Hear Failure: Diagnosis and Severity Estimation through machine learning

Sourav Kumar
Philips Summer Intern
Undergrad, EE, IITD

July 12, 2021

1 Introduction

Hear Failure is a serious medical condition.It prevents the heart from fulfilling the circulatory demands of the body, since it impairs the ability of the ventricle to fill or eject blood. It is characterized by symptoms, such as breathlessness, ankle swelling and fatigue that may be accompanied by signs, for example elevated jugular venous pressure, pulmonary crackles, and peripheral edema, caused by structural and/or functional cardiac or non-cardiac abnormalities. HF is associated with high morbidity and mortality rates. As such we require machine learning models which can accurately diagnose and estimate progress/severity of this condition. This would help the doctor whether experienced or inexperienced to diagnose early with confidence and save lives

1.1 Epidemiology and Medical Diagnosis

Medical diagnosis is supported by several tests-Blood tests, chest radiography, electrocardiography and echocardiography The combination of data from above tests result in the formulation of several criteria that determine the presence of HF.

The Framingham Heart study state establishes the presence of congestive heart failure by the simultaneous presence of at least two of the major criterias or one major and two minor criteria given in table below:

Table 1. Criteria for Congestive Heart Failure*

Major Criteria
Paroxysmal nocturnal dyspnea
Neck vein distension
Rales
Radiographic cardiomegaly (increasing heart size on chest X-ray film)
Acute pulmonary edema
Third sound gallop
Increased central venous pressure (>16 cm water at the right atrium)
Circulation time ≥25 s
Hepatojugular reflux
Pulmonary edema, visceral congestion or cardiomegaly at autopsy
Weight loss ≥4.5 kg in 5 days in response to treatment of CHF
Minor Criteria
Bilateral ankle edema
Nocturnal cough
Dyspnea on ordinary exertion
Hepatomegaly
Pleural effusion
Decrease in vital capacity by 33% from maximal value recorded
Tachycardia (rate ≥120 beats/min)

Figure 1: Criteria for Congestive Heart Failure

2 Previous Works

Most models in the literature for HF detection focus on the use of heart rate variability(HRV) .They make use of classical machine learning methods-SVMs, Logistic regression, Random Forests Previous existing models using SVMs and random Forests already achieve good results on their existing data sets. The Cleveland data set is the one which has been most extensively used by machine learning researchers to experiment on heart failure techniques.

Very few methods use imaging techniques (CNNs) and Sequence models (RNNs). Thus deep learning is little applied in this area.

3 Methodology : Detection

3.1 Experiments with classical machine learning

3.1.1 Dataset

Two publicly available data set were used for applying traditional machine learning approaches.

- Cleveland Heart Disease UCI: It has a total of 303 records with 13 features namely- age sex, chest pain type (4 values), resting blood pressure, serum cholestoral in mg/dl, fasting blood sugar $\leq$ 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise induced angina, oldpeak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy, thal: 3 = normal; 6 = fixed defect; 7 = reversable defect.
- BMC medical informatics dataset: It contains a total of 299 examples. The dataset contains 12 features labelled to predict mortality.

3.1.2 Results

- The train-test split was 80-20
- SVM with different kernels, random forests with varying number of estimators were tried and kNN techniques for different K values were tried

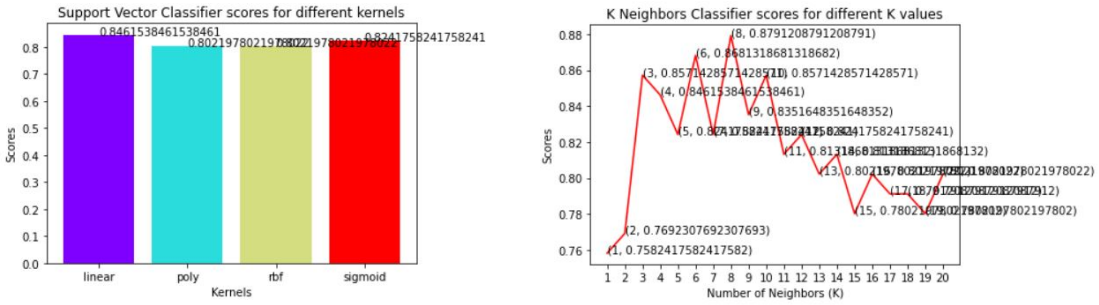

Figure 2: Classical machine learning approaches on cleveland dataset

- The best results on this dataset was observed for KNN with 8 neighbours where an accuracy of 87.9% was achieved on the test set.

