Heart Failure Prediction

João Alves [a], Luís Pinto [a] and Paulo Alves [a]

[a] University of Minho, Campus Gualtar, Braga 4710, Portugal

11 June 2021

Abstract

The aim of this study is to predict, through data mining processes, the death event due to heart failure. This serious problem, which is one of the deadliest diseases in the world, entails several consequences and can be caused by various reasons. Clinical Decision Support Systems, with the information collected from patient's electronic health records combined with data mining techniques, might decrease the incidence of heart failure during people's live.

Keywords: Data Mining; Heart Failure; Decision Support Systems.

1 Introduction

Heart failure is the state in which muscles in the heart wall get fade and enlarge, limiting heart pumping of blood. Physiologically, the ventricles of the heart can get inflexible and do not fill properly between beats. For that reason, the heart fails in fulfilling the proper demand of blood in body and consequently, the person starts feeling difficulty in breathing and his well-being is worsened [1].

The main reason behind heart failure includes coronary heart disease, diabetes, high blood pressure and other diseases like HIV, alcohol abuse or cocaine, thyroid disorders, excess of vitamin E in body, radiation or chemotherapy.

Cardiovascular diseases (CVDs) are the number one cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide [2].

Heart failure is a common event caused by CVDs and the dataset used in this paper contains 12 features that can be utilized to predict mortality by heart failure. Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies. People with cardiovascular disease or with high cardiovascular risk need early detection and management and for that purpose a well-trained machine learning model or algorithm can be a great help.

Data mining, which is a set of approaches that allows the extraction of information from data through its analysis, is a descriptive or a predictive technique. It is the technique used to do this project. The predictive (or explanatory) processes, such as classification or scoring for qualitative data and regression for quantitative, anticipate new information based on the present facts. The current project was a classification problem, as the main purpose was to know if a person will die due to heart failure or not.

2 Background and Related Work

The clinical community groups heart failure into two types based on the ejection fraction value, that is the proportion of blood pumped out of the heart during a single contraction, given as a percentage with physiological values ranging between 50% and 75%. The former is heart failure due to reduced ejection fraction ($HFrEF$), previously known as heart failure due to left ventricular systolic dysfunction or systolic heart failure and characterized by an ejection fraction smaller than 40%. The latter is heart failure with preserved ejection fraction ($HFpEF$), formerly called diastolic heart failure or heart failure with normal ejection fraction [3].

A vital organ, such as the heart, is extremely important and for that reason predicting heart failure has become a priority for medical doctors and physicians, but to date forecasting heart failure-related events in clinical practice usually has failed to reach high accuracy.

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, and to detect the most important clinical features (or risk factors) that may lead to heart failure [4].

2.1 Related Work

Nowadays, it is tremendously important to improve healthcare services. One of the major improvements is the accurate prediction of bad events. Some works about data mining in healthcare field, especially those regarding heart failure or heart disease, have been improving these predictions during the last few years. It's the case of the article from 2020, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone" [1]. This article explores data mining to predict if a patient will develop or not heart failure based on serum creatinine and ejection fraction. This study is important in healthcare services because it predicts Heart Failure, using Logistic Regression, which obtained a 83.3% of Accuracy, showing that Data Mining Models are very useful for supporting the decision-making and for preventing future critical events.

3 Methodology, Materials and Methods

During the Data Mining Process, the Cross Industry Standard Process for Data Mining (*CRISP-DM*) Methodology was followed, which is a hierarchical process model that divides the process of data mining into six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. This methodology was followed due to its advantages such as increasing the success of data mining projects and allowing the implementation of data mining models in real environments.

