Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2023; Vol 12
Open Access

"eHealth Informatics using EHR: A Comparative Analysis of Data Mining Classification Techniques"

Dr. Vaishali S Parsania
Assistant Professor, Computer Department,
Christ College, Rajkot, Gujarat, India

Cite this paper as: **Dr. Vaishali S Parsania** (2023). **"eHealth Informatics using EHR: A Comparative Analysis of Data Mining Classification Techniques"** . *Frontiers in Health Informatics*, *Vol.12(2023), 597-605 DOI: 10.5281/zenodo.19394781*

Abstract

eHealth informatics has emerged as a transformative domain that integrates healthcare, information technology & data analytics to improve clinical decision-making. With the rapid growth of Electronic Health Records (EHRs), there is an increasing need for intelligent systems capable of extracting meaningful insights from complex and heterogeneous medical data. This paper presents analysis of machine learning classification techniques applied to 5000 EHR datasets to enhance decision support in healthcare systems.

The research utilizes a structured EHR dataset comprising 5000 patient records across multiple diseases and healthcare attributes. Various classification algorithms, including Naïve Bayes, BayesNet, Random Forest, JRip, OneR, and PART, are implemented and evaluated using performance metrics. A 10-fold cross-validation approach is employed to ensure reliability and robustness of the results.

The findings reveal that probabilistic models, particularly Naïve Bayes and BayesNet, demonstrate comparatively stable and efficient performance across multiple evaluation parameters. The study highlights the importance of selecting appropriate algorithms based on dataset characteristics and healthcare requirements.
This research contributes to the eHealth informatics by providing insights into the effectiveness of machine learning approaches in healthcare data analysis. It also reveal the potential of integrating analytics into healthcare systems to achieve improved accuracy, efficiency, and quality of care.

Keywords

eHealth Informatics, Electronic Health Records (EHR), Machine Learning, Data Mining, Healthcare Analytics, Classification Algorithms, Naïve Bayes, Random Forest, Predictive Modeling

Introduction:

Distribution of EHR dataset with 5000 Records:

The distribution of database of EHR is shown to have an insight about the sub data type dispersion in the EHR used. The EHR distribution shown here under contains 5000 datasets.

Distribution According to Disease:

The data shown in the figure show how the data are grouped in each attribute per disease. Different color reflects the types of data accommodated by each of the attribute.