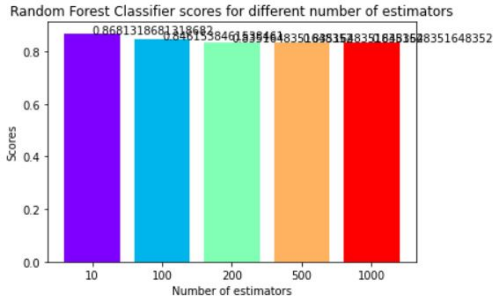

Figure 3: Random forest result on cleveland dataset

- On the second dataset all the above approaches were tried again and it was found that best results are obtained with random forest 100 or more estimators and an accuracy of 87% was achieved. KNN performed poorly so did SVM except for linear kernel.

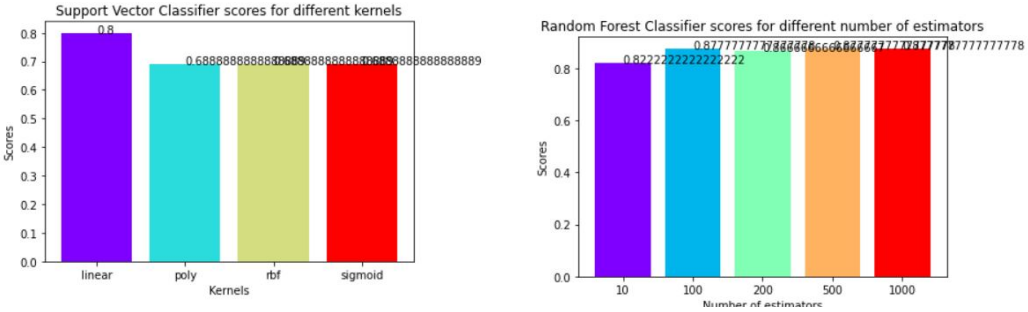

Figure 4: Results of traditional radnom forest and svm on the BMC dataset

3.2 Experiments with longitudinal EHR data; Sequence Modelling

Two approaches were tried with this type of dataset- Classical Ml techniques i.e SVMs, KNN, random forest and Sequence Modelling using **LSTMs**.

3.2.1 Dataset Description

- The data set consists of a cohort of 500 patients (control and cases). The data set is imbalanced and the cases comprise only about 15% of the total number of examples.
- There are total of 59 features belonging to three categories- Diagnosis(ICD9 Codes)(6 features), medications(51 features) and vitals(2 features).
- The source data were from the Geisinger Clinic which is a multispecialty group with a large primary care practice

3.2.2 Feature Construction

- **For SVM,KNN; Non sequential modelling**: Only data points falling within the observation period are used for feature construction. If there multiple entries of a feature then an aggregation technique (averaging, count, boolean) is applied.

