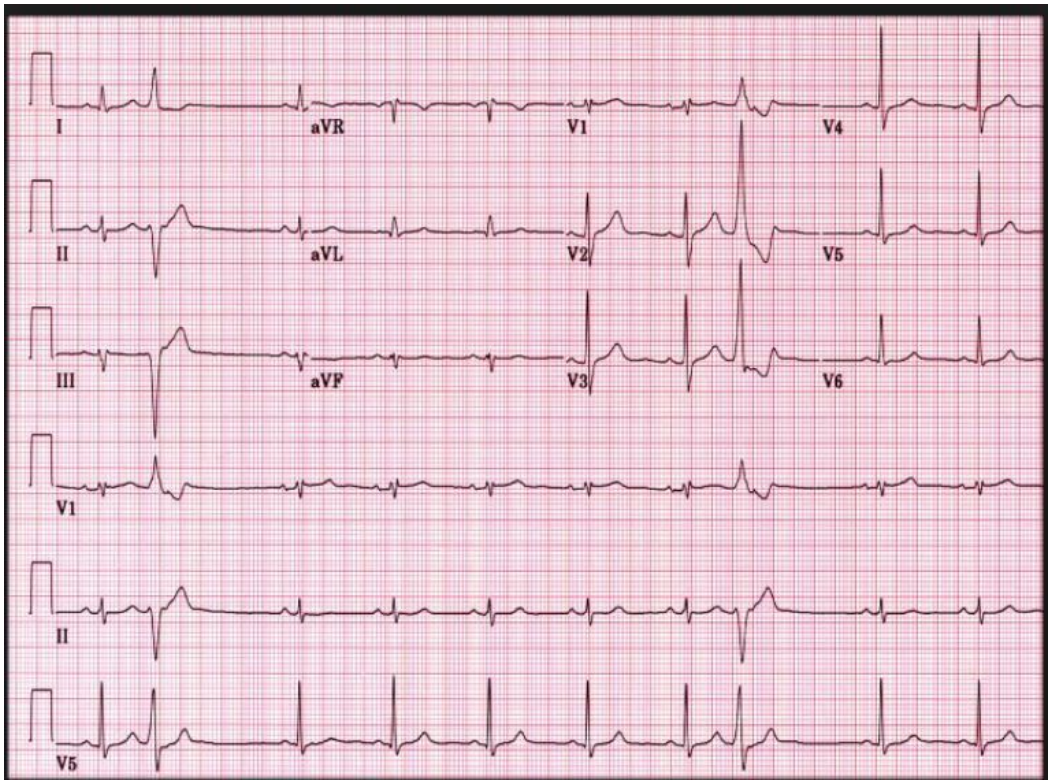
Developing a probabilistic program to analyze combinations of risk factors for arrhythmia detection

Shouri Bochetty

RISE Online STEM Research Institute



Abstract/Summary:

Arrhythmia is a disease characterized by abnormal electrical signals in the heart that results in ineffective pumping. Stethoscopes, as well as EKG's, are used to diagnose arrhythmia. Modern Stethoscopes diagnose arrhythmias with a 40% accuracy rate. EKG's, on the other hand, diagnose with an high accuracy rate but can be very expensive. Sometimes, money on EKG's can be wasted if you had a benign arrhythmia (heart simply skips a beat). This application/program analyzes combinations of certain risk factors that a person may have and gives the probability that a person develops a life-threatening arrhythmia. The program will be used with the numerical results of the stethoscope instead of a doctor's diagnosis. The application will significantly reduce the cost of medical EKG's for families by correcting for inaccuracies in the diagnoses of the doctor's stethoscope, therefore reducing the number of "useless" EKG-related arrhythmia tests taken. The program has been tested using published data on arrhythmia patients from various websites, but it is not possible to get data on the probability that a person could develop arrhythmia. The program has been able to predict (probability was greater than 70%) that a person develops a life-threatening arrhythmia with an 80% accuracy rate.