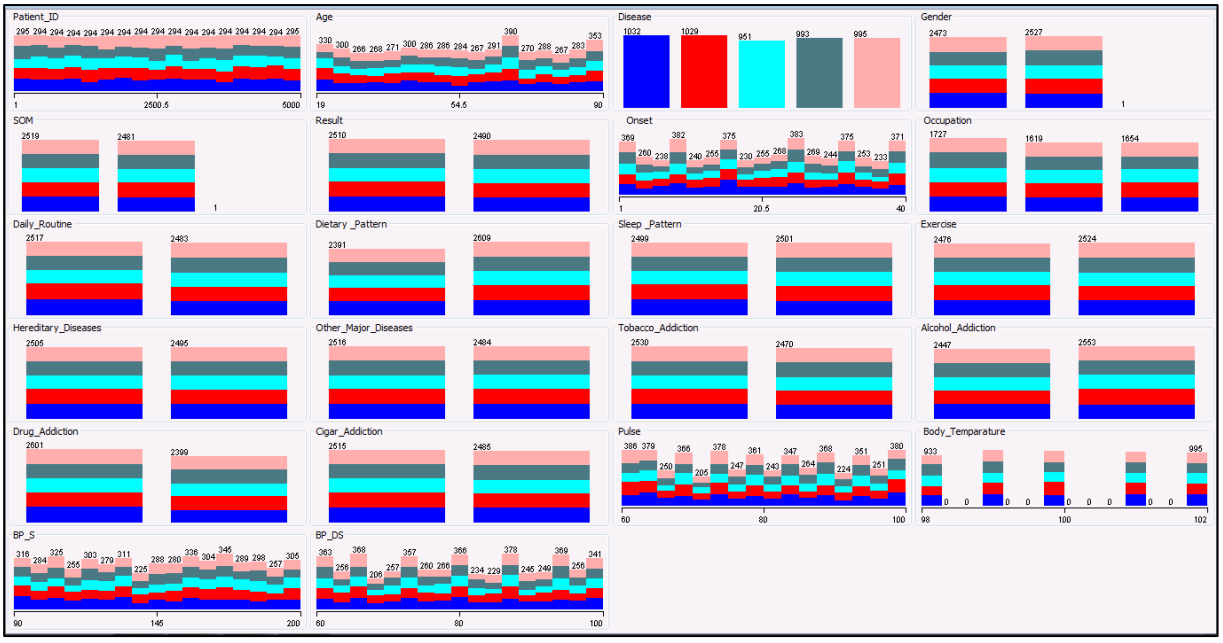


Figure: Distribution of EHR Datasets of size 5000 for each attribute Disease wise

This figure underneath indicates insight about how EHR Datasets of size 5000 is distributed for each disease (Gastritis, RA, Allergy, Jaundice, Fever). There are 1032 records of Gastritis, 1029 records of RA, 951 records of Allergy, 993 records of Jaundice and 995 records of Fever.

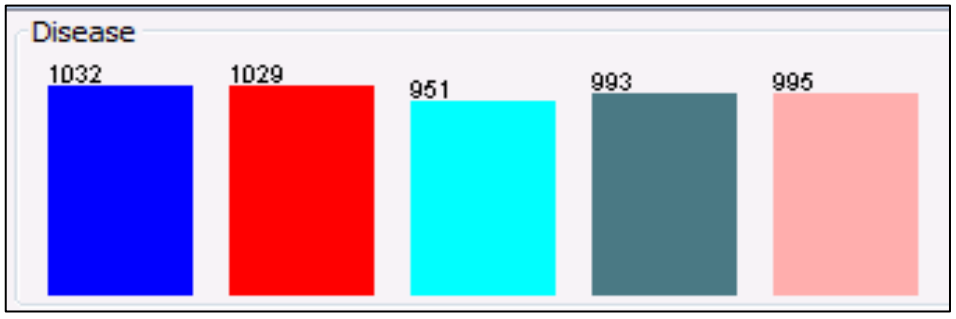


Figure: Distribution of EHR Datasets of size 5000 Disease wise

For the purpose of how the distribution of data is done according to disease in the EHR dataset of 5000 records, is illustrated in detail by taking one attribute Dietary Pattern. There are 2391 records of irregular dietary pattern and 2609 records of regular dietary pattern. Blue color indicates the disease Gastritis, Red color indicates disease RA, Bottle Green color indicates disease Allergy, Grey color indicates disease Jaundice, Pink color indicates disease Fever. Similar distributions are indicated for remaining 21attributes in the above figure.

Distribution According to SOM

The data shown in the figure show how the data are grouped in each attribute per system of medicine. Different color reflects the types of data accommodated by each of the attribute.

