An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction

Project

CSE 4238

Soft Computing Lab

Submitted by

Farhan Sharukh Hasan	**170204066**
Rahat Kader Khan	**170204074**
Shweta Bhattacharjee Porna	**170204111**

Submitted To

Nibir Chandra Mandal
Mr. H M Zabir Haque

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

March 8, 2022

ABSTRACT

In this project,we have implemented a paper named "An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction" where we adopt a two-step approach: first, we employ least absolute shrinkage and selection operator (LASSO) based feature weight assessment followed by majority-voting based identification of important features. Next, the important features are homogenized by using a fully connected layer, a crucial step before passing the output of the layer to successive convolutional stages. The data is curated from the National Health and Nutritional Examination Survey (NHANES) with the goal of predicting the occurrence of Coronary Heart Disease (CHD).z

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Heart disease is a leading cause of death today, with coronary heart disease (CHD) being the most common form of cardiovascular disease that accounts for approximately 13% of deathsin the US (Benjamin, 2019). Timely diagnosis of heart disease is crucial in reducing health risk and preventing cardiac arrests. An American Heart Association study projects an almost 100% increase in CHD cases by 2030 (Benjamin, 2019).Numerous risk factor variables often make the prediction of CHD difficult, which in turn, increases the cost of diagnosis and treatment. In order to resolve the complexities and cost of diagnosis, advanced machine learning models are being widely used by researchers to predict CHD from clinical data of patients. In this project,we have implemented a paper named "An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction" where we adopt a two-step ap- proach: first, we employ least absolute shrinkage and selection operator (LASSO) based feature weight assessment followed by majority-voting based identification of impor- tant features. Next, the important features are homogenized by using a fully connected layer, a crucial step before passing the output of the layer to successive convolutional stages. The data is curated from the National Health and Nutritional Examination Sur- vey (NHANES) with the goal of predicting the occurrence of Coronary Heart Disease (CHD).

We propose an efficient neural network with convolutional layers using the NHANES dataset to predict the occurrence of CHD. A complete set of clinical, laboratory and examination data are we used in the analysis along with a feature selection technique by LASSO regression. Data preprocessing is performed using LASSO followed by a feature voting and elimination technique. The performance of the network is compared to several existing traditional ML models in conjunction with the identification of a set of important features for CHD. Our architecture is simple in design, elegant in concept, sophisticated in training schedule, effective in outcome with far-reaching applicability in problems with unbalanced datasets. Our research contributes to the existing studies in three primary ways: 1) our model uses a variable elimination technique using LASSO and feature voting as preprocessing steps;

2) we leverage a shallow neural network with convolutional layers, which improves CHD prediction rates compared to existing models with comparable subjects (the 'shallowness' is dictated by the scarcity of class-specific data to prevent overfitting of the network during training); 3) in conjunction with the architecture, we propose a simulated annealing-like training schedule that is shown to minimize the generalization error between train and test losses.It is important to note that our work is not intended to provide a sophisticated architecture using a neural network. We also do not focus on providing theoretical explanation on how our network offers resistance to data imbalance. Instead, our goal is to establish that under certain constraints one can apply convolutional stages despite the scarcity of data and the absence of well-defined data augmentation techniques and to show that the shallow layers of convolution indeed offer resilience to the data imbalance problem by dint of a training schedule. The proposed pipeline contributes to improving CHD prediction rates in imbalanced clinical data, based on a robust feature selection technique using LASSO and shallow convolutional layers. This serves to improve prediction algorithms included in smart healthcare devices where sophisticated neural algorithms can learn from past user data to predict the probability of heart failure and strokes. Prediction rates could be integrated in healthcare analytics to provide real time monitoring which not only benefits the patients but also medical practitioners for efficient operations. The present research also focuses on a systematic training schedule which can be incorporated in smart devices to improve tracking of different predictor variable levels for heart failure

Chapter 2

Motivation

Being the first leading causes of death, different types of heart diseases result in serious illness and disability, decreased quality of life, and hundreds of billions of dollars in economic loss every year. The burden of Coronary Heart Disease (CHD) is disproportionately distributed across the population. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. In our society, many people (specially old-aged) suffer from Coronary Heart Disease (CHD) due to imbalance of cholesterol level, glucose level, irregularities in blood pressure, etc. Our goal is to efficiently guess the possibilities of Coronary Heart Disease of people of different ages using a proposed Neural Network architecture. The challenge is to make prediction of the occurrence of Coronary Heart Disease (CHD) at a low cost, using the best architecture. These results can be achieved by employing appropriate computer based information and decision support system. It enables significant knowledge, e.g, relationships between medical factors related to heart disease and patterns, to be established. The CNN classifier that we are going to use results in high specificity and test accuracy along with high values of recall and area under the curve (AUC).

