Predict survival of patients with heart failure using Various Machine Learning Algorithms

A Technical Report on the Mini Project

submited by

Sathya Pramod D.S
01JST19SE016

To

Department of Information Science & Engineering
JSS Science and Technology University (SJCE), Mysuru

Certificate

This is to certify that the project entitled "Predict survival of patients with heart failure using Various Machine Learning Algorithms" is the result of my research work at the department of ISE. I certify that no part of this work is either copied or submitted else where.

(Sathya Pramod DS)

Contents

1 INTRODUCTION **4**

2 LITERATURE SURVEY **6**

3 PROPOSED METHODOLOGY **7**

3.1 Data Source 7

3.2 Architecture Diagram 8

4 Description of Algorithms **10**

4.1 Support vector machines (SVM) 10

4.2 Random forest 10

4.3 Decision tree 10

4.4 Multi-layer Perceptron (MLP) 11

5 Results and Discussions **11**

5.1 **Results Of Machine Learning Algorithm** 13

6 Conclusion **14**

References **15**

List of Figures

1 Type of Heart Disease 5

2 Data Sample 7

3 Dataset Description 8

4 Architecture Diagram 9

5 Result of Support Vector Machine 13

6 Result of Random Forest 13

7 Result of Decision Tress 14

8 Result of Multi-layer Perceptron 14

1 INTRODUCTION

Cardiovascular diseases kill approximately 17 million people globally every year, and they mainly exhibit as myocardial infarctions and heart failures. Heart failure (HF) occurs when the heart cannot pump enough blood to meet the needs of the body.Available electronic medical records of patients quantify symptoms, body features, and clinical laboratory test values, which can be used to perform biostatistics analysis aimed at highlighting patterns and correlations otherwise undetectable by medical doctors. Machine learning, in particular, can predict patients' survival from their data and can individuate the most important features among those included in their medical records.

Data mining is extracting information and knowledge from huge amount of data. Data mining is an essential step in discovering knowledge from databases. There are numbers of databases, data marts, data warehouses all over the world. Data Mining is mainly used to extract the hidden information from a large amount of database. Data mining is also called as Knowledge Discovery Database (KDD). The data mining has four main techniques namely Classification, Clustering, Regression, and Association rule. Data mining techniques have the ability to rapidly mine vast amount of data. Data mining is mainly needed in many fields to extract useful information from a large amount of data. The fields like the medical field, business field, and educational field have a vast amount of data, thus these fields data can be mined through those techniques more useful information. Data mining techniques can be implemented through a machine learning algorithm. Each technique can be extended using certain machine learning models.

In this system, a heart disease data set is used. The main aim of this system is to predict the possibilities of occurring heart disease of the patients in terms of percentage. This is performed through data mining classification techniques. Machine learning technique is applied to the dataset through the machine learning machine learning algorithm namely Decision tree classification, Random Forest, Support vector machine and MLP Classification models. These models are used to enhance the

accuracy level of the machine learning technique. This model performs both the classification and prediction methods. These models are performed using python Programming Language.

Health care field has a vast amount of data, for processing those data certain techniques are used. Data mining is one of the techniques often used. Heart disease is the Leading cause of death worldwide. This System predicts the arising possibilities of Heart Disease. The outcomes of this system provide the chances of occurring heart disease in terms of percentage.