The data presented in this work has information about 299 patients, 105 women and 194 men, who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features. These patients were admitted to the Institute of Cardiology and Allied Hospital Faisalabad-Pakistan during April-December (2015). All the patients were more than 40 years old, having left ventricular systolic dysfunction and falling in NYHA class III and IV [5]. Follow-up time was 4–285 days, with an average of 130 days. Disease was diagnosed by cardiac echo report or notes written by physician. Age, serum sodium, serum creatinine, gender, smoking, Blood Pressure (BP), Ejection Fraction (EF), anaemia, platelets, Creatinine Phosphokinase (CPK) and diabetes were considered as potential variables explaining mortality caused by coronary heart disease. Age, serum sodium, CPK and Time after the heart failure event are continuous variables whereas EF, serum creatinine, platelets, high blood pressure, anaemia, diabetes, sex, smoking and Death Event were taken as categorical variables. EF was divided into three levels (i.e. EF30, 3045) and platelets was also divided into three level on the basis of quartiles. Serum creatinine greater than its normal level (1.5) is an indicator of renal dysfunction. Anaemia in patients was assessed by their haematocrit level. Following McClellan et al. [6] the patients with haematocrit less than 36 (minimum normal level of haematocrit) were taken as anaemic. The information related to risk factors were taken from blood reports, while smoking status and blood pressure were taken from physician's notes [7].

4 Knowledge Discovering Process

4.1 *Business Understanding*

The main objective of this study is to estimate death rates due to heart failure and to investigate its link with some major risk factors by choosing Faisalabad, the third most populous city of Pakistan, as study area. This research considers patient's characteristics with heavier focus on the follow-up period and, although not one hundred percent accurate, data mining process allows to compare the risk factors in an equitable manner and predict which ones are impactful in the death event due to heart failure.

4.2 *Data Understanding*

The data collected and presented for this study comes from patients who were admitted to Institute of Cardiology and Allied hospital Faisalabad-Pakistan during April-December (2015). The dataset consists of 13 attributes: *age* - age of the patient measured in years; *anaemia* - decrease of red blood cells or hemoglobin (values are boolean: 0 for no anaemia and 1 for anaemic); *creatinine_phosphokinase* - level of the CPK enzyme in the blood in mcg/L; *diabetes* - if the patient has diabetes (values are boolean: 0 for no diabetes and 1 for diabetic); *ejection_fraction* - percentage of blood leaving the heart at each contraction; *high_blood_pressure* - if the patient has hypertension (values are boolean: 0 for no and 1 for yes); *platelets* - number of platelets in the blood in kiloplatelets/mL; *serum_creatinine* - level of serum creatinine in the blood in mg/dL; *serum_sodium* - level of serum sodium in the blood in mEq/L; *sex* - woman or man (binary: 1 and 0 respectively); *smoking* - smoker (values are boolean: 0 for non-smoker and 1 for smoker); *time* - follow-up period measured in days after the heart failure; *DEATH_EVENT* - if the patient died during the follow-up period (values are Boolean: 0 for No and 1 for Yes).

The following chart, **figure 1**, shows how many death events resulted in death (Yes) or not (No), according to the dataset:

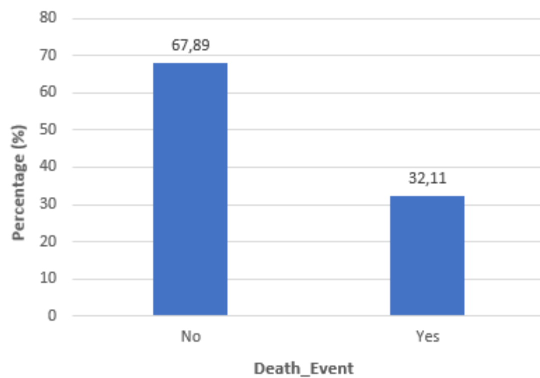

Figure 1: Labelling classification

4.3 *Data Preparation*

This phase of the data mining process involved the selection and preparation of the data to be used by the data mining models in Rapidminer. After selecting the data exposed before, a pre-processing phase started. In this phase, all the data with null and noise values must be removed. The current dataset does not have missing values, so the next step was to inspect each individual attribute. The attribute Death_Event had to be changed from integer to binomial. Furthermore, the "Remap Binomials" was applied in Rapidminer to change the meanings of the binomial values of this attribute for better understanding. 0 (zero) now means that the patient didn't die, and 1 (one) means that a death event has occurred. At last, an outlier was applied that is further explored in the next subchapter. The first data mining models didn't present satisfactory results using the attributes which, by analyzing the correlation matrix, **figure 2**, correlated the most with the label *Death_Event*. Four attributes were identified as uncorrelated with the *Death_Event*, these were: *smoking, sex, platelets, diabetes*. These results had a correlation of less than 0.05 with the attribute *Death_Event* and will be considered when building the Scenarios and discussing the data.

