Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2020, 7(11):165-169**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Detection of Cardiovascular Disease using Machine Learning Techniques

Ide, Mercy Azibaye.[1], Mercy Nwanyanwu[2], Chinenyeze, Evelyn C.[3]

[1]International Institute of Tourism and Hospitality Yenagoa, Bayelsa State, Nigeria
[2]Department of Computer Science, Captain Elechi Amadi Polytechnic Rumuola, Port Harcourt, Rivers State, Nigeria
[3]Department of Computer Science, Rivers State University, Port Harcourt, Rivers State, Nigeria
arikawei_4real@yahoo.com, mercynthia201@gmail.com, ivytranslate@outlook.com

Abstract One of the leading causes of death is cardiovascular disease. Identifying and predicting this disease in patients is the first step towards stopping their progression. We evaluated the capabilities of machine learning models in detecting cardiovascular disease. Our research utilizes supervised machine learning models to identify patients with such disease. Using an open-source dataset found on Kaggle, we conducted an exhaustive search of all available feature variables within the data to develop models for cardiovascular detection. Using different time-frames and feature sets for the data, XGBoost and RandomForest Classifiers were evaluated on the classification performance. We concluded a machine learned model based on examination and social history to provide an automated identification mechanism for patients at risk of cardiovascular diseases. The developed XGBoost for cardiovascular disease achieved accuracy of 0.74% while RandomForest achieved accuracy of 0.73%. The models identified age, height, weight, gender, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake and physical activity as key contributors.

Keywords Cardiovascular Disease, Diagnosis, machine learning techniques

Introduction

The leading cause of death in the United States is cardiovascular disease. It is estimated that the overall direct medical cost of cardiovascular disease will rise from $273 billion in 2010 to $818 billion in 2030. A system dynamics model for cardiovascular disease was developed by Hirsch *et al* [1]. They used the model to measure the efficacy of different treatments. Their model however, was unable to capture the impact of heterogeneous populations on the efficacy of various interventions, thus restricting the generalizability of the results to other populations.

There are great advantages to using data analytics in the health care system to provide insights, increase diagnosis, optimize outcomes, and minimize costs in the world of ever-growing data where hospitals are slowly implementing big data systems [2]. Efficient application of machine learning in particular, strengthens the work of medical professionals and improves the health care system's effectiveness. Through the success of machine learning models along with clinicians, major advances in diagnostic accuracy have been seen. Since then, machine learning models have been used in the prediction of several common diseases [3], including diabetes prediction [4], hypertension identification in diabetic patients, and Cardiovascular Disease (CVD) patient classification [5].

Machine learning models may be helpful in recognizing cardiovascular disease patients. Sometimes, several variables contribute to the detection of patients at risk for these common diseases. Machine learning techniques may help detect hidden patterns that might otherwise be overlooked in these variables.

In this paper, we use supervised machine learning models to predict cardiovascular disease. In turn, we are able to identify the feature of the disease which affects the prediction. The cardiovascular disease dataset found on Kaggle is used to train and test models for the prediction of CVD.

Related Work

Agent-based Simulation of Disease Spread Abroad Ship, suggested by Gutierrez [6]. He alluded to the catastrophic consequences that previous pandemics had caused. He points out that instruments that can clearly show the path and trajectory of the outbreak of the disease can enable the medical sector to take appropriate action and determine the efficacy of the precautionary measures used in the critical situation. He referred to previous models that displayed similar characteristics, such as the mathematical model [7] and stochastic models [8]. He believed that the deterministic model requires pre-existing data and that it cannot be extended to non-existent diseases and those that are likely to cause a potential pandemic. He also states that there is no chance in mathematical models. He notes that he used the same principles as the previous model for modeling human behavior, but argues that, in comparison to the closed environment used by the author, the previous model used an open environment. It also argues that precautionary steps are not applied in the second model and no airborne pathogens are modelled in the third model. The author stated that numerous simulations were conducted to adjust the results of precautionary steps, the percentage of vaccinated individuals and the methods of transmission. He also offers tables that display the effects of adjusting each parameter for various diseases. The author believed that while he was effective in modeling a disease simulation specific to navy ships, he agrees to the degree that there is plenty of space for future work to be done in the field of epidemic multi-agent based systems. In order to estimate cardiovascular parameters using a multi-agent scheme, Al-Jaafreh & Al-Jumaily [9] proposed a combination of two separate methods: Pulse Wave Velocity (PWV), heart rate and artery resistance. The principles of multi-agent collaboration were used. They have observed, reducing the error percentage and increasing measurement accuracy of cardiovascular parameters. Mehdizadeh *et al.* [10] presented a brief survey of some biological and biomedical applications of ABMS in biomedical engineering. These studies are related to fields of cancer, tissue engineering, angiogenesis, lung disease, clinical, morphogenesis, bone and epidemiology. In order to model the process of sprouting angiogenesis (blood vessel formation) inside polymeric porous scaffolds used for regenerative medicine, they have developed a multi-layered agent dependent system. Faber *et al* [11] simulated blood flow with the additional involvement of foreign bodies in capillary vessels of the human body and estimated the extent and usefulness of the data obtainable from such an environment. Cardiovascular flow modelled by Mabry *et al* [12] with a migrating agent method. They identified the fundamental problem of cardiovascular flow, followed by a connection using the SimAgent system between the model of the migrating agent and physiological processes. The agents that represent blood flows migrate along emanating paths from one area to another. They also addressed how SimAgents offer special cardiovascular simulation computing approaches. In their studies, they have used approaches to parallel processing. Yazdanbod & Marcus [13] have presented a 3-dimensional, interactive, agent-based model of the blood coagulation process within the Lindsay Composer (LC) computational framework, which can be used to simulate and visualize physiological processes inside the human body.