Chapter 3

Methodology

3.1 LASSO Shrinkage and Majority Voting

LASSO or least absolute shrinkage and selection operator is a regression technique for variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model it produces. In LASSO, data values are shrunk toward a central point, and this algorithm helps in variable selection and parameter elimination. This type of regression is wellsuited for models with high multicollinearity. LASSO regression adds a penalty equal to the absolute value of the magnitude of coefficients, and some coefficients can become zero and are eventually eliminated from the model. This results in variable elimination and hence models with fewer coefficients. LASSO solutions are quadratic problems and the goal of the algorithm is to minimize

$$\sum_{i=1}^{n}\left(y_i - \sum_{j} x_{ij}\gamma_j\right)^2 + \lambda \sum_{j=1}^{p} |\gamma_j|$$

which is the same as minimizing the sum of squares with constraint $|\gamma i|s$ Some of the γ values are shrunk to exactly zero, resulting in a regression model that's easier to interpret. A tuning parameter, which is the amount of shrinkage, controls the strength of the regularization penalty. When $\gamma = 0$, no parameters are eliminated. The estimate is equal to the one found with linear regression. As increases, more coefficients are set to zero and eliminated.

As increases, bias increases and as decreases, variance increases. The model intercept is usually left unchanged. The γ value for a variable (factor) can be interpreted as the importance of the variable in terms of how the it contributes to the underlying variation in the data. The variable with a zero γ considered unimportant. It is to note that LASSO shows misleading results in case of data imbalance, which may prompt incorrect selection of important variables if we perform LASSO on the entire dataset. Note that the variables in our dataset are mixed data type – a subset of them are categorical. In this work, as a standard practice, we use group LASSO and we refer it as LASSO for simplicity. In order to mitigate the effect of imbalance, we adopt a strategy to randomly subsample the dataset and iterate LASSO multiple times. Majority voting is performed on the set of γ values to identify the variable that are nonzero in major number of iterations. Let us assume, that LASSO is performed N times on N randomly subsampled dataset, where each instance has equal number of examples in case of CHD and no-CHD.

$$\chi(\gamma) = \begin{cases} 1 & if\ \gamma \neq 0 \\ 0 & otherwise \end{cases}$$

$$[\chi(\gamma_{1,c})\ \chi(\gamma_{2,c}) \dots\dots\ \chi(\gamma_{N,c})]\mathbf{1} \geq \frac{\boldsymbol{N}}{\boldsymbol{\alpha}} \Longrightarrow \mathbf{c}\ \text{is selected}$$

3.2 CNN Architecture

The architecture is a sequential one-input-one-output feedforward network. For simplicity, we assume the class of subjects with presence of CHD as class '1' and the subjects with absence of CHD as class '0'. As mentioned in the previous section, the number of active phenotypes of CHD obtained from majority voting is 50. Let the number of training examples be N, which indicates that the input layer in Fig. 2 has dimension of R,power(N*50) . The dense or fully connected layers, consisting of 64 neurons collectively act as a linear combiner of the 50 variables and bias, which effectively homogenizes different variable types before nonlinear transformation. The nonlinear transformation is carried out by rectified linear unit (ReLU). Dropout with 20% probability is performed to reduce overfitting. Following the fully connected layer, there is a cascade of convolution layers. In the first convolution layer, there are two filters of kernel width 3 and stride 1. The layer is not provided

with external zero-padding. In the pooling layer, we rigorously experiment with different pooling strategies and find average pooling working marginally better than max pooling under all constraints. The first convolution layer converts the output of fully connected block ∈R,Power($N \times 64$) to a tensor of dimension R,Power($N \times 64 \times 1$) . The tensor is then subjected to batch normalization, nonlinear transformation and average pooling with an output tensor of dimension R,Power($N \times 32 \times 2$) . The filters in the last no-zero-padded convolution layers are taken with kernel 5 and stride 1, delivering an output tensor of R,Power($N \times 13 \times 4$) to the next dense layer after the average pooling layers. The categorical output is observed at the end of the softmax layer, where we set the loss function as the categorical cross-entropy loss. The bias in each layer is initialized with random numbers drawn from a truncated normal distribution with variance $1 \div \sqrt{n}$, where n is the number of 'fan-in' connections to the layer. We use Adam optimizer with learning rate 0.005, $\beta 1 = 0.9$, $\beta 2 =$ 0.999 and zero decay. Our proposed architecture consists of 32,642 trainable and 1,164 non-trainable parameters. We experiment with several hyperparameters that are associated with our model to obtain consistent class-wise accuracy. We provide results by varying subsampling of input data, epochs, class-weights, the number of neurons in each dense layer except the last one, and the number of filters in each convolution layer during training

