Death Prediction of Heart Disease Patients using Machine Learning

Ranjani Mudaliar
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, Tamil Nadu, India
mudaliar.ranjani2019@vitstudent.ac.in

Sonal Rajurkar
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, Tamil Nadu, India
sonal.rajurkar2019@vitstudent.ac.in

Sumit Tile
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, Tamil Nadu, India
sumit.tile2019@vitstudent.ac.in

Prof. Vijayarajan V
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, Tamil Nadu, India
Vijayrajan.v@vit.ac.in

Abstract: As the world is stepping forward in the field of technology, people are getting addicted towards it. They are more concerned about their future and career but not giving importance to their health. As a result, there is a tremendous increase in health issues. Due to excess impact of technology on the lifestyle of human beings the percentage of early deaths is increasing enormously. Technology can ruin the life of the people; on the other side it is saving lives. In this paper we are going to process and analyze the data generated by past early deaths. With the help of these data we will predict the enduring life span of the patient having some severe disease with the help of Machine Learning Techniques. This will be helpful for taking the extra precautions to increase their life span by medical treatment.

I. INTRODUCTION

In the medical field there are many technologies used already for diagnosing disease but there is no technology that can help us to detect the death of the person i.e. to estimate time how long a person can survive after diagnosing the disease so that it can be helpful to doctors try to cure them using different medication. In this paper we tried to detect the death of the person due to heart failure using some Machine Learning Techniques. So, this will be helpful to cure the patients and reduce the death rate.

There are many Machine Learning Techniques that can be used but as it is related to human life it should be more accurate. So, to reach good accuracy there is one technique i.e. Ensemble Learning. In Ensemble learning, we try to combine the output of different models fit over the same dataset which helps to find the best accuracy from all the different models tried over the Ensemble.

II. LITERATURE SURVEY

- **"Early Stage Prediction of sudden cardiac death"**
  The cardiac is an adequate reason for sudden death. This paper gives the information about the study carried out on the sudden cardiac death that happened depending upon the data gathered from various patients and their ECG reports. They have used different machine learning techniques like Multilayer Perceptron (MLP), Support Vector Machine (SVM) and k-nearest neighbor (KNN). With use of these machine learning algorithms in a given method the system can predict the sudden death one hour before the occurrences which will be useful to save lives of the patients in hospitals. [1]

- **"An Automated Strategy for early risk identification of sudden cardiac death by using machine learning approach on measurable arrhythmic risk markers"**
  To increase the endurance chances of a patient having cardiac disease, the early risk prediction of an unexpected sudden cardiac death can play

a major role. It can increase the possibilities of survival of lives by using automated prediction strategies based on measurable arrhythmic markers. This strategy includes the set of conduction and arrhythmic conduction-repolarization markers and these parameters are calculated by using the ECG (Electrocardiogram) reports and QRS markers. This data is used by algorithms like “Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF), and k-Nearest Neighbor (KNN)” for classification of normal and sudden cardiac death risk groups. [2]

- **“The Prediction of a Potentially Fatal Cardiac Event in the Next 2 to 24 Hours and The Prediction of a Myocardial Infraction Related Death or Sudden Death”**
  The prediction base is a combination of markers such as patients who had sudden death, who had Post Myocardial Infraction and the normal patients. This can predict the MI related death in the next 2 to 24 hours. This study held on the data gathered from the 39 patients who died by sudden death and 44 post MI patients. The method comprises various eleven GWS predictive markers and the results in case of sudden death are 7% false negative and 3% false negative in case of potentially fatal cardiac event. [3]

- **“Death Prediction and Analysis Using Web Mining Techniques”**
  The research includes the data gathered from news from Queensland Government information of the past eighteen years, digital histories of patients from electronic patient records and data from freely available web resources which will help together in accurate real time prediction of death. Classification technique is used for the prediction and the multilayer perceptron algorithm, linear regression, MLP regressor and SMO regression is used in it. In all methods the linear regression is proven to be the best to predict the deaths with highest accuracy. [4]