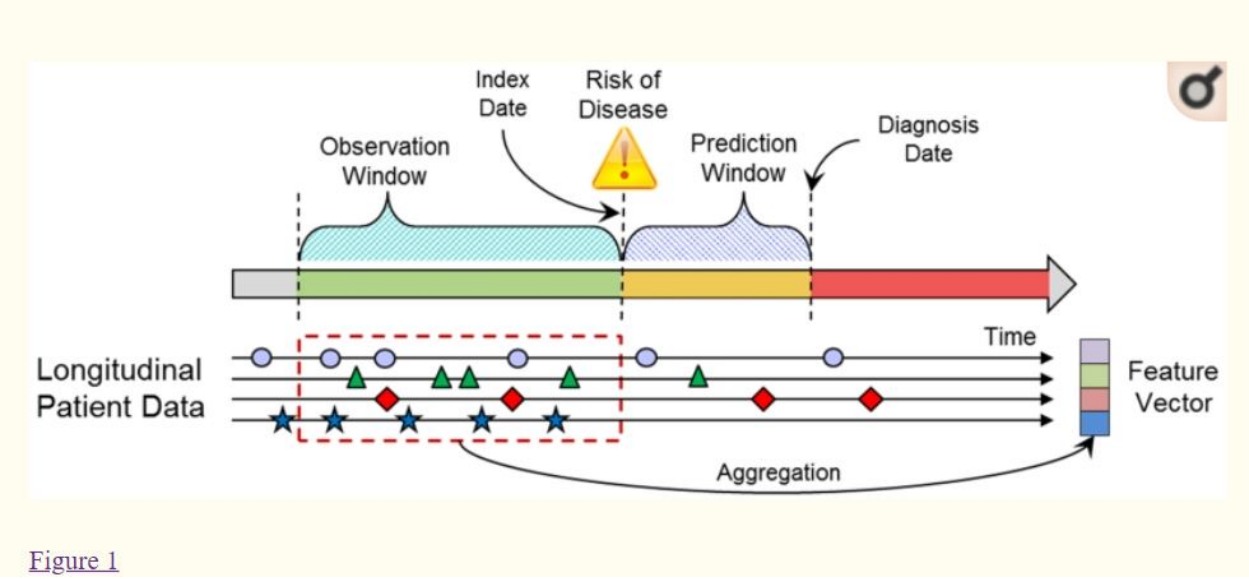

Figure 5: Caption

- **For sequence Modelling-LSTMs** : A sequence was created using the visitations of a patient during the observation period.So for each patient we have a sequence of features.This sequence was later converted into a one hot representation to be fed to an LSTM unit.Padding and masking was used to make each sequence of same length

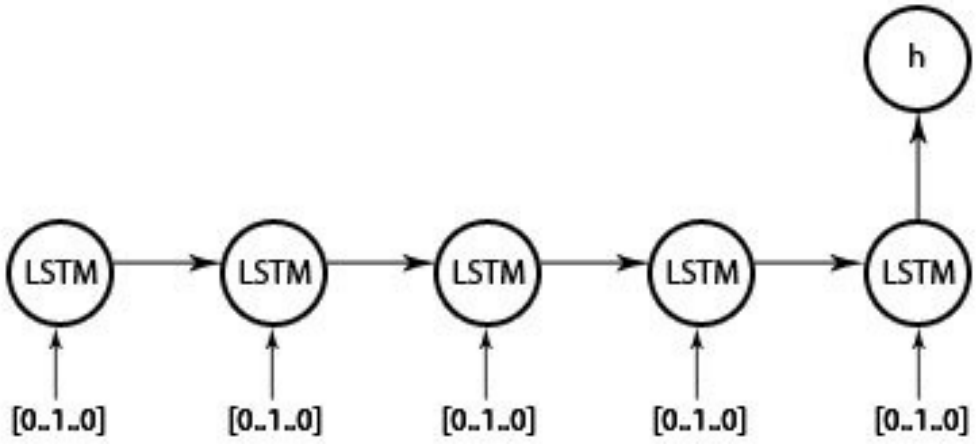

Figure 6: Caption

3.2.3 Results with Non Sequential modelling-SVMs,KNNs

SVM, KNN and random forests all gave an accuracy of 92 % on the test set.