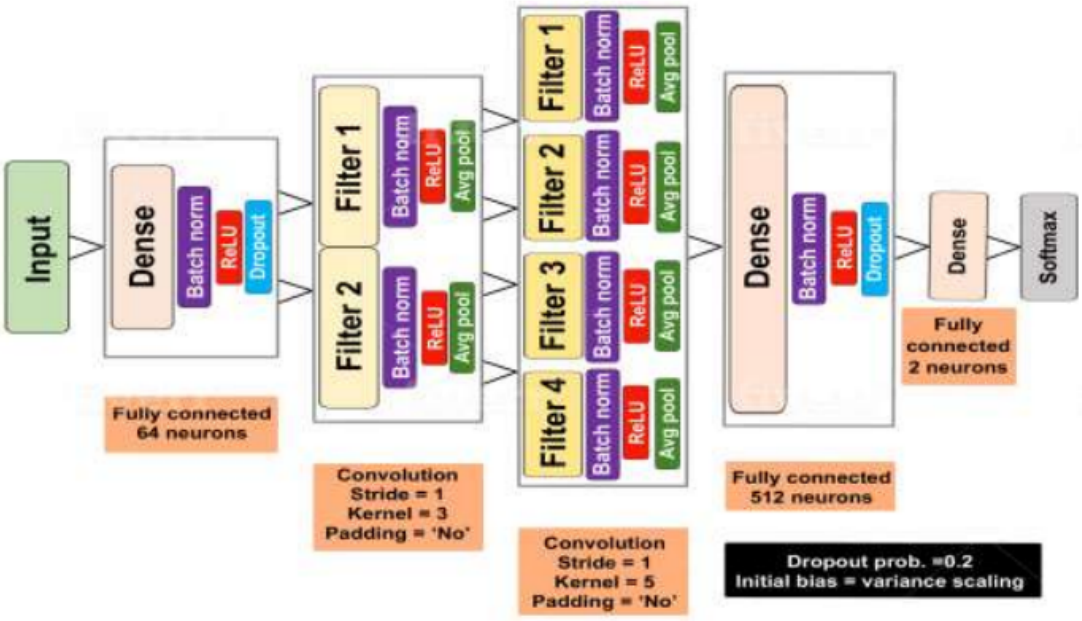

Figure 3.1: Proposed convolutional neural network architecture. The 'Input' is a 1D numerical array corresponding to all the factors/variables from LASSO-Majority Voting preprocessing stage. The 'Dense' layer, immediately after the 'Input', combines all the factors and each neuron (computing node) at the output of 'Dense' layer is a weighted combination of all the variables, indicating a homogeneous mix of different variable types. The next two convolution layers seek representation of the input variables via the 'Dense' layer. The next two 'Dense' layers are followed by the 'Softmax' layer. The last two 'Dense' layers (before the 'Softmax' layer) can be retrained for transfer learning in case new data is obtained. The associated training parameters, such as dropout probability, number of neurons, activation function (we used ReLU), pooling types, and convolution filter parameters are shown in the above figure. Owing to the large number of parameters that can lead to overfitting of training data points,

Chapter 4

Experiments

4.1 Dataset Collection

Our study uses the NHANES data from 1999-2000 to 2015-2016. The dataset is compiled by combining the demographic, examination, laboratory and questionnaire data of 37,079 (CHD– 1300, Non-CHD – 35,779) individuals as shown in Figure 1. Demographic variables include age and gender of the survey participants at the time of screening. Participant weight, height, blood pressure and body mass index (BMI) from the examination data are also considered as a set of risk factor variables to study their effect on cardiovascular diseases. NHANES collects laboratory and survey data from participants once in every two years depending on their age and gender.

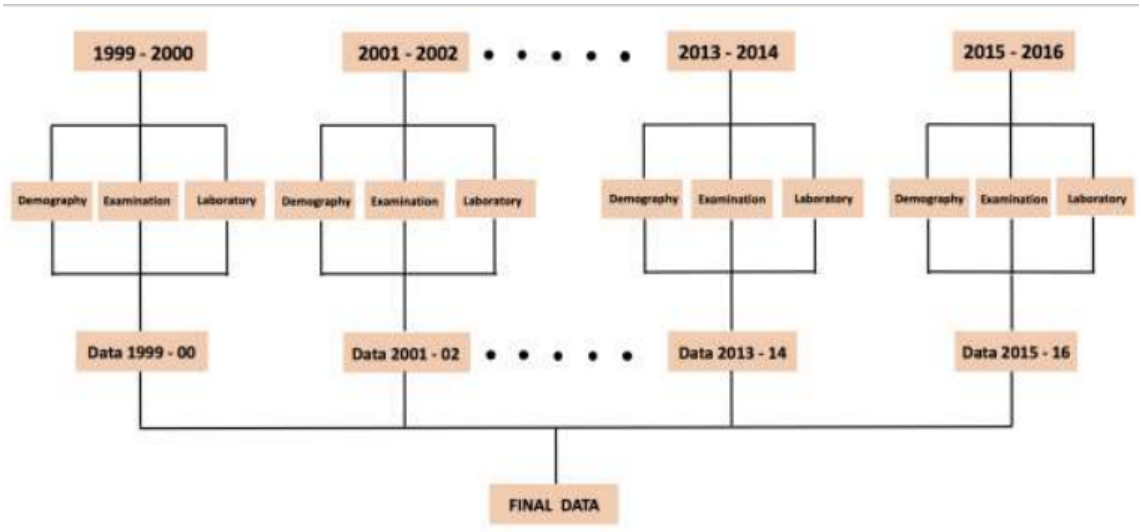

Figure 4.1: Data compilation from National Health and Nutritional Survey (NHANES). The data is acquired from 1999 to 2016 in three categories – Demography, Examination and Laboratory. Based on the nature of the factors that are considered, the dataset contains both the quantitative and the qualitative variables

4.2 Dataset Description

There are a total of 37079 rows and 51 columns in the dataset. Out of the 51 columns, 50 are features and the remaining one is the target column.