- **“From evaluation of the patient to evaluation of the intensive care unit”**
  Tried to collect the data related to the risk that caused death to the patients in ICU, so that new model can help in order to prevent those deaths in future. Data collected includes diagnostic information, chronic conditions. All the data were being tested using kappa statistics and correlation coefficients. This study was not able to give the appropriate results; the prediction was poor. The development of this new model can help critically ill patients by adjusting some parameters. [5]

- **“Assessing contemporary intensive care unit outcome: development and validation of the Australian and New Zealand Risk of Death admission model.”**
  Basically, this prediction was particularly for Australia and New Zealand hospital patients where the data was collected of the patients in ICU after 24 hrs., they collected the old mortality and predict the new mortality that can happen. They used a linear regression model for this approach, the prediction was somewhat correlated with the previous deaths, model accepted the data properly and model was predicting similar to what was to be obtained. The performance was better for future prediction so that the patients can be saved. [6]

- **“An Artificial Neural Network to Predict Mortality in ICU Patients and Application of Solar Physics Analysis Methods”**
  This paper predicts the mortality of the patient according to the data given here it will only consider Heart related problems. Data contains Heart rate, blood pressure through those vital signs the model will predict whether the mortality rate of that person is high or low. The technique used is the Artificial Neural Network for training the data. They were provided with

two data sets; they had to work with both the data sets and predict the model for that dataset. The result was most successful but for clinically adapting the model more work has to be carried out. [7]

- **“A Machine Learning Approach to predict cause of death from Verbal Autopsy Data”**
In this data visualization is being done using the WEKA platform, with this the decision tree can be used for prediction and different classification techniques available in the WEKA tool and the results were being compared, J48 decision tree algorithm gave the best results for the prediction of death from the verbal autopsy dataset. The technique is Artificial Neural Network. [8]

- **“DeepDeath: Learning to Predict the Underlying Cause of Death with Big Data”**
In predicting the cause of disease the most important as well as necessary thing needed is the data. This paper highlights the same issue. The authors designed two different classes of models for the analysis purpose, first one is Hadoop based ensemble technique and the other one is DeepDeath, which is the deep classifier based on recurrent neural networks (RNN). They worked on both the classes and compared the results between them. They concluded that both the classes performed better than the random classifier. [9]

- **“Automated 5- Year Mortality Prediction Using Deep Learning and Radiomics Features from Chest Computed Tomography”**
Through this paper the authors proposed prognostic methods like chest computed tomography (CT) that predicted 5- year mortality in elderly individuals. They did the prediction using a set of features extracted from the CT image. They worked on two approaches one being the unified framework approach and the other one being the multi stage framework approach. The performance was measured by calculating the accuracy of both. 5-year mortality gave the accuracy in range of 64.5% to 66.5% whereas radiomics generated the accuracy as 64.6%. [10]

- **“Multiple task transfer learning with small sample sizes”**
In this paper, authors proposed a framework for classification of data with small numbers of samples. Their framework is the hybrid of both multitask and transfer learning which takes the advantage of both, as in transfer learning data samples can be employed from source tasks and all tasks can be considered as in multi task learning. Every task is modelled jointly with the other tasks. They performed the experiments on 3 real world datasets of healthcare, handwritten data and face data. They formulated the optimization cost function in which the source data influence the target task. This optimization problem was solved using an efficient algorithm named ADMM. It gives better performance as compared to recent state of art MTL methods. [11]

III. DATASET

Death Prediction of any human is a novel idea, there is no dataset available directly to fulfill the requirements of death prediction. So, we tried to club different dataset that may help to predict the death of the human having heart disease. The attributes that we involved in the model for death prediction are age group, Glucose level, Blood Pressure which includes systolic Blood pressure and diastolic Blood Pressure, stroke, diabetes level, Heart rate, WBC, Hemoglobin, Drug addiction level in blood, Obesity, BMI rate, Chest Pain, Cancer, Temperature, Sodium, Potassium, Chloride level in blood. These different attributes are selected according to their priorities that can help to predict the heart disease death prediction. There are a total 4200 samples collected from a varied available dataset. The attribute Stroke is taken from deltex medical site which has unit as mL.

