2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Unsegmented Heart Sound Classification Using Hybrid CNN-LSTM Neural Networks

Drishti Ramesh Megalmani[1], Shailesh B G[1], Achuth Rao M V[2], Satish S Jeevannavar[1], and Prasanta Kumar Ghosh[2]

Abstract— **Cardiac Auscultation, an integral part of the physical examination of a patient, is essential for early diagnosis of cardiovascular diseases (CVDs). The ability to accurately diagnose the heart sounds requires experience and expertise, which is lacking in doctors in the early years of clinical practice. Thus, there is a need for an automatic diagnostic tool that would aid medical practitioners with their diagnosis. We propose novel hybrid architectures for classification of unsegmented heart sounds to normal and abnormal classes. We propose two methods, with and without the conventional feature extraction step in the classification pipeline. We demonstrate that the F score using the approach with conventional feature extraction is 1.25 (absolute) more than using a baseline implementation on the Physionet dataset. We also introduce a mechanism to tag predictions as unsure and compare results with a varying threshold.**

I. INTRODUCTION

Non-communicable Disorders (NCDs) kill 41 million people every year. Cardiovascular Diseases (CVDs) are one of the leading causes of death in the world and account for most NCD deaths. There was an estimated 422.7 million cases of CVDs globally and 17.92 million deaths due to CVDs in 2015. Cardiac auscultation using stethoscope still remains the central tool for diagnosis of valvular and other structural heart abnormalities. It may reveal many pathologic cardiac conditions including arrhythmias, valve diseases and heart failure. Heart sounds provide important initial clues in disease evaluation, serve as a guide for further diagnostic examination, and, thus, play an important role in the early detection for CVDs. As opposed to an experienced Cardiologist, Pulmonologist (specialists), many fresh medical students, graduates and allied healthcare staff face difficulties in hearing and identifying the normal/abnormal characteristic sounds leading to a serious concern of potential missed diagnosis and/or early detection. Another persistent issue is the low doctor patient ratios in low and middle income countries. In such low resource settings, a patient's first point of contact in majority of the cases is at a primary care or community clinic manned by a nurse or a primary care physician. It requires 2-7 years of training and experience (learning curve) before a newly trained healthcare graduate can identify clinically significant auscultatory sounds and become proficient in identifying normal/abnormal heart and lung sounds. Therefore, an automated assistive detection technology is needed to help medical professionals make informed decisions and detect abnormalities at the early screening stage.

The human cardiac cycle consists of two phases- systole and diastole. These phases are defined by the two main fundamental heard sounds S1 and S2 ('lub' and 'dub'). Heart murmurs, caused due to turbulent blood flow in or near the heart, can either be harmless or abnormal. Based on the presence, position and extent of a murmur in either of the phases of the cardiac cycle, it can be classified into pan, mid or late systolic and diastolic murmurs. Over the years, a lot of developmental work has been going on with respect to classification and segmentation of heart sounds involving latest machine learning and signal processing techniques. Heart sounds can be classified into two classes- normal and abnormal. The basic workflow of the two-class heart sound classification includes 1) Pre-processing 2) Segmentation 3) Feature Extraction 4) Classification.

Pre-processing includes steps like filtering, denoising, enhancement, etc. Segmentation acts as an essential step in the automatic analysis of phonocardiogram (PCG), used to extract the main heart sound states- S1, systole, S2, diastole. For this purpose, various hand-crafted spectro-temporal features [1] are extracted. Extraction of such elaborate hand crafted features not only requires expert domain knowledge but also requires manual labour in many cases. Classification pipeline involves segmentation which needs to be robust to real world scenarios. This, in turn, requires a large number of annotations from medical experts, which is, however, cumbersome and time-consuming. Previous works showcased the extensive use of segmentation and a wide array of feature extraction techniques to perform the task at hand. In [2], the authors proposed a technique using Markov chain analysis to model temporal changes in the signal to extract relevant features along with other spectral and statistical features. Rubin et al. [3] converted the audio waveforms into time-frequency domain spectrograms that are later fed into a deep neural network to perform classification. In [4], authors extracted 124 time-frequency features and used an ensemble of Adaboost and CNN based network to classify the heart sounds. This was the top scorer for the Physionet challenge, with an overall score of 86.05% (average of sensitivity and specificity that are 94.24% and 77.81% respectively). Classification using features extracted from unsegmented heart sounds eliminates the time taken and computational overhead for a separate segmentation task and provides effective results. Many of the earlier studies [5, 6, 7] have

