Clinical Data Analytics in Healthcare using Machine Learning: Mortality Prediction

[1]Patel Krupal I, [2]Dhaval A. Parikh, [1]ME Student, [2]Associate Professor
Lalbhai Dalpatbhai College of Engineering Ahmedabad,
Gujarat Technological University, Ahmedabad, India.

Abstract: Big data and data analytics changed the way we manage, visualize and analyze data. Now days so much of money is spent for medical treatment and cost incurred during stay at hospital is very high. Predicting stays and possibilities for diseases will reduce this cost. Data analytics also create opportunity in healthcare to and insight from all clinical data captured from many sources like electronics device installed at hospitals and from notes generated from nurses and doctors. Solving problem using machine learning require problems domain knowledge, different machine learning algorithms and understanding of statistics. Accuracy of machine learning algorithm varies and depends on many factor like the nature, size and quality of the data. In this paper, we propose a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of Mortality Prediction. The prediction model is introduced with different combinations of features and several known classification techniques. Many useful reports and predictive results will be generated as an outcomes which will be useful for doctors to diagnose patients.

Keywords – **Machine learning, mortality prediction, feature selection, prediction model, classification Algorithms.**

I. INTRODUCTION

It is difficult to identify mortality because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Various techniques in data mining and neural networks have been employed to and out the severity of heart disease among humans. The severity of the disease is classified based on various methods like K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naïve Bayes (NB). The perspective of medical science and data mining are used for discovering various sorts of metabolic syndromes. Data mining with classification plays a significant role in the prediction of Mortality and data investigation. Various methods have been used for knowledge abstraction by using known methods of data mining for Mortality Predication. In this work, numerous readings have been carried out to produce a prediction model using not only distinct techniques but also by relating two or more techniques. These amalgamated new techniques are commonly known as hybrid methods.

The rest of the paper is organized as follows, Section II discuss the basic of machine learning. Section III discusses machine learning methods. Section IV framework of machine learning. Section V Mortality related works, existing methods and techniques available. Section VI discusses proposed method. Section VII mimic data set. Section VIII discusses the health parameter. Section IX Data pre-processing followed by feature selection, classification modeling and performance measure. Section X gives the algorithms used and the experimental setup. Section XI shows the evaluation of datasets and experimental setup. It also shows how the experiment was conducted and the results that were achieved. Section XII comparison of hybrid algorithm with other existing algorithm. Section XIII ends with a conclusion of current work and some notes on future enhancement.

II. Machine Learning

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

III. Machine Learning Methods

Two of the most widely adopted machine learning methods are supervised learning which trains algorithms based on example input and output data that is labeled by humans, and unsupervised learning which provides the algorithm with no labeled data in order to allow it to find structure within its input data.

Supervised learning, the computer is provided with example inputs that are labeled with their desired outputs. The purpose of this method is for the algorithm to be able to "learn" by comparing its actual output with the "taught" outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabeled data.

Unsupervised learning, data is unlabeled, so the learning algorithm is left to find commonalities among its input data. As unlabeled data are more abundant than labeled data, machine learning methods that facilitate unsupervised learning are particularly valuable. The goal of unsupervised learning may be as straightforward as discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify raw data.

Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabeled data generally doesn't require additional resources.

Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

IV. Framework for machine learning

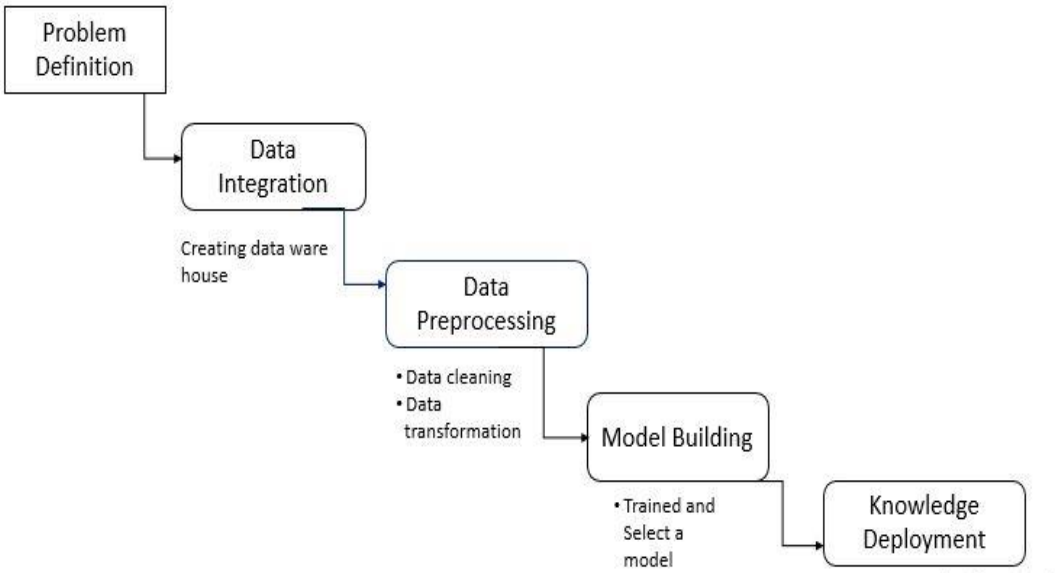

Figure 1: the generator is in a feedback loop with the discriminator and generator [1]

V. Closely related work

There is ample related work in the fields directly related to this paper. ANN has been introduced to produce the highest accuracy prediction in the medical field [6]. The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear Model[2].we have analysed the early prediction of diabetes by taking into account various risk factors related to this disease using machine learning techniques. Extracting knowledge from real health care dataset can be useful to predict diabetic patients [5].

Support vector machine and use it to predict lifestyle diseases. The simulated model will prove to be an intelligent low-cost alternative to detect possible genetic disorders caused by unhealthy lifestyles [6]. Machine learning techniques as missing value imputation. The results are compared with traditional mean/mode imputation. Machine learning techniques may be the best approach to imputing missing values for better classification outcome[3].

This study of feature selection algorithms of large survey shows that the feature selection algorithm consistently improves the accuracy of the classifier [7].Random Forest Algorithm outperforms all the other algorithms in terms of accuracy and when the size of the dataset is reduced to almost half of the original, then Naïve Bayes Algorithm shows the best results in terms of accuracy [9]. In this paper, predictive models by using machine learning methods including KNN, SVM, LR, and decision tree classifiers to predict chronic kidney disease. From the experimental results, SVM classifier gives the highest accuracy [4].

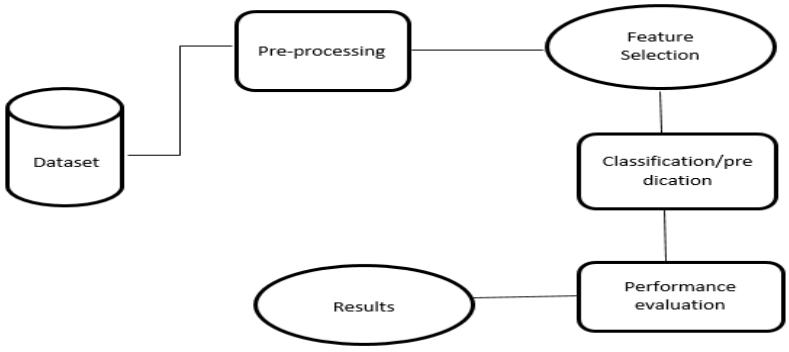

Figure 2: experiment overflow [5]

VI. Proposed method

Hybrid ML Algorithm- clustering method and Random Forest. Hybrid method has stronger capability to predict mortality compared to existing methods. This hybrid algorithm uses bagging of random forest to improve the stability, reduce variance and accuracy.