Dataset is imbalanced

They had poor AUC score.So SVM, knn or random forests do not capture the class imbalance

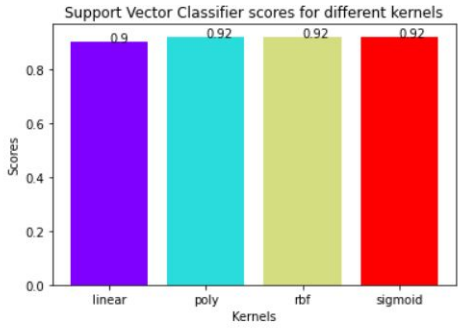

Figure 7: Results with Svm on EHR data

3.2.4 Results with LSTMs

There are three types of results in this category depending upon the choice of features:

- **Using Only Diagnosis features:**

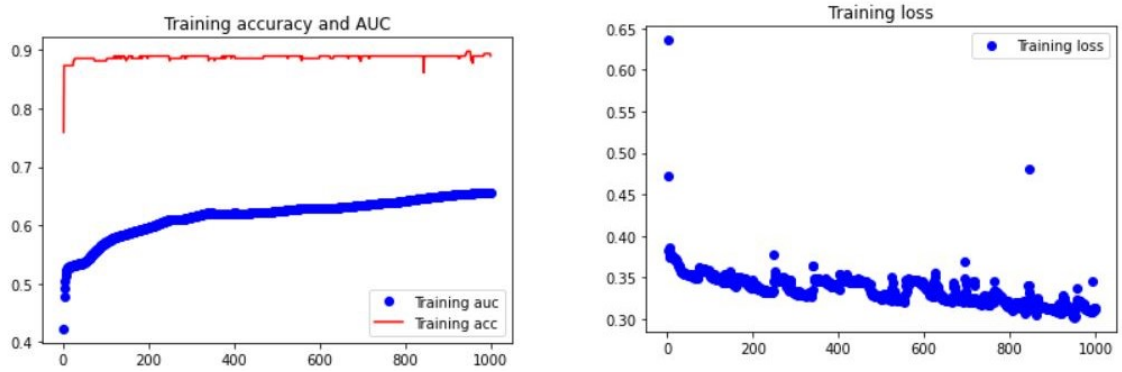

Figure 8: Training curves when only using diagnosis features

- After 1000 epochs the AUC achieved is 73% for both training and test set Accuracy achieved is around 90%
- **Using Only Medication features:**

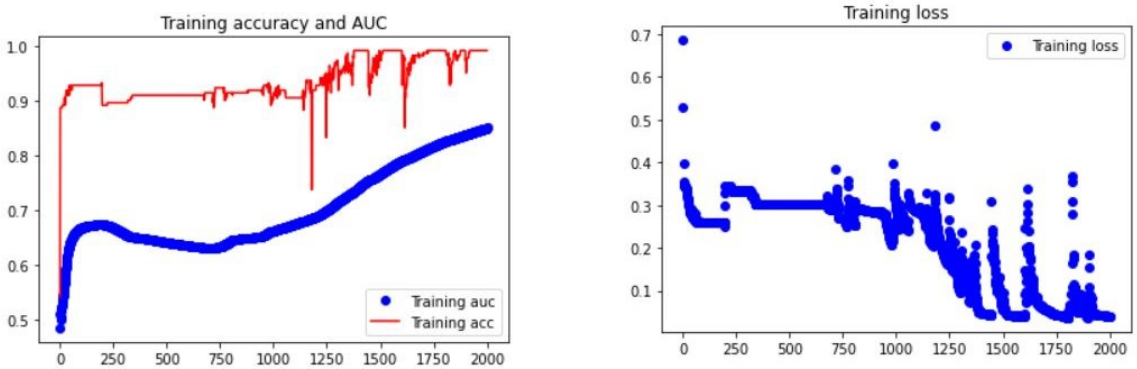

Figure 9: Training curves using only medication data

- There are a total of 51 medications in the dataset.Just training with medications features gave better results then diagnosis AUC achieved was around 85%