#	Column	Non-Null Count	Dtype
0	SEQN	37079 non-null	int64
1	Gender	37079 non-null	int64
2	Age	37079 non-null	int64
3	Annual-Family-Income	37079 non-null	int64
4	Ratio-Family-Income-Poverty	37079 non-null	float64
5	X60-sec-pulse	37079 non-null	int64
6	Systolic	37079 non-null	int64
7	Diastolic	37079 non-null	int64
8	Weight	37079 non-null	float64
9	Height	37079 non-null	float64
10	Body-Mass-Index	37079 non-null	float64
11	White-Blood-Cells	37079 non-null	float64
12	Lymphocyte	37079 non-null	float64
13	Monocyte	37079 non-null	float64
14	Eosinophils	37079 non-null	float64
15	Basophils	37079 non-null	float64
16	Red-Blood-Cells	37079 non-null	float64
17	Hemoglobin	37079 non-null	float64
18	Mean-Cell-Vol	37079 non-null	float64
19	Mean-Cell-Hgb-Conc.	37079 non-null	float64
20	Mean-cell-Hemoglobin	37079 non-null	float64
21	Platelet-count	37079 non-null	float64
22	Mean-Platelet-Vol	37079 non-null	float64
23	Segmented-Neutrophils	37079 non-null	float64
24	Hematocrit	37079 non-null	float64
25	Red-Cell-Distribution-Width	37079 non-null	float64
26	Albumin	37079 non-null	int64
27	ALP	37079 non-null	int64
28	AST	37079 non-null	int64
29	ALT	37079 non-null	int64
30	Cholesterol	37079 non-null	float64
31	Creatinine	37079 non-null	float64
32	Glucose	37079 non-null	float64
33	GGT	37079 non-null	int64
34	Iron	37079 non-null	float64
35	LDH	37079 non-null	int64
36	Phosphorus	37079 non-null	float64
37	Bilirubin	37079 non-null	float64
38	Protein	37079 non-null	float64
39	Uric.Acid	37079 non-null	float64
40	Triglycerides	37079 non-null	float64
41	Total-Cholesterol	37079 non-null	float64
42	HDL	37079 non-null	float64
43	Glycohemoglobin	37079 non-null	float64
44	Vigorous-work	37079 non-null	int64
45	Moderate-work	37079 non-null	int64
46	Health-Insurance	37079 non-null	int64
47	Diabetes	37079 non-null	int64
48	Blood-Rel-Diabetes	37079 non-null	int64
49	Blood-Rel-Stroke	37079 non-null	int64
50	CoronaryHeartDisease	37079 non-null	int64

dtypes: float64(31), int64(20)

Figure 4.2: Names of the features column.

The description of the dataset are given below:

	count	mean	std	min	25%	50%	75%	max
SEQN	37079.0	48901.041236	26753.636441	2.000	26120.500	50065.000000	71173.500	93702.000
Gender	37079.0	1.513282	0.499830	1.000	1.000	2.000000	2.000	2.000
Age	37079.0	48.943661	18.010440	20.000	33.000	48.000000	63.000	85.000
Annual-Family-Income	37079.0	7.358208	3.994083	1.000	4.000	7.000000	10.000	15.000
Ratio-Family-Income-Poverty	37079.0	2.559026	1.624789	0.000	1.140	2.180000	4.130	5.000
X60-sec-pulse	37079.0	72.579250	12.242108	32.000	64.000	72.000000	80.000	224.000
Systolic	37079.0	124.090078	19.254741	0.000	111.000	121.000000	134.000	270.000
Diastolic	37079.0	69.919253	13.575804	0.000	62.000	70.000000	78.000	132.000
Weight	37079.0	80.988276	20.678734	32.300	66.500	78.200000	92.100	371.000
Height	37079.0	167.389601	10.122908	129.700	160.000	167.100000	174.600	204.500
Body-Mass-Index	37079.0	28.824588	6.608982	13.180	24.220	27.800000	32.100	130.210
White-Blood-Cells	37079.0	7.269524	2.478754	1.400	5.700	6.900000	8.400	117.200
Lymphocyte	37079.0	30.225459	8.590412	2.700	24.400	29.700000	35.500	94.500
Monocyte	37079.0	7.915710	2.324364	0.600	6.400	7.700000	9.100	66.900
Eosinophils	37079.0	2.833415	2.116698	0.000	1.500	2.300000	3.500	37.300
Basophils	37079.0	0.700906	0.474369	0.000	0.400	0.600000	0.900	13.900
Red-Blood-Cells	37079.0	4.668632	0.509368	2.260	4.320	4.660000	5.010	8.300
Hemoglobin	37079.0	14.139073	1.541599	5.800	13.100	14.100000	15.200	19.700
Mean-Cell-Vol	37079.0	89.534540	5.745514	50.800	86.700	89.900000	93.000	125.300
Mean-Cell-Hgb-Conc.	37079.0	30.365387	2.336812	14.600	29.300	30.600000	31.700	60.800
Mean-cell-Hemoglobin	37079.0	33.897452	0.932481	27.800	33.300	33.861696	34.500	44.900
Platelet-count	37079.0	253.012886	67.403298	4.000	208.000	246.000000	290.000	1000.000
Mean-Platelet-Vol	37079.0	8.196637	0.923022	4.700	7.600	8.100000	8.800	15.100
Segmented-Neutrophils	37079.0	58.372284	9.581765	0.800	52.400	58.800000	64.800	96.600
Hematocrit	37079.0	41.695175	4.374323	19.700	38.700	41.800000	44.800	59.000
Red-Cell-Distribution-Width	37079.0	13.082278	1.304517	9.700	12.300	12.800000	13.500	37.800
Albumin	37079.0	42.528116	3.585254	19.000	40.000	43.000000	45.000	57.000
ALP	37079.0	70.789611	26.073559	7.000	55.000	67.000000	82.000	729.000

Figure 4.3: Description of the features and target column.

