An Accurate CNN Architecture For Atrial Fibrillation Detection Using Neural Architecture Search

Najmeh Fayyazifar
School of Science, Edith Cowan University, Perth, Australia
Fayyazifar1@ecu.edu.au

***Abstract*— The accurate and timely diagnosis of Atrial Fibrillation (AF), a common condition presenting as an abnormal heartbeat that often results in serious disease, would assist in reducing morbidity. In this study, we make use of Electrocardiogram (ECG) data in order to create an automatic method for detecting AF. For this purpose, we employed a neural architecture search (NAS) algorithm. The efficiency of NAS algorithms on image classification tasks has been well established, however, studies on using NAS methods for ECG classification are very limited. Our experiments show that our automatically designed neural model performs very well and arguably outperforms currently available deep learning models. This model achieved the accuracy and F1-score of 84.15%± 0.6 and 82.45± 0.2 on the publicly available subset of PhysioNet challenge 2017 dataset, respectively.**

Keywords— atrial fibrillation, neural architecture search, CNN architecture, deep learning

I. INTRODUCTION

Cardiac arrhythmias are one of the most common conditions associated with cardiovascular diseases. Atrial Fibrillation (AF) is one of the arrhythmias which associates with higher cardiovascular risk and mortality: such as sudden cardiac death [1], stroke, and heart failure [2]. AF affects about 30 million people worldwide [3] and is estimated to increase significantly by 2030 [4].

An electrocardiogram (ECG) is used for measuring cardiac electrical activity and abnormality detection. In practice, ECGs are reviewed by clinicians, at different time intervals, to detect abnormal cardiac activities. However, this is a labour-intensive task and requires expert knowledge. The development of automatic abnormality detection systems can assist medical staff to diagnose anomalies more efficiently. Furthermore, the availability of wearable ECG-monitoring devices, which normally provide single-lead ECG data, has become widespread. This has facilitated an increase in the early detection of arrhythmias. The development of an automatic arrhythmia detection system is the essential task for analysing these long-term ECGs provided by wearable devices.

Machine learning algorithms, and more specifically deep neural networks, are artificial intelligence tools that can be utilized to automate learning processes. Deep learning algorithms have provided accurate results in many computer vision tasks: such as image recognition and video analysis. Specifically, the development of Convolutional Neural Networks (CNNs) has reduced required pre-processing by relying on filters to extract essential information rather than manually designing feature extraction methods. In addition, feature extraction in CNNs has been proven to be noise insensitive [5], which makes CNN structures suitable for classification of noisy data. However, selection of proper CNN structures based on the problem at hand, is still an unresolved issue. In the last few years, Neural Architecture Search (NAS) methods [6, 7] have been developed to automatically design neural network architectures for image classification tasks. Despite the efficiency of these search methods in finding a suitable architecture on ImageNet and/or CIFAR10 dataset, work on discovering an appropriate neural network architecture on medical data, and more specifically, on biological signals is very limited.

In this paper, we present a robust CNN structure for the detection of AF, from single-lead ECG data, found by means of neural architecture search. The rest of the paper is organized as follows: In Section II, we review recent research on atrial fibrillation detection. Section III describes our proposed methodology, where our results are summarized in Section IV. We will conclude our findings in Section V.

II. RELATED WORK

The "PhysioNet computing in cardiology challenge 2017" (PhysioNet 2017) on AF detection [8] was a ground-breaking attempt to address the question of how accurate can atrial fibrillation be diagnosed among other arrhythmias and normal sinus rhythm. Within the challenge, different approaches were proposed, to detect AF more accurately. All four teams [9-12] who jointly secured first place in the competition (F1-score of 83%), applied some sort of pre-processing on ECGs: such as different noise removal techniques; and then extracted various types of hand-crafted features (such as: morphological, Heart Rate Variability (HRV), frequency domain, time-frequency domain, and statistical features). In order to reduce the number of features used in the classification stage, some feature selection algorithms like Maximal Information Coefficient (MIC) and minimum Redundancy Maximum Relevance (mRMR) have been applied by [9]. AdaBoost [9], Random forest [10], Tree Gradient Boosting- XGBoost [11, 12], and Linear Discriminative Analysis (LDA) [11] classifiers were applied for the final classification of ECGs.