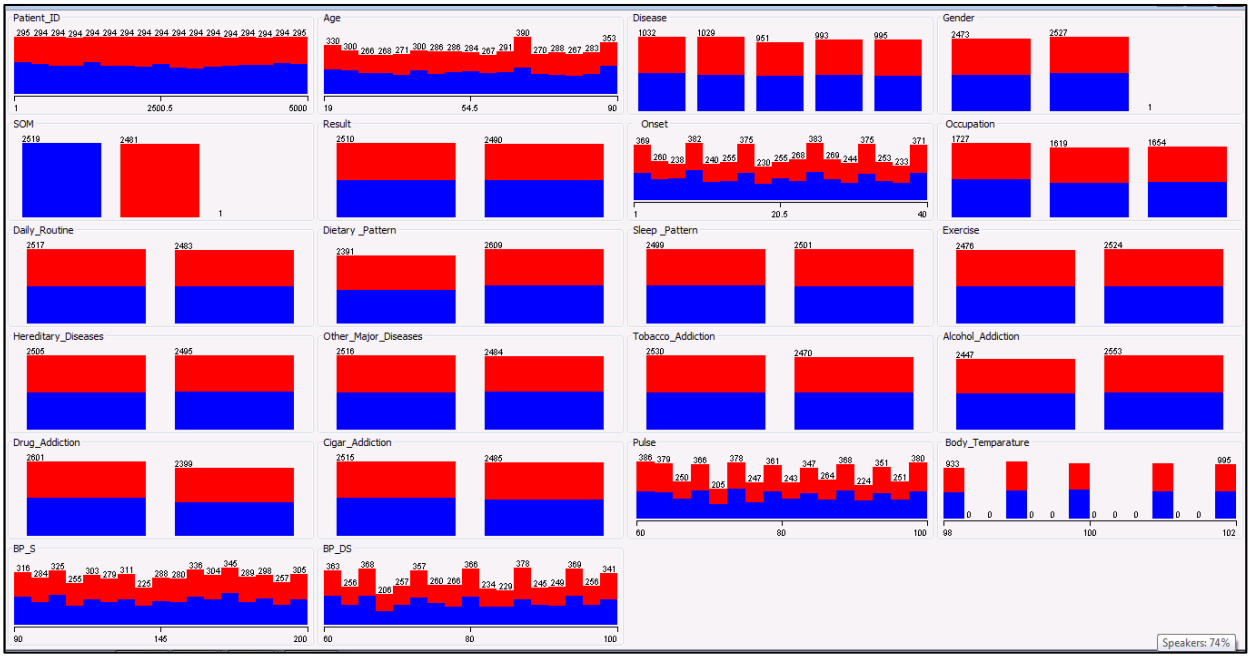


Figure: Distribution of EHR Datasets of size 5000 for each attribute SOM wise

The figure given below indicates insight about how EHR Datasets of size 5000 is distributed for each System of Medicine (Ayurvedic, Allopathic).

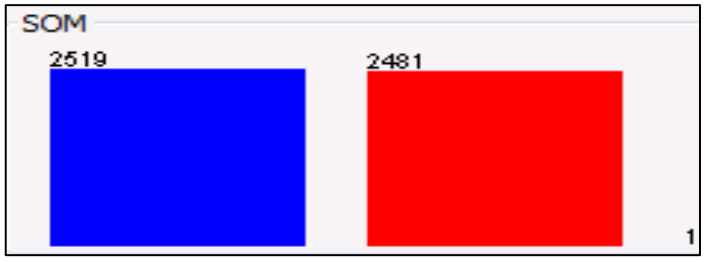


Figure: Distribution of EHR Datasets of size 5000 SOM wise

The figure here under indicates insight about how EHR Datasets of size 5000 is distributed Occupation wise for each System of Medicine (Ayurvedic, Allopathic).

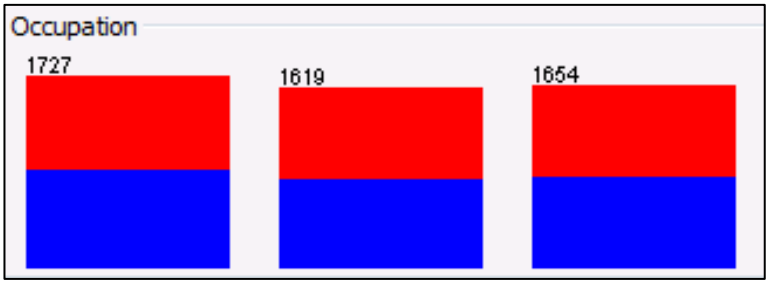


Figure: Distribution of EHR Datasets of size 5000 SOM wise

For the purpose of how the distribution of data is done according to System of Medicine (SOM)

in the EHR dataset of 5000 is illustrated in detail by taking attributes Occupation. Here 1727 records are of Stressful Work, 1619 Physically Hard Work and 1654 Light Work in the EHR dataset. Blue color indicates the SOM AYU and Red color indicates SOM ALO.
Similar distributions are indicated for remaining all the attributes in the above figure.

Performance Measures of Classifiers:

Measuring performance of specified classification algorithms has significant effect on evaluation and analysis.
For measuring performance of Naïve bayes, BayesNet, Random Forest, PART, JRip and OneR classification techniques the following parameters are taken. In Classification techniques parameters to be examined are accuracy, sensitivity, precision, specificity, f-measure, ROC Area, Kapa Statistics, Mean absolute error, Relative absolute error and confusion matrix [2].