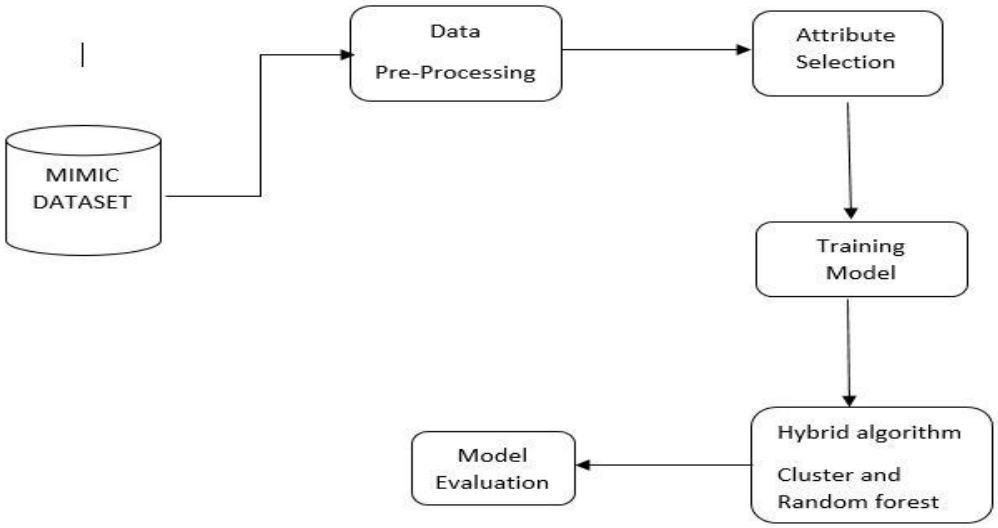

Figure 3: Overall proposed architecture and training [2]

VII. Dataset

- Performance of proposed model is evaluated on critical care database of MIMIC-III.(https://physionet.org/content/mimiciii/1.4/)
- MIMIC-III is a relational database consisting of 26 tables in which five tables are used to define and track patient clinical data.
- chartevents table (30,712,483 Record)
- labevents table (27,854,055 Record)
- patients table (46,520 Record)
- d_items table and d_labitems table

Patient table contains important column like Gender, DOB, DOD and Expire Flag. Chart table contains column like charttime, storetime, CGID, value, label etc. But we need only Label and their values. Lab table also has same schema as chart table.

VIII. Health Parameters

Chart data: Contains all the charted data available for a patient. During ICU stay, the primary repository of a patient's information is their electronic chart. i.e. Patient controlled analgesia (PCA), Chest PT, Protonix etc.

Lab data: Data contains information regarding laboratory based measurements. i.e. Blood from an arterial line, urine from a catheter etc.

Healthcare is changing from traditional practice to new modern evidence based practice. In this new practice, patient data are collected from different devices which are installed at hospitals site. The amount of data generated at hospitals are very huge. So, we have to apply data processing methods to reduce that data. Reducing data without losing information is biggest challenge. In ICU, different types of data are generated. Some of the data are periodic which chart events data. Some data are collected manually which are lab events data. We first uses lab events data for our implementation. Then, we used chart events data for evaluation of machine learning algorithm accuracy.

The lab test value from test is very important parameters for diagnosing patient condition. If particular lab test value is not in specified range, then that patient may face some serious problem. In our experiment we uses average value of tests performed on the patient. That average test value can be used by doctor for assistance in diagnosis.

IX. Dataset Preparation

The data in database are stored in very normalized way. Because there are more than 500 lab and chart events are defined in medical field. Not all lab tests are performed on every patient. Most lab tests are performed according to the diseases. Many tests have to be performed frequently on patient. So, we uses average lab test result and chart result value for our experiment. So per patient one

instance with all lab test and chart value. If lab test was not performed on patient then 0 assigned to that lab test result same as with chart value.

- **Data import –** MIMIC dataset are in excel format. So we have to import excel data into MySQL.
- **Data Pre-Processing** – Spreading Data , NA/Null values removal, Redundant data removal, Normalizing attribute value
- **Attribute selection –** Applying Attribute selection method using Weka

X. Training Model

After Data preparation data apply to Hybrid model. Here is an algorithm

Algorithm 1: Clustering of dataset

Require:

Input: Datasets with patient information
for ∀apply the k-means algorithm **do**
on the dataset R(d1,d2,d3,. . . . . dn)
end for

Output: Clustered datasets with K-means C(c1, c2,. . . . . .cn)

Algorithm 2: Radom forest on Clusters

Require:

Input: Clustered datasets C(c1, c2, c3,....cn)
for ∀Apply random forest algorithm **do**
on the clustered dataset C(c1, c2, c3,....cn)
end for
Output: Trained models RF(rf1, rf2,...rfn)

For training samples of data R, We are applying K-Means to identify the clusters. This cluster can be created on the basis of type of data i.e. lab events or chart events data. The clustering of datasets is done on the basis of the variables and criteria of K-means. Then, the Random forest classifiers are applied to each clustered dataset in order to estimate its performance. For Testing, we are partitioning the same cluster into train and test data. The best performing models are identified.