Although the approaches that utilized hand-crafted features performed well during the PhysioNet 2017, manual feature engineering in ECGs has some drawbacks. For example, due to the dynamic, non-linear and non-stationary nature of ECG signals [13], manual feature extraction methods are time-consuming, and also require expert knowledge [14]. Additionally, since ECGs contain a considerable amount of noise, noise removal is a preliminary step in AF detection systems which do not use CNN structures. Therefore, our focus is on models which used the convolution-based networks for automatic feature extraction and classification.

The second winner of the challenge (F1-score of 82.1%) [15] promoted the efficiency of CNNs and Convolutional Recurrent Neural Networks (CRNNs) for the detection of AF without any need for noise removal. Their proposed CNN network, with 24 convolutional layers, achieved an F1-score and accuracy of 79% and 81.2%, respectively, using data augmentation and five-fold cross validation. The accuracy and F1-score of their CRNN network were reported as 79.2% and 82.3%, respectively.

In the other research that used deep learning (presented at PhysioNet 2017) [16] , in each signal, QRS complexes were detected, and then the quality of each complex was assessed in order to discard noisy samples in the dataset. The ECG signals were transformed into spectrograms and were fed into a CNN model with 40 layers. In order to increase the performance of the proposed model, some post-processing was applied. For each sample, if the difference between the predicted probability of belonging to "normal rhythm" and "other arrhythmia" was lower than 0.4, the sample was re-classified by extracting hand-crafted features and using an AdaBoost classifier. In the other deep learning approach for AF detection, Andreotti et al. [17], compared the performance of feature-based approaches (time domain, frequency domain and HRV features) with a CNN model. Their proposed CNN structure outperformed the bag trees classifier (using hand-crafted features).

Efforts to develop a deep learning model for ECG classification have not been constrained to the PhysioNet 2017. Hannun et al. [18] proved that deep learning algorithms can exceed cardiologists' performance in arrythmia classification . This study was conducted on a new dataset collated by the authors, containing 14 classes of cardiac rhythms. A committee of nine board-certified cardiologists annotated the data. They also evaluated the performance of their proposed deep neural network on PhysioNet 2017 dataset. Their model consisted of with 33 convolutional layers plus a linear dense layer and skip connections. They claimed an F1-score of 83% on the PhysioNet 2017 dataset. The other deep learning approach for AF detection (after the PhysioNet challenge) [19] incorporated domain knowledge (R peak position and P wave) in their deep learning algorithm. Their residual network, consisted of 16 convolution blocks, achieved the accuracy and F1-score of 83.14% and 79.59% on the publicly available subset of PhysioNet 2017 dataset, respectively.

Although convolutional based neural networks have provide the opportunity to automatically extract features, manually designing of network's structure is a time-demanding task. In the last few years, neural architecture search has become a popular topic in the computer vision community, helping the community to automatically generate suitable architecture with higher accuracy and less complexity for a specific problem. However, there are very few studies that have investigated the effect of NAS algorithms on discovering accurate neural architecture on medical data, and more specifically on one-dimensional ECG signals. In this paper, we have used the Efficient Neural Architecture Search (ENAS) algorithm proposed by Pham et al. [6] to design an accurate neural architecture which can be used to address the problem of AF detection from single-lead ECG dataset released by the PhysioNet 2017.

III. METHODODLOGY

A. Data Transformation

A time-frequency representation is used to capture the information of existing frequency components in any given time of the signal, as well as the temporal behaviour of the signal over time [20]. In this paper, we have computed the logarithmic spectrogram of input raw ECG signals to be able to extract both temporal and frequency information. A Tukey window with a length of 64 (corresponds to 213ms of signals with 300 Hz sampling frequency) , 50% overlap, and the shape parameter of 0.25 , is used to compute the spectrogram.

B. ENAS Search

We employed ENAS algorithm [6] to discover an accurate model for the problem of cardiac arrhythmia classification. Firstly, the search space (allowed operations) was designed. The possible operations are convolution (with filter sizes of 3*3 and 5*5), depth-wise separable convolution (with filter sizes of 3*3 and 5*5), max-pooling, and average-pooling. Then an RNN controller is trained to make two sets of decisions namely; which operation should be performed at each layer, and which previous layers should be connected to the current layer.

