Study On ECG Data Dependency For Atrial Fibrillation Detection Based On Residual Networks

Hyo-Chang Seo
Asan Medical Institute of Convergence Science and Technology, University of Ulsan College of Medicine
Seok Oh
Asan Medical Institute of Convergence Science and Technology, University of Ulsan College of Medicine
Segyeong Joo (sgjoo@amc.seoul.kr)
Asan Medical Institute of Convergence Science and Technology, University of Ulsan College of Medicine

Research Article

Keywords: Atrial fibrillation, PhysioNet, electrocardiogram, neural network

Posted Date: May 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-439203/v1

Abstract

Atrial fibrillation (AF) is an arrhythmia that can cause blood clot and may lead to stroke and heart failure. To detect AF, deep learning-based detection algorithms have recently been developed. However, deep learning models were often trained with limited datasets and were evaluated within the same datasets, which makes their performance generally drops on the external datasets, known as data dependency. For this study, three different databases from PhysioNet were used to investigate the data dependency of deep learning-based AF detection algorithm using the residual neural network (Resnet). Resnet 18, 34, 50 and 152 model were trained with raw electrocardiogram (ECG) signal extracted from independent database. The highest accuracy was about 98–99% which is evaluation results of test dataset from the own database. On the other hand, the lowest accuracy was about 53–92% which was evaluation results of the external dataset extracted from different source. There are data dependency according to the train dataset and the test dataset. However, the data dependency decreased as a large amount of train data.

Introduction

Atrial fibrillation (AF) is the most common cardiac arrhythmia which is irregular or rapid heartbeat. The number of AF patients is expected to increase by 12.1 million[1] and the related cost of AF is estimated at USD 6–26 billion per year[2] in US. Furthermore, AF can not only form thrombosis, which is the can cause stroke, but also affect heart failure and other heart disease[3]. AF rises the risk of stroke five times[4] and the risk of death twice[5], compared to healthy a person. Therefore, considering the social cost of healthcare and the quality of life, early and accurate detection of AF is important and beneficial. In the clinical environment, the detection of AF is manually done with visual inspection of the electrocardiogram (ECG) recordings. Cardiologists inspect the ECG recordings collected about 24 hours by ambulatory ECG device (Holter monitor). However, manually inspecting large amounts of ECG recordings can be tedious and time-consuming[6,7]. Also, time and frequency components of ECG are very subtle for accurate and consistent manual inspection[7]. A study showed that the manual inspections of many primary care practitioners are insufficient for accurate detection of AF[8]. This implies that there are limitations in detecting hidden patterns of AF and extensive training of clinician is necessary to find AF effectively.

Recently, with the emerging research on artificial intelligence (AI), automatic AF detection algorithms have been developed to resolve above problems. Reported AI based AF detection algorithms generally utilizes machine learning or deep learning techniques. Machine learning based AF detection algorithms employ features, which are measured or calculated by original ECG signal[9–17]. This feature extraction step is important for the machine learning based AF detection algorithms. However, it is generally the most time-consuming process in developing those algorithms. Recent year, deep learning-based AF detection algorithms have been developed. Deep learning is an AI algorithm that automatically train the computational model to solve complex problems. Model learns a representation of the data through training the multiple processing layers. Afterward, this trained model can be used to predict events on new data with performance beyond human-level. Due to these advantages, deep learning techniques is

widely used nowadays in various healthcare applications such as medical imaging, drug discovery, and genomics. However, with the high performance of the most deep learning-based algorithm developed in healthcare field, they suffer from data dependency, which means that the developed algorithm generally works well within the database used for the development but the performance generally drops when the algorithm was used in other database. Unlike general applications of deep learning, healthcare data is highly heterogeneous, ambiguous, noisy, and incomplete. Furthermore, healthcare data collected from different medical institutions, hospitals, or devices is uneven and no uniform which can lead to worthless analysis[24]. To avoid adverse effect on patient, thorough validation is necessary before applying deep learning-based algorithm to healthcare data. The validation using external data collected from various devices or institutions is important to evaluate the generalization performance of deep learning-based algorithm. However, deep learning-based algorithm is generally validated by the internal database used for the development. For example, in the medical imaging application included radiology, ophthalmology, and pathology diagnostic analysis, most deep learning-based algorithms did not employ the validation using external database[25].