- **Using only Vitals as features**

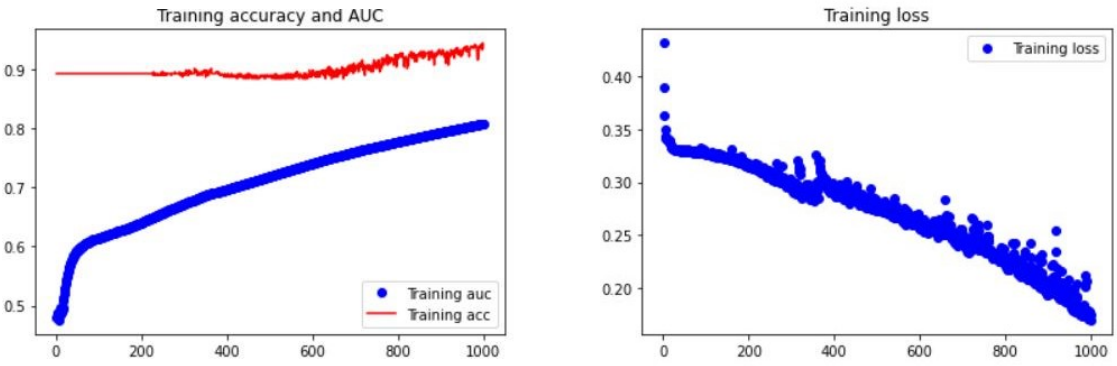

Figure 10: Training curves with only vitals as features

- There are two vitals features-diastolic and systolic pressure The results in terms of AUC were not as good as when training with medication data AUC achieved was 80% Training accuracy was 95%

- **Combining all Features**

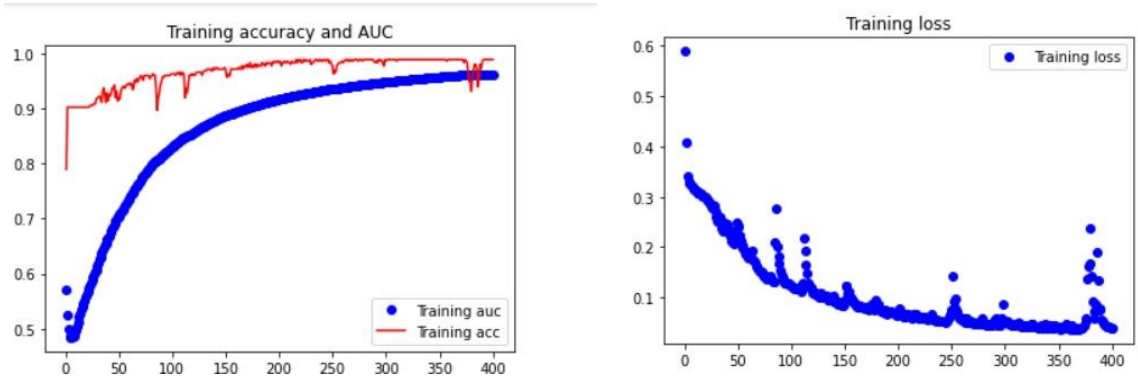

Figure 11: Training curves when combining all features

- When combining all features we get a total of 59 features spanning over a maximum of 130 timesteps for some patients It was observed that AUC growth was faster After only 400 epochs the data was overfit- 95% auc

4 Severity Estimation of HF

The most commonly employed classifications for HF severity are NYHA and ACC/AHA stages of HF They both classify patients into four categories Here we have dealt with predicting the NYHA class.

Class	Patient Symptoms
I	No limitation of physical activity. Ordinary physical activity does not cause undue fatigue, palpitation, dyspnea (shortness of breath).
II	Slight limitation of physical activity. Comfortable at rest. Ordinary physical activity results in fatigue, palpitation, dyspnea (shortness of breath).
III	Marked limitation of physical activity. Comfortable at rest. Less than ordinary activity causes fatigue, palpitation, or dyspnea.
IV	Unable to carry on any physical activity without discomfort. Symptoms of heart failure at rest. If any physical activity is undertaken, discomfort increases.