Confusion Matrix:

A confusion matrix is a simple methodology for displaying the degree of accuracy of classification results. The confusion matrix is defined by labeling the desired classification on the rows and the predicted classification on the columns [2]. Confusion matrix is described as follow.
Table 1: Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	**a**	**b**
	Positive	**c**	**d**

- *a* is the number of **right** presumption that an instance is **negative**,
- *b* is the number of **wrong** presumption that an instance is **positive**,
- *c* is the number of **wrong** of presumption that an instance **negative**, and
- *d* is the number of **right** presumption that an instance is **positive**.

Accuracy:

Accuracy is the percentage of predictions those are correct. The accuracy is a measurement method for the amount of imminence of measurements of an amount to that amount correct value. The overall accuracy is the ratio between the total number of correctly classified instances and the test set size [3].
Accuracy = (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)

Sensitivity:

Sensitivity is numerical measures of the concert of a binary categorization test identified as classification function in statistics. Sensitivity is also known as recall rate which procedures percentage of real positives which are rightly recognized. Sensitivity represents probability of correctly labeled members of the target class [4].

Sensitivity = positives correctly classified / total positives = TP / (TP+FN)

Specificity:

Specificity is numerical measures of the performance of a binary classification test, also known in statistics as classification function. Specificity procedures the amount of negatives which are rightly recognized. The specificity is a statistical measure of how accurately a binary classification test correctly identifies the negative cases [4].

Specificity = negatives correctly classified / total negatives = TN / (TN+FP)

Precision:

Precision is probability that a positive prediction is correct. It is also called positive predictive value and is the portion of fetched instances that are appropriate. High precision means that an algorithm returns substantially more relevant results than irrelevant. [5]

Precision = True Positives / (True Positives + False Positives)

False Positive Rate:

The false positive rate (FP) is the percentage of negatives cases that were mistakenly classify as positive. A false positive error is an outcome which shows a given condition has been satisfied; when in reality has not been satisfied it means incorrectly a positive effect has been tacit. [2]

F-Measure:

The F-Measure computes average of the information retrieval precision and recall metrics. It can be used to determine a performance of algorithm. The F-measure is defined as a harmonic mean of precision (P) and recall (R).

F-measure = 2×precision×recall / precision + recall

ROC Curve

Receiver Operating Characteristics (ROC) graphs have long been used in signal detection theory to depict the tradeoff between hit rates and false alarm rates over noisy channel. A ROC graph depicts relative tradeoffs between benefits (true positives) and costs (false positives). [6]

Kapa Statistics

Kapa statistics is an index that evaluates the agreement against which might be expected by chance. Kappa statistic is a generic term for several similar measures of agreement used with classified data. The kappa measure of agreement is the ratio

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

Where P (A) is the probability of times the k evaluators agree, and P (E) is the probability of times the k evaluators are expected to agree by chance alone. [7]

Mean Absolute Error:
Mean absolute error is the average of the difference between predicted and the actual value in all test cases; it is the average prediction error [7].

Relative Absolute Error:
Relative Absolute Error is the total absolute error made relative to what the error would have been if the prediction simply had been the average of the actual values [7].

Cross Validation:
In the task of implementation of classifier on EHR dataset 10-fold cross validation is used to have accurate results. Each fold will serve as a new test set and thus can be consider identical with hold-out estimations.

Implementation of Algorithms on EHR Dataset of size 5000
Selected classifiers are implemented in Weka environment and parameters like Accuracy, Sensitivity, Precision, Specificity are calculated for measuring performance of those classifiers when applied on EHR Dataset of size 5000.
Calculation for Naïve Bayes Classifier:
TP=1326 ,TN=1231 , FP=1259 , FN=1184
Precision = TP / (TP + FP) = 1326/ (1326+ 1259) = 0.513
Sensitivity = TP / (TP + FN) = 1326/ (1326+ 1184) = 0.528
Specificity = TN / (TN + FP) = 1231/ (1231 + 1326) = 0.512
Accuracy = (TP + TN) / (TP + TN + FP + FN) = (1326 + 1231)/ (1326 +1231 +1259 +1184) = 0.511