Deep learning model build with AF data collected from the different setting, such as sampling frequency, resolution, and acquisition environment, may suffer from data dependency. There are several open databases for studying heart related research. Many previously reported papers for making AF detection algorithm utilized these databases. However, most research does not consider the data dependency, which can be problem when the algorithms are used in real environment. In this study, to quantity this data dependency, we experimentally investigated the data dependency of deep learning model of AF classification build with those open databases.

Results

The experiments were executed on python package with Keras, along with the computer environments of Intel(R) Core(TM) i7-6900k CPU 3.20GHz, NVIDIA Geforce GTX 1080 Ti of GPU and Windows 10 operating system.

We evaluated different Resnet models (Resnet 18, 34, 50 and 152) as shown in Table 3. In each trained model, the highest accuracy is about 98 ~ 99% on the test dataset of internal database. On the other hand, the lowest accuracy is about 53 ~ 92% on the test dataset of external database. Initially, the differences between highest and lowest accuracy in Resnet 18 model are 17.26% on the training model of LTAFDB, 18.87% on the AFDB and 44.59% on the MITDB, respectively. Secondly, those of Resnet 34 model are 17.97% on the LTAFDB, 21.24% on the AFDB and 44.10% on the MITDB, respectively. Those of Resnet 50 model are 16.90% on the LTAFDB, 20.44% on the AFDB and 45.89% on the MITDB. Those of Resnet 152 model are 16.42% on the LTAFDB, 19.34% on the AFDB and 42.92% on the MITDB. There is no significant performance difference according to the number of layer, so that following experiments were executed with Resnet 50 model which has medium depth in the models used in the experiment.

Table 3
Performance results of different Resnet models

		Test data		
Trained model		LTAFDB	AFDB	MITDB
Model	Train data			
Resnet 18	LTAFDB	**99.04**	92.03	82.10
	AFDB	85.01	**99.41**	86.68
	MITDB	74.48	70.98	**99.89**
Resnet 34	LTAFDB	**98.70**	92.19	81.14
	AFDB	84.94	**99.27**	78.14
	MITDB	63.65	65.55	**99.78**
Resnet 50	LTAFDB	**98.66**	92.13	83.90
	AFDB	84.35	**99.20**	80.48
	MITDB	68.79	65.61	**99.82**
Resnet 152	LTAFDB	**98.53**	92.00	84.67
	AFDB	83.87	**99.21**	81.14
	MITDB	66.23	64.41	**99.56**

The confusion matrices of Resnet 50 model are represented Figure. 1. In the case of the trained model on LTAFDB, the evaluated case by internal database(LTAFDB) shows the highest true positive rate 98.76% and true negative rate 98.55%. However, the highest false negative rate 5.60% is reported on the evaluated case by the external database (AFDB) and the highest false positive rate 20.23% is reported on the evaluated case by the external database(MITDB). Secondly, in the case of the trained model on AFDB, the highest true positive rate 99.11% and true negative rate 99.43% are resulted on the evaluated case by the internal database (AFDB). However, the highest false negative rate 33.12% is resulted on the evaluated case by the external database (LTAFDB) and the highest false positive rate 21.26% is resulted on the evaluated case by the external database (MITDB). Thirdly, in the case of the trained model on MITDB, the highest true positive rate 98.27% and true negative rate 99.84% are resulted on the evaluated case by the internal database (MITDB). However, the highest false negative rate 76.86% and false positive rate 7.03 % are resulted on the evaluated case by the external database (AFDB).

