CS236781: Deep Learning on Computational Accelerators

Detect Atrial Fibrillation using long short-term memory networks (LSTM) with RR interval signals

Nadav Orzech[1]

Roni Englender[2]

ID: 311549455, nadav.or@campus.technion.ac.il [1]
ID: 312168354, roni.en@campus.technion.ac.il [2]

1. Abstract

Atrial fibrillation (AF), permanent or intermittent (paroxysmal AF) is a heart arrhythmia which affects a large population worldwide. Correct diagnosis of this heart condition is essential to enable treatment and prevent strokes. In this work we want to explore machine learning methods to identify and classify this arrhythmia. This report is based on a study that aims to detect Atrial fibrillation using long short-term memory network (LSTM) with RR interval signals, which was trained and tested with data from the MIT-BIH Atrial fibrillation Database. We aim to improve the studies model by adding a soft attention layer which will focus on relevant features of the LSTM output. Our system achieved 83.4% accuracy with the baseline model and 81.3% accuracy with the modified model.

2. Intro

2.1 Problem Domain

Atrial fibrillation (AF) is one of the most common types of arrhythmias, which are irregular heart rhythms [1]. AF occurs when the upper chambers of the heart (atria) beat out of rhythm and as a result, blood is not pumped efficiently to the rest of the body, causing an unusually fast heart rate, quivering, or thumping sensations in the heart [2]. Often episodes of AF are asymptomatic [3]. AF is the most common sustained cardiac arrhythmia and as of 2020, 33 million people are affected by this disease worldwide [4]. AF patients are at moderate-to-high risk of stroke and the disease is a common factor of heart failure [5]. As such, establishing an effective monitoring system for early AF detection along with an effective approach to treating AF is essential [5].

AF is often transient or paroxysmal in nature, and the correct diagnosis of AF can be challenging in patients with paroxysmal AF [6]. The main characteristic of AF disorder is the irregular rhythm of the heartbeat or more specifically when a varying period is observed in Electrocardiogram (ECG) signal between R–R peaks [7]. The disease is hard to diagnose, since patients suffering from AF may not have symptoms at early onset, and there is spontaneous termination of arrhythmia. Thus using machine learning to detect AF can be very beneficial.

Heart disease prediction using machine learning has become common in the last few decades. There are numerous studies using deep learning techniques to detect heart arrhythmias in general and AF in particular. Machine learning algorithms have the potential to improve patient outcomes and reduce the workload of clinicians particularly where diagnoses are made from large volumes or complex patterns of data such as in AF.

2.2 Existing approaches and drawbacks

We based our project on a study that aims to detect Atrial Fibrillation using long short-term memory network (LSTM) with RR interval signals [8]. Their proposed Computer-Aided Diagnoses (CAD) system can be used for long-term monitoring of the human heart. The system achieved 98.5% accuracy with 10-fold cross-validation (20 subjects) and 99.77% accuracy with blindfold validation (3 subjects).

We plan to address and improve the proposed system's architecture by modifying the model's layers and adding an attention layer to our system (further explained in section 3.2)

3. Methods

This section introduces the approach used by the paper we rely on [8] and our modifications and improvements.

3.1 Original Approach

The model implemented in the study [8] is based on Recurrent Neural Network (RNN). To identify RR intervals as AF it is necessary to examine each RR interval in relation to other intervals over time. So, in order to classify AF it is crucial to be able to put an RR interval in its context. RNNs allow the network to retain and utilize state information, meaning information on what has happened in previous time steps. The RNNs have a "memory" that captures information about all elements of the same input sequence.

LSTM is an improvement on a standard RNN since it incorporates a gating mechanism and is able to deal with the vanishing gradient problem that can be encountered when training a traditional RNN. LSTM has the ability to control which information is remembered and which is forgotten. The model from the study [8] used a Bidirectional LSTM, which utilizes past and future data from the input sequence. This enables the network to make a more accurate prediction because it is given a wider context.

A global max pooling in one dimension was used in the study's [8] model, after the bidirectional LSTM layers. The goal of a max pooling layer is to down sample the input representation by reducing its dimensionality and allowing for assumptions to be made according to the features with the max values.

3.2 Modifications and Improvements

The main modification that was made was changing the one-dimension max pooling layer to an attention layer. The attention mechanism is widely used in deep learning networks, in fields such as Natural Language Processing (NLP) and Computer Vision, and is inspired by the ability of the human brain to direct our focus and pay greater attention to certain factors when processing data.

Since paroxysmal AF occurs as sporadic periods of AF between normal heart rhythms, a soft attention mechanism can put more emphasis (attention) on beats with higher prevalence of AF. Our proposed modification is changing the max pooling layer which chooses the max values, and replacing it with a soft attention layer that its output will be a combination from all the values, with bigger emphasis on relevant values, and not only the maximum.

3.3 Data used

The experiments conducted by the study [8] and by us were based on data from the MIT-BIH Atrial Fibrillation Database [9][10]. The database includes 23 long-term ECG recordings of human subjects with atrial fibrillation (mostly paroxysmal). The individual recordings are each 10 hours in duration and contain two ECG signals each sampled at 250 samples per second. The R peaks are labeled and the RR intervals were extracted according to these labels.