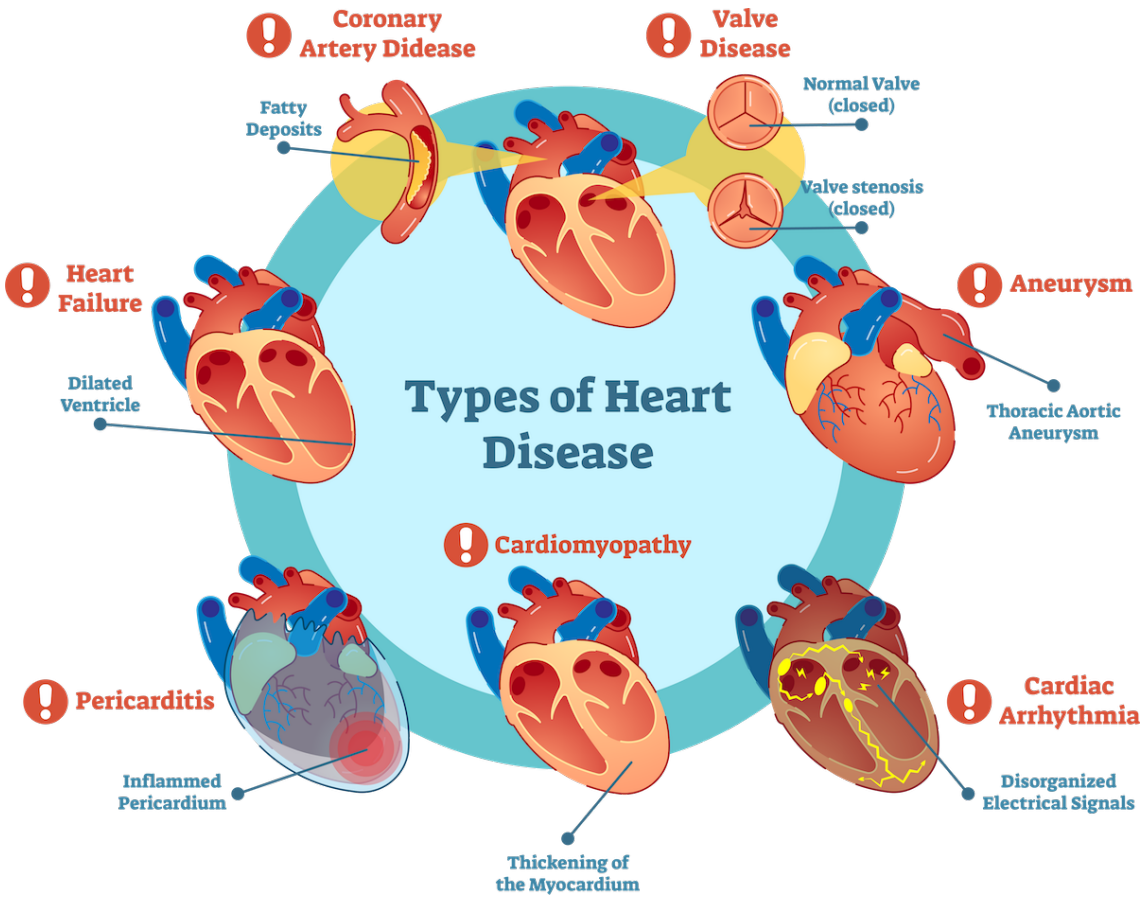

Figure 1: Type of Heart Disease

2 LITERATURE SURVEY

Marjia Sultana, Afrin Haider and Mohammad ShorifUddin[1] have illustrated about how the datasets available for heart disease are generally a raw in nature which is highly redundant and inconsistent. There is a need of pre-processing of these data sets; in this phase high dimensional data set is reduced to low data set. They also show that extraction of crucial features from the data set because there is every kind of features. Selection of important features reduces work of training the algorithm and hence resulted in reduction in time complexity.

Through this paper the information about Data Mining and heart diseases has been gathered. The detailed information about heart diseases, symptoms of heart attack and heart disease types are presented in this paper, the three main data mining techniques namely Decision Tree, Neural Networks and Naive Bayes Classifier are used. The main task of data Prediction is done using these three techniques.The Advantages and Disadvantages of each technique can be known using this paper [2].

The core concept of this paper is predicting heart disease using data mining Techniques. The main Methodology used for prediction is KNN Algorithms, Decision Trees like CART, C4.5, CHAID, J48, ID3 Algorithms, and Naive Bayes Techniques. This system uses 13 medical attributes as input and with that input, Data sets it to process the data mining techniques and shows the most accurate one. [3]

M.A.Jabbar, B.L Deekshatulu, Priti Chndra [5], an optimisation of feature has been done to achieve higher classification efficiency in Decision Tree . It is an approach for early detection of heart disease by utilizing variety of feature. These kind of approach can also be utilize for other sphere of research. Other than decision tree various other approach where adopt for achieving the goal of perfect detection of heart disease in human Yogeswaran Mohan et.al [6] have collected raw data form EEG device and used to train neural network for pattern classification .