For a model with M layers, the selection of operations in each layer is independent, amounting to 6^M possible combinations. Each layer m can be connected to m-1 previous layers, so the possible number of connections among the layers is M(M-1)/2. Therefore, the RNN controller makes $2^{M(M-1)/2}$ decisions to form the final connections in the model. The overall search space for the controller is $6^M * 2^{M(M-1)/2}$.

In this model, the controller is a Long-Short Term Memory (LSTM) which makes the decisions using a softmax classifier. The training is performed in two stages. In the first stage, the controller's policy is fixed, and parameters of the network are trained by means of minimizing the loss function (using Stochastic Gradient Descent-SGD). In the next stage, the parameters of the network are fixed, and controller parameters are trained to maximize the expected reward function. In this model, the reward function is the accuracy of the model on an unseen validation dataset (Section IV. A). Based on the reward function that the controller has received at each epoch, the RNN controller learns to sample more accurate networks in the next epochs of the algorithm.

C. Train from Scratch

After the search algorithm is performed, different architectures are sampled at each epoch of the ENAS algorithm. The accuracy of each model on an unseen dataset (test dataset, Section IV. A) is computed. The model with the highest test accuracy is selected as the final model and is trained from scratch. All these steps are driven from within EANS approach. The details of parameter setting for training from scratch are provided in Section IV. B

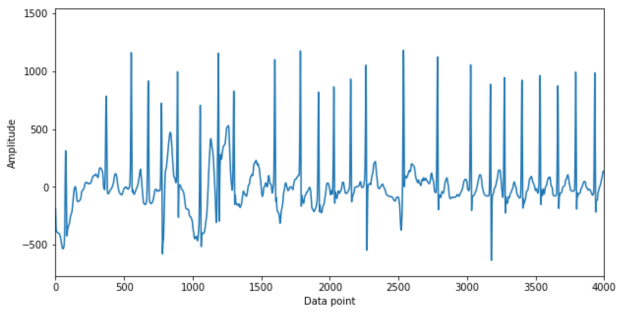
IV. EXPRIMENTAL RESULTS

A. Dataset

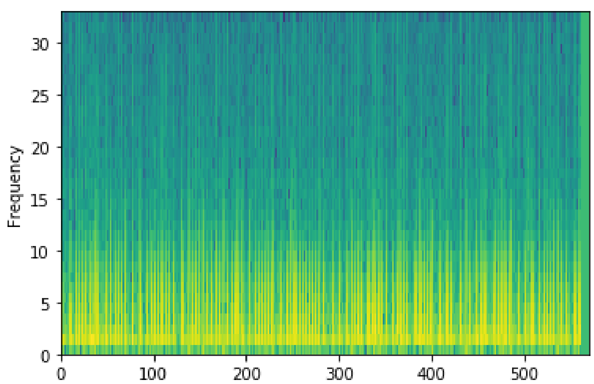
The dataset used in this research was released by "PhysioNet computing in cardiology challenge", 2017 [8]. It consists of 12186 samples, collected from individual subjects wearing an AliveCor single-channel ECG device. Samples are divided into a training set (8528 samples) and a test set (3658 samples). The test dataset is not publicly available. The sampling frequency of ECG recordings is 300 Hz, and the length of samples varies between 2714 (9 sec) and 18286 (61 sec) data points. Each ECGs' bandwidth is between 0.5Hz and 40Hz with a dynamic range of a ± 5 mV. The dataset contains four classes namely: Atrial Fibrillation, Normal Sinus Rhythm, Other arrhythmias and Noise. The number of

TABLE I. PROFILE OF DATASET

Type	Number of recordings (percentage)
Normal	5076 (59.5)
Atrial Fibrillation	758 (8.9)
Other arrhythmias	2415 (28.3)
Noise	279 (3.3)



a) Time domain representation of AF signal



b) Spectrogram representation of a

Fig. 1: a) Time domain representation of AF signal- b) Spectrogram representation of a

recordings of each class and a sample of AF and its spectrogram representation are shown in Table I and Fig.1, respectively. As it can be seen in Fig.1(a), in ECGs, the AF represents as irregularities in heartbeat (the interval between each consecutive peak corresponds to a single heartbeat).