4. Implementation and Experiments

4.1 Data Preprocessing

The data preprocess is based on the preprocessing methods described in the study [8]. The relevant data from the 23 ECG recording files is represented in the p_signals. The p_signals are read and split into RR intervals according to the R peaks from the file record. Each interval is labeled as negative or positive for AF according to the ".atr" file. Then samples are created by splitting RR interval sequences into overlapping sequences of 100 beats for each HR trace, where there are 99 beats overlapping with the adjacent sample. A beat sequenced is labeled as AF if it contains one or more beats that were classified as showing signs of atrial fibrillation. The other sequences are labelled as normal, i.e. negative.

Since we encountered memory limit problems when trying to save all the samples created into one dataset, data processed from each file is saved into a subset dataset and then the final dataset is created by concating all the subset datasets. Also, we created sequences of 100 beats with overlap of 95 beats instead of 99 beats which caused memory allocation errors.

4.2 Model Architecture

The system architecture we used is based on the study [8] with some improvements made by us. The details of the architecture of the study are shown in Figure 1. According to the study's approach the sequences of 100 beats are sampled into the bidirectional LSTM. The number of cells in each of the layers of the bidirectional LSTM was set to twice the input sequence length. This section of the model architecture is implemented with the same structure as the study's.

In our model architecture an attention layer is used between the bidirectional LSTM layers and a fully connected layer as opposed to a max pooling layer that is used in the study's architecture. Both techniques, the attention and max pooling, reduce the dimension of the sequence size from 100 to 1.

The output of this layer is then inserted into two fully connected layers that are used to produce the final classification. These fully connected layers are implemented with the same structure as the original study. The model was implemented using PyTorch.

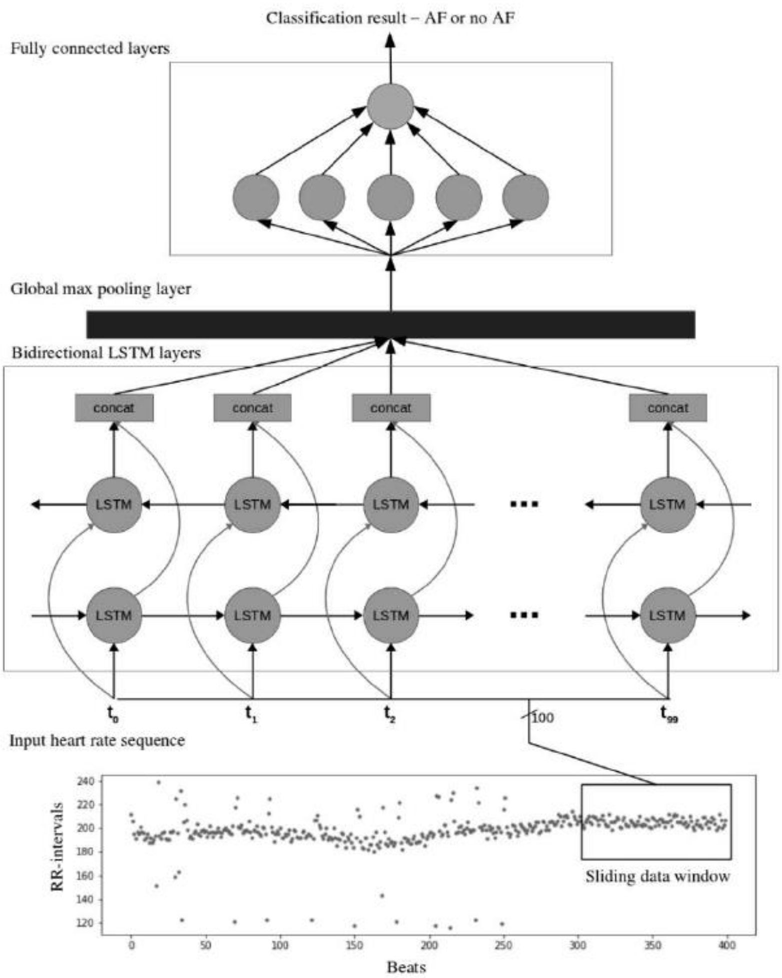

Figure 1: The bidirectional LSTM architecture used in the paper [8]

Layer Type	**Output Shape**
Bidirectional LSTM	100, 400
Attention	400
Fully connected Rectified Linear Unit (ReLU)	50
Dropout	50
Fully connected (Sigmoid)	2

Table 1: Modified Bidirectional LSTM model architecture

4.3 Experiments and configuration

All the experiments we conducted were given the same following parameters:

- Batch size: 1024
- Hidden dim size: 200
- Bidirectional LSTM dropout: 0.1
- Dropout: 0.1
- Learning rate: 5e-4
- Maximum number of epochs: 40
- Early stopping: 8

The loss function used in the training was Cross Entropy loss, and the Optimizer used was Adam optimizer.