Figure 12: NYHA HF severity classification

Stage A: Patients at risk for heart failure who have not yet developed structural heart changes (i.e. those with diabetes, those with coronary disease without prior infarct)

Stage B: Patients with structural heart disease (i.e. reduced ejection fraction, left ventricular hypertrophy, chamber enlargement) who have not yet developed symptoms of heart failure

Stage C: Patients who have developed clinical heart failure

Stage D: Patients with refractory heart failure requiring advanced intervention (i.e. biventricular pacemakers, left ventricular assist device, transplantation)

Figure 13: ACA/AHA HF severity Classification

4.1 Dataset Description

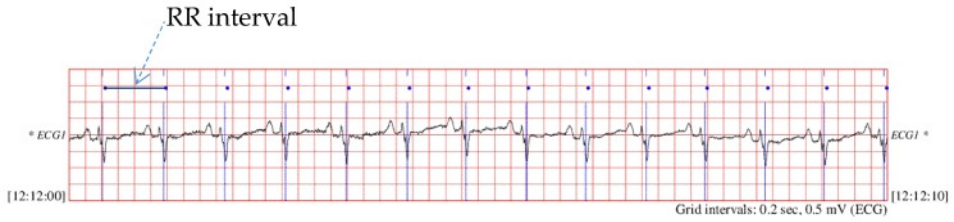

Figure 14: ECG signal and beat annotation

The data set contains ECG samples.Five open-source databases were used. For CHF patients, the BIDMC congestive heart failure database (BIDMC-CHF) and congestive heart failure RR interval database (CHF-RR), available on PhysioBank, were used. The BIDMC-CHF dataset has 15 subjects (11 men, aged 22–71 years, and four women, aged 54–63) with severe CHF (NYHA class 3–4), and the CHF-RRI dataset includes 29 recordings of subjects aged 34–79 with CHF (NYHA class 1–3). For normal subjects, the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) normal sinus rhythm (NSR) , the Fantasia database (FD) and the normal sinus rhythm RR interval database (NSR-RR) were used.

The dataset was divided into a total of 4 classes (0-Normal,1-HF prone,2-moderate,3-Severe)

Database	Total Segments		
	N = 500	N = 1000	N = 2000
BIDMC congestive heart failure database (CHF)	3214	1607	803
Congestive heart failure RR interval database (CHF)	6622	3311	1655
MIT-BIH normal sinus rhythm database (NSR)	3579	1739	869
Normal sinus rhythm RR interval database (NSR)	11,583	5791	2895
Fantasia dataset (NSR)	500	250	125

Figure 15: Dataset sources

4.2 Feature Construction

The rr interval of the ecg signal were used as features. The entire sequence was split into segments of 500 length each. This was then used for sequence modelling.