Next, the ROC curve of Resnet 50 model is shown as Figure. 2. Initially, Figure. 2(a) shows the ROC curves of trained model on LTAFDB. The highest AUC score is 0.9994 on the evaluated case by the internal database (LTAFDB) and the lowest AUC score is 0.9494 on the evaluated case by external database (MITDB). Figure. 2(b) shows the results of trained model on AFDB. The highest AUC score is 0.9993 on

the evaluated case by internal database (AFDB) and the lowest AUC score is 0.9190 on the evaluated case byexternal database (MITDB). Figure. 2(c) shows the results of trained model on MITDB. The highest AUC score is 0.9999 on the evaluated case by the internal database (MITDB) and the lowest AUC score is 0.5296 on the evaluated case by the external database (AFDB).

In order to estimate the data dependency only healthy subjects, we evaluated the trained models on MIT-BIH Normal Sinus Rhythm database (NSRDB)[35] composed to only normal sinus rhythm (“Non-AF” class) recorded from patients had no significant arrhythmias. It is good estimate of the false positive rate in healthy subjects[20]. The trained 50 models on LTAFDB, AFDB and MITDB were performed with the specificity of 97.16, 97.60, and 95.46 and false positive rate of 2.84, 2.40 and 4.55, respectively. It is listed in Table 4.

Table 4
Results of the Resnet 50 model on NSRDB

	NSRDB	
Train data	Sp (%)	Fpr (%)
LTAFDB	97.16	2.84
AFDB	97.60	2.40
MITDB	95.45	4.55

Discussion

We experimentally investigate the data dependency of deep learning-based AF classification using Resnet and raw ECG signal. As indicated in Table 3, the highest accuracy of each trained model was resulted on evaluating the test dataset extracted from own database within all Resnet model used in this study (Resnet 18, 34, 50 and 152). In contrast, the model accuracy was decreased on evaluating the external dataset extracted from different source. Resnet generally shows a good performance without gradient exploding or gradient vanishing even if model is much deeper network. However, the data dependency occurs regardless of the depth in Resnet architecture. Therefore, the deeper network cannot resolve the data dependency. On the unseen data, when the true positive rate increases, the false positive rate also tends to occur higher. Also, the true negative rate and false negative rate also show the same trend. Unlike the evaluation results of own database, if model show a high sensitivity for external data, specificity oppositely is low. Similarly, the high specificity for external data lead to low sensitivity in trained models. These results imply that the trained model may biasedly predicts the external data to be positive or negative.

The MITDB has the imbalanced and smallest amount of data among the databases used in this study. The trained models of those show a largest data dependency in the experiment results. On the contrary,

the LTAFDB has the largest amount of data among them. In the trained model with LTAFDB, the data dependency is lower than other trained models. Also, these results imply that training the deep learning model using the large amount and balanced data can decrease the data dependency on the AF detection algorithm. However, the acquisition and use of large amount AF data may be difficult because of the patient privacy and legal issues for healthcare data.

When evaluate the external data extracted from NSRDB, all trained Resnet 50 models with the LTAFDB, AFDB, and MITDB showed specificity more than 95% and false positive rate about 2–4%. The specificities of the trained models tested with NSRDB was higher than that of the trained models with other databases except the database used for building the model. These results implies that normal sinus rhythm of healthy patients less suffers from data dependency.

The data dependency in building AI models can be caused by several aspects. Data imbalance is one of the most common cause. If a database has AF events far lesser than normal rhythms, this is common in medical databases in general, the performance can be biased and the performance can be drop when tested with external database. Another problem is noise in ECG signals by motion artifact or other reasons. Physical movement of patient when measuring ECG can cause wandering of baseline of ECG or unwanted noises. These noises can be minimized from digital filtering or other signal processing. However, during these processes, distortion or losing characteristic waveform of the original ECG can occur and the processed signals can have different characteristics according to the method of the preprocessing. The performance of AI models can lower due to the difference in preprocessing method of the database. The other problem is discrepancy in measuring hardware. There are several companies making devices for measuring ECG. These devices have different in hardware settings, such as amplifier configuration, filters, and gain, and software settings, such as sampling frequency and resolution. The waveforms from approved ECG devices do not differ largely, and resampling or normalization technique can reduce these problems but not perfectly resolve.