Methodology

Dataset Preprocessing

Data mining methods and techniques for transforming raw patient records to an appropriate format for training and testing machine learning models are part of the first stage of the pipeline. At this stage, the raw patient data was extracted from the dataset of cardiovascular diseases to be interpreted in the preprocessing stage as records. Any undecipherable values (errors in datatypes and standard formatting) from the database have also been converted to null representations in the preprocessing stage.

In the feature extraction stage, the patient records were then represented as a data frame of characteristics and a class mark. The characteristics are an array of patient data gathered by demographic, examination and social history. The class mark is a categorical attribute, represented as a binary patient classification: 0-Non-cases, 1-Cases. For research, categorical features were encoded with numerical values.

Feature Selection

We assessed the feature dependency of the models for cardiovascular disease prediction in order to build an accurate model based on a restricted set of available features, i.e. features that did not require excessive testing of patients. The study was performed based on XGBoost classifier (based on the output of the model), where an error rate metric was used to rank features. More precisely, feature importance scores for each decision tree are determined in the XGBoost model by how much the split-point(s) for each feature increases the rate of binary classification error. The error rate is estimated over the total amount of data as the number of misclassified data.

Classification

Answer varies for all the cardiovascular symptoms. For instance; if age = 60, height = 163, weight = 78, gender = male, systolic blood pressure = 143, diastolic blood pressure = 125, cholesterol = above normal, glucose = normal, smoking = no and alcohol intake = yes and physical activity no, then the subject was labeled as "You have a high chance of having a Cardiovascular Disease. Please contact a Cardiologist as soon as possible", thus label = 1, otherwise label = 0.

Dataset

The cardiovascular disease dataset is an open-source dataset found on Kaggle. The data consists of 70,000 patient records (34,979 presenting with cardiovascular disease and 35,021 not presenting with cardiovascular disease) and contains 11 features (4 demographic, 4 examinations, and 3 social history): Age, Height, Weight, Gender, Systolic blood pressure, Diastolic blood pressure, Cholesterol, Glucose, Smoking, Alcohol intake and Physical activity.

Table 1: Feature importance for cardiovascular disease classifier with demography, examination and social history. This table shows the most important features for predicting cardiovascular disease. In gender column, male = 1 and female = 0

Id	Age	Gender	Height	Weight	Ap_hi	Ap_lo	Cholesterol	Gluc	Smoke	Alco	Active	Cardio	Bmi
1	62	1	155	69.0	130	80	2	2	0	0	1	0	28.720
2	40	1	163	71.0	110	70	1	1	0	0	1	1	26.722
3	60	1	165	70.0	120	80	1	1	0	0	1	0	25.711
4	40	0	165	85.0	120	80	1	1	1	1	1	0	31.221
5	64	1	155	62.0	120	80	1	1	0	0	1	0	25.806

Results

Table 2: Prediction of Bp and Bmi Level

Id	Bp_level	Age_level	Bmi_level
1	high	6	Overweight
2	normal	2	Overweight
3	normal	6	Overweight
4	normal	2	Obese
5	normal	6	overweight

Table 3: Compares the performance metrics of XGBoost Classifier and RandomForest

Model	Precision	Recall	F1-score	Accuracy
XGBoost Classifier	0.76	0.70	0.73	0.74
RandomForest Classifier	0.75	0.70	0.72	0.73

The developed XGBoost for cardiovascular disease achieved F1-score of 0.73% and accuracy of 0.74% while RandomForest achieved F1-score 0.72% and accuracy of 0.73%. The models identified age, height, weight, gender, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake and physical activity as key contributors.