Machine learning applied to medical records, in particular, can be an effective tool both to predict the survival of each patient having heart failure symptoms [7], and to detect the most important clinical features (or risk factors) that may lead to heart

failure. Scientists can take advantage of machine learning not only for clinical prediction, but also for feature ranking. Computational intelligence, especially, shows its predictive power when applied to medical records, or coupled with imaging. Further, deep learning and meta-analysis studies applied to this field have also recently appeared in the literature, improving on human specialists' performance, albeit showing lower accuracy (0.75 versus 0.59).

3 PROPOSED METHODOLOGY

3.1 Data Source

The dataset used here for predicting heart disease is taken from UCI Machine learning repository. UCI is a collection of databases that are used for implement machine learning algorithms. The dataset used here is real dataset. The dataset consists of 300 instance of data with the appropriate 14 clinical parameters. The clinical parameter of dataset is about tests which are taken related to the heart disease as like blood pressure level, chest pain type, electrocardiographic result and etc.This means that categorical data must be converted to a numerical form. One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. For that purpose, One-Hot Encoding will be used to convert these two columns to one-hot numeric array. The categorical value represents the numerical value of the entry in the dataset. This encoding will create a binary column for each category and returns a matrix with the results.

	age	anaemia	creatinine_phosphokina…	diabetes	ejection_fraction	high_bloc
0	75	0	582	0	20	
1	55	0	7861	0	38	
2	65	0	146	0	20	
3	50	1	111	0	20	
4	65	1	160	1	20	

Figure 2: Data Sample

Feature	Explanation	Measurement	Range
Age	Age of the patient	Years	[40,..., 95]
Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0, 1
High blood pressure	If a patient has hypertension	Boolean	0, 1
Creatinine phosphokinase	Level of the CPK enzyme in the blood	mcg/L	[23,..., 7861]
(CPK)			
Diabetes	If the patient has diabetes	Boolean	0, 1
Ejection fraction	Percentage of blood leaving	Percentage	[14,..., 80]
	the heart at each contraction		
Sex	Woman or man	Binary	0, 1
Platelets	Platelets in the blood	kiloplatelets/mL	[25.01,..., 850.00]
Serum creatinine	Level of creatinine in the blood	mg/dL	[0.50,..., 9.40]
Serum sodium	Level of sodium in the blood	mEq/L	[114,..., 148]
Smoking	If the patient smokes	Boolean	0, 1
Time	Follow-up period	Days	[4,...,285]
(target) death event	If the patient died during the follow-up period	Boolean	0, 1

mcg/L: micrograms per liter. mL: microliter. mEq/L: milliequivalents per litre

Figure 3: Dataset Description

To suppress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling with MinMaxScaler. The final step on data preprocessing is the training and testing data. The dataset will be split into two datasets, the training dataset and test dataset. The data usually tend to be split inequality because training the model usually requires as much data-points as possible.The common splits are 70/30 or 80/20 for train/test.

The training dataset is the initial dataset used to train ML algorithms to learn and produce right predictions.

The test dataset, however, is used to assess how well ML algorithm is trained with the training dataset. You can't simply reuse the training dataset in the testing stage because ML algorithm will already "know" the expected output, which defeats the purpose of testing the algorithm.

3.2 Architecture Diagram

In the first step, we collect the data from the repository and categorized the dataset. Second step we preprocess the data like cleaning the data, handling missing values, converting data using label encoder and transform the data. Third step model is trained to predict the unseen data. The training model used 80% of the data and