IV. METHODOLOGY

Ensemble Learning was being chosen for death prediction. Ensemble Learning is the Machine Learning technique which is mostly used to improve accuracy. Ensemble Learning takes different models and works with the training set and gives different predictions for each model and those predictions are being combined to final prediction using the meta-regression. These combining models using meta-regression is called stacking which is an Ensemble technique.

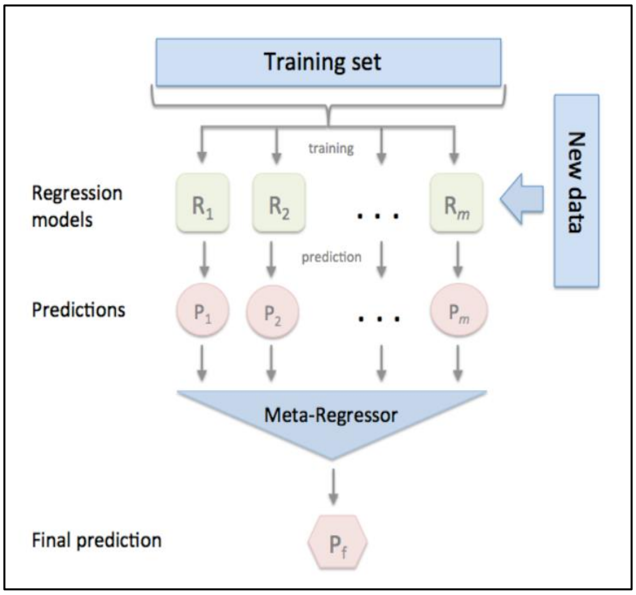

Figure 1

In Ensemble Learning bagging and boosting is also done to improve the predictions. Bagging can also be called as bootstrap aggregation. To reduce the variance, bagging is used i.e. bagging uses the bootstrap sampling which helps in obtaining the subsets of the data for training the base learners. Now in bagging the base learners' predictions have to be combined. To achieve that, voting is used for classification models and average method is used for regression models.

Boosting which is another most important technique of Ensemble Learning. Boosting helps to change the slow learners into strong learners. It tries to fit the series of slow learners to some weighted version of the data i.e. maximum weight is given to the data which are misclassified. And then those predictions are being combined as voting in classification models and averaging in regression models.

V. IMPLEMENTATION

In this real time problem of death prediction, Ensemble Learning is being used for the prediction. The different models used as inputs here are SVM, Linear regression, ElasticNet regularization and ridge and three different predictions are obtained. These predictions are combined by meta-regression. In this, the predictions are actually averaged into the final prediction because it is a regression model.

```python
from sklearn.linear_model import ElasticNet

lr = LinearRegression()
elasticnet =ElasticNet(random_state=1)
svr_lin = SVR(kernel='linear')
svr_rbf = SVR(kernel='rbf')

stregr = StackingRegressor(regressors=[svr_lin, lr, elasticnet],
                           meta_regressor=svr_rbf)
```

Figure 2

According to the figure shown, there are three different models used for prediction namely Linear Regression, ElasticNet regularization and SVM.

Linear Regression is a normal linear mathematical model which helps to find the value of the dependent variable over the independent variable.

$$Y = MX + C$$

Y is the Output Prediction
M is the Slope
C is the y-intercept
X is the data point through which prediction is made.

ElasticNet Regularization is the combination of Ridge and Lasso regularization i.e. L2 and L1 respectively. Regularization helps in generalizing the models to get more accurate predictions. Lasso regularization uses absolute value function and drives coefficient parameters to zero. Ridge Regularization uses square function on the coefficients.

SVM i.e. Support Vector Machine is a supervised Machine Learning model. This is used for classification for two-group classification problems. The main goal of SVM is to discover hyperplanes in N-dimensional space which helps to classify the data points. It just tries to separate the two classes of data points. We need to find the maximum margin which helps for future predictions.

Ridge regression is some variation of Linear regression. It is mostly used for analyzing the multiple regression data. In this type of data, the error and the variance come very high so ridge regression helps to reduce errors and maintain the predictions to be more correct. Misclassification can also be reduced.

