare and Biomedicine: A Survey of the Literature. *Springer Science+Business Media*, LLC 2011.
3. DE1.0codebook,Availablefrom:https://www.cms.gov/Research-Statistics-DataandSystems/ DownloadablePublicUseFiles/ SynPUFs/ Downloads/SynPUF_Codebook.pdf
4. Machine Learning Group of University of Waikato, Downloading and InstallingWeka,Availablefrom:http//www.cs.waikato.ac.nz/ml/weka/downloading.html
5. Adler Perotte and George Hripcsak. ―Temporal Properties of Diagnostic Code Time Series in Aggregate. IEEE Biomedical and Health Informatics, Vol. 17, No. 2, pp: 477-483, March 2013.
6. GurbuzE,Biligisayar Muhendisilgi, Ondokuz, ―Diagnostic of diabetes using Adaptive SVM and feature selection, *Signal Processing and Communications Applications (SIU),1EEE* 19th conference, pp:42-45, April 2011.
7. K.R.Lakshmi,Y.Nagesh,M.VeeraKrishna, ―Performance Comparison of three data mining techniques for predicting kidney dialysis survivalility, *International Journal of Advances in Engineering &Technology*,, Vol.7, Issue 1, pp.242-254, Mar 2014.
8. Divya Jain , Vijendra Singh, "Feature selection and classification systems for chronic disease prediction: A review", ELSEVIER, Egyptian Informatics Journal, 2018.
9. D. Ashok Kumar and R. Govindasamy, "Performance and Evaluation of Classification Data Mining Techniques in Diabetes", International Journal of Computer Science and Information Technologies, Vol. 6 (2), 2015.
10. F. H. D. Araújo, A. M. Santana and P. Santos Neto, "Evaluation of Classifiers Based on Decision Tree for Learning Medical Claim Process", *IEEE Latin America Transactions*, vol. 13, no. 1, Jan. 2015.
11. Shen Song, Jim Warren, Patricia Riddle, "Developing High Risk Clusters for Chronic Disease Events with Classification Association Rule Mining", *in the Proceedings of the Seventh Australasian Workshop on Health Informatics and Knowledge Management,* vol. 153, 2014.
12. Varun Chandola, Sreenivas R. Sukumar, Jack Schryver, "Knowledge Discovery from Massive Healthcare Claims Data", 19th*ACM conference on knowledge discovery and data mining*, 2013.
13. Chih-Hung Jen , Chien-Chih Wangb, Bernard C. Jiang , Yan-Hua Chu , Ming-Shu Chen, "Application of classification techniques on development an early-warning system for chronic illnesses", *ELSEVIER Expert Systems with Applications*, 2012.
14. Asha Gowda Karegowda, Punya V, M.A.Jayaram, A.S .Manjunath,"Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5", International Journal of Computer Applications (0975 – 8887), Volume 45– No.12, May 2012.
15. Duygu Çalis , Esin Dog antekin , "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier", *ELSEVIER Expert Systems with Applications* Vol.38 pp. 8311–8315, 2011.
16. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: COLT 1992: *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152. ACM Press, New York, 1992.
17. *Werbos, P.J, "The roots of backpropagation: from ordered derivatives to neural networks and political forecasting", Wiley Interscience, New York, 1994.*
18. Freund, Y., Schapire, R.E, "A decision-theoretic generalization of on-line learning and an application to boosting", In: Vit´anyi, P.M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg, 1995.
19. Freund, Y., Schapire, R.E "Experiments with a new boosting algorithm", In: ICML, pp.148–156, 1996.
20. Schapire, R.E, "The strength of weak learnability", Mach. Learn.,vol 5, 197–227, 1990.
21. https://www.cms.gov/
22. Stephanie J. Hickey, "Naive Bayes Classification of Public Health Data with Greedy Feature Selection", Communications of the IIMA, Vol. 13, pp: 87-97, 2013.
23. Vinitha Dominic, Deepa Gupta, Sangita Khare."An Effective Performance Analysis of Machine Learning Techniques for CardioVascular Disease‖", Applied Medical Informatics. vol 36, No 1, pp: 23-32, March 2015.

AUTHORS PROFILE

Mohan Kumar K N, research scholar in HOD in Dept. of Information Science & Engineering at Adichunchanagiri Institute of Technology, Chikkamagaluru, Karnataka. Presently working as Assistant Professor at SJB Institute of Technology. His area of research includes Machine learning, Deep learning, Pattern Recognition, Game Theory, and Predictive analysis. Authored few International publications which include Journals and Peer-reviewed Conferences.

Dr.S.Sampath presently working as Professor & HOD in Dept. Of Information Science & Engineering at Adichunchanagiri Institute of Technology, Chikkamagaluru, Karnataka. His fields of interest are High Performance computing, Data Mining & Machine Learning. He has been recognized as reviewer for many reputed International Journals in the field of Computer Science including Journal of Computational Science from Elsevier. He has published many research papers in good international journals and conferences.

.

Dr. Mohammad Imran received Ph.D. in computer Science in year 2013 from university of msyore. Presently working as Senior Data Scientist, His area of research includes Machine learning, Deep learning, Pattern Recognition, Computer Vision, Biometrics, Image Processing, Predictive analysis. Authored 35 International publications which include Journals and Peer-reviewed Conferences.