Acknowledgements

First, I would like to thank Mrs. Jacklyn Naughton, my online RISE instructor, who advised and helped me throughout my research. She also taught me a lot about the scientific method, statistics, and research in general. Next, I would like to thank my classmates for being very supportive, proofreading my work, and for giving me great suggestions. Finally, I would like to thank my parents for providing me with the tools to run my program.

Purpose and Hypothesis

Purpose: To develop an algorithm that diagnose harmful arrhythmia more accurately than a doctor's diagnosis with a stethoscope, which gives the probability a patient develops life-threatening arrhythmia. However, the program will be used with the numerical results of the stethoscope. The algorithm should help save money spent on EKG's by determining which people need to get an EKG test.

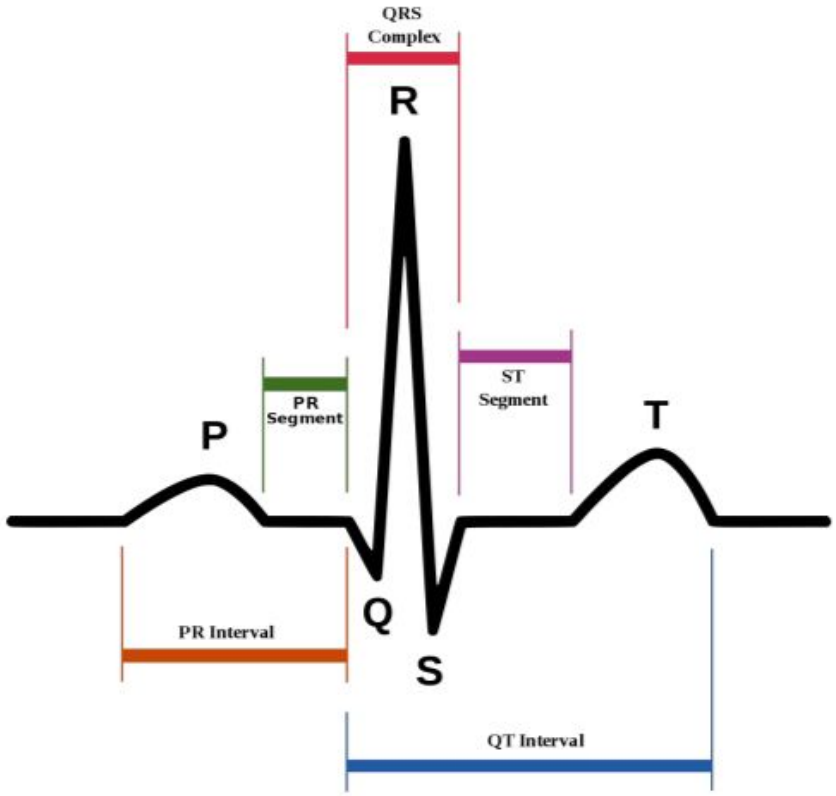
Hypothesis: If the program is applied to patients with arrhythmia with the numerical results from a stethoscope, it will diagnose life-threatening arrhythmias with a higher accuracy rate than a doctor's diagnosis with a stethoscope.

Introduction

Having arrhythmia can mean you potentially have heart disease, the wrong balance of electrolytes in the blood, injury from a heart attack, etc. Detecting arrhythmia more accurately can significantly help with finding out if you have any of the causes of arrhythmia, which can be deadly. Currently doctors use a stethoscope and their ears to detect arrhythmia. This method has a 40% accuracy rate. A mathematical model/computer algorithm would be much more efficient to determine if one has arrhythmia.

Background Information/Introduction cont.

The cycle of one contraction/pumping of the heart contains five stages/waves. These waves are the P, Q, R, S, and T waves. These waves have attributes like a specific duration and an amplitude that extends below and above the baseline. The P wave cycle signifies the depolarization of the atrias. The QRS complex signifies the depolarization of the ventricles. The T wave signifies the repolarization of the ventricles. The durations of the waves can be observed using a stethoscope, however, you need an EKG test to determine the amplitudes of these waves.