Attributes	age	anaemia	creati...	diabetes	ejection...	high_...	platelets	serum_...	serum_...	sex	smoking	time	DEATH_EVENT
age	1	0.088	-0.082	-0.101	0.060	0.093	-0.052	0.159	-0.046	0.065	0.019	-0.224	0.254
anaemia	0.088	1	-0.191	-0.013	0.032	0.038	-0.044	0.052	0.042	-0.095	-0.107	-0.141	0.066
creatinine_phosphokinase	-0.082	-0.191	1	-0.010	-0.044	-0.071	0.024	-0.016	0.060	0.080	0.002	-0.009	0.063
diabetes	-0.101	-0.013	-0.010	1	-0.005	-0.013	0.092	-0.047	-0.090	-0.158	-0.147	0.034	-0.002
ejection_fraction	0.060	0.032	-0.044	-0.005	1	0.024	0.072	-0.011	0.176	-0.148	-0.067	0.042	-0.269
high_blood_pressure	0.093	0.038	-0.071	-0.013	0.024	1	0.050	-0.005	0.037	-0.105	-0.056	-0.196	0.079
platelets	-0.052	-0.044	0.024	0.092	0.072	0.050	1	-0.041	0.062	-0.125	0.028	0.011	-0.049
serum_creatinine	0.159	0.052	-0.016	-0.047	-0.011	-0.005	-0.041	1	-0.189	0.007	-0.027	-0.149	0.294
serum_sodium	-0.046	0.042	0.060	-0.090	0.176	0.037	0.062	-0.189	1	-0.028	0.005	0.088	-0.195
sex	0.065	-0.095	0.080	-0.158	-0.148	-0.105	-0.125	0.007	-0.028	1	0.446	-0.016	-0.004
smoking	0.019	-0.107	0.002	-0.147	-0.067	-0.056	0.028	-0.027	0.005	0.446	1	-0.023	-0.013
time	-0.224	-0.141	-0.009	0.034	0.042	-0.196	0.011	-0.149	0.088	-0.016	-0.023	1	-0.527
DEATH_EVENT	0.254	0.066	0.063	-0.002	-0.269	0.079	-0.049	0.294	-0.195	-0.004	-0.013	-0.527	1

Figure 2: Correlation Matrix

4.4 *Modeling*

This phase consisted of inducing the Data Mining Models (DMM) in *Rapidminer* using the prepared data. Since Classification was the chosen Approach (A), there were 6 different DM techniques (DMT) that were used: *Random Forest*, *J48*, *OneR*, *JRip*, *PART*, *Decision Tree*. These algorithms were used with the standard configurations in Rapidminer.

For each DM technique, two sampling methods (SM) were tested: Percentage Split, with 67% of the data used for training and the remaining amount for testing, and Cross Validation, using 10 folds and where all data is used for testing. In addition, there were two data approaches (DA) tested: with outliers and without outliers, with a detection value defined as 20. There was only one target variable, which was the death_event variable, and the considered scenarios (S), in order to evaluate which attributes were the most relevant to predict whether a patient dies or not, were: S1: {All attributes}; S2: {creatinine_phosphokinase, serum_creatinine, serum_sodium, time}; S3: {age, ejection_fraction, serum_creatinine, anaemia, high_blood_pressure}; S4: {age, time}.

Scenario 1, which includes all attributes, was chosen as the control scenario. Scenario 2's attributes were aggrouped after running several tests on Rapidminer, ending with a collective consensus on the best results. Scenario 3 was set by referencing two other papers that are related to the used dataset [3, 7]. Scenario 4, which only includes two attributes, was built by following the logic and real life expertise that time plays a huge factor on the death event after a heart attack.

In **figure 3**, it's represented one of the many schemes used in Rapidminer software. In this example, there are seven different operators. With the first operator, "Select attributes" operator, different attributes within the dataset can be removed or not. Then, "Remap Binominals" operator was applied to change the labels in order to put the label with the event of death to 1 and the label without it to 0. The "Set Role" operator was utilized to select the label and its values which the algorithm will try to predict and the "Split Data" operator was employed to split the data into 67% for training and 33% for testing. In this example, the machine learning algorithm used was Random Forest, so there is the "Random Forest" operator. Finally, there is also the "Apply Model" operator which applies the algorithm with the training and the testing sets. To be able to see the performance of the algorithm, the operator "Performance" was, also, used to visualize the confusion matrix and metrics such as accuracy, sensitivity, precision and specificity.