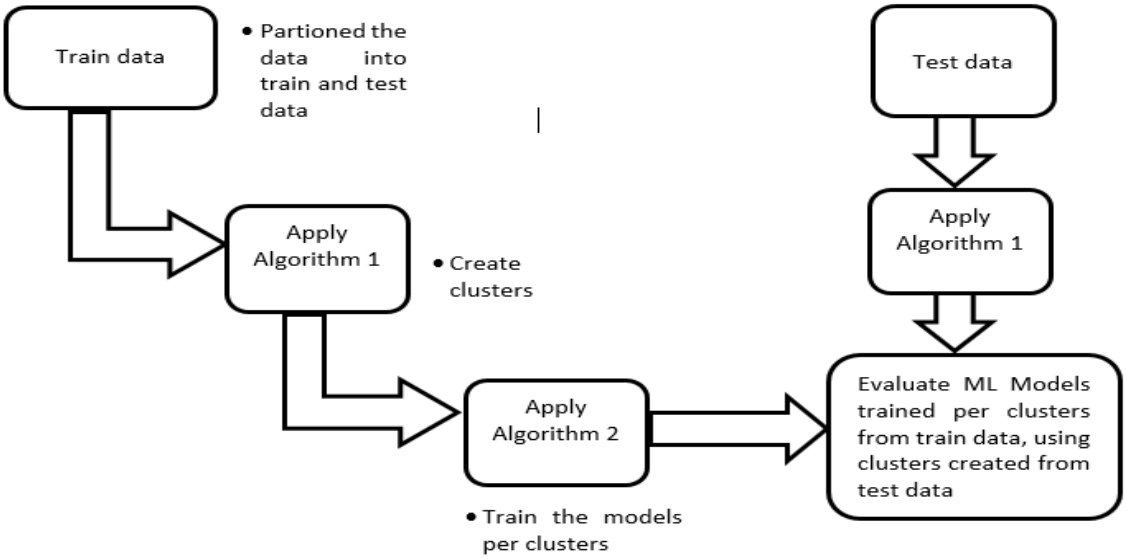

Figure 4: algorithm explanation

XI. Model Evaluation

Model Evaluation done by confusion matrix. A confusion matrix is a table that is often used to describe the performance of a classification model

Results of hybrid algorithm shown in table. Performance of our model is calculated by taking average of all clusters performances.

Algorithm	Perfomance Evaluation Parameters		
	Accuracy	Precision	F-Measure
Hybrid	0.78	0.83	0.85

Table 1: Results of hybrid algorithm

XII. Comparison

Comparison graph indicates that hybrid algorithm gave batter results then other algorithms.