Materials

- A computer with Pyspark downloaded in it as well as python (the model of the computer doesn't matter)
- ALS algorithm downloaded into computer
- Excel spreadsheet
- K-means algorithm downloaded into computer
- Data from patients from various websites about the intervals and amplitudes (must be a CSV file)

-- Note all of the websites that data was obtained from are listed in the references section.

How the program works/Design Plan

1. The program uses the K-means algorithm to cluster any data into who could potentially have arrhythmia and normal people. It checks the durations of the P-wave and the T-wave. These two factors most significantly contribute to arrhythmia, which will be explained later in my paper as per the reason why.
2. Next, in the cluster of people who could potentially have/develop arrhythmia, the ALS algorithm was used to fill in any missing information of data.
3. The program then goes through the weighted percentage risk factors (based on how much each risk factor contributed to arrhythmia) and assigns full or half weightage depending on how much the patient's value is outside of the normal range. In this step only the 10 risk factors out of the 15 that can be calculated without an EKG test were considered. For most of the risk factors, if it was outside of the normal range by between 0 to 10%, it was

assigned 50% of that risk factor's total weightage. If it was outside of the normal range by greater than 10%, the risk factor's full weightage was assigned. In the case of the durations of the P-value and the T-value, if the patient's value was outside of the normal range by 0 to 20%, it was assigned half weightage; if it was outside of the normal range by greater than 20%, it was assigned full weightage. This was determined, by creating a graph for each of the patient's risk factors and the percentage that it was outside of the normal range. The best horizontal of best fit was found, and this was the percentage of importance (weightage) given to the respective risk factor.

4. Based on the results of the 10 tested risk factors, the patient would take an EKG test if the algorithm said the probability he develops arrhythmia is greater than 70%. The EKG test would be useful in finding the amplitudes of the waves (P, Q, R, S, and T waves of the heartbeat).
5. Finally, the algorithm would give a combined probability based on the test involving the EKG and the test not involving the EKG test. The test involving the EKG had an 30% importance and the test not involving the EKG had an 70% importance. These percentages were established based on the weights given to the risk factors, which will later be discussed. The weighted average of these was done to find the total probability that the patient develops arrhythmia.

--Note that anything outside of this range would receive the full weightage of that risk factor in the program

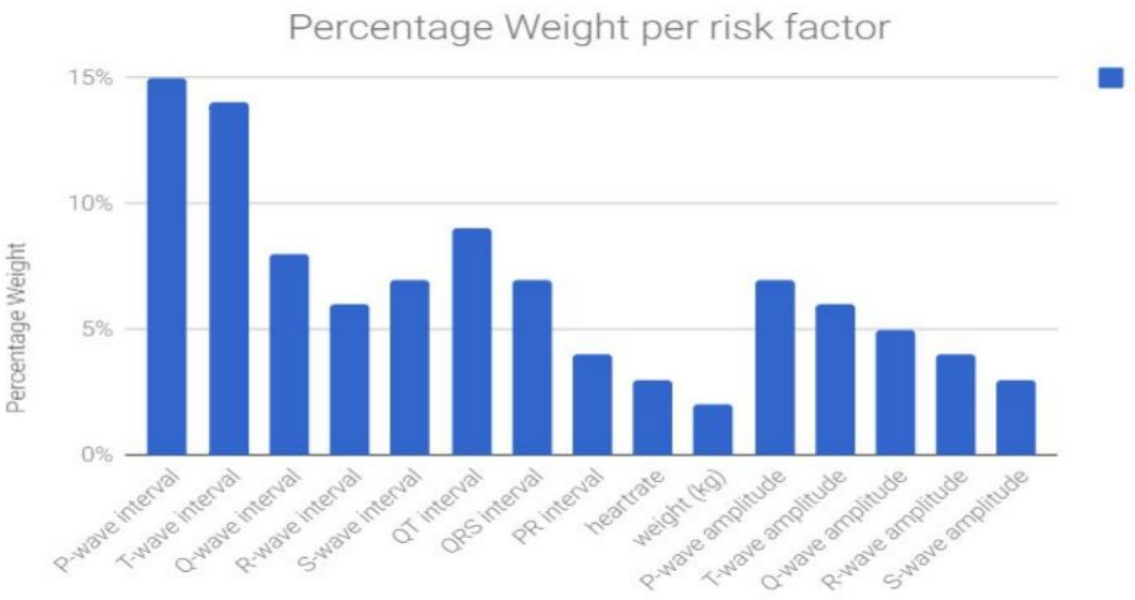
QRS(dur)-76-135 ms P(dur) male and female-80-120 ms QT(dur)-345-434 ms P-R(dur)-110-210 ms T(dur)-140-180 ms Q (dur)-20-50 ms R (dur)-25-55 ms S (dur)-10-20 ms P (amp)- 0.1 to 0.3