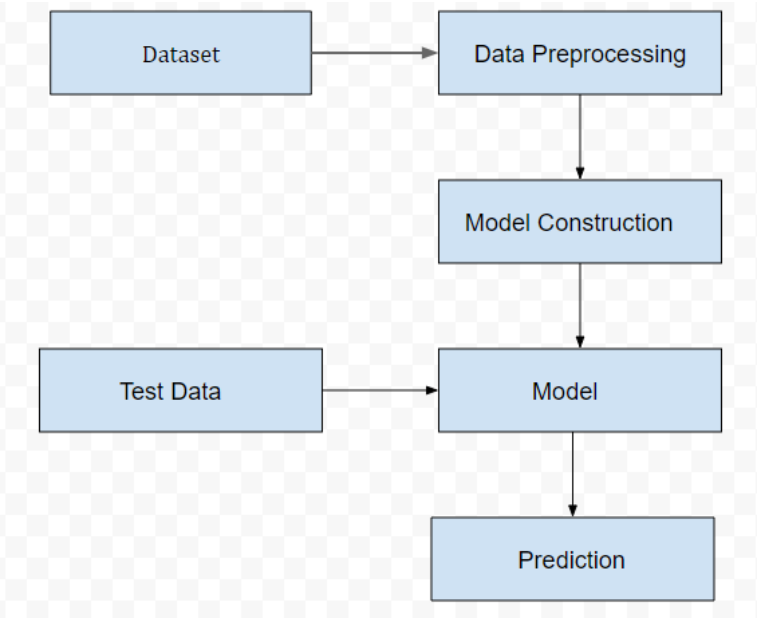

Figure 4: Architecture Diagram

20% is for test model. In the final step the we predict the accuracy of survival of heart patients.

During preprocessing step we spit the dataset into training and testing dataset. Train dataset to detect the heart failure present in the dataset using appropriate supervised learning algorithms.

Apply the machine learning techniques which are helpful for finding heart failure for any of new data occurred in the data. After this data acquisition suitable machine learning algorithm must be applied to compute efficiency and capability of the model, here we have applied various machine learning algorithms. Metrics like accuracy, precision will be calculated for the proposed model. This system architecture focuses 3 parts such as flow data, Machine learning techniques, and modules for detecting heart failure and feature selection modules

4 Description of Algorithms

4.1 Support vector machines (SVM)

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. In 1960s, SVMs were first introduced but later they got refined in 1990. SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.

4.2 Random forest

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

4.3 Decision tree

In general, Decision tree analysis is a predictive modelling tool that can be applied across many areas. Decision trees can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions. Decisions trees are the most powerful algorithms that falls under the category of supervised algorithms.

4.4 Multi-layer Perceptron (MLP)

Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function by training on a dataset, where is the number of dimensions for input and is the number of dimensions for output. Given a set of features and a target , it can learn a non-linear function approximator for either classification or regression. It is different from logistic regression, in that between the input and the output layer, there can be one or more non-linear layers, called hidden layers. Figure 1 shows a one hidden layer MLP with scalar output.

5 Results and Discussions

In this section, we first describe the results we obtained for the survival prediction on the complete dataset ("Survival machine learning prediction on all clinical features" section), the results obtained for the feature ranking ("Feature ranking results" section), and the results on the survival prediction when using only the top two most important features of the dataset ("Survival machine learning prediction on serum creatinine and ejection fraction alone" section and "Serum creatinine and ejection fraction linear separability" section), all independently from the follow-up time. We then report and discuss the results achieved by including the follow-up time of each patient in the survival prediction and feature ranking.

For methods that needed hyper-parameter optimization (neural network, Support Vector Machine, and k-Nearest Neighbors), we split the dataset into 60% (179 randomly selected patients) for the training set, 20% (60 randomly selected patients) for the validation set, and 20% (the remaining 60 patients) for the test set. To choose the top hyper-parameters, we used a grid search and selected the models that generated the highest Matthews correlation coefficient.

For the other methods (Random Forests, One Rule, Linear Regression, Naïve Bayes, and Decision Tree), instead, we split the dataset into 80% (239 randomly selected patients) for the training set, and 20% (the remaining 60 patients) for the

test set.

For each of the 100 executions, our script randomly selected data instances for the training set and for the test (and for the validation set, in the case of hyper-parameter optimization) from the complete original dataset. We trained the model on the training set (and validated it on the validation set, in the case of hyper-parameter optimization). We then applied the script to the test set. Given the different selections of data instances for the dataset splits, each execution led to slightly different results.

Our prediction results showed that Random Forests outperformed all the other methods, by obtaining the top MCC (+0.384), the top accuracy (0.740), and the top ROC AUC (0.800). The Decision Trees obtained the top results on the true positives (sensitivity = 0.532) and on the F1 score (0.554), and was the only classifier able to predict correctly the majority of deceased patients. The linear Support Vector Machines achieved an almost perfect prediction score on the negative elements (specificity = 0.961), but a poor score on the positive elements (sensitivity = 0.072). The Artificial Neural Network perceptron, instead, obtained the top value on the Precision-Recall AUC (0.750).

Because of the imbalance of the dataset (67.89% negative elements and 32.11% positive elements), all the methods obtained better prediction scores on the true negative rate, rather than on the true positive rate . These results occur because the algorithms can see more negative elements during training, and therefore they are more trained to recognize deceased patient profiles during testing.