Table2: Outcomes of Classification Algorithms applied on EHR Dataset of 5000

Performance Measures	**Classifier Applied on EHR Dataset of size 5000**					
	BayesNet	**Navie bayes**	**Random Forest**	**JRip**	**OneR**	**PART**
Accuracy	0.511	0.511	0.499	0.504	0.502	0.507
Sensitivity	0.527	0.528	0.61	0.523	0.51	0.51
Precision	0.513	0.513	0.501	0.506	0.502	0.509
Specificity	0.495	0.512	0.388	0.485	0.495	0.504
F-measure	0.520	0.521	0.55	0.514	0.504	0.51
ROC Area	0.514	0.514	0.497	0.5	0.502	0.506
Kapa Statestics	0.022	0.023	-0.0021	0.008	0.004	0.014
Mean Absolute Error	0.499	0.499	0.501	0.499	0.497	0.494
Relative Absolute Error	0.998	0.990	1	0.99	0.99	0.988
TP Rate	0.527	0.528	0.61	0.523	0.51	0.51
FP Rate	0.505	0.506	0.612	0.515	0.505	0.508

In the above table classifiers output taken on the 5000 EHR dataset is shown. The statistical parameters like Accuracy, Sensitivity, Specificity, F-measure, Precision etc can be considered for the analytical purpose from the specified classifiers output.