[1]R&D/Machine Learning at Ai Health Highway India Private Limited drishti.m@aisteth.com, satish.sj@aihighway.org

[2]Electrical Engineering, Indian Institute of Science, Bangalore 560012, India achuthr@iisc.ac.in,prasantg@iisc.ac.in

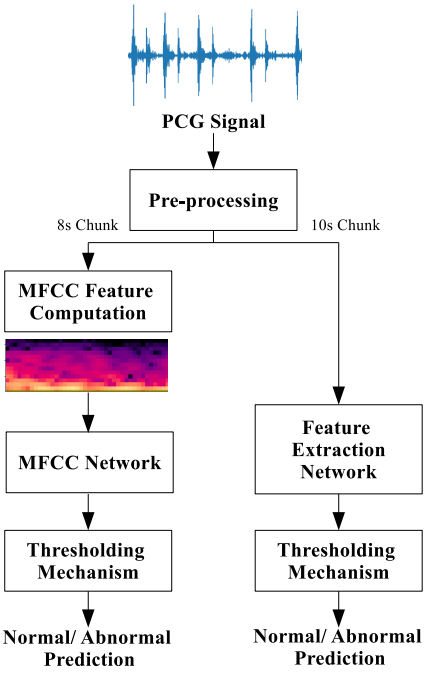

Fig. 1. The block diagram of the proposed heart sound classification mechanism.

used unsegemented input data to extract various features like wavelet entropy, hilbert's envelope, power spectral density, scalograms etc. Various classifiers like k nearest neighbour (kNN) and deep neural networks have also been used.

We aim to perform binary classification of heart sounds into normal and abnormal by eliminating the segmentation step, thereby reducing the complexity and computational overhead of performing segmentation. Our contributions are two-fold. First, we propose novel architectures for unsegmented heart sound classification comprising both convolutional neural network (CNN) layers and long short-term memory (LSTM) [10] layers. The CNN layers learn the contextual local information, thereby acting as filters, whereas the LSTM layers extract the the temporal information. Second, we provide a comparative analysis of utilizing feature engineering and extraction in the modeling pipeline. We showcase the performance changes while using a hybrid network for both extracted mel-frequency cepstral coefficients (MFCCs) features and raw time domain waveform input. We also introduce a post-processing step that evaluates the confidence of classifying a particular audio file as normal or abnormal. We use a thresholding mechanism for finding unsure predictions thereby reducing the number of misclassifications and boosting the results.

II. DATASET

A. Dataset Description

For this study, a publicly available heart sound dataset, Physionet[8], was used. It consists of 3240 heart audio files of varying length ranging from about 5 seconds to 122 seconds. The dataset is collected by various teams using heterogeneous sensing equipment from multiple countries under different clinical and nonclinical (e.g., home visits) environmental conditions. We can thereby notice both noisy and clean heart sound recordings. The dataset contains a binary annotation of normal vs abnormal. The recordings were collected both from healthy subjects and patients with a variety of heart conditions, especially coronary artery disease and heart valve disease. The subjects were from different age groups including children, adults and elderly. The dataset consists of 665 abnormal and 2575 abnormal recordings. The dataset comprises 6 different sets, namely a to f, which vary with respect to the recording instrument, environment, pathalogical conditions, recording positions etc. All the audio files were sampled to 2000Hz. Bandpass Butterworth filtering of the frequency range, 20Hz to 500Hz was employed on each audio file. This was done to remove both high-frequency noise and unwanted low-frequency artifacts. To overcome the differing durations of the heart sound recordings, random chunks of constant sizes were extracted and used as input to the classifiers.