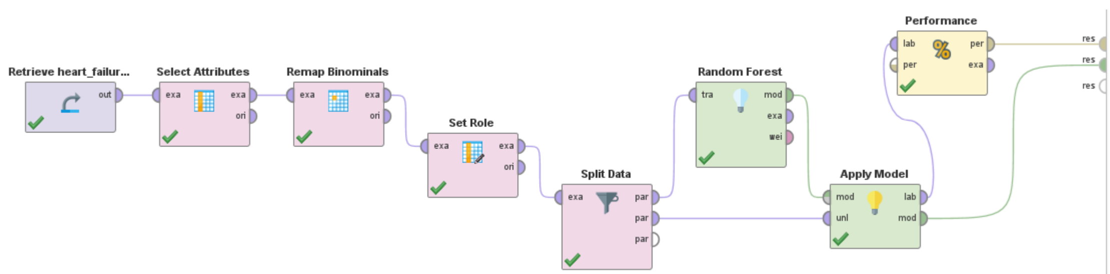

Figure 3: Example of a scheme that was used in modelling

The data mining model (DMM) can be described through an equation where it can be characterized by the approach (A), a set of scenarios (S), a sampling method (SM), a data approach (DA), a data mining technique (DMT) and a target (TG):

$$DMM_n = A_f \times S_i \times DMT_Y \times SM_C \times DA_b \times TG_t$$

This equation is described by:

A_f = {Classification}; S_i = {S1, S2, S3, S4}; DMT_Y = {Random Forest, J48, OneR, JRip, PART, Decision Tree}; SM_C = Percentage Split, Cross Validation; DA_b = {With Outliers, Without Outliers}; TG_i = {Death Event or no Death Event}.

Therefore, in total, 96 models were induced using DMM = {Classification, 4 Scenarios, 6 DM Techniques, 2 Sampling Methods, 2 Data Approaches, TG_1}.

4.5 *Evaluation*

The performance of each DMM was assessed through its confusion matrix (CMX), which presents the number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). With these results, it is possible to calculate sensitivity, specificity, precision and accuracy in order to evaluate the algorithms' performance.

Tables 1, 2, 3 and **4** present the models that, for each DM technique, achieved the best accuracy, precision, sensitivity and specificity results, respectively.

In order to select the best models, a threshold was used. Ideally, this threshold should have combined the four calculated metrics in order to find the most appropriate model – sensitivity 90% – with a satisfactory accuracy and precision to prevent a high number of false negatives, because it is better to predict that a person will have heart failure and prepare him and try to avoid that than predicting the inverse situation, which means not taking any measures, and the person ends up dead due to heart failure. However, the results obtained showed lower accuracy and sensitivity values.

Table 1: The 5 models with the highest accuracy

DM Technique	Scenario	Sampling Method	Data Approach	Accuracy
Random Forest	S1	Percentage Split	Without Outliers	88,04%
J48	S1	Percentage Split	Without Outliers	86,96%
OneR	S1, S2, S4	Percentage Split	Without Outliers	86,96%
JRip	S2, S4	Percentage Split	Without Outliers	86,96%
Decision Tree	S4	Percentage Split	Without Outliers	86,96%
PART	S2	Percentage Split	Without Outliers	85,87%

Table 2: The 5 models with the highest precision

DM Technique	Scenario	Sampling Method	Data Approach	Precision
PART	S2	Percentage Split	With Outliers	93,33%
Random Forest	S2	Percentage Split	With Outliers	90,91%
J48	S1	Percentage Split	With Outliers	90,91%
OneR	S1, S2, S4	Percentage Split	With Outliers	90,00%
JRip	S1, S2, S5	Percentage Split	With Outliers	85,71%
Decision Tree	S2, S4	Percentage Split	With Outliers	85,71%