It can be concluded that it is necessary to validate the deep learning based AF detection algorithm using the various external databases at the developing step to avoid the data dependency.

Limitation There are some limitations in this study. Initially, this study was implemented with only one deep learning architecture, Resnet. Evaluation using other deep learning architectures will be helpful in investigating the data dependency. Secondly, “Non-AF” classes in the MITDB and LTAFDB are composed of normal sinus rhythm but “Non-AF” class in the AFDB is composed of all other rhythm because of the absence of normal sinus rhythm annotation. This limitation may lead to low specificity when evaluate the AFDB. However, the performance of sensitivity could be effectively reflected. Thirdly, the data used in this study is from three open-source databases, LTAFDB, AFDB, and MITDB. The using more databases collected from various location, device, and setting will be helpful to effective research results.

Methods

We train the deep learning model using three different AF databases and evaluate the data dependency using not used for training. The training method is described in Figure. 3

Open database The three open databases on Physionet, Long-Term Atrial Fibrillation database (LTAFDB)[26], MIT-BIH Atrial Fibrillation database (AFDB)[27], MIT-BIH Arrhythmia database (MITDB), are used[28]. The LTAFDB consist of 84 subjects with paroxysmal or sustained atrial fibrillation, which is two-channel ECG signal digitized at 128 Hz with 12-bit resolution over 20 mV range for about 24 to 25h[26]. The annotated diseases are normal sinus rhythm (N), supraventricular tachyarrhythmia (SVTA), ventricular tachycardia (VT), atrial fibrillation (AF), ventricular bigeminy (B), ventricular trigeminy (T), idioventricular rhythm (IVR), and atrial bigeminy (AB), sinus bradycardia (SBR). The AFDB is composed of 25 subjects with atrial fibrillation (mostly paroxysmal), which is two-channel ECG signals each sampled at 250 samples per second with 12-bit resolution over a range ± 10 mV for 10 hours. The rhythm annotation files were prepared manually. The rhythm annotations of types are atrial fibrillation (AF), atrial flutter (AFL), atrioventricular junctional rhythm (J), and other rhythms (N)[27]. In this study, two ECG recordings of AFDB (records 00735 and records 03665) were excluded because they are unavailable. The MITDB with 48 half-hour two-channel ECG recordings are included in 47 subjects. The 23 recordings were collected from a mixed population of inpatient (about 60%) and outpatient (about 40%). The remaining 25 recordings were collected from the same set to include less common but clinically significant arrhythmias. The recordings were digitized at 360 samples per second per channel with 11-bit resolution over a 10 mV range[28]. In the LTAFDB and MITDB, we used AF rhythm as “AF” class and normal sinus rhythm as “Non-AF”. In the AFDB, since there is no normal sinus rhythm annotation, we used AF type rhythm as “AF” class and other rhythms type rhythm as “Non-AF” class. All the two-channel ECG signals in each database are used to training and test datasets. The detailed descriptions about the databases are shown on Table 1.

Table 1. Description of three different databases, LTAFDB, AFDB, and MITDB

	LTAFDB	AFDB	MITDB
No. of recording	84 records	25 records	48 records
Channel	2 Ch	2 Ch	2 Ch
Duration	24 to 25 hours	10 hours	Half-hour
Sapling rate	128 Hz	250 Hz	360 Hz
Resolution	12 bit	12 bit	11 bit
Voltage range	20 mV	10 mV	10 mV
Acquisition location	Not reported	Boston’s Beth Israel Hospital	Boston’s Beth Israel Hospital