III. METHODOLOGY

A. Proposed Architectures

Heart sounds are a combination of quasi-periodic fundamental heart sounds like S1 and S2, psuedo-periodic murmurs and other noises. Due to the overlap of frequency ranges of these different components, it is difficult to identify them in the time or frequency domain alone. For physiological time series data like heart sounds, we need to extract the information with respect to the chronological changes in the signal as well as its repeating nature which depends largely on historical information. Thus, analysis in the time-frequency domain is suitable for heart sound classification. We thereby propose a combination of both CNN and LSTM layers. CNN layers are used to extract local correlations and frequency characteristics from the input heart sound data and LSTM layers help us extract long term dependencies from the learned local features. The combination of layers is proposed with the aim that the network exploits temporal information better, which is essential for heart sound classification. A conventional model pipeline for audio classification involves feature extraction and engineering as a prerequisite, usually extracting spectrograms, MFCCs and chroma features. In this work, we aim to provide a comparative study between the conventional modelling approach with hand-crafted features, i.e., MFCC, and a network that takes time-domain waveforms as inputs, thereby eliminating the feature engineering step. For the latter, we aim to provide a hybrid neural network that attempts to jointly perform feature extraction as well as classification instead of separating the two tasks. Fig. 1 shows the block diagram of the proposed methodology for heart sound classification.

1) MFCC Network: The model given by Table I is a hybrid CNN and LSTM network with 2 convolutional layers, a max pool layer and two LSTM layers followed by a series of time distributed layers. The 1D convolutional layers

Mel-frequency Cepstral Coefficients Features (799 x 13)
1-D CNN 5 x 1 @64, ReLU
Max-Pooling 5 x 1, Dropout 0.4
1-D CNN 5 x 1 @64, ReLU
LSTM @ 64
LSTM @ 32, Dropout 0.3
Time Distributed Dense @ 64, ReLU
Time Distributed Dense @ 32, ReLU
Time Distributed Dense @ 16, ReLU
Time Distributed Dense @ 8, ReLU
Flatten
Dense @ 2, softmax

TABLE I
TOPOLOGY OF MFCC NETWORK

10 sec Raw Audio Chunk (20000,1)
1-D CNN 30 x 1 stride 5 @32, ReLU
1-D CNN 5 X 1 @32, ReLU, Dropout 0.3, BN
Max-Pooling 2 x 1 stride 2
1-D CNN 5 X 1 @64, ReLU, Dropout 0.3, BN
Max-Pooling 2 x 1 stride 2
1-D CNN 5 X 1 @64, ReLU, Dropout 0.3, BN
Max-Pooling 2 x 1 stride 2
1-D CNN 5 X 1 @128, ReLU, Dropout 0.3, BN
Max-Pooling 2 x 1 stride 2
LSTM @ 128
LSTM @ 64
LSTM @ 32, Dropout 0.4
Time Distributed Dense @ 32, ReLU
Time Distributed Dense @ 16, ReLU
Time Distributed Dense @ 8, ReLU
Flatten
Dense @ 2, softmax

TABLE II
TOPOLOGY OF FEATURE EXTRACTION NETWORK

have a kernel size of 5. The pool size for the max pooling layer is also 5. A dropout regularizer is introduced to avoid overfitting. Rectified Linear Unit (ReLU) activation is used for all the CNN layers as well as the fully connected layers. The last fully connected layer utilizes a softmax activation to obtain class probabilities.

2) Feature Extraction Network: Instead of using specific hand-crafted features to train the neural network, we experimented with raw data as the input with the aim to attain a network that would inherently learn complex high-level and relevant features from the raw audio inputs. Table II explains the architecture of the model. A 10 second raw audio chunk is the input to this network. We experimented with various kernel sizes for the first CNN layer from 10 to 40 with an increment of 10. A repeating block consisting of a 1D CNN layer with kernel size 5 and a max pool layer for dimensionality reduction with pool size of 2 and stride of 2 is considered. Batch Normalization and Dropout are introduced in each block. The output of the fourth block is reshaped to 150 $\times$ 128, after which three LSTM layers and three time distributed dense layers are used. All CNN and fully connected layers use a ReLU activation function except the last dense layer where we use a softmax activation function. The initial weights of the convolutional layers are set using Xavier initialization [2]. The bias terms are set to zero.