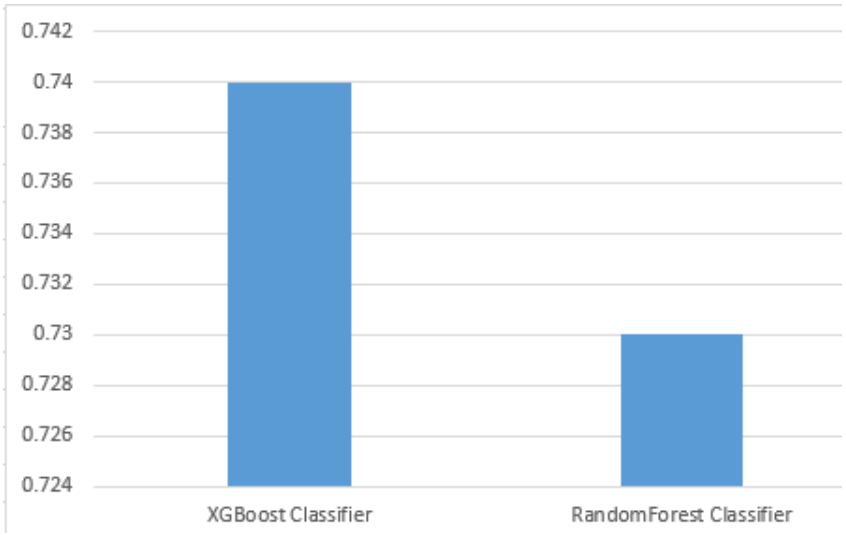

Figure 1: Prediction accuracy rate summary

Conclusion

Our research utilizes supervised machine learning models to identify patients with such disease. Using an open-source dataset found on Kaggle, we conduct an exhaustive search of all available feature variables within the data to develop models for cardiovascular detection. Using different time-frames and feature sets for the data, XGBoost and RandomForest Classifiers were evaluated on the classification performance. We conclude machine learned model based on examination and social history to provide an automated identification mechanism for patients at risk of cardiovascular diseases.

References

[1]. Hirsch, G., Homer, J., Evans, E. & Zielinski A. (2010). A system dynamics model for planning cardiovascular disease interventions. Am J Public Health, 100(4), 616–22.

[2]. Raghupathi, W. & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science System, 2(1), 3.

[3]. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol Journal, 13, 8–17.

[4]. Semerdjian, J. & Frank, S. (2017). An Ensemble Classifier for Predicting the Onset of Type II Diabetes. ArXiv e-prints. 1708.07480.

[5]. Parthiban, G. & Srivatsa, S. K. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. International Journal Appl. Information System (IJAIS), 3, 2249–0868.

[6]. Gutiterrez, L. (2005). Agent-based Simulation of Disease Spread Abroad Ship, Masters thesis, Naval Postgraduate School California.

[7]. Wang, W. & Ruan, S. (2004). Simulating the SARS Outbreak in Beijing with Limited Data. Journal of Theoretical Biology. 227(3). 369-79.

[8]. Muller, G., Grebaut, P. & Gouteux, J.P. (2004). An Agent Based Model of Sleeping Sickness: Simulation Trials of a Forest Focus in Southern Cameroon. CRAS, 327(1). 1-11.

[9]. Al-Jaafreh, M. & Al-Jumily, A. (2005). Multi Agent System for Estimation of Cardiovascular Parameters. 1st International Conference on Computers, Communications, & Signal Processing with Special Track on Biomedical Engineering, 269-299.

[10]. Mehdizadeh, H., Artel, A., Brey, E. M. & Cinar, A. (2011). Multi-agent systems for biomedical simulation: Modeling vascularization of porous Scaffolds, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7047, 113-128.

[11]. Faber, L., Boryczko, K. & Kisiel-Dorohinicki, M. (2014). Hybrid Architecture for Simulation of Blood Flow with Foreign Bodies. 28th European Conference on Modelling and Simulation, ECMS Proceedings, ISBN: 978-0-9564944-8-1.

[12]. Mabry, S. L., Bic, L. F. & Baldwin, K. M. (2000). Modling Cardiovascular Flow with a Migrating Agent System. International Conference on Health Sciences Simulation, Proceedings of Western Multi Conference, Medical Sciences Simulation.

[13]. Yazdanbod, I. & Marcus, S. (2011). An Agent-Based Simulation of Blood Coagulation Processes, Journal of Undergraduate Research in Alberta (JURA), 1.