AST	37079.0	25.722511	19.695625	7.000	19.000	23.000000	27.000	1672.000
ALT	37079.0	25.601850	25.889693	4.000	16.000	21.000000	28.000	1997.000
Cholesterol	37079.0	5.077399	1.079629	0.155	4.319	4.991000	5.740	14.611
Creatinine	37079.0	78.632276	39.157384	17.700	61.880	73.370000	88.400	1573.520
Glucose	37079.0	5.595013	2.059786	1.050	4.718	5.110000	5.662	34.250
GGT	37079.0	29.459667	43.576787	3.000	14.000	20.000000	31.000	2274.000
Iron	37079.0	15.266181	6.416872	0.900	10.900	14.500000	18.800	99.800
LDH	37079.0	132.045632	31.961662	4.000	113.000	128.000000	146.000	1539.000
Phosphorus	37079.0	1.203049	0.182223	0.484	1.098	1.195000	1.324	2.648
Bilirubin	37079.0	11.801173	5.276652	0.000	8.550	10.260000	13.680	224.010
Protein	37079.0	72.050158	4.967992	47.000	69.000	72.000000	75.000	113.000
Uric.Acid	37079.0	321.723326	86.129723	23.800	261.700	315.200000	374.700	1070.600
Triglycerides	37079.0	1.695405	1.283654	0.102	0.903	1.344000	2.066	34.559
Total-Cholesterol	37079.0	5.081713	1.072682	1.530	4.320	5.020000	5.740	14.090
HDL	37079.0	1.370344	0.415985	0.160	1.070	1.290000	1.600	5.840
Glycohemoglobin	37079.0	5.676496	1.050223	2.000	5.200	5.400000	5.800	18.800
Vigorous-work	37079.0	1.783840	0.448324	1.000	2.000	2.000000	2.000	3.000
Moderate-work	37079.0	1.598856	0.511199	1.000	1.000	2.000000	2.000	3.000
Health-Insurance	37079.0	1.218587	0.461102	1.000	1.000	1.000000	1.000	9.000
Diabetes	37079.0	1.907333	0.349674	1.000	2.000	2.000000	2.000	3.000
Blood-Rel-Diabetes	37079.0	1.549502	0.497550	1.000	1.000	2.000000	2.000	2.000
Blood-Rel-Stroke	37079.0	1.796165	0.402853	1.000	2.000	2.000000	2.000	2.000
CoronaryHeartDisease	37079.0	0.040670	0.197527	0.000	0.000	0.000000	0.000	1.000

Figure 4.4: Description of the features and target column.

The exhaustive list of variables is: gender, age, annual-family-income, ratio-family-incomepoverty, 60sec pulse rate, systolic, diastolic, weight, height, body mass index, white blood cells, lymphocyte, monocyte, eosinophils, basophils, red blood cells, hemoglobin, mean cell volume, mean concentration of hemoglobin, platelet count, mean volume of platelets, neutrophils, hematocrit, red blood cell width, albumin, alkaline phosphatase (ALP), aspartate aminotransferase (AST), alanine aminotransferase (ALT), cholesterol, creatinine, glucose, gamma-glutamyl transferase (GGT), cholesterol, creatinine, glucose, iron, iron, lactate dehydrogenase (LDH), phosphorus, bilirubin, protein, uric acid, triglycerides, total choles-

terol, high-density lipoprotein (HDL), glycohemoglobin, vigorous-work, moderate-work, health-Insurance, diabetes, blood related diabetes, and blood related stroke. However, in this list of variables, there are a couple of linearly dependent variables in terms of their nature of acquisition or quantification and some uncorrelated variables (annual family income, height, ratio of family income-poverty, 60 sec pulse rate, health insurance, lymphocyte, monocyte, eosinophils, total cholesterol, mean cell volume, mean concentration of hemoglobin, hematocrit, segmented neutrophils). We do not consider these variables for subsequent processing and analysis.

4.3 Statistic of your dataset

A few statistical analysis on the dataset are shown below:

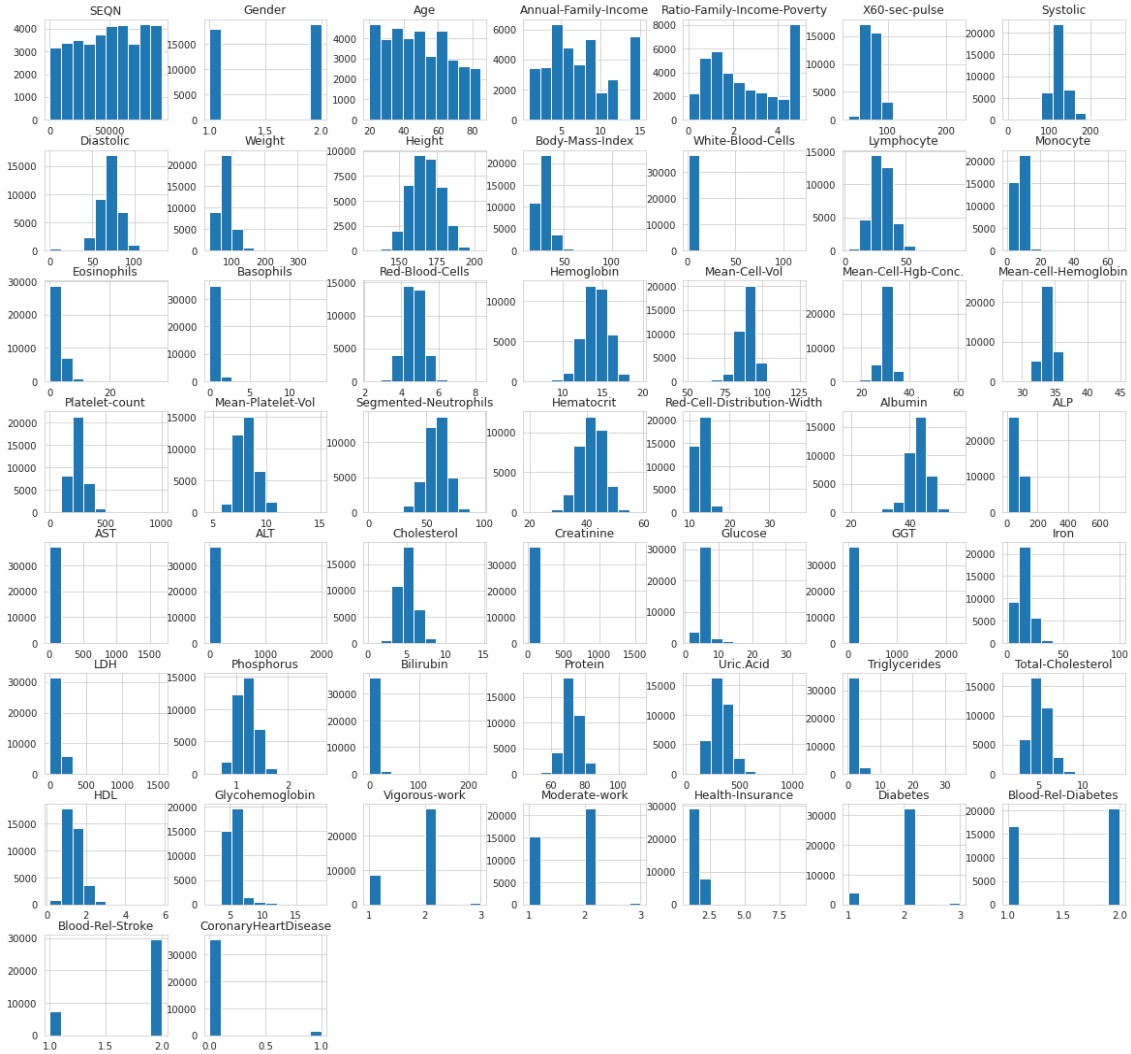

Figure 4.5: Histogram of the dataset.

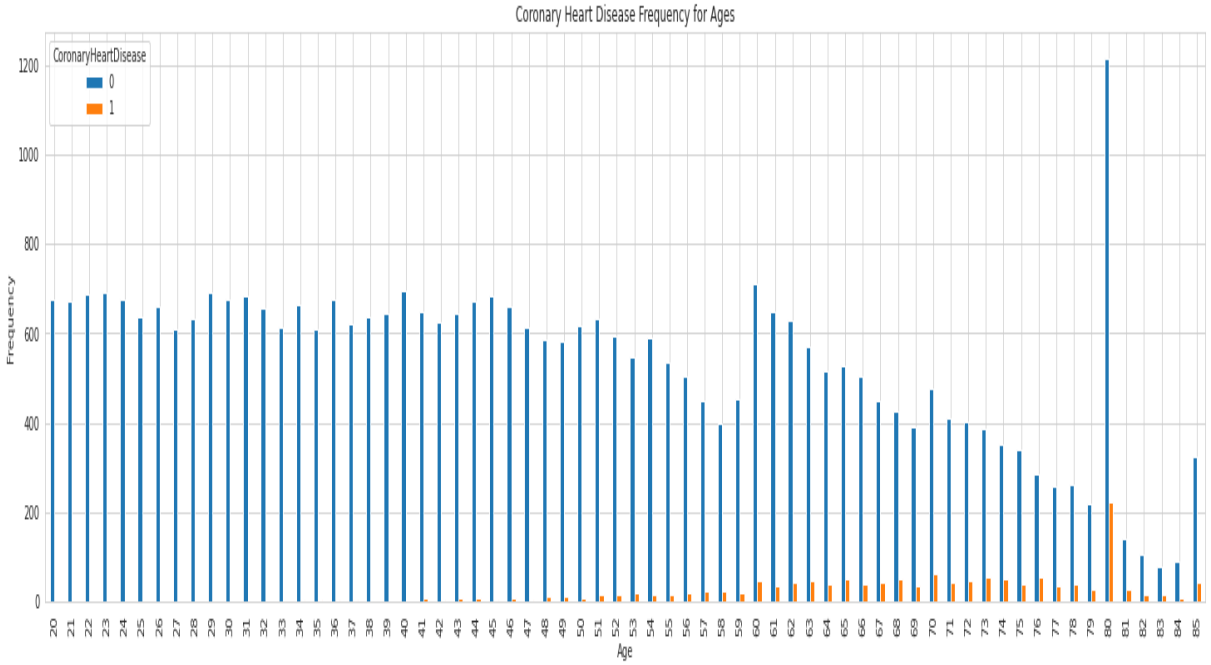

Figure 4.6: Coronary Heart Disease Frequency for Ages.