IV. EXPERIMENTS & RESULTS

A. Experimental Setup

That dataset considered in this work consists of recordings of varying durations. Hence, chunks of constant sizes were extracted. Experiments were done with chunk sizes from 2s to 10s in increments of 2s. The larger chunk sizes, namely, 8s and 10s, were found to be optimal for the MFCC network and the feature extraction network, respectively. The smaller chunk sizes do not encapsulate enough data for extracting cues specific to abnormality. The dataset consists of audio files with varying durations, with set b having recordings only upto 8 seconds long. To maintain a fixed duration of audio files to train and evaluate all the models, we considered audio files greater than 10s. Thus, among the 3240 audio files, 2578 files are considered. The dataset was split three ways- training, validation and testing. The training set consisted of 2087 files with 262 abnormal and 1825 normal files. The validation set comprised 252 files, of which 136 were abnormal and 116 were normal. Finally the test set had 239 files with 123 normal and 116 abnormal files. The split was in terms of audio files as a whole and not just with respect to chunks to avoid overlap among train, validation and test sets. From the training set, 30000 random chunks with almost an equal number of normal and abnormal cases were considered. This was done to avoid any prevalence to the majority (normal) class due to data imbalance. A total of 6000 chunks were extracted separately from validation and test sets with an equal number representation from both classes.

For the MFCC network, 8 second chunks were considered. A 25 ms window size with a step size of 10 ms was used, on which a fast Fourier transform of size 64 was performed. The power spectrum is passed through Mel-scale 26 filterbank channels. We then decorrelated the filter bank energies by applying a Discrete Cosine Transform (DCT), and retained the first 13 mel-frequency cepstral coefficients.

For the feature extraction network, we extracted 10 second chunks randomly from the training set. We experimented with various kernel sizes for the first layer of the network. By keeping all the other parameters fixed, we trained models with kernel sizes from 10 to 40 with an increment of 10. We noticed an increase in accuracies as we increased the kernel size, but a slight decrease with a size of 40. The chunk level accuracies on the validation set were 80.43%,

Model	Accuracy	Specificity	Sensitivity	F1 Score
C. Potes et.al.[4]	91.60	**90.52**	92.68	91.94
MFCC Network	**93.2**	89.43	**96.85**	**93.187**
Feature Extraction Network	91.63	88.79	94.31	91.62

TABLE III

COMPARISON OF RESULTS OBTAINED BY PROPOSED ARCHITECTURES AND BASELINE

Set	MFCC Network		Feature Extraction		C. Potes et.al.[4]	
	Confusion Matrix	F1 Score	Confusion Matrix	F1 Score	Confusion Matrix	F1 Score
Set a	[[1 5] [4 67]]	**87.82**	[[1 5] [6 65]]	86.21	[[3 3] [8 63]]	87.55
Set c	[[0 0] [0 7]]	**100**	[[0 0] [0 7]]	**100**	[[0 0] [0 7]]	**100**
Set d	[[0 2] [0 4]]	**53.33**	[[0 2] [0 4]]	**53.33**	[[0 2] [0 4]]	**53.33**
Set e	[[103 3] [0 36]]	**97.91**	[[101 5] [0 36]]	96.55	[[102 4] [1 35]]	96.52
Set f	[[1 1] [0 5]]	**83.98**	[[1 1] [1 4]]	71.43	[[0 2] [0 5]]	59.52

TABLE IV

SET WISE CLASSIFICATION RESULTS

Threshold	No. of files Eliminated	C. Potes et.al.[4]		MFCC Network	
		Accuracy	F1 Score	Accuracy	F1 Score
0.25	13	93.35	93.51	94.51	94.51
0.20	11	92.96	93.16	94.56	94.55
0.15	9	93.03	93.22	94.19	94.18
0.10	7	92.66	92.89	94.24	94.23

TABLE V

RESULTS OBTAINED FOR MFCC NETWORK AND BASELINE (WITH THRESHOLD)

Threshold	No. of files Eliminated	C. Potes et.al.[4]		Feature Extraction Network	
		Accuracy	F1 Score	Accuracy	F1 Score
0.25	23	94.89	95.07	95.37	95.36
0.20	21	94.02	94.22	94.50	94.49
0.15	16	94.16	94.32	93.72	93.71
0.10	13	93.35	93.51	92.92	92.91

TABLE VI

RESULTS OBTAINED FOR FEATURE EXTRACTION NETWORK BASELINE (WITH THRESHOLD)

89.33%, 90.87% and 89.18% for kernel sizes 10, 20, 30 and 40 respectively. Hence, the first 1D CNN layer in the proposed network has a filter size of 30 and a stride of 5. The first layer has a larger filter size in comparison to the others so as to have a global view of the audio raw data and extract the spectral features over time. We used normalized raw audio data for the feature extraction network and did not perform any feature engineering beforehand.