Now these models will predict the individual output and they are combined using the meta-classifier namely rbf i.e. Radial Basis Function that produces positive distance which is an absolute value.

$$y(\mathbf{x}) = \sum_{i=1}^{N} w_i \, \phi(\|\mathbf{x} - \mathbf{x}_i\|),$$

Figure 3

VI. RESULT

The predictions from individual models were obtained (as discussed in previous section) and these were merged with the meta regressor. As a result, the accuracy obtained is 32.285 % which is greater than the individual model accuracy.
For this purpose, the concept of Mean Squared Error and variance score is used. MSE gives the average value of the squared error. If the MSE value is smaller, then the model is more accurate. The variance score shows how the data points are varied from the mean value. Higher the variance, the model is more efficient.

```
correctly predicted: 1356
Mean Squared Error: 6.3750
Variance Score: 0.8439
Accuracy by correct predictions: 32.2857
```

Figure 4

VII. CONCLUSION

In this paper, death prediction of heart disease patient is done using Machine Learning Techniques. Ensemble Learning technique is used which helps to give more accurate prediction by using two or more Machine Learning Techniques. This was done as accuracy matters more in any Machine Learning Model.

VIII. REFERENCES

[1] "Devi, Reeta, Hitender Kumar Tyagi, and Dinesh Kumar. "Early stage prediction of sudden cardiac death." In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 2005-2008. IEEE, 2017."

[2] "Lai, D., Zhang, Y., Zhang, X., Su, Y. and Heyat, M.B.B., 2019. An automated strategy for early risk identification of sudden cardiac death by using machine learning approach on measurable arrhythmic risk markers. *IEEE Access*, *7*, pp.94701-94716."

[3] "Wood, N. B. "The prediction of a potentially fatal cardiac event in the next 2 to 24 hours and the prediction of a myocardial infarction related death or sudden death." In *Computers in Cardiology 2001. Vol. 28 (Cat. No. 01CH37287)*, pp. 509-512. IEEE, 2001."

[4] "Aqlan, Hesham Abdo Ahmed, Shoiab Ahmed, and Ajit Danti. "Death prediction and analysis using web mining techniques." In *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 1-5. IEEE, 2017."

[5] "Moreno, Rui P., Philipp GH Metnitz, Eduardo Almeida, Barbara Jordan, Peter Bauer, Ricardo Abizanda Campos, Gaetano Iapichino et

al. "SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission." *Intensive care medicine* 31, no. 10 (2005): 1345-1355."

[6] "Paul, E., M. Bailey, J. Kasza, and D. V. Pilcher. "Assessing contemporary intensive care unit outcome: development and validation of the Australian and New Zealand Risk of Death admission model." *Anaesthesia and intensive care* 45, no. 3 (2017): 326-343."

[7] "Pollard, Tom J., Louise Harra, David Williams, Steve Harris, Demetrio Martinez, and Kevin Fong. "2012 PhysioNet Challenge: An artificial neural network to predict mortality in ICU patients and application of solar physics analysis methods." In *2012 Computing in Cardiology*, pp. 485-488. IEEE, 2012."

[8] "Flaxman, Abraham D., Alireza Vahdatpour, Sean Green, Spencer L. James, and Christopher JL Murray. "Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards." *Population health metrics* 9, no. 1 (2011): 29."

[9] "Hassanzadeh, Hamid Reza, Ying Sha, and May D. Wang. "DeepDeath: Learning to predict the underlying cause of death with Big Data." In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3373-3376. IEEE, 2017."

[10] "Carneiro, Gustavo, Luke Oakden-Rayner, Andrew P. Bradley, Jacinto Nascimento, and Lyle Palmer. "Automated 5-year mortality prediction using deep learning and radiomics features from chest computed tomography." In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 130-134. IEEE, 2017."

[11] "Saha, Budhaditya, Sunil Gupta, Dinh Phung, and Svetha Venkatesh. "Multiple task transfer learning with small sample sizes." *Knowledge and information systems* 46, no. 2 (2016): 315-342."