The data consists of 23 ECG recording files, with every file representing a different subject. 20 of the recording files were used for the training set, and the remaining 3 files were used for the test set. This ensures that the evaluation is done not only to unseen data but also to unseen patients as well.

The experiments we conducted were all with the same parameters as explained above:

1. Baseline Model with overlap between sequences of 0 beats
2. Improved Model (with attention) with overlap between sequences of 0 beats
3. Baseline Model with overlap between sequences of 33 beats
4. Improved Model (with attention) with overlap between sequences of 33 beats
5. Baseline Model with overlap between sequences of 95 beats
6. Improved Model (with attention) with overlap between sequences of 95 beats

We evaluated the results based on the loss rate, total accuracy, and accuracy of the positive and negative samples.

5. Results and discussion

5.1 Results

From now on we will refer to our improved model with the attention layer as "Attention model".
The results of the experiments conducted can be seen in the following table:

Model	TP	TN	FP	FN	Pos Acc	Neg Acc	Accuracy
Baseline 0	860	347	61	316	93.3%	52.4%	82.6%
Attention 0	831	457	90	206	90.2%	**69%**	80.9%
Baseline 33	1328	653	45	337	96.7%	65.9%	**83.4%**
Attention 33	1354	556	19	434	**98.6%**	56.2%	81.3%
Baseline 95	9131	8147	4936	4410	67.4%	62.3%	64.9%
Attention 95	5462	12047	1036	8078	40.3%	92.1%	65.8%

Table 2: Results table (Model number represents overlap size)

Experiments with Models with overlap size = 0:

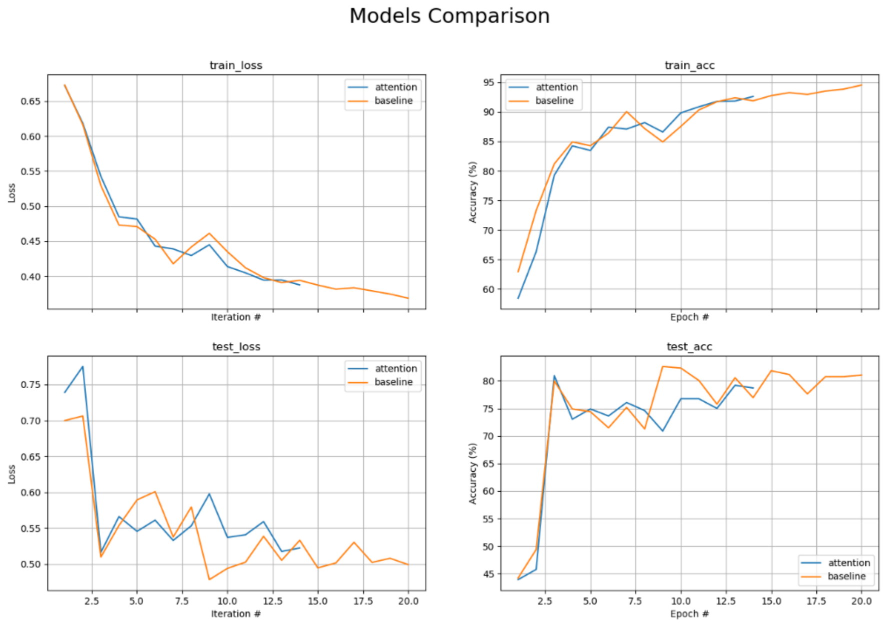

Figure 2: Comparison between Baseline and Attention Models with overlap 0

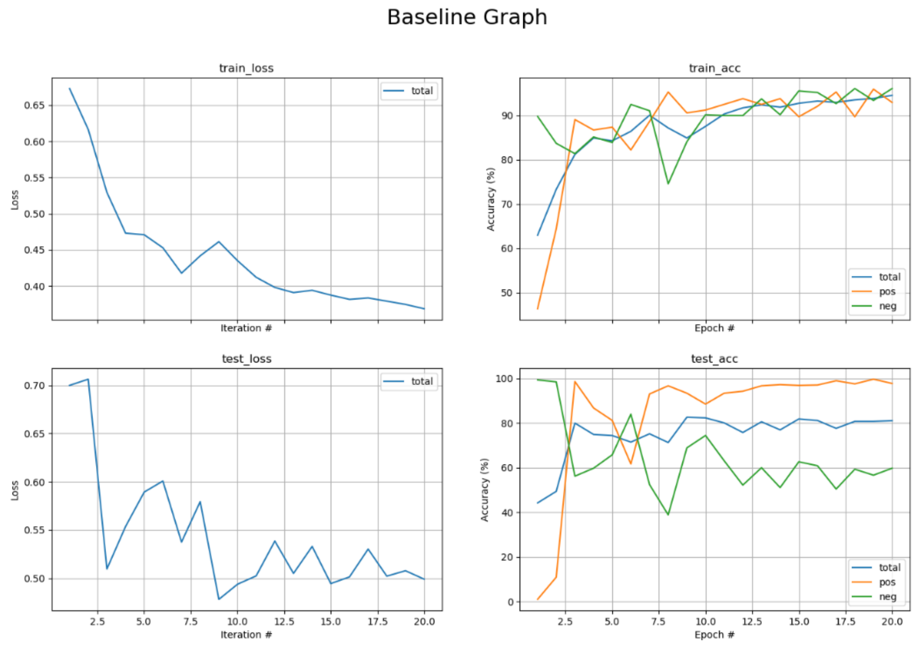

Figure 3: Results of Baseline Model with overlap 0

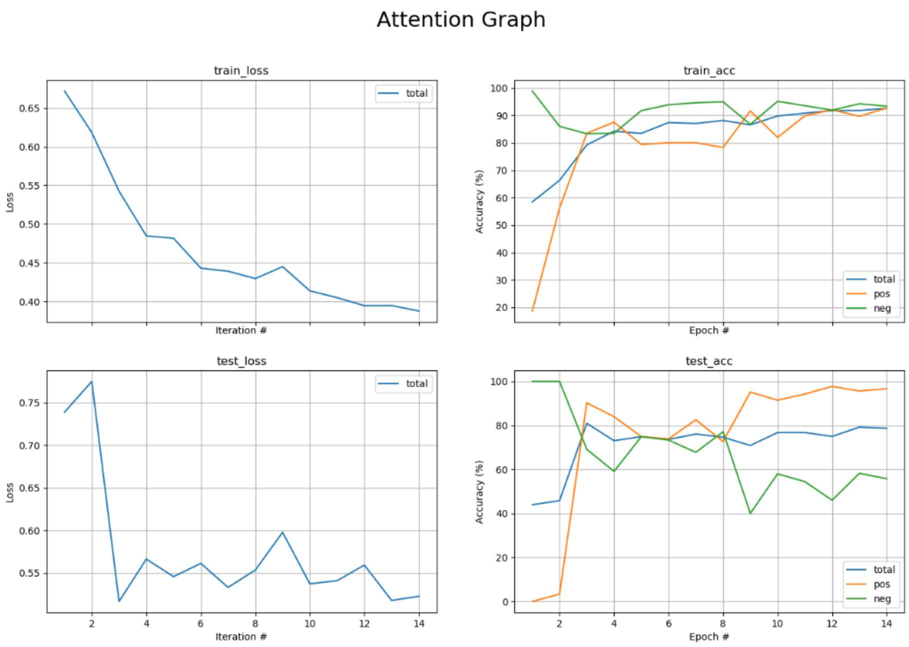

Figure 4: Results of Attention Model with overlap 0

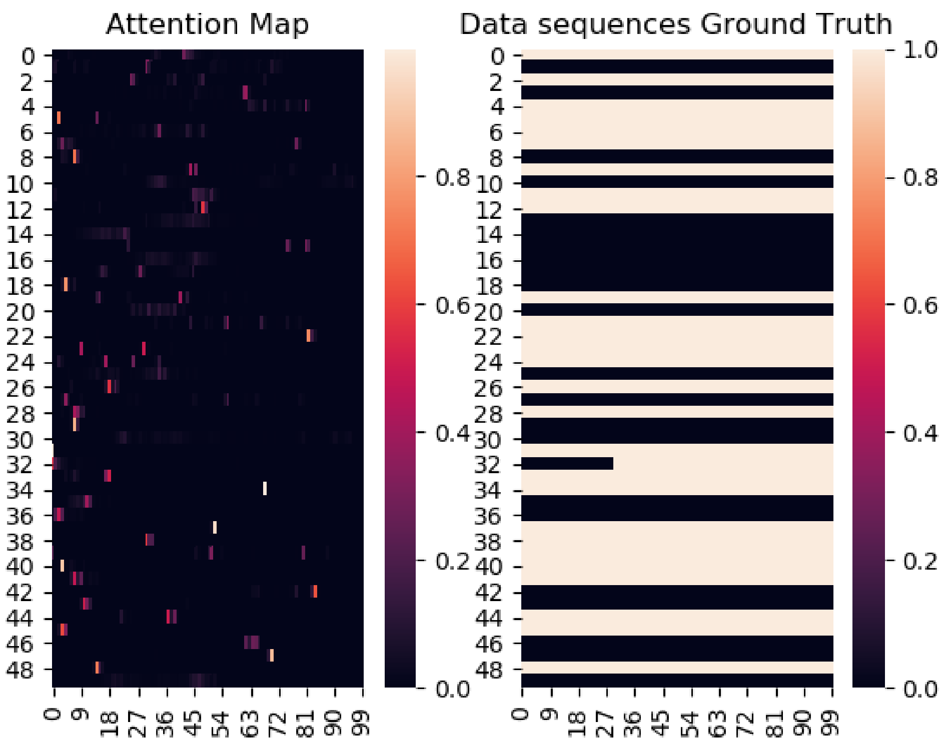

Figure 5: Attention map and ground truth sequences from Attention Model with overlap 0

Experiments with Models with overlap size = 33:

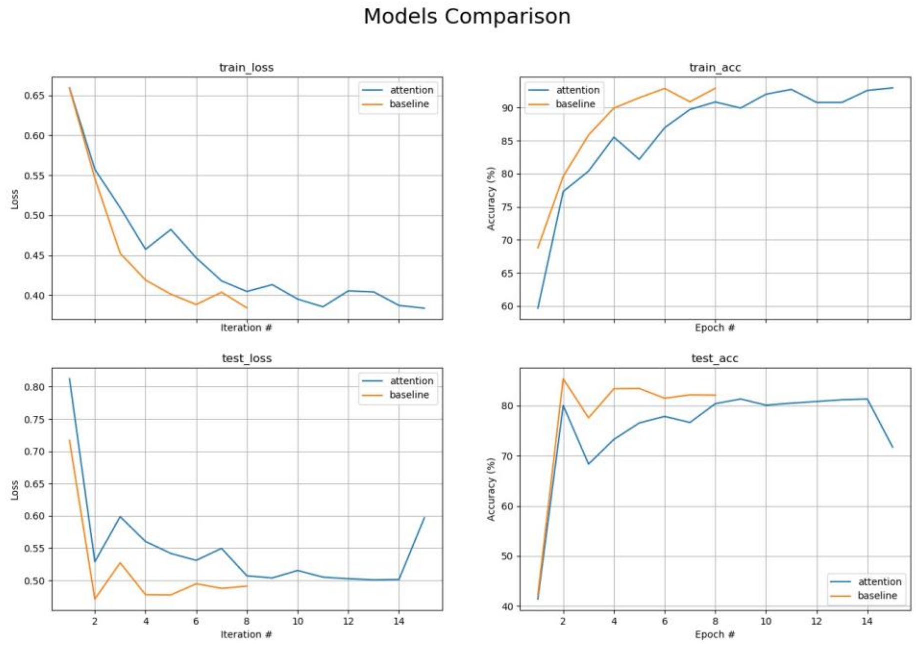

Figure 6: Comparison between Baseline and Attention Models with overlap 0

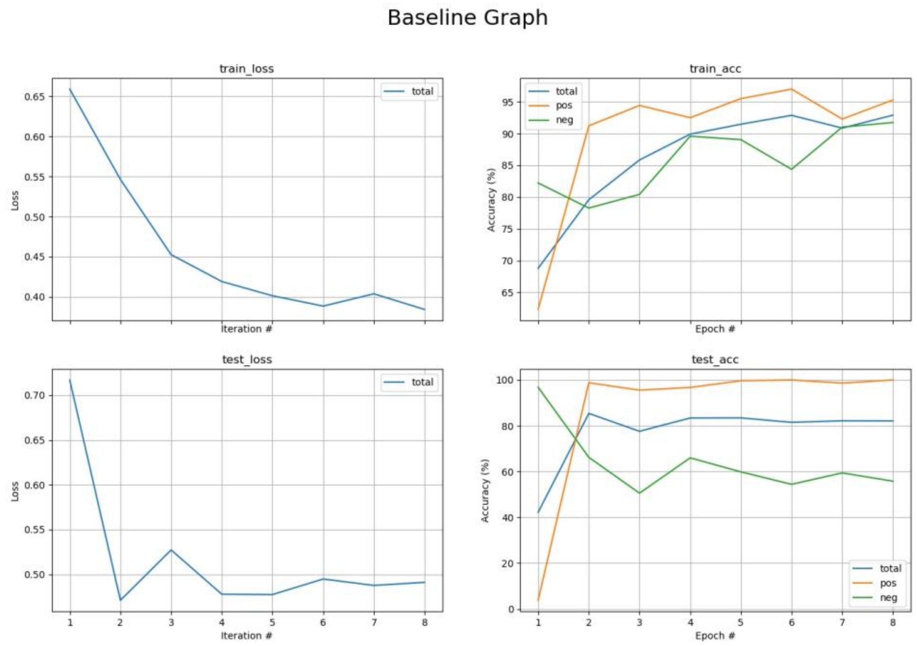

Figure 7: Results of Baseline Model with overlap 33

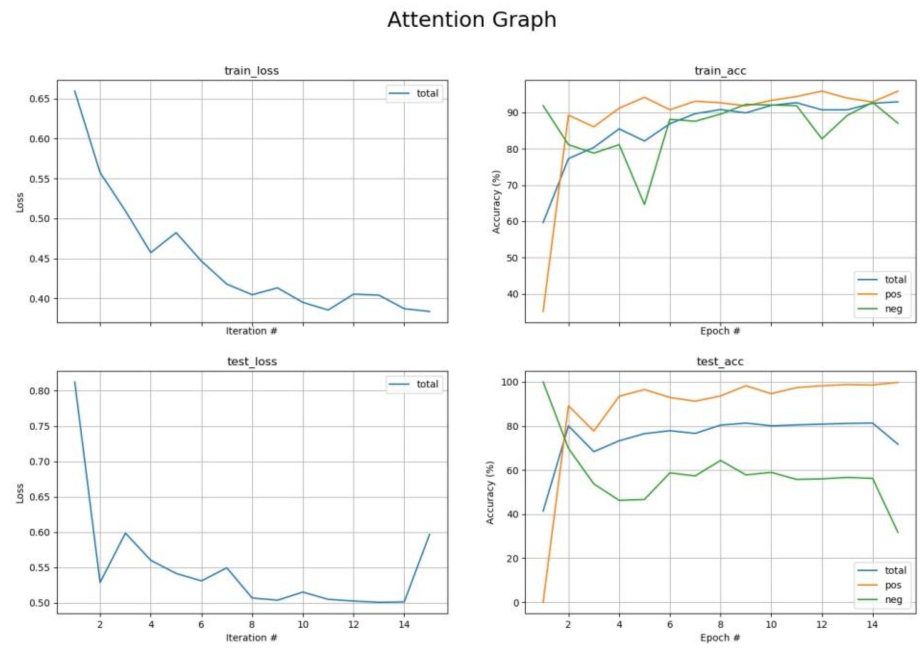