The categorical cross entropy function is used to train both the models. The objective function of both the networks are optimized with the Adam optimizer [20] with a learning rate of 0.001, β_1 of 0.9, β_2 of 0.999, ε of 1e-07 and a decay of 0.004. A batch size of 64 was used for both training and validation. Training was done for 50 epochs with an early stopping criteria on the validation loss with a patience of 5.

B. Results & Discussion

In our experiments, we trained 2 different hybrid models with inputs- MFCC features and raw audio input, respectively. We have used the pretrained model, proposed by C. Potes [8] for the heart sound classification task on the same corpus, as the baseline model. We haven't considered the group of 300 files set aside as validation set in the Physionet challenge. Our test dataset comprises randomly selected audio files from the entire Physionet dataset. This gives the baseline model an advantage in the sense that the recordings in test set used in this work may already been used for training the baseline model. The proposed models were trained and tested on chunks of audio files. The predictions for the entire audio files were aggregated through majority voting over individual chunks. The results for our proposed approaches and the baseline are using reported in Table III. The results are reported using four evaluation metrics-accuracy, sensitivity, specificity and F1 score.

The proposed MFCC network performs better than the baseline model. We also notice that both the MFCC and feature extraction network have high sensitivity (correctly identifying the abnormal patient) scores, which is clinically an important feature to have in such a critical application. We also observe that the specificity values of the proposed

networks are lower than the baseline, which signifies that there are more false positives.

To form the test dataset, we have considered files from each set in the same ratio as present in the original entire dataset. Since, each set in the dataset represents the heart sound files collected from various types of environments and stethescopes, we have evaluated the set-specific performance for all the three models (Table IV). We notice that the MFCC model performs better or on par with the baseline model in all the sets. The feature extraction model on the other hand performs better than the baseline for sets e and f, equally well for c and d and has a lower performance for set a.

We introduce a thresholding mechanism to tag certain predictions for audio files as unsure. The posterior class probabilities from the last softmax layer are considered for all the chunks per audio file. Based on the predictions for each chunk, we collect the list of probabilities for the chunks with abnormal and normal predictions separately. The median normal and abnormal probabilities are found from the collected list of probabilities. We find the relative change between the two median values, which we assign as a grade to each audio file prediction. This grade signifies how sure or unsure the classifier is about the prediction of the model. We vary a threshold from 0.1 to 0.25 in intervals of 0.05. If the grading for a particular audio file is lower than the threshold, its tagged as "unsure" and eliminated for evaluation. The scores with thresholding are reported in Table V and VI. We see an improvement in the performance over the baseline for both the networks. There is a steady increase in both accuracy and F1 score as we increase the threshold, especially in the case of the feature extraction model. This shows that the audio files that are marked as unsure are indeed the ones misclassified, resulting an improved result. The predictions made by the model done with a low confidence level (grading) are the ones eliminated by the thresholding mechanism. Since there is a rise in the performance, it suggests that many of the wrong classifications are with a lower confidence. With a threshold of 0.25, we see that both the networks perform better than the baseline model.

V. CONCLUSIONS

In this work, we propose two novel hybrid networks for unsegmented heart sound classification with and without a feature engineering step. We provide a comparative study between the proposed approaches and a baseline model. The MFCC network results in a sensitivity of 97.478%, accuracy of 94.51% and an F1 score of 94.509%, which are better than those using the baseline model. The feature extraction model performs better than the baseline model as well. With the proposed thresholding mechanism to detect "unsure" cases, the feature extraction model performs the best among the three at a threshold of 0.25 resulting in an accuracy of 95.37% and an F1 score of 95.36%. Our proposed method, feature extraction model, thereby provides an approach to classify abnormal and normal heart sounds effectively without the overhead of two intermediate steps-segmentation and feature extraction. This not only reduces the time taken and computational overhead but also does not require expert knowledge which would have been essential for extensive feature engineering.

REFERENCES

[1] H. Tang, H. Chen, T. Li and M. Zhong, Classification of normal/abnormal heart sound recordings based on multi-domain features and back propagation neural network, Computing in Cardiology Conference (CinC), Vancouver, BC, pp.593-596, 2016.