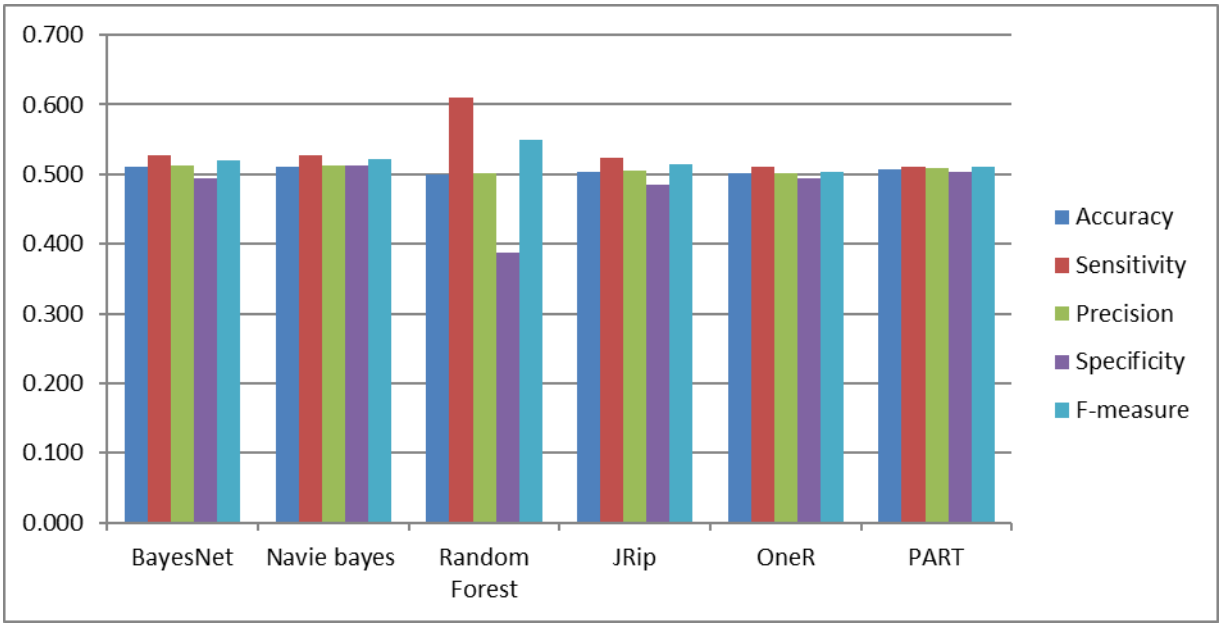


Figure1: Performance Measures (1) of Classifiers applied on EHR of 5000

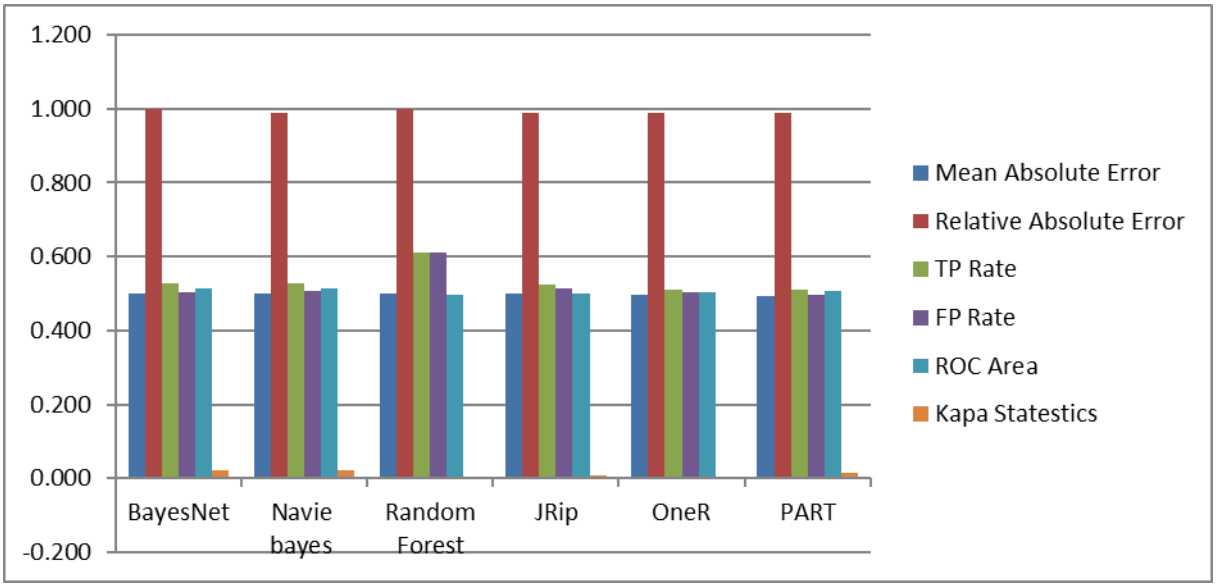


Figure: Performance Measures (2) of Classifiers applied on EHR of 5000

From the above table and charts it can be exposed that Accuracy and Precision is better in Naïve bayes and BayesNet as compared to other classifier. Specificity result is also better in Naïve bayes and PART. Sensitivity and F-measure are better in Random Forest and Naïve Bayes. Mean absolute error and relative absolute error are less found in PART. TP Rate is good in Naïve bayes and Random Forest. FP Rate is good in OneR, Naïve bayes and BayesNet. Kapa statestics is better found in Naïve bayes and BayesNet. [8]

Naïve bayes, BayesNet and Random Forest are generating good results. Considering majority of parameters, it can be concluded Naïve bayes is better as compared to the other data mining classifier when applied on the selected EHR of size 5000.

The above charts are based on the table shown to exhibit outcome of classification algorithms applied on EHR Datasets of different sizes. Comparative analysis between specified classifications algorithms applied on EHR Dataset is made easy by these pictographic representations.

Statistical comparison of results

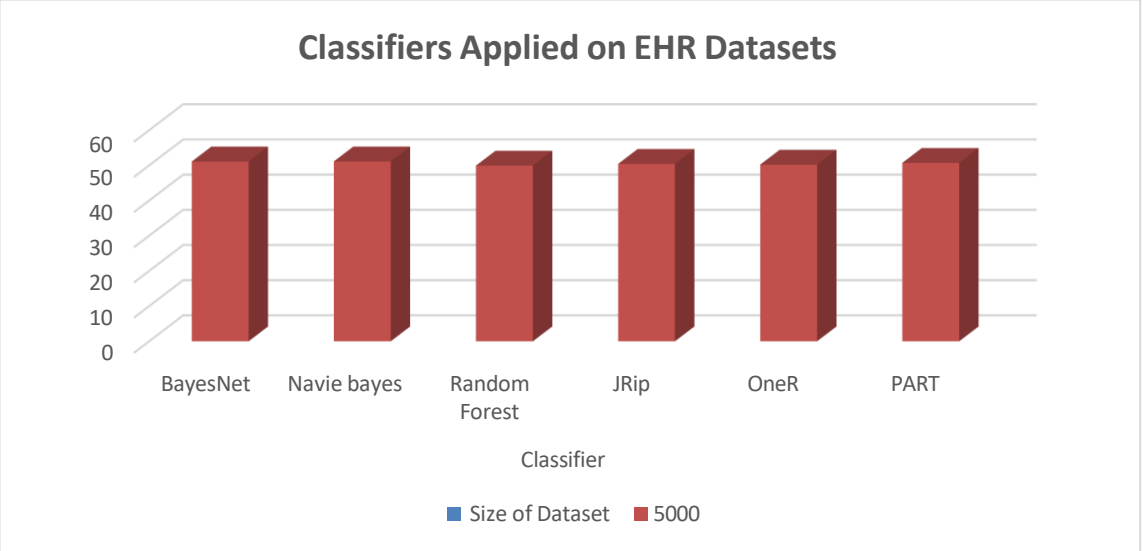
The implementation of above selected classification algorithm in Weka environment is done with 4 EHR datasets of 500, 1000, 2000 and 5000.