Table 3: The 5 models with the highest sensitivity

DM Technique	Scenario	Sampling Method	Data Approach	Sensitivity
JRip	S1	Percentage Split	Without Outliers	86,21%
Random Forest	S2	Percentage Split	Without Outliers	79,31%
J48	S1	Percentage Split	Without Outliers	79,31%
PART	S1	Percentage Split	Without Outliers	79,31%
Decision Tree	S2	Percentage Split	Without Outliers	79,31%
OneR	S1, S2, S4	Percentage Split	Without Outliers	75,86%

Table 4: The 5 models with the highest specificity

DM Technique	Scenario	Sampling Method	Data Approach	Specificity
PART	S2	Percentage Split	With Outliers	98,51%
Random Forest	S1, S2	Percentage Split	With Outliers	97,01%
J48	S1	Percentage Split	With Outliers	97,01%
OneR	S1, S2, S3, S4	Percentage Split	With Outliers	97,01%
JRip	S3	Percentage Split	With Outliers	97,01%
Decision Tree	S2, S4	Percentage Split	With Outliers	95,52%

As it can be observed, overall the best achieved accuracy, precision, sensitivity and specificity values were 88.04%, 93.33%, 86.21% and 98.51%, respectively. In order to choose the most suitable model, a threshold was established, and the models were ranked according to their sensitivity results. The defined threshold was sensitivity >70%, accuracy >85%, precision >70% and specificity >85%.

Table 5 presents the best four models that achieved the threshold by their ranking order.

Table 5: Best models achieving the highest values of sensitivity

DM Technique	Scenario	Sampling Method	Data Approach	Sensitivity	Accuracy	Precision	Specificity
JRip	S1	Percentage Split	Without Outliers	86,21%	85,87%	73,53%	85,71%
J48	S1	Percentage Split	Without Outliers	79,31%	86,96%	79,31%	90,48%
Random Forest	S2	Percentage Split	Without Outliers	79,31%	85,87%	76,67%	88,89%
OneR	S1, S2, S4	Percentage Split	Without Outliers	75,86%	86,96%	81,48%	92,06%

It is observed that all four best models used Percentage Split as the sampling method and used the data without outliers. It can also be seen that there was a tie on the sensitivity values between the second and the third ranked models. Despite this tie, the first ranked model was, from these two, the one with the highest accuracy value.

5 Discussion

From the analysis of the best results, Scenario 1 comes out as the best scenario, where all variables are present, as within the best four models, three of them were from the Scenario 1. This happened, possibly because all attributes of the dataset are important to classify whether someone will die from heart-failure or not. On the other hand, the worst scenario, as any model of the Scenario 3 isn't present in the last table, thus showing that only using attributes like *age, ejection_fraction, serum_creatinine, serum_sodium, anaemia* and *high_blood_pressure* in our models is not enough to correctly classify the prediction of death due to heart failure. This set of attributes was suggested by other related articles, however with this data preparation and sample approaches, the results obtained were not the best. So, to understand if using other attributes could get better results, other sets were formed. Regarding Scenario 4, the attribute of *time* is of major importance, returning viable results through data mining processes. Its importance is also noticed in a healthcare context, where medical staff prioritize patients whose follow-up period is higher. In addition, with only these two attributes from Scenario 4, the healthcare professional can predict death by heart failure, without needing to wait for exam's results.

It is also notorious that the best results were achieved using percentage split rather than cross-validation. The last technique is normally better, however in this case that didn't happen, probably because of the small size of the dataset. Cross-validation uses all cases for training and testing over several iterations. Since the dataset isn't too large, perhaps using a well-balanced percentage for training and consequently for testing as well, such as 67% and 33%, respectively, gets a superior metric over cross-validation.

Besides all of this, it is also possible, through the analysis of the table containing the best results, to understand that by using outliers detection we can get better results overall. This happens because outliers are data that differ drastically from all others, that is, it is a value that escapes normality and can (and probably will) cause anomalies in the results obtained through algorithms and analysis systems. So, through its elimination, we can get more suitable results.