In our work, the fact that our traditional biostatistics analysis selected ejection fraction and serum creatinine as the two most relevant features confirmed the relevance of the feature ranking executed with machine learning. Moreover, our approach showed that machine learning can be used effectively for binary classification of electronic health records of patients with cardiovascular hearth diseases.

As a limitation of the present study, we have to report the small size of the dataset (299 patients): a larger dataset would have permitted us to obtain more reliable results. Additional information about the physical features of the patients (height,

weight, body mass index, etc.) and their occupational history would have been useful to detect additional risk factors for cardiovascular health diseases. Also, if an additional external dataset with the same features from a different geographical region had been available, we would have used it as a validation cohort to verify our findings.

Regarding future developments, we plan to apply our machine learning approach to alternative datasets of cardiovascular heart diseases and other illnesses (cervical cancer, neuroblastoma, breast cancer , and amyotrophic lateral sclerosis).

5.1 Results Of Machine Learning Algorithm

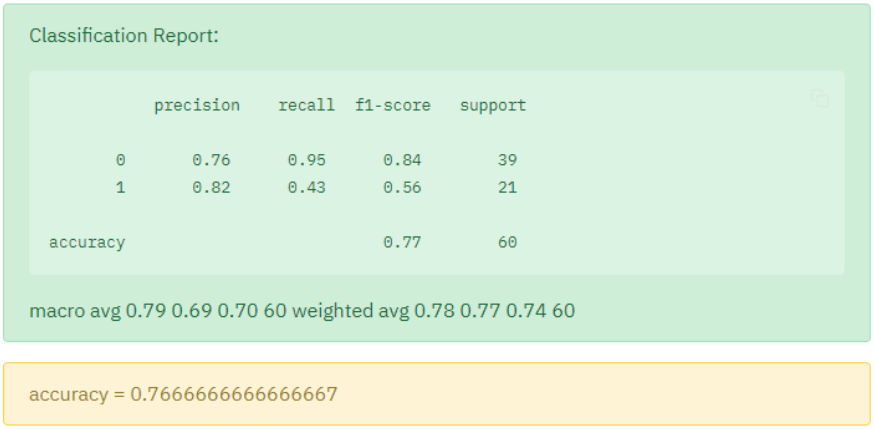

Figure 5: Result of Support Vector Machine

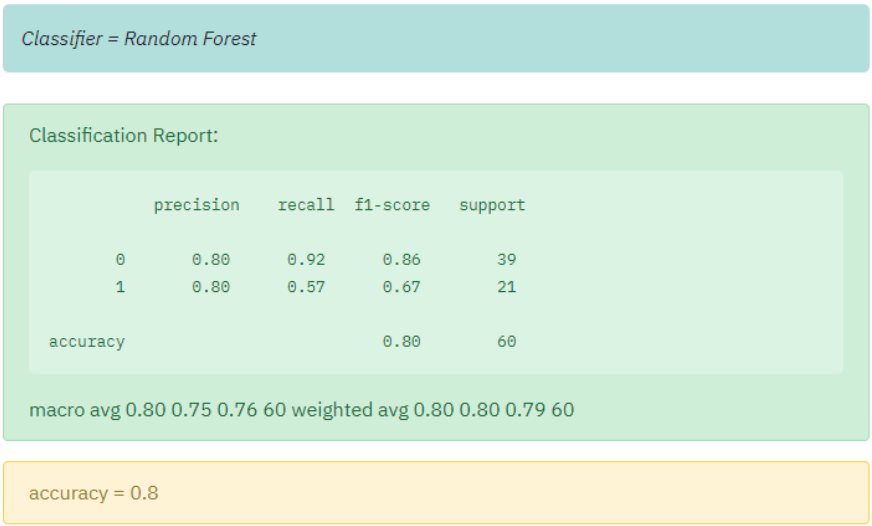

Figure 6: Result of Random Forest

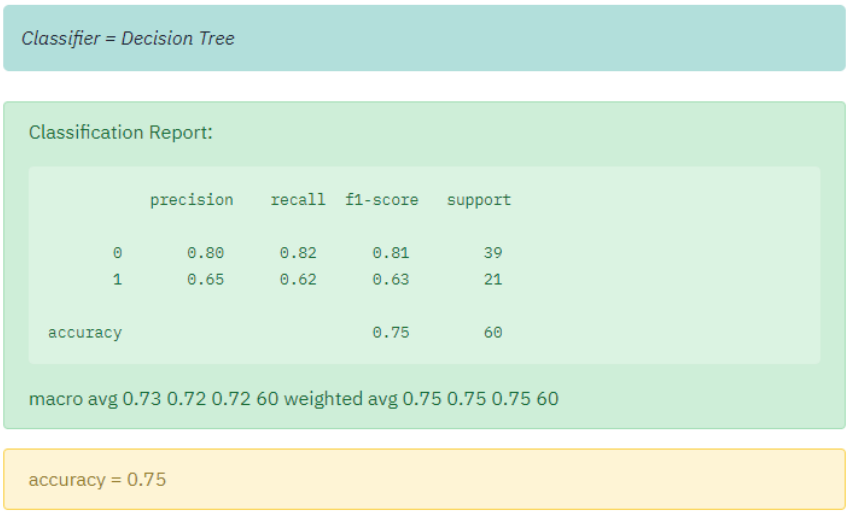

Figure 7: Result of Decision Tress

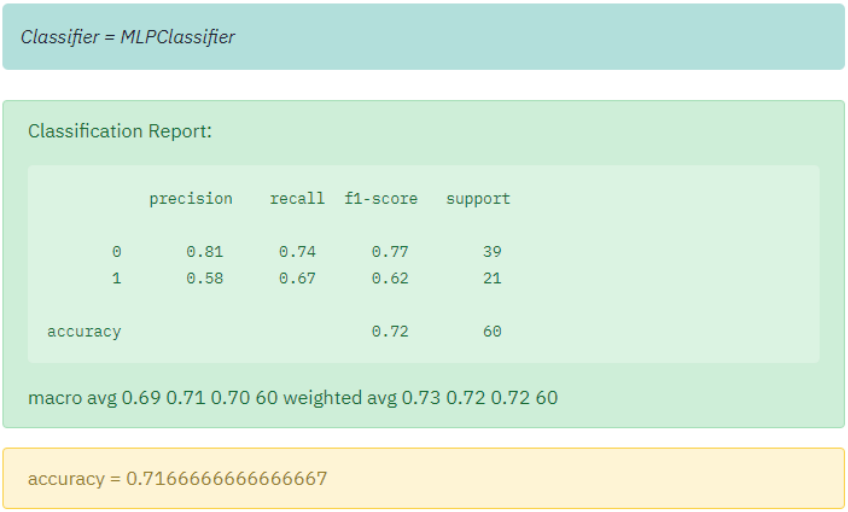

Figure 8: Result of Multi-layer Perceptron

6 Conclusion