The EHR dataset is subjected to BayesNet, Naïve Bayes, Random Forest, JRip, OneR and PART algorithms. The results of these classification algorithms are displayed in above section in detail with many performance measures. The table here under shows the correctly classified instances of specified classification algorithms applied on the EHR datasets of different sizes.

Table: Correctly Classified Instances of Classifiers Applied on EHR Datasets

	Classifier					
	BayesNet	**Navie bayes**	**Random Forest**	**JRip**	**OneR**	**PART**
Size of Dataset						
5000	**51.1**	**51.14**	**49.94**	**50.42**	**50.22**	**50.72**

The table shows comparative pictographic view of correctly classified instances of different classification algorithms applied on EHR datasets.



Conclusion:

The above table and chart shows that varied statistical outcomes taken with different classifier. In BayesNet, Random forest, JRip, OneR and PART results are fluctuated with data size. When

data size grows or shrinks the steadiness is not maintained. In Naïve Bayes as compared to other algorithms results are better. The average result of Naïve Bayes is also comparative good.

By further modifying and altering this basic Classification algorithm the better algorithms can be designed. This algorithm may give the better results in terms of Accuracy and other statistical parameters. In the subsequent paper improved level of algorithms can be designed with the intention of getting the better result.

References:

[1] Dr. Vaishali S Parsania, "Retrieving Information from Special EHR of Multiple Systems of Medicine by Applying Naïve Bayes, BayesNet, PART, JRip and OneR Classification Algorithms", International Journal of Creative Research Thoughts (IJCRT), ISSN 2320-2882, Impact Factor: 7.1, Volume 3, Issue 2 May 2015

[2] A. Dean Forbes, "Classification-algorithm evaluation: Five performance measures based on confusion matrices", Journal of Clinical Monitoring , May 1995, Volume 11, Issue 3, pp 189-206

[3] Andy Neely, Huw Richards, John Mills, Ken Platts and Mike Bourn, "Designing performance measures: a structured approach", University of Cambridge, Cambridge, UK

[4] "Sensitivity & Specificity", Lippincott Williams & Wilkins, "Studying a Study and Testing a Test", 2005, 157- 159

[5] Ranawana, R. Palade, V., "Optimized Precision - A New Measure for Classifier Performance Evaluation", Evolutionary Computation, 2006. CEC 2006. IEEE Congress on, 2254- 2261, 2006

[6] Ma S. and Huang J. Regularized roc method for disease classification and biomarker selection with microarray data. Bioinformatics, 21(24):4356, 2005

[7] Olaiya Folorunsho, "Comparative Study of Different Data Mining Techniques Performance in knowledge Discovery from Medical Database", IJARCSSE, Volume 3, Issue 3, March 2013

[8] Vaishali V Kaneria et. al., "Applying Naïve Bayesian Classifier for Getting Probability Based Result for E-Knowledge Services In Healthcare", International Journal of Engineering Research & Technology (IJERT)

[9] Dr. Vaishali S Parsania et. al., "Applying Naïve bayes, BayesNet, PART, JRip and OneR Classification Techniques on Hypothyroid Database for Comparative Analysis", (IJDI-ERET) (ISSN 2320-7590), Vol. 3, No. 1, June- 2014