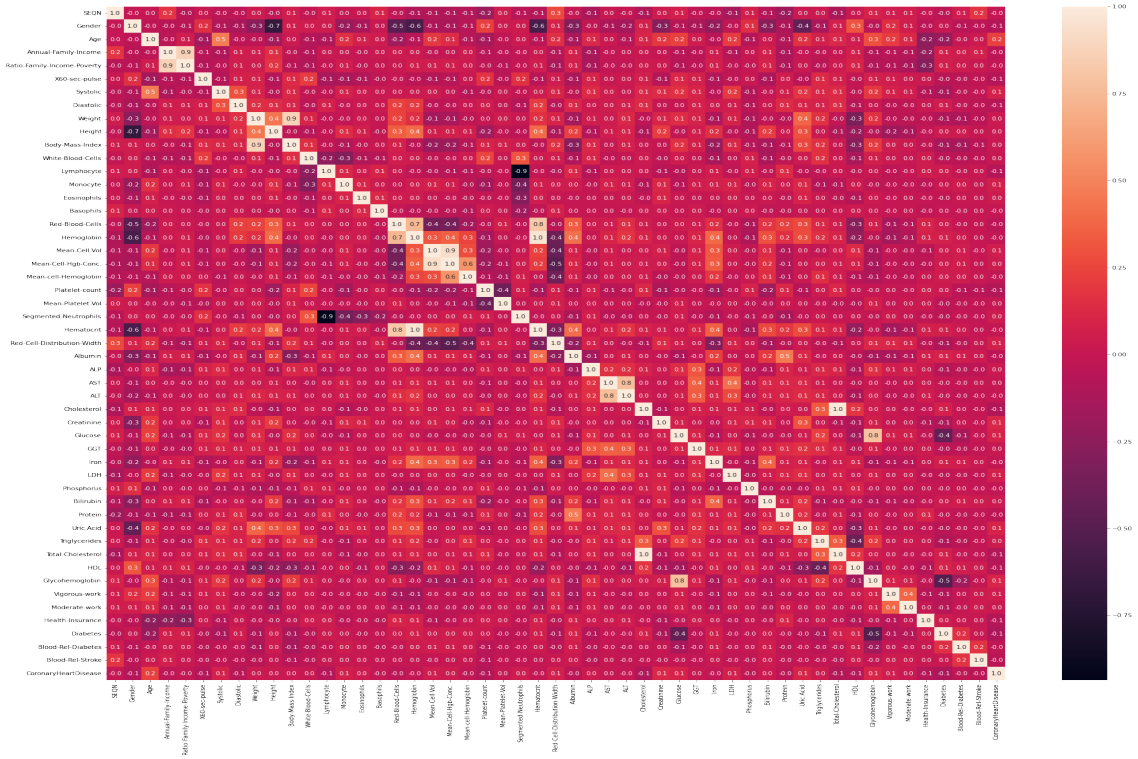

Figure 4.7: Correlation Matrix of all the columns.

Chapter 5

Results/Evaluation

5.1 Performance of your model.

We calculated the performance matrix based on Accuracy, Recall, AUC, Specificity. The result of these performance are provided below:

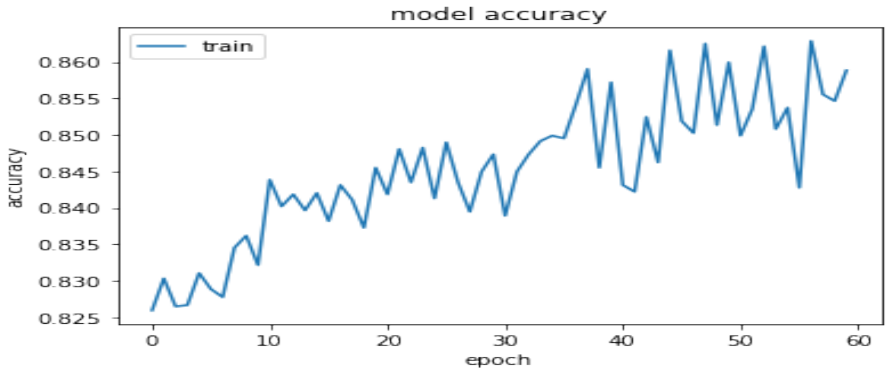

Figure 5.1: Model accuracy per epochs.