We randomly divided the publicly available dataset into three sets: training (80%), validation (10%), and test (10%) in the neural architecture search algorithm (section III.B), and used 5-fold cross validation with 80% training and 20% testing for training the discovered architecture from scratch (section III.C).

B. Experiment Setup

In both the “neural architecture search (NAS) algorithm” and “train from scratch”, the parameters of neural models were trained using Nesterov momentum[21] .The learning rate was initialized at 0.05 and a decay rate of 0.97 was applied at every epoch to ensure a monotonous decrease in the training loss. The minimum learning rate was set to 0.0001 which prevents the learning rate from reaching zero. An L2 regulizer of 0.00025 in the NAS algorithm and 0.0003 in training from scratch was applied to reduce the impact of overfitting. We initialized the parameters of the network with the ‘He’ initializer [22] and ran the algorithm for 300 epochs using early stopping with patience of 10 which ensures the algorithm will run until the validation loss stops decreasing. A dropout rate of 0.3 was set after each convolutional layer in training from scratch. An Adam optimizer with a learning rate of 0.00035 was used to train the policy parameters. On a single Nvidia GTX 1080 GPU, NAS algorithm took 40 hours and the required time for training from scratch was 24 hours.

C. Derived Architecture

Our proposed CNN architecture is shown in Fig.2. The numbers in the graph correspond to standard neural network operations, as shown on the legend. As the graph illustrates, the proposed model consists of three sections (8 layers in each section), separated by a max-pooling layer of both pooling size and stride of 2. At layers 8 and 16, the outputs of all previous layers in that section are concatenated and passed through the max-pooling layer. Then, the output of max pooling is fed as an input to all layers in the subsequent section. In our proposed architecture, selection of operations at each layer and skip connections are independent of other layers, therefore, the types of operations and the number of skip connections differ in each section.

There are four different convolutional operations namely: convolution with filter sizes of 3*3 and 5*5, and separable-convolution with filter size of 3*3 and 5*5. The stride for all convolution layers was set to 1. After each convolution layer; batch normalization, a Relu activation function, and a dropout with the rate of 0.3 was applied. The pooling size of both max-pool and average-pool operations was 3*3 with the stride of 1. In the first section, the number of output filters was set to 48 and was doubled at the end of sections one and two. The output of the last layer was passed to a sotfmax layer with four nodes.

D. Comparison of Performance

The performance of our proposed architecture in terms of accuracy and F1-score metrics on publicly available subset of PhysioNet 2017 dataset, using 5-fold cross validation is reported in Table II. To compute these metrics, we trained our proposed architecture five times, each with a different random initialization seed, and computed the mean and standard deviation of these models. The F1-score of each class and the total F1-score are computed as defined by PhysioNet 2017 [8].

Table II shows that the “normal sinus rhythm” class has the highest score. This could be due to the larger number of training samples of this class (almost 60%). The “other arrhythmia” class has the lowest score since the samples of this class have a high variability (samples from a wide range of arrhythmias with different patterns).

In the PhysioNet challenge 2017, the proposed models were ranked using a test dataset which is not publicly available. Models need to be uploaded into the PhysioNet website, in order to be scored on the test dataset. However, only models with a limited size (maximum 50Mb) can be uploaded to the website. Our model is larger than this limitation (600 Mb) and has not been uploaded to the website in order to be ranked on test dataset. Moreover, the ranking website is closed now. However, in order to compare our results with the state-of-the art deep learning models, we used their reported F1-score on publicly available subset of

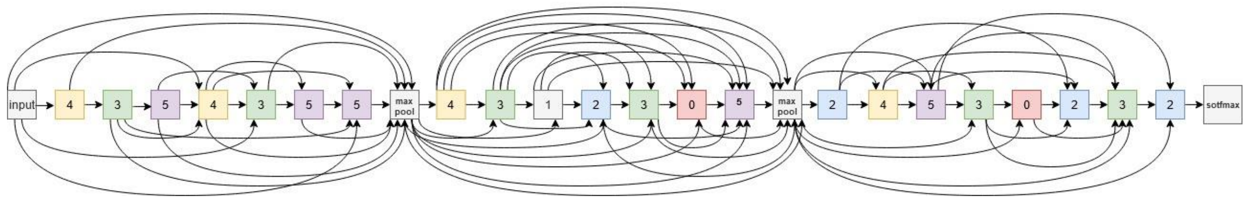


*Fig.2. Our proposed architecture- numbers in the layer corresponds to 0: Conv 3*3, 1: Separable conv 3*3, 2: Conv5*5, 3: Separable conv5*5, 4: Average pooling, 5: Max pooling*