In our work, the fact that our traditional biostatistics analysis selected ejection fraction and serum creatinine as the two most relevant features confirmed the relevance of the feature ranking executed with machine learning. Moreover, our approach showed that machine learning can be used effectively for binary classification of electronic health records of patients with cardiovascular hearth diseases.

As a limitation of the present study, we have to report the small size of the dataset (299 patients): a larger dataset would have permitted us to obtain more reliable re-

sults. Additional information about the physical features of the patients (height, weight, body mass index, etc.) and their occupational history would have been useful to detect additional risk factors for cardiovascular health diseases. Also, if an additional external dataset with the same features from a different geographical region had been available, we would have used it as a validation cohort to verify our findings.

Regarding future developments, we plan to apply our machine learning approach to alternative datasets of cardiovascular heart diseases and other illnesses (cervical cancer, neuroblastoma, breast cancer , and amyotrophic lateral sclerosis).

References

[1] M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," 2016

[2] T. M. Lakshmi, A. Martin, R. M. Begum, and V. P. Venkatesan, "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data," Int. J. Mod. Educ. Comput. Sci., vol. 5, no. 5, pp. 18–27, 2017.

[3] J. Schmidhuber, "Deep Learning in neural networks: An overview," 2018.

[4] Santhana Krishnan.J, Geetha.S , "Prediction of Heart Disease Using Machine Learning Algorithms ," 2019.

[5] M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, "Prediction of risk score for heart disease using associative classification and hybrid feature subset selection," Int. Conf. Intell. Syst. Des. Appl. ISDA, pp. 628–634, 2016.

[6] Davide Chicco Giuseppe Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone",2019.