mv Q (amp)- 0.1 to 0.4 mv R (amp)-1.3 to 1.75 mV S (amp)-1.0 to 1.3 mv T (amp)- 0.2mv-0.5 mv --Note: 1mv = 10mm

Redesign and Retest

There wasn't much redesigning or retesting that was done in the algorithm/program, except where initially in the program the percentage weights to the 15 risk factors were assigned based on basic intuition, but then later, the best line of best fit was calculated to minimize the Root Mean Square error (distance between the predicted percentage that a patient develops arrhythmia and the actual percentage the patient develops arrhythmia on a Euclidean plane). Since it is impossible to get data on the actual percentage that a patient develops arrhythmia, it was assumed that a patient who had arrhythmia has a greater than 70% chance of developing arrhythmia. Additionally, the program was scrutinized for any bugs and they were removed. Also, in an effort to minimize the usage of the ALS algorithm, the ALS algorithm was only used once to fill in the missing data of the category of patients who are likely to develop arrhythmia. This is largely due to the fact that ALS is very costly to use and uses up a lot of resources. Previously, the ALS algorithm was used on all patients. Now, the algorithm checks the P-wave duration and the T-wave duration (most significantly contributing risk factors) first before placing a patient into the category of likely to develop arrhythmia and not likely to develop arrhythmia.

Weights given per risk factor



Analysis was done on how common it was for a risk factor to be out of the normal range in someone with arrhythmia. Existing data was checked to see which factors were normally in range and which weren't. Initially, extra weightage was assigned to the P-wave and T-wave durations and amplitude. It was hypothesized that they would be extra important as the T-wave repolarizes the ventricles, returning the heart to its normal stage, setting the stage for the next cycle of the heart pumping blood. Also, the T-wave depolarizes the atria, and in arrhythmia the charge of ions that exists in the ventricles and atria are especially important. It was observed that in patients with arrhythmia, these values were consistently out of the normal range.