AF detection model The number of ECG segments used for experiments is listed in Table 2. Since the sampling rate of the LTAFDB, the AFDB and the MITDB are different as 128 Hz, 250 Hz and 360 Hz respectively, the AFDB and MITDB are downsampled at 128 Hz. In the addition, Each ECG data are

normalized by Z-score normalization. The normalized ECG data are divided into a duration of 10 seconds (1280 samples) for input size. The residual network (Resnet) model developed by He[29] is used for AF detection because this Resnet model has recently been used for a lot of studies on cardiac arrhythmia classification[30–32]. Resnet has a good performance without gradient vanishing because of the shortcut of the previous layer X to layer ahead F(X) as shown in Figure. 4(a). If the dimension of X and F(X) are not match, the convolution layer (Conv) and Batch Normalization (BN) are used to match the spatial resolution as shown in Figure. 4 (b). In this study, we employed different 1-D Resnet models to detect AF. The architecture of original Resnet model is converted 2-D to 1-D for training 1-D ECG signal. Additionally, we trained the 1-D Resnet 18, 34, 50, and 152 layers at each three database and compared with each other. For instance, Resnet 50 architecture as shown in Figure. 4(c). We used the cross-entropy which is well known cost function on classification problem. Subsequently, the cost function was minimized by Stochastic Gradient Descent optimizer. The initial learning rate was set to 0.001. A momentum was set to 0.9 and a weight decay set to 0.0001 based on [29]. Mini-batch was size of 32[33]. The learning rate was divided by 10 when error plateaus, and the weight of networks was initialized as in [34]. The early stopping technique was implemented over 10 epochs to avoid overfitting. We train and test different Resnet with 18, 34, 50, and 152 layers composed of training dataset 80% and test dataset 20% using LTAFDB, AFDB, and MITDB, independently. The validation consists of 20% of each train dataset.

Table 2. The number of data segments for AF classification

	Train data		Test data		Total data	
	Non-AF	AF	Non-AF	AF	Non-AF	AF
LTAFDB	460,740	588,990	115,615	146,816	576,355	735,806
AFDB	79,960	53,535	19,944	13,428	99,904	66,963
MITDB	9,696	1,213	2,435	291	12,131	1,504
NSRDB	-	-	314,982	-	314,982	-

Performance evaluation The confusion matrix visualizes the summary of classification results and reports the number of true positive, true negative, false positive and false negative. The confusion matrix is used to visualize the performance of trained model datasets composed of independent databases. The Receiver Operating Characteristic curve (ROC curve) illustrate the Sensitivity against the false positive rate for various decision thresholds. The Area Under Curve (AUC) is a populate evaluation metric which measures the area under entire ROC curve. We used the ROC curve and AUC to present the data dependency according to dataset. Finally, statistics is used to report the data dependency according to different Resnet models (Resnet 18, 34, 50 and 152). They are accuracy defined as the proportion of correctly classified segments among the total number of segments, sensitivity defined as the proportion of true positive among the total number of positive segments, and specificity defined as the proportion of true negative among the total number of negative segments.

Declarations

Data availability

Open database used in this study is illustrated in Methods.

Competing interests

The authors declares he has no competing interest.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2021R1A2C1013755)