TABLE II. PEEFORMANCE OF OUR PROPSED MODEL

$F1_{Normal}$	$F1_{AF}$	$F1_{Other}$	F-total ± std	Accuracy ± std
91.39	83.87	72.34	82.45 ± 0.2	84.15 ± 0.6

Physionet 2017 dataset, using cross validation. Table III summarizes the F1-score of top three deep learning models during the challenge) as well as the model proposed by Li et al. [19] after the challenge.

Table III shows that our proposed neural model performs very well and arguably outperforms currently available deep learning models on the task of AF detection from single-lead ECGs. However, although we, and the authors we cite, are using cross validation on the publicly available subset of PhysioNet 2017 dataset, it does not mean that the experimental protocol is identical in all aspects. Therefore, we cannot be 100% confident that if we ran their model under our protocol, we would report exactly the same realist as they did.

One of the strengths of our approach is that it does not require any pre-processing or post-processing. It can be seen that Rubin's [16] result is close to our model; however, as explained in Section II, they applied pre-processing (evaluating the quality of signals in order to discard noisy samples from training set, and QRS detection) and post-processing (re-classification of samples of "normal rhythm and "other arrhythmia" classes using hand-crafted features).

In Table IV, we compare our model with the second winner of PhysioNet challenge [15] in more detail, as they didn't apply pre-processing, and they transformed the input signals into spectrograms, similar to our approach. The CNN model proposed by Zilhmann et al. [15] achieved an F1-score of 75.8% (without data augmentation) and 79.0% (with data augmentation). Data augmentation improved the score since it made the dataset larger (by adding synthetic data), and also reduced the impact of imbalanced dataset. Using LSTM blocks (memory units) improved their model's efficiency. Zilhmann' highest F1-score (79.2%) was reported by their proposed CRNN model (LTSM blocks) and using data augmentation.

TABLE III. COMPARISON WITH THE EXISTING DEEP LEARNING MODELS FOR AF DETECTION

Method	F1-score (%)- cross validation
Rubin et al. [16]	82.00
Zihlmann et al.[15]	79.2
Andreotti et al. [17]	74.00
Li et al.[19]	79.59
Our proposed model	**82.45 ± 0.2**

TABLE IV. COMPARISON OF OUR PROPOSED MODEL WITH SECOND WINNER OF PHYSIONET CHALLENGE 2017

Method	Data augmentation	F1-score %	Accuracy %
Zilhmann-CNN[15]	×	75.8	80.5
Zilhmann-CNN[15]	✓	79.0	81.2
Zilhmann-CRNN [15]	×	74.6	79.2
Zilhmann-CRNN[15]	✓	79.2	82.3
Our proposed model	×	**82.45± 0.2**	**84.15± 0.6**

Our proposed model's F1-score and accuracy are 82.45% and 84.15%, respectively, without adding synthetic data. Our experiments support the conclusion that our model outperformed Zilhmann's best model; however, we believe modification of our search space (adding memory units) might further improve our current results.

Our proposed model seems to outperform other existing deep learning methods for Atrial fibrillation detection; however, the model proposed by PhysioNet 2017 winners had a higher score than our model (they used manual feature engineering with decision tree). Nevertheless, using hand-crafted features on AF detection is highly dependent on the pre-processing steps such as: noise removal and QRS detection and therefore, is vulnerable to error. It is our contention that further investigation into modification of our NAS search space, would lead to outperforming approaches that utilizing engineered features.

V. CONCLUSION

Accurate and early detection of atrial fibrillation is an essential task for reduction of development of cardiac disease. In order to design a robust and automatic atrial fibrillation detection model, we have employed a neural architecture search algorithm (ENAS), which designs different neural models and determines the most accurate CNN architecture among those models. The accuracy and F1-score of our proposed model on publicly available subset of "PhysioNet 2017" dataset are 84.15% ± 0.6 and 82.45% ± 0.2, respectively. These results (albeit with larger size of the model) support the conclusion that our model outperforms other currently available deep learning methods. However, we didn't validate the efficacy of the ENAS approach (ENAS could have discarded some models that perform equally or better than the selected model). In future work, it would be desirable to confirm the efficacy of the ENAS approach on the task of atrial fibrillation detection by sampling a set of discarded architectures and evaluating their performance. Additionally, our model is expected to further improve by adding LSTM blocks to our search space, as they incorporate memory units which are beneficial for feature extraction from time series data.