Variables

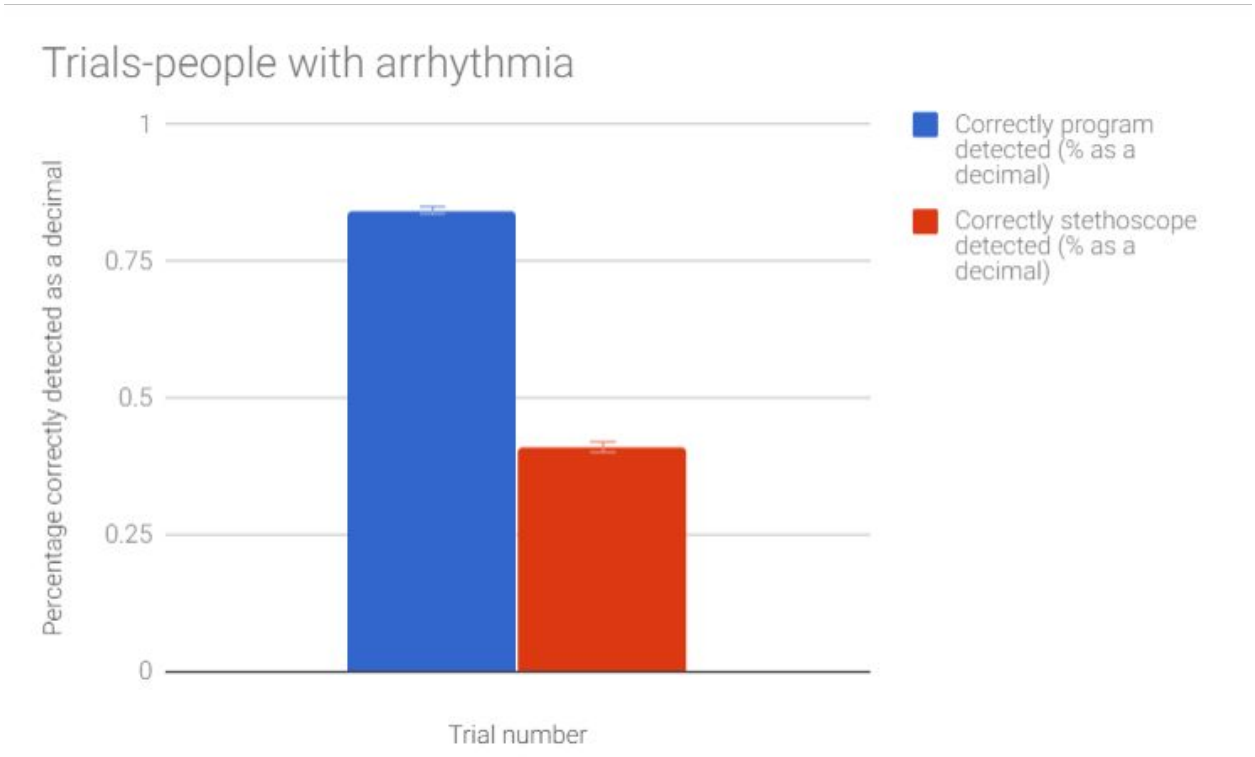
Independent variable: The design of the software program and its algorithms

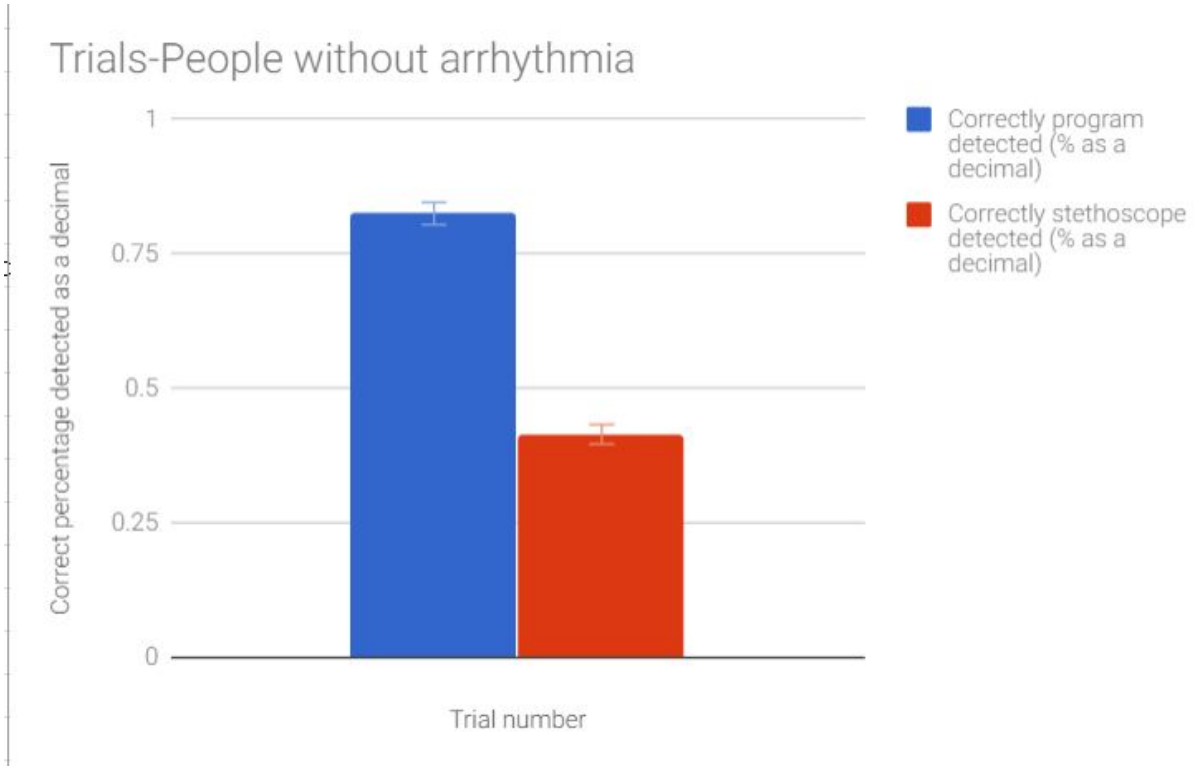
Dependent variable: The number of patients correctly diagnosed as having arrhythmia or as normal (accuracy rate of the program)