Figure 8: Results of Attention Model with overlap 33

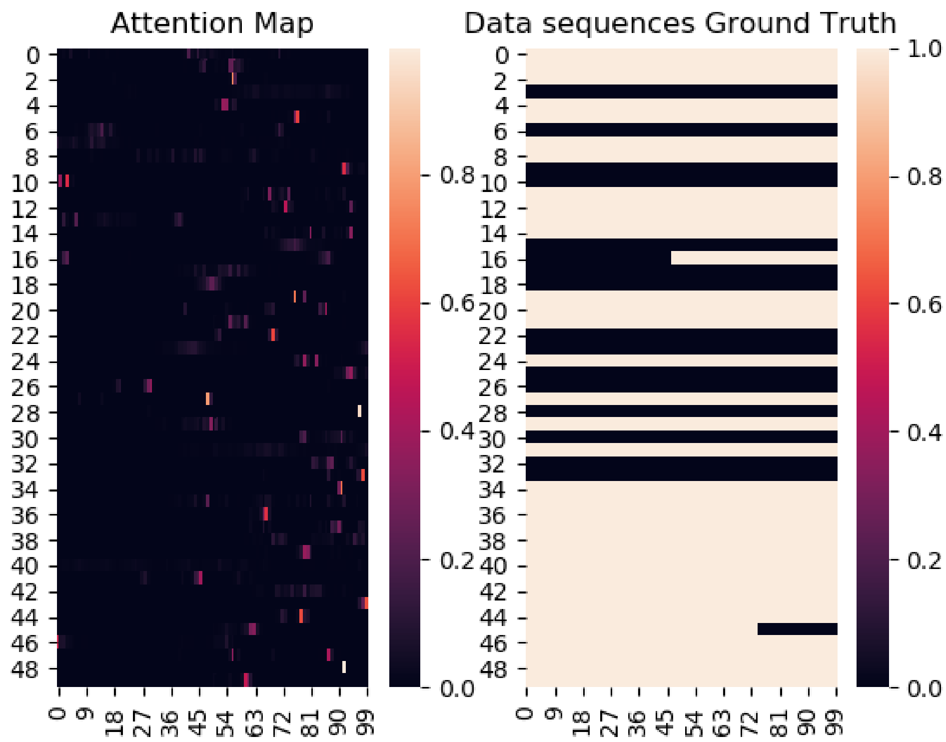

Figure 9: Attention map and ground truth sequences from Attention Model with overlap 33

Experiments with Models with overlap size = 95:

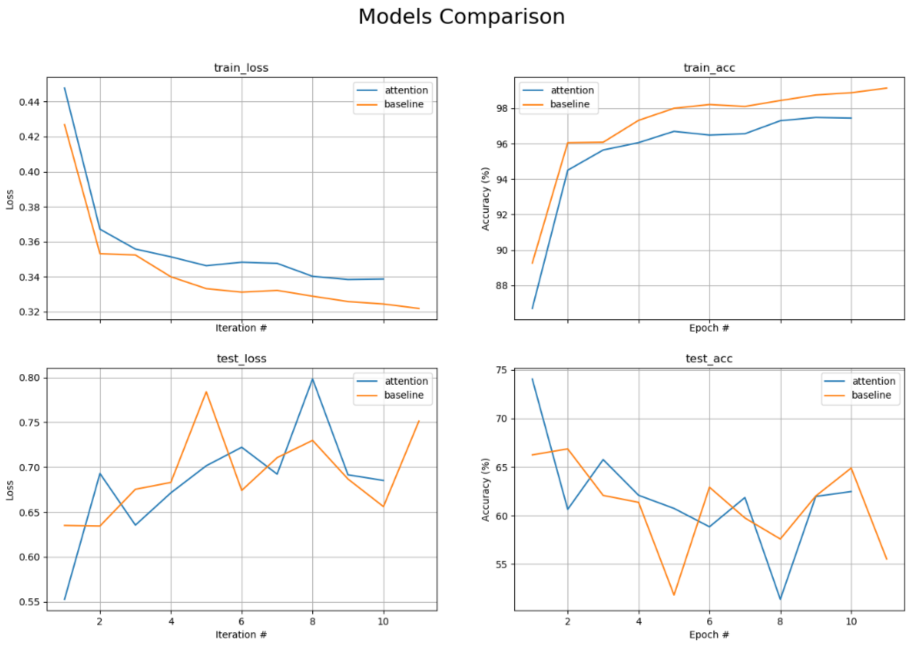

Figure 10: Comparison between Baseline and Attention Models with overlap 95

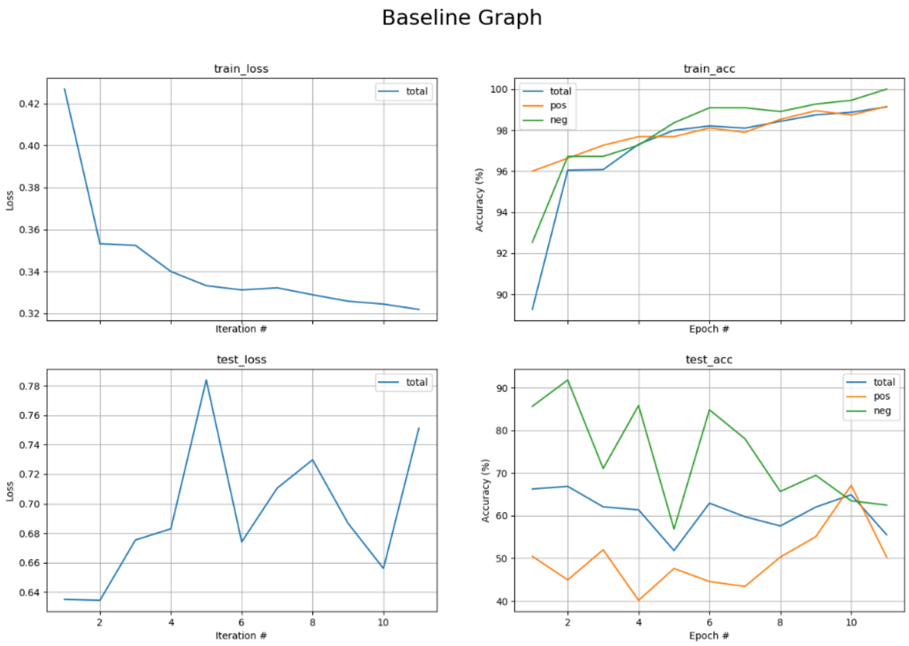

Figure 11: Results of Baseline Model with overlap 95

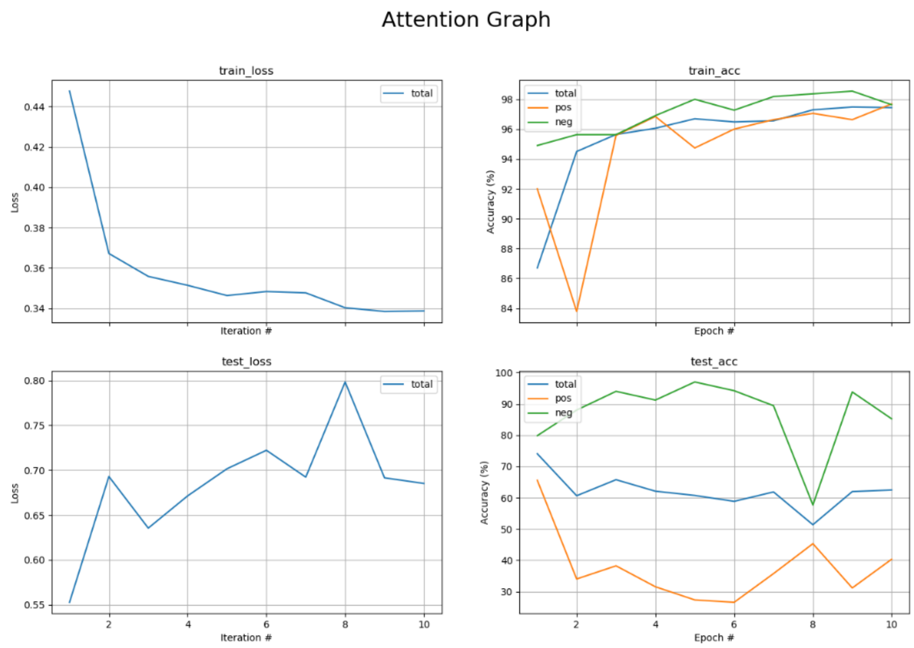

Figure 12: Results of Attention Model with overlap 95

5.2 Analyze and explain results

As explained in the previous sections, with memory restrictions we reconstructed the baseline model with overlap of 95 beats, and hyperparameters as defined in the study [8]. The results portrayed in the graphs and results table (figure 2) show that both the Baseline model and Attention model with overlap 95 achieved the lowest results of all overlaps. When examining the graphs of these models, we can see that the training accuracy keeps ascending while the test accuracy descends, shown in figure 11. This can be explained because of overfitting on the train set. When creating a dataset with an overlap of 95, more samples are being created however it creates multiple samples that are very similar to one another with only a few beats different. Our assumption is that the overfitting happens because throughout the training we encounter sequences that are almost identical, which reduces the model's generalization. The results of the Baseline model and Attention model are similar, both heavily affected from the overfit.