REFERENCES

[1] Rattanawong, P., et al., *Atrial fibrillation is associated with sudden cardiac death: a systematic review and meta-analysis.* Journal of Interventional Cardiac Electrophysiology, 2018. **51**(2): p. 91-104.

[2] Kirchhof, P., et al., *2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS.* European journal of cardio-thoracic surgery, 2016. **50**(5): p. e1-e88.

[3] Chugh, S.S., et al., *Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study.* Circulation, 2014. **129**(8): p. 837-847.

[4] Björck, S., et al., *Atrial fibrillation, stroke risk, and warfarin therapy revisited: a population-based study.* Stroke, 2013. **44**(11): p. 3103-3108.

[5] Qian, Y., et al., *Very deep convolutional neural networks for noise robust speech recognition.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016. **24**(12): p. 2263-2276.

[6] Pham, H., et al., *Efficient neural architecture search via parameter sharing.* arXiv preprint arXiv:1802.03268, 2018.

[7] Miikkulainen, R., et al., *Evolving deep neural networks*, in *Artificial Intelligence in the Age of Neural Networks and Brain Computing*. 2019, Elsevier. p. 293-312.

[8] Clifford, G.D., et al. *AF Classification from a short single lead ECG recording: the PhysioNet/Computing in Cardiology Challenge 2017*. in *2017 Computing in Cardiology (CinC)*. 2017. IEEE.

[9] Datta, S., et al. *Identifying normal, AF and other abnormal ECG rhythms using a cascaded binary classifier*. in *2017 Computing in Cardiology (CinC)*. 2017. IEEE.

[10] Zabihi, M., et al. *Detection of atrial fibrillation in ECG hand-held devices using a random forest classifier*. in *2017 Computing in Cardiology (CinC)*. 2017. IEEE.

[11] Teijeiro, T., et al. *Arrhythmia classification from the abductive interpretation of short single-lead ECG records*. in *2017 Computing in Cardiology (CinC)*. 2017. IEEE.

[12] Hong, S., et al. *ENCASE: An ENsemble ClASsifiEr for ECG classification using expert features and deep neural networks*. in *2017 Computing in Cardiology (CinC)*. 2017. IEEE.

[13] Koelstra, S., et al., *Deap: A database for emotion analysis; using physiological signals.* IEEE transactions on affective computing, 2011. **3**(1): p. 18-31.

[14] Längkvist, M., L. Karlsson, and A. Loutfi, *A review of unsupervised feature learning and deep learning for time-series modeling.* Pattern Recognition Letters, 2014. **42**: p. 11-24.

[15] Zihlmann, M., D. Perekrestenko, and M. Tschannen. *Convolutional recurrent neural networks for electrocardiogram classification*. in *2017 Computing in Cardiology (CinC)*. 2017. IEEE.

[16] Rubin, J., et al. *Densely connected convolutional networks and signal quality analysis to detect atrial fibrillation using short single-lead ECG recordings*. in *2017 Computing in Cardiology (CinC)*. 2017. IEEE.

[17] Andreotti, F., et al. *Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ECG*. in *2017 Computing in Cardiology (CinC)*. 2017. IEEE.

[18] Hannun, A.Y., et al., *Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network.* Nature medicine, 2019. **25**(1): p. 65.

[19] Li, X., et al. *Domain Knowledge Guided Deep Atrial Fibrillation Classification and Its Visual Interpretation*. in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2019.

[20] Gröchenig, K., *Foundations of time-frequency analysis*. 2013: Springer Science & Business Media.

[21] Nester ov, Y.E. *A method for solving the convex programming problem ith convergence rate O (1/k^ 2)*. in *Dokl. akad. nauk Sssr*. 1983

[22] He, K., t al. *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*. in *Proceedings of the IEEE international conference on computer vision*. 2015.