In addition, with this study, it is possible to identify *smoking, sex, platelets* and *diabetes* as irrelevant variables, having no effect on the mortality among heart failure patients. Most studies conclude on the male gender as a predictor of this death event. However, data mining processes concluded that the gender is not significant, which is also backed by other papers such as Roma'n et al. [8]. Regarding the variables: diabetes and smoking, there are reports of concerns on the effects of these variables on heart problems at initial stages, highlighting the side effects of medication [9, 10]. The chosen dataset only concerns patients of NYHA class III and IV, which are advanced stages of heart failure, thus we find no direct correlation between these patients' heart disease and those variables.

Thus, it is possible to claim that the most suitable model, from all the 96 induced models, is DMM = {Classification, S1, JRip, Percentage Split, Without Outliers, TG_1}.

6 Conclusion and Future Work

Healthcare improvement is visible with the growing application of data mining processes. Powerful patterns from the big data that flows through the medical infrastructure are uncovered when using such algorithms. These results can help predict diseases or outcomes of complex health systems such as pandemic outbreaks [11] or type 2 diabetes mellitus [12]. Through data mining, the prediction of death due to heart failure is becoming more precise and, from this study, it's evident that the follow-up time from a heart attack plays a huge factor, while basic knowledge is contradicted and uncovered as false dogmas, such as the correlation between smoking and heart diseases. In this paper, JRip returned the best sensitivity results compared to other algorithms, especially because it was the only algorithm that had a sensitivity above 80%. In addition, Scenario 1, which englobes all attributes, had the best results compared to more specific scenarios. This could be a consequence of the limitation of the present study. We must report the small size of the dataset (299 patients). A larger dataset would have permitted us to obtain more reliable results. In the future, additional information about the physical features of the patients – such as height, weight, body mass index – and the encompassing of another international region could help validate and verify our findings.

Acknowledgements

This work has been supported by José Machado, professor of Learning and Knowledge Extraction Systems and Director of Centro ALGORITMI, University of Minho and Hugo Peixoto, professor of Learning and Knowledge Extraction Systems.

References

[1] American Heart Association. ”Forging papers to sell fake art,” Accessed May 22, 2021. [Online]. Available: https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure

[2] Who. ”Cardiovascular diseases (CVDs),” Accessed May 22, 2021. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[3] Chicco D, Jurman G, “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone,” 2020. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5

[4] Willyard, C. ”Can AI Fix Medical Records?,” Accessed May 23, 2021. [Online]. Available: https://www.nature.com/articles/d41586-019-03848-y

[5] American Heart Association. ”Classes of Heart Failure,” Accessed Jun 1, 2021. [Online]. Available: https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/classes-of-heart-failure

[6] McClellan WM, Flanders WD, Langston RD, Jurkovitz C, Presley R., “Anemia and renal insufficiency are independent risk factors for death among patients with congestive heart failure admitted to community hospitals: a population-based study.” Jul. 13, 2002. [Online]. Available: https://jasn.asnjournals.org/content/13/7/1928

[7] Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA, “Survival analysis of heart failure patients: A case study.” Jul. 20, 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0181001

[8] Román AV, Shamagian LG, Diéguez MAB, Veloso PR, González-Juanatey JR., "Influence of sex on mortality in hospitalized patients with congestive heart failure and preserved or depressed systolic function." 2005. [Online]. Available: https://doi.org/10.1016/S1885-5857(06)60396-4

[9] Otero-Raviña F, Grigorian-Shamagian L, Fransi-Galiana L, Názara-Otero C, Fernández-Villaverde JM, del Álamo-Alonso A, et al., "Morbidity and mortality among heart failure patients in Galicia, N.W. Spain: the GALICAP study," Jul. 24, 2009. [Online]. Available: https://doi.org/10.1016/j.ijcard.2008.04.025

[10] U.S. Department of Health and Human Services, "A Report of the Surgeon General.The Health Consequences of Smoking — 50 Years of Progress," Rockville, MD, Tech. Rep., 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK179276/pdf/Bookshelf_NBK179276.pdf

[11] Ayyoubzadeh SM, Zahedi H, Ahmadi M, R Niakan Kalhori S., "Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study." Apr. 14, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7159058/

[12] Kazerouni, F., Bayani, A., Asadi, F. et al., "Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression: a comparison of four data mining approaches." Aug. 27, 2020. [Online]. Available: https://doi.org/10.1186/s12859-020-03719-8