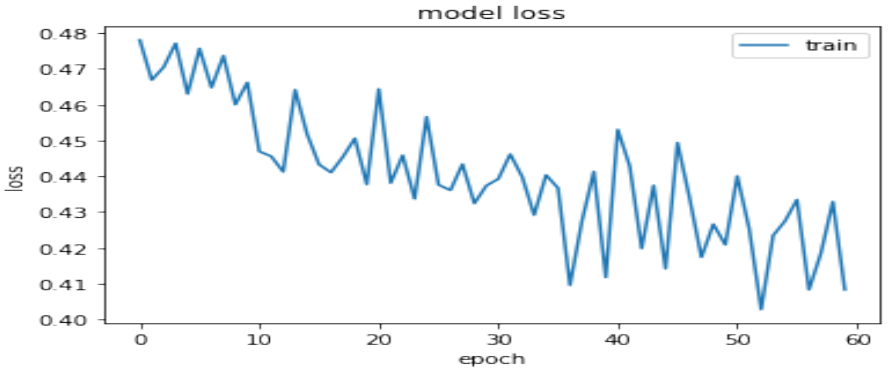

Figure 5.2: Model loss per epochs.

The table below shows the calculated result of our model:

Accuracy	Recall	AUC	Specificity
0.7764631509780884	0.746	0.8562962621271898	0.8047711937577947

Table 5.1: Performance metrics of our proposed model.

5.2 Comparison between our result and actual paper results

Model	Accuracy	Recall	AUC	Specificity
Paper's Model	0.8178	0.773	0.7678	0.818
Our Model	0.7765	0.746	0.855	0.8048

Table 5.2: Comparsion of the two models.

5.3 Discussion about these two results

From the above two comparisons, we see that in case of Accuracy, Recall and Specificity, the paper's model gave a better result but in terms of Area Under the Curve(AUC), our the performance of our model was improved. Some of the results came really close to that of the author's model. The author may have used some other normalization techniques during the data preprocessing phase to achieve a better accuracy and specificity. A highly specific test means that there are few false positive results. Incase of Recall, a high area under the curve represents both high recall and high precision, where high recall relates to a low false negative rate.

Chapter 6

Conclusion And Future Work

The proposed model initiates with the application of LASSO regression in order to identify the contribution of significant variables or attributes in the data variation. Using multiple instances of randomly subsampled datasets, LASSO is performed repeatedly to check the consistency of the variable contribution, which is a crucial step in our algorithm to control the true-negatives of variable selection.A possible future direction of this work is to consider nutrition and dietary data recorded by NHANES as additional predictor variables for CHD prediction.

A possible future direction of this work is to consider nutrition and dietary data recorded by NHANES as additional predictor variables for CHD prediction. Dietary factors play an important role in CHD occurrence (Masironi, 1970, Bhupathiraju 2011) and the prediction accuracy of CHD by including additional dietary variables could be explored. For example, until very recently, several prospective studies concluded that total dietary fat was not significantly associated with CHD mortality (Skeaff 2009, Howard 2006). However, according to American Heart Association (AHA), it is the quality of fat which determines CHD risk (Lichtenstein 2006, USDA 2010). Individual experiments performed with NHANES dietary data have discussed the association of cholesterol, LDL, HDL, amino acids and dietary supplements with CHD (references). However, individual consumption of nutrients takes place collectively in the form of meals consisting of combination of nutrients (Hu 2002, Sacks 1995). This may lead to multi-collinearity among factors and thus a more complex dietary pattern analysis, controlling for multicollinearity of CHD associated significant nutrients could lead to a more comprehensive approach to CHD prevention. Additionally, some of the clinical predictor variables included in the classification model of CHD prediction may themselves be impacted by certain dietary habits of patients. Thus, inclusion of dietary data of patients along with clinical predictor variables, in prediction of CHD, can also lead to potential endogeneity issues. However, with appropriate treatment of endogeneity, dietary data inclusion is expected to provide further insights and improved accuracy of CHD diag-

nosis. Finally, the preferred selection between data augmentation and data subsampling is much debated and demands attention in this section. Our argument in favor of subsampling is as follows: as observed from the t-SNE figures in the result section, the CHD and no-CHD classes are densely interspersed. Moreover, the class-specific clusters are highly non-convex and extremely hard to separate using naïve nonlinear classifiers. Synthetic data samples using strategies, such as a random sampling on the line connecting an arbitrary pair of data samples (used in SMOTE, ADASYN) might receive the wrong label. It is because of the fact that the newly sampled data sample has the likelihood to be labeled as "0" (for training) if the pair of data samples belongs to class "0". However, the data sample may be biologically implausible or, in case of potential plausibility, may actually be a sample from class "1" as both the classes are densely mixed. Especially, when the data is significantly imbalanced, such as in the case of our data, the number of synthesized data samples of the minority class is large. A countable fraction of such newly synthesized, incorrectly labeled data imposes a large bias on the trained network and increases the probability of misclassification. Therefore, we prefer to adopt the sub-sampling strategy, where the authenticity of data is preserved, barring the measurement and acquisition noise. It is an interesting avenue to explore if the extension of shallow CNN models, in terms of architecture and data sub-sampling, to implementation of neural net-based learning on similar clinical datasets, improves the prediction accuracy of the classification process. As explained earlier, our model can also be used as a transfer learning model and the last two dense layers can be retrained for new data. Thus, a significant future research direction would be to implement CNN for predictions from similar clinical datasets where such imbalanced number of positive and negative classifications exist.