[2] S. Vernekar, S. Nair, D. Vijaysenan and R. Ranjan, A novel approach for classification of normal/abnormal phonocardiogram recordings using temporal signal analysis and machine learning, Computing in Cardiology Conference (CinC), Vancouver, BC, pp. 1141-1144, 2016.

[3] Jonathan Rubin, , Rui Abreu, Anurag Ganguli, Saigopal Nelaturi, Ion Matei, and Kumar Sricharan, Recognizing Abnormal Heart Sounds Using Deep Learning, CoRR, 2017.

[4] C. Potes, S. Parvaneh, A. Rahman and B. Conroy, Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds, Computing in Cardiology Conference (CinC), Vancouver, BC, pp. 621-624, 2016.

[5] Krishnan, P.T., Balasubramanian, P. and Umapathy, S., Automated heart sound classification system from unsegmented phonocardiogram (PCG) using deep neural network, Phys Eng Sci Med 43:505–515, 2020.

[6] Singh, Sinam Ajitkumar, and Swanirbhar, Mujumder, Classification of unsegmented heart sound recording using KNN classifier, Journal of Mechanics in Medicine and Biology 19, no.04, 2019

[7] Philip L, Alan M, Heart sound classification from unsegmented phonocardiograms, Physiol Meas 38:165870, 2018.

[8] C. Liu, D. Springer, Q. Li, B. Moody, R. A. Juan, F. J. Chorro, F. Castells, J. M. Roig, I. Silva, A. E. W. Johnson, Z. Syed, S. E. Schmidt, C. D. Papadaniil, L. Hadjileontiadis, H. Naseri, A. Moukadem,A. Dieterlen, C. Brandt, H. Tang, M. Samieinasab, M. R. Samieinasab, R. Sameni, R. G. Mark, and G. D. Clifford, An open access database for the evaluation of heart sound algorithms, Physiological Measurement, vol. 37, no. 12, pp. 2181–2213, 2016.

[9] Xavier Glorot, Yoshua Bengio, Understanding the difficulty of training deep feedforward neural networks, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, PMLR 9:249-256, 2010.

[10] Sepp Hochreiter and Jürgen Schmidhuber, Long Short-Term Memory, Neural Comput. 9:1735–1780, 1997.

[11] li, Fan & Tang, Hong & Shang, Shang & Mathiak, Klaus & Cong, Fengyu. . Classification of Heart Sounds Using Convolutional Neural Network. Applied Sciences 10: 3956, 2020.

[12] Simarjot Kaur Randhawa, Mandeep Singh, Classification of Heart Sound Signals Using Multi-modal Features, Procedia Computer Science, Volume 58, Pages 165-171, 2015.

[13] Springer D, Tarassenko L, Clifford G, Logistic regression-HSMM-based heart sound segmentation, IEEE Trans Biomed Eng 63:82232, 2015.

[14] Hamidi M, Ghassemian H, Imani M, Classification of heart sound signal using curve fitting and fractal dimension, Biomed Signal Process Control 39:3519, 2018.

[15] Langley P, Murray A, Abnormal heart sounds detected from short duration unsegmented phonocardiograms by wavelet entropy, Comput Cardiol 545, 2016.

[16] Dominguez-Morales JP, Jimenez-Fernandez AF, Dominguez-Morales MJ, JimenezMoreno G, Deep neural networks for the recognition and classification of heart murmurs using neuromorphic auditory sensors, IEEE Trans Biomed Circuits Syst 12:24–34, 2018.

[17] Abdollahpur M, Ghaffari A, Ghiasi S, Molla Kazemi MJ, Detection of pathological heart sounds, Physiol Measurement 38:1616–1630, 2017.

[18] Maglogiannis I, Loukis E, Zafiropoulos E, Stasis A, Support vectors machine-based identification of heart valve diseases using heart sounds, Comput Methods Programs Biomed 95:47–61, 2009.

[19] Li, F., Liu, M., Zhao, Y. et al. Feature extraction and classification of heart sound using 1D convolutional neural networks, EURASIP J. Adv. Signal Process., 59, 2019.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Proc. of 3rd Int. Conf. for Learn. Representations (ICLR), San Diego, CA, USA, 2015.