Control group: The number of patients correctly diagnosed as having arrhythmia or as normal using a stethoscope (accuracy rate of the stethoscope)

Controlled Variables: the type of stethoscope used and brand was constant throughout all of the test runs, the same sets of data were used for both my program and for the stethoscope

Data





Trial number-people with arrhythmia	Correctly program detected (% as a decimal)	Correctly stethoscope detected (% as a decimal)
Trial 1	0.8426966292	0.404494382
Trial 2	0.84	0.412
Trial 3	0.8477366255	0.4156378601
Average	0.8434777516	0.4107107474
St DEV	0.003927016386	0.005682508696
ST Error	0.002267263968	0.003280797925
CI	0.006620379044	0.009579884011

Trial number-people without arrhythmia	Correctly program detected (% as a decimal)	Correctly stethoscope detected (% as a decimal)
Trial 1	0.8108108108	0.427027027
Trial 2	0.8316326531	0.4081632653
Trial 3	0.8325123153	0.4088669951
Average	0.8249852597	0.4146857625
ST DEV	0.01228330992	0.01069363909
ST Error	0.007091772291	0.006173975409
CI	0.02070787581	0.01802792176

-Note:

In the with arrhythmia category trial runs:

Trial 1: the program was applied to 267 patients

Trial 2: the program was applied to 250 patients

Trial 3: the program was applied to 243 patients

In the without arrhythmia category trial runs:

Trial 1: the program was applied to 185 patients

Trial 2: the program was applied to 196 patients

Trial 3: the program was applied to 203 patients

Data Analysis

The difference between correctly program detected and stethoscope detected for both people without arrhythmia and people with arrhythmia trial runs is significant. None of the error bars overlap. The program is about 40% higher in both cases than a diagnosis with a stethoscope. There is a lot of error when using a stethoscope, especially due to the fact that a doctor cannot numerically or statistically assess the patient's chance of developing arrhythmia, as he/she doesn't know how much each factor contributes to one having/developing arrhythmia. The error bars were calculated using the confidence interval, which was calculated using the standard error multiplied by the t-distribution critical value of (n-1), where n is the sample size at theta equal to 0.05.

Experimental Error

In the algorithm, there were some sources of error. Many were due to the simplifications that were made. However, these simplifications do not render the results of the algorithm useless. One of the simplifications was that the program only accounted for 15 risk factors, and did not account for risk factors such as, depression, high BP, medical history, etc. There were almost 200+ risk factors and most of them had very little effect in determining if a patient develops arrhythmia. Additionally, as it was a high school research project, it was not feasible to write a program that accounts for 200+ risk factors. Although this simplification was made, the 15 risk factors that were chosen most heavily contributed to arrhythmia. Also, only 1,344 patients were in the trial runs. This is a relatively low amount considering that the program could be applied to all potential arrhythmia patients in the world. However, most of the patients had largely varying numbers of intervals, durations, and amplitudes of the 5 waves (P,Q,R,S, and T). In addition to this, the ALS algorithm was used in the program only once where it could have been used in other places to improve accuracy. This is largely due to the fact that ALS is a very costly algorithm to run and takes up a lot of resources. Furthermore, as this is a computer research project error can come from unknown bugs in the software. However, my program was scrutinized for any bugs, and all identified bugs were fixed. Each part of the program currently functions without any errors.