We also decided to conduct the same experiments on both models with sequences overlapping of 0 and 33 beats, to observe the model's behavior when dealing with sequences that don't have overlapping beats or have a small overlap.

In the results from the two experiments with overlap 33 (shown in figures 7-9) we can see that both of the models achieved higher accuracy than the matching models with overlap 95, correlating with our explanations above. As opposed to our assumptions, the Baseline model performed slightly better than the Attention model.

In figure 10 the right plot shows for randomly sampled sequences their ground truth, and the left plot shows the attention weights for those samples. From this plot we can see which beats in the sequence the attention layer focus's on. The attention map graph reinforces our results that the attention model did not succeed in finding relevant patterns in the sequences, thus it did not improve on the baseline model.

The experiments on both models with overlap 33 produced very similar results to the experiments on both models with overlap 0 (shown in figures 3-6), with the overlap 33 receiving slightly better results. The dataset created with overlap of 33 is larger than with overlap of 0, but still maintains the diversity of the sequences. So thus models with overlap of 33 trained on larger datasets, which can explain the higher results.

In conclusion, the proposed change of adding an attention layer did not improve on the results of the baseline model, but the change in the preprocess section of downsizing the overlap between the sequences improved our results.

5.3 Conclusion

As stated in previous sections, our baseline structure differed slightly from the original study [8] structure, since in the data pre-process we created sequences with overlap of 95 as opposed to the study that created sequences with overlap of 99, resulting in fewer samples being made. Our models achieved lower accuracy rates with each overlap we tested with the highest being **83.4%** than the accuracy rates achieved in the study [8] which were **99.7%**.

Based on our experiments we can see that using a max pooling layer after the bidirectional LSTM is better than using an attention layer. This might be explained with the fact that for most sequences created, all the beats of the sequence are labeled the same, with only a few sequences having beats labeled normal and also beats labeled as AF, making it hard for the attention layer to "focus" on relevant beats in the sequence.

Acronyms

AF	Atrial Fibrillation
CAD	Computer-Aided Diagnosis
ECG	Electrocardiogram
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
RNN	Recurrent Neural Network

References

[1] - https://www.nhlbi.nih.gov/health-topics/atrial-fibrillation

[2] - https://www.medtronic.com/us-en/patients/conditions/atrial-fibrillation-afib.html

[3] - Munger, TM; Wu, LQ; Shen, WK (January 2014). "Atrial fibrillation". *Journal of Biomedical Research*.

[4] - Chung, MK; Eckhardt, LL; Chen, LY; Ahmed, HM; Gopinathannair, R; Joglar, JA; Noseworthy, PA; Pack, QR; Sanders, P; Trulock, KM; American Heart Association Electrocardiography and Arrhythmias Committee and Exercise Cardiac Rehabilitation, and Secondary Prevention; Committee of the Council on Clinical Cardiology; Council on Arteriosclerosis, Thrombosis and Vascular Biology; Council on Cardiovascular and Stroke Nursing; Council on Lifestyle and Cardiometabolic Health (March 2020). "Lifestyle and Risk Factor Modification for Reduction of Atrial Fibrillation: A Scientific Statement From the American Heart Association".

[5] - Atrial Fibrillation Recurrence and Peri-Procedural Complication Rates in nMARQ vs. Conventional Ablation Techniques: A Systematic Review and Meta-Analysis
Ka H. C. Li1,2,3†, Mei Dong4†, Mengqi Gong5, George Bazoukis6, Ishan Lakhani2,3, Yan Y. Ting2,3, Sunny H. Wong2,3, Guangping Li7, William K. K. Wu8, Vassilios S. Vassiliou8, Martin C. S. Wong10, Konstantinos Letsas5, Yimei Du11, Victoria Laxton12, Bryan P. Yan1, Yat S. Chan1, Yunlong Xia12, Tong Liu2*, Gary Tse2,3* and International Health Informatics Study (IHIS) Network

[6] - Improved Detection of Silent Atrial Fibrillation Using 72-Hour Holter ECG in Patients With Ischemic Stroke
Martin Grond, Marek Jauss, Gerhard Hamann, Erwin Stark, Roland Veltkamp, Darius Nabavi, Markus Horn, Christian Weimar, Martin Köhrmann, Rolf Wachter, Ludger Rosin, and Paulus Kirchhof (Oct 2013)

[7] - Early Detection of Atrial Fibrillation Based on ECG Signals
Nuzhat Ahmed1 and Yong Zhu2,* (March 2020)

[8] - Automated detection of atrial fibrillation using long short-term memory network with RR interval signals
OliverFaust1,AlexShenfield1,MurtadhaKareema,Tan RuSanb,HamidoFujitac,U. RajendraAcharyadef

[9] Moody GB, Mark RG. A new method for detecting atrial fibrillation using R-R intervals. Computers in Cardiology. 10:227-230 (1983).

[10] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000).