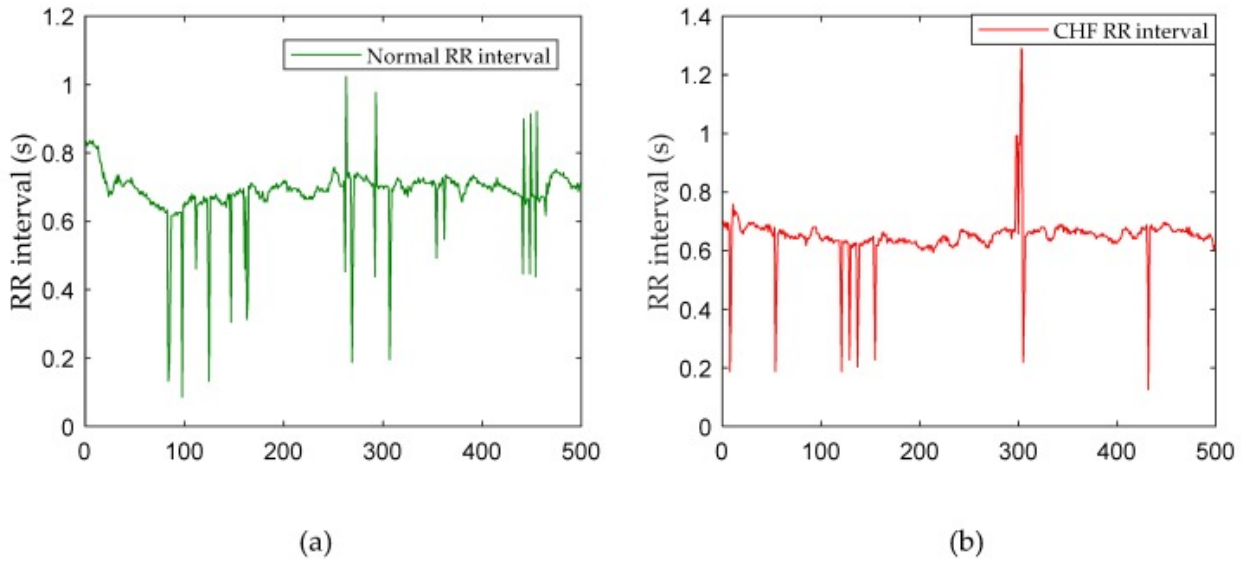

Figure 16: RR interval from ECG signals

4.3 Model Description

LSTM-Based Deep Convolutional Neural Network Structure:A combination of BIdirectional LSTM and Inception from Google's Inception Net was developed. Two such units were used. Dropout regularization with value 0.4 was used after flattening the layers and Activation function used was softmax. The loss function used was categorical cross entorpy. Adam optimizer was used.

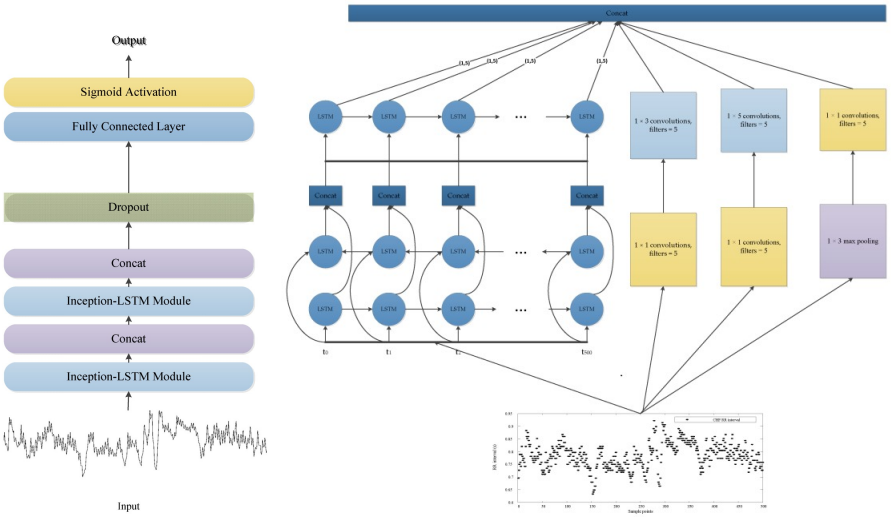

Figure 17: Network Structure(left) and LSTM Inception Module(right)