References

1. Colilla, S. *et al.* Estimates of current and future incidence and prevalence of atrial fibrillation in the U.S. adult population. *Am J Cardiol.* **112**, 1142–1147 https://doi.org/10.1016/j.amjcard.2013.05.063 (2013).
2. Kim, M. H., Johnston, S. S., Chu, B. C., Dalal, M. R. & Schulman, K. L. Estimation of total incremental health care costs in patients with atrial fibrillation in the United States. *Circ Cardiovasc Qual Outcomes.* **4**, 313–320 https://doi.org/10.1161/circoutcomes.110.958165 (2011).
3. Gillis, A. M., Krahn, A. D., Skanes, A. C. & Nattel, S. Management of atrial fibrillation in the year 2033: new concepts, tools, and applications leading to personalized medicine. *Can J Cardiol.* **29**, 1141–1146 https://doi.org/10.1016/j.cjca.2013.07.006 (2013).
4. Wolf, P. A., Abbott, R. D. & Kannel, W. B. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. *Stroke.* **22**, 983–988 https://doi.org/10.1161/01.str.22.8.983 (1991).
5. Benjamin, E. J. *et al.* Impact of atrial fibrillation on the risk of death: the Framingham Heart Study. *Circulation.* **98**, 946–952 https://doi.org/10.1161/01.cir.98.10.946 (1998).
6. Lyon, A., Mincholé, A., Martínez, J. P., Laguna, P. & Rodriguez, B. Computational techniques for ECG analysis and interpretation in light of their contribution to medical advances. *J R Soc Interface.* **15**, https://doi.org/10.1098/rsif.2017.0821 (2018).
7. Martis, R. J., Acharya, U. R. & Adeli, H. Current methods in electrocardiogram characterization. *Comput Biol Med.* **48**, 133–149 https://doi.org/10.1016/j.compbiomed.2014.02.012 (2014).
8. Mant, J. *et al.* Accuracy of diagnosing atrial fibrillation on electrocardiogram by primary care practitioners and interpretative diagnostic software: analysis of data from screening for atrial fibrillation in the elderly (SAFE) trial. *Bmj.* **335**, 380 https://doi.org/10.1136/bmj.39227.551713.AE (2007).
9. Babaeizadeh, S., Gregg, R. E., Helfenbein, E. D., Lindauer, J. M. & Zhou, S. H. Improvements in atrial fibrillation detection for real-time monitoring. *J Electrocardiol.* **42**, 522–526

https://doi.org/10.1016/j.jelectrocard.2009.06.006 (2009).

10. Yaghouby, F., Ayatollahi, A., Bahramali, R., Yaghouby, M. & Alavi, A. H. Towards automatic detection of atrial fibrillation: A hybrid computational approach. *Comput Biol Med.* **40**, 919–930 https://doi.org/10.1016/j.compbiomed.2010.10.004 (2010).
11. Kennedy, A. *et al.* Automated detection of atrial fibrillation using R-R intervals and multivariate-based classification. *J Electrocardiol.* **49**, 871–876 https://doi.org/10.1016/j.jelectrocard.2016.07.033 (2016).
12. Martis, R. J., Acharya, U. R., Prasad, H., Chua, C. K. & Lim, C. M. Automated detection of atrial fibrillation using Bayesian paradigm. *Knowl. Based Syst.* **54**, 269–275 https://doi.org/10.1016/j.knosys.2013.09.016 (2013).
13. Annavarapu, A. & Kora, P. ECG-based atrial fibrillation detection using different orderings of Conjugate Symmetric–Complex Hadamard Transform. *International Journal of the Cardiovascular Academy.* **2**, 151–154 https://doi.org/10.1016/j.ijcac.2016.08.001 (2016).
14. Kora, P., Annavarapu, A., Yadlapalli, P., Sri Rama Krishna, K. & Somalaraju, V. ECG based Atrial Fibrillation detection using Sequency Ordered Complex Hadamard Transform and Hybrid Firefly Algorithm. *Engineering Science and Technology, an International Journal.* **20**, 1084–1091 https://doi.org/10.1016/j.jestch.2017.02.002 (2017).
15. Daqrouq, K., Alkhateeb, A., Ajour, M. N. & Morfeq, A. Neural network and wavelet average framing percentage energy for atrial fibrillation classification. *Comput Methods Programs Biomed.* **113**, 919–926 https://doi.org/10.1016/j.cmpb.2013.12.002 (2014).
16. Asgari, S., Mehrnia, A. & Moussavi, M. Automatic detection of atrial fibrillation using stationary wavelet transform and support vector machine. *Comput Biol Med.* **60**, 132–142 https://doi.org/10.1016/j.compbiomed.2015.03.005 (2015).
17. Tripathy, R. K., Paternina, M. R. A., Arrieta, J. G., Pattanaik, P. & AUTOMATED DETECTION OF ATRIAL FIBRILLATION ECG SIGNALS USING TWO STAGE VMD AND ATRIAL FIBRILLATION DIAGNOSIS INDEX. *Journal of Mechanics in Medicine and Biology.* **17**, 1740044 https://doi.org/10.1142/S0219519417400449 (2017).
18. Faust, O. *et al.* Automated detection of atrial fibrillation using long short-term memory network with RR interval signals. *Comput Biol Med.* **102**, 327–335 https://doi.org/10.1016/j.compbiomed.2018.07.001 (2018).
19. Xia, Y., Wulan, N., Wang, K. & Zhang, H. Detecting atrial fibrillation by deep convolutional neural networks. *Comput Biol Med.* **93**, 84–92 https://doi.org/10.1016/j.compbiomed.2017.12.007 (2018).
20. Andersen, R. S., Peimankar, A. & Puthusserypady, S. A deep learning approach for real-time detection of atrial fibrillation. *Expert Systems with Applications.* **115**, 465–473 https://doi.org/10.1016/j.eswa.2018.08.011 (2019).
21. Cao, X., Yao, B. & Chen, B. Atrial Fibrillation Detection Using an Improved Multi-Scale Decomposition Enhanced Residual Convolutional Neural Network. *IEEE Access.* **7**, 89152–89161 https://doi.org/10.1109/ACCESS.2019.2926749 (2019).