Conclusion

In conclusion, the hypothesis was supported. The algorithm/program had an accuracy rate of about 80% while the stethoscope had an accuracy rate of about 40%. In the trials with arrhythmia, the program detected people who have arrhythmia with an 84.3% accuracy rate while the stethoscope had an 41.1% accuracy rate. The algorithm/program performed about 40% higher than the stethoscope. However, the program is not accurate enough to be an accurate measure of if someone has arrhythmia or not. The algorithm must have an 95% accuracy rate in order for it to be commonly used in a doctor's office. Although, some simplifications were made, most of them were insignificant or were somewhat addressed to. This program has many applications, as it could potentially be used to diagnose arrhythmias instead of relying on a doctor's diagnosis. It will help save money spent on EKG's by reducing the number of useless EKG tests taken (heart simply skips a beat during the stethoscope test). It could potentially save many deaths caused by arrhythmias. There are many directions that this project can progress in the future. One could be to make the program account for other risk factors besides the 15 that were used. A more intelligent algorithm could reduce the number of iterations in the algorithm/program. This is especially important as the ALS algorithm is very costly to use and takes up a lots of resources. Current research has been done on arrhythmia, however most do not take a similar approach to what was done. However, some make use of probabilistic models to detect arrhythmia.

Reference List

Srivastava, Nitish, Hinton, Geoffrey E, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1): 1929–1958, 2014.

Turakhia, Mintu P, Hoang, Donald D, Zimetbaum, Peter, Miller, Jared D, Froelicher, Victor F, Kumar, Uday N, Xu, Xiangyan, Yang, Felix, and Heidenreich, Paul A. Diagnostic utility of a novel leadless arrhythmia monitoring device. The American journal of cardiology, 112 (4):520–524, 2013.

Clifford, GD, Liu, CY, Moody, B, Lehman, L, Silva, I, Li, Q, Johnson, AEW, and Mark, RG. Af classification from a short single lead ecg recording: The physionet computing in cardiology challenge 2017. 2017.

Shah, Atman P and Rubin, Stanley A. Errors in the computerized electrocardiogram interpretation of cardiac rhythm. Journal of electrocardiology, 40(5):385– 390, 2007.

Guglin, Maya E and Thatai, Deepak. Common errors in computer electrocardiogram interpretation. International journal of cardiology, 106(2):232–237, 2006.

Pan, Jiapu and Tompkins, Willis J. A real-time QRS detection algorithm. IEEE transactions on biomedical engineering, (3):230–236, 1985.

Li, Cuiwei, Zheng, Chongxun, and Tai, Changfeng. Detection of ECG characteristic points using wavelet transforms. IEEE Transactions on biomedical Engineering, 42(1):21–28, 1995. Mart´ınez, Juan Pablo, Almeida, Rute, Olmos, Salvador, Rocha, Ana Paula, and Laguna, Pablo. A waveletbased ECG delineator: evaluation on standard databases. IEEE Transactions on biomedical engineering, 51(4): 570–581, 2004.

Moody, George B and Mark, Roger G. A new method for detecting atrial fibrillation using RR intervals. Computers in Cardiology, 10(1):227–230, 1983.

Artis, Shane G, Mark, RG, and Moody, GB. Detection of atrial fibrillation using artificial neural networks. In Computers in Cardiology 1991, Proceedings., pp. 173– 176. IEEE, 1991.

https://archive.ics.uci.edu/ml/datasets/arrhythmia

https://www.physionet.org/physiobank/database/mitdb/

http://datamining.togaware.com/survivor/Cardiac_Arrhythmia.html