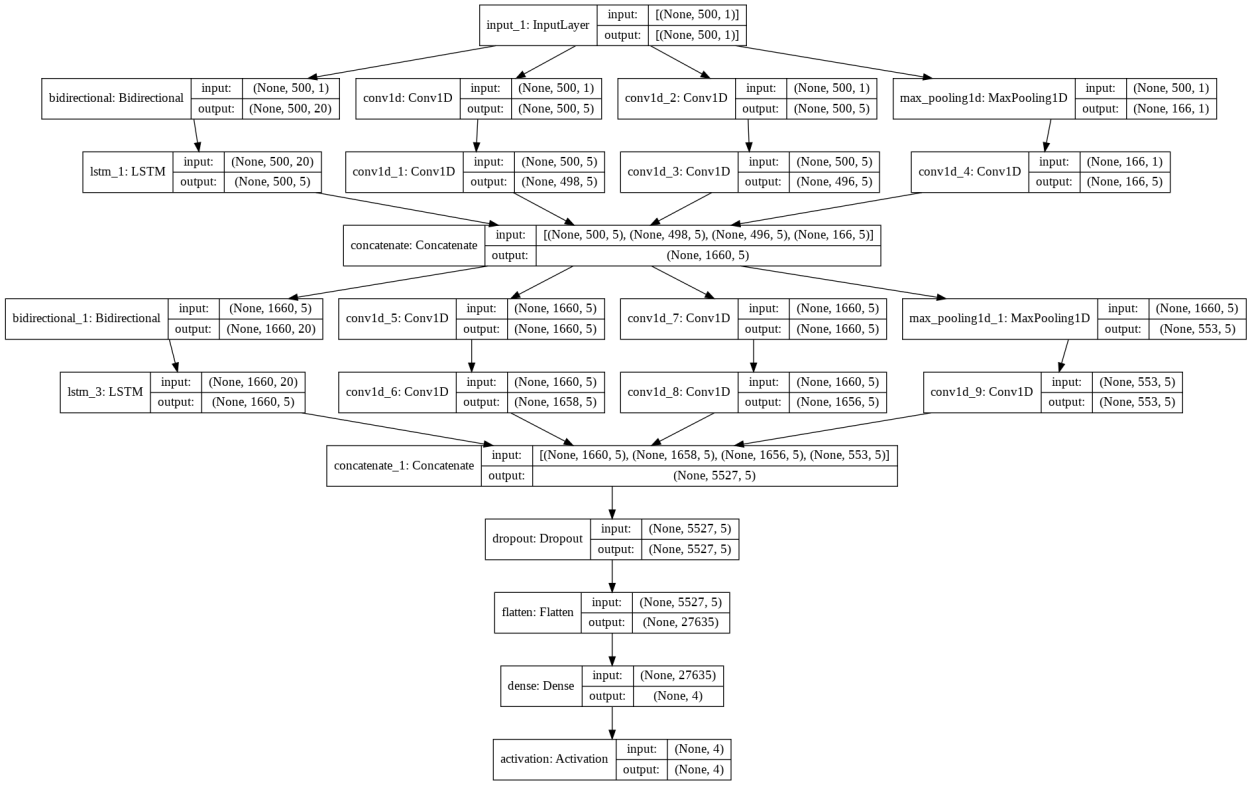

Figure 18: Complete model description

4.4 Results

- After 100 epochs the training accuracy achieved was 93% and validation accuracy was 78%
- Training and val AUC achieved was equal to 95%

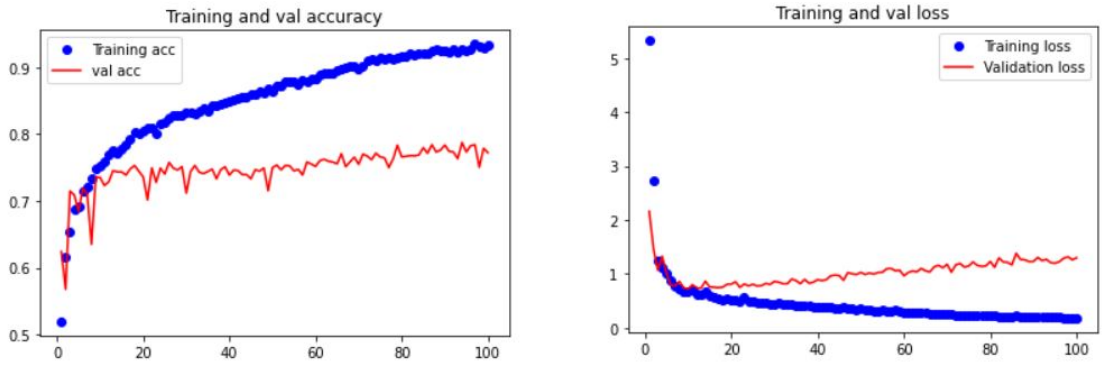

Figure 19: Training curves

5 References

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6480269/#B22-sensors-19-01502
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5341145/
- https://pubmed.ncbi.nlm.nih.gov/31610714/
- https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f
- https://www.sciencedirect.com/science/article/pii/S2001037016300460