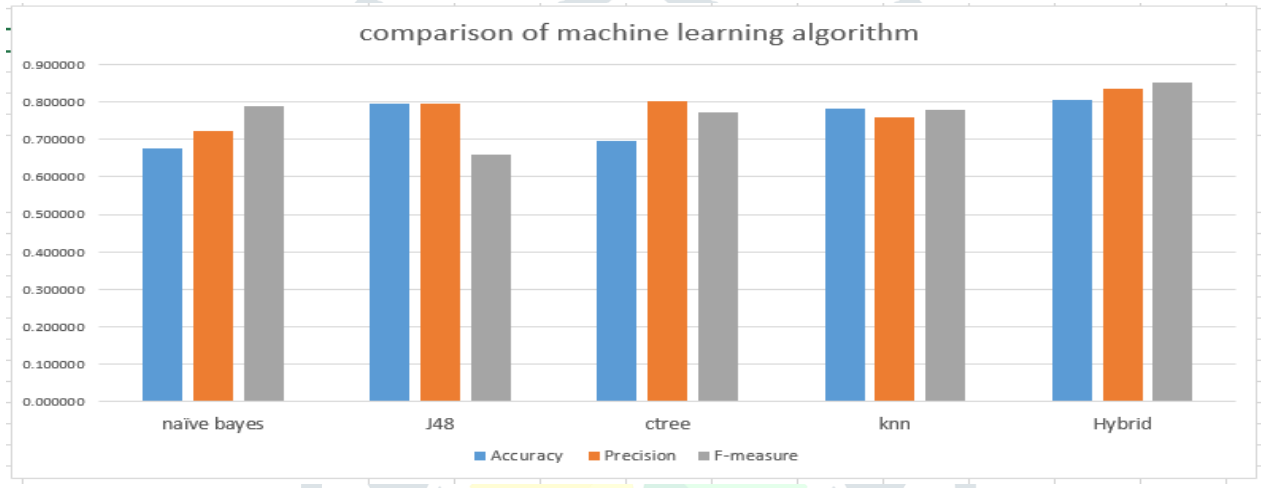

Figure 5: comparison of algorithms

XIII. Conclusions

Identifying the processing of raw healthcare data of Mortality information will help in the long term saving of human lives and early detection of abnormalities in health conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards Mortality. Mortality prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if Mortality is predicted at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world datasets instead of just theoretical approaches and simulations. The proposed hybrid approach is used combining the characteristics of Random Forest (RF) and Cluster. Hybrid algorithm proved to be quite accurate in the prediction of Mortality prediction. The future course of this research can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Furthermore, new features election methods can be developed to get a broader perception of the significant features to increase the performance of Mortality prediction.

XIV. References

1. Microsoft Azure, “How to choose algorithms for Microsoft Azure Machine Learning.” https://docs.microsoft.com/en-in/azure/machine-learning/machine-learning-algorithm-choice.

2. Senthilkumar Mohan, Chandrasegar Thirumalai1, And Gautam Srivastava “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”publication June 19, 2019

3. M. Mostafizur Rahman and D. N. Davis “Machine Learning-Based Missing Value Imputation Method for Clinical Datasets” IAENG Transactions on Engineering Technologies 2015

4. AnusornCharleonnan, ThipwanFufaung, TippawanNiyomwong "Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques" MITiCON-2016.

5. Md. Faisal Faruque, Iqbal H. Sarker "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus" international Conference on Electrical, Computer and Communication Engineering (ECCE) IEEE 2019

6. MrunmayiPatil, Vivian Brian Lobo , Pranav Puranik "A Proposed Model for Lifestyle Disease Prediction Using Support Vector Machine" 9th ICCCNT 2018 July 10-12, 2018, IISC

7. S. Vanaja and K. R. kumarBharathiar University Coimbatore Tamil Nadu India, "Analysis of Feature Selection Algorithms on Classification: A Survey". International Journal of Computer Applications 2014

8. R. Venkatesh, C. Balasubramanian & M. Kaliappan "Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique" Journal of Medical Systems Springer (2019)

9. Indu Kumar, KiranDogra, ChetnaUtreja, Premlata Yadav "supervised machine learning algorithms for stock market trend prediction" ICICCT 2018

10. Ripon Patgiri, HemanthKatari, Ronit Kumar, and Dheeraj Sharma "Empirical Study on Malicious URL Detection Using Machine Learning" Springer Nature Switzerland AG 2019

11. Geeta Chhabra, VasudhaVashisht "A Classifier Ensemble Machine Learning Approach to Improve Efficiency for Missing Value Imputation" GUCON Sep 28-29, 2018

12. Sahil Dhankhad, Emad A. Mohammed, Behrouz Far "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection" IEEE International Conference on Information Reuse and Integration for Data Science 2018

13. Sergio Ledesma , Mario-Alberto Ibarra-Manzano, Eduardo Cabal-Yepez, "Analysis of Data Sets With Learning Conflicts for Machine Learning" Department of Electrical and Computer Engineering, ACCESS IEEE 2018

14. Xianchuan Wang, Zhiyi Wang, Jie Weng, Congcong Wen, "A New Effective Machine Learning Framework for Sepsis Diagnosis" Department of Computer Science ACCESS IEEE 2018

15. Laxmi Narayana Pondhu, Govardhani Kummari "Performance Analysis of Machine Learning Algorithms for Gender Classification" International Conference on Inventive Communication and Computational Technologies IEEE 2018

16. Riya Roy, Thomas George K "Detecting Insurance Claims Fraud Using Machine Learning Techniques" International Publishing Switzerland IEEE 2017

17. "Machine learning flow chart" https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94.