22. Faust, O., Kareem, M., Shenfield, A., Ali, A. & Acharya, U. R. Validating the robustness of an internet of things based atrial fibrillation detection system. *Pattern Recognit. Lett.* **133**, 55–61 https://doi.org/10.1016/j.patrec.2020.02.005 (2020).
23. Cao, P. *et al.* A novel data augmentation method to enhance deep neural networks for detection of atrial fibrillation. *Biomedical Signal Processing and Control.* **56**, 101675 https://doi.org/10.1016/j.bspc.2019.101675 (2020).
24. Xu, J. *et al.* Federated Learning for Healthcare Informatics. *J Healthc Inform Res.* 1–19 https://doi.org/10.1007/s41666-020-00082-4 (2020).
25. Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y. & Park, S. H. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. *Korean J Radiol.* **20**, 405–410 https://doi.org/10.3348/kjr.2019.0025 (2019).
26. Petrutiu, S., Sahakian, A. V. & Swiryn, S. Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. *Europace.* **9**, 466–470 https://doi.org/10.1093/europace/eum096 (2007).
27. Moody, G. B. & Mark, R. G. *Computers in Cardiology.* **Vol. 10**, 227–230 (1983).
28. Moody, G. B. & Mark, R. G. The impact of the MIT-BIH arrhythmia database. *IEEE Eng Med Biol Mag.* **20**, 45–50 https://doi.org/10.1109/51.932724 (2001).
29. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778, doi:10.1109/CVPR.2016.90 (2016).
30. Hannun, A. Y. *et al.* Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med.* **25**, 65–69 https://doi.org/10.1038/s41591-018-0268-3 (2019).
31. Li, Z., Zhou, D., Wan, L., Li, J. & Mou, W. Heartbeat classification using deep residual convolutional neural network from 2-lead electrocardiogram. *J Electrocardiol.* **58**, 105–112 https://doi.org/10.1016/j.jelectrocard.2019.11.046 (2020).
32. Han, C., Shi, L. & ML-ResNet: A novel network to detect and locate myocardial infarction using 12 leads ECG. *Comput Methods Programs Biomed.* **185**, 105138 https://doi.org/10.1016/j.cmpb.2019.105138 (2020).
33. Kandel, I. & Castelli, M. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express.* **6**, 312–315 https://doi.org/10.1016/j.icte.2020.04.010 (2020).
34. He, K., Zhang, X., Ren, S. & Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification 2015 *IEEE International Conference on Computer Vision (ICCV)*, 1026–1034, doi:10.1109/ICCV.2015.123 (2015).
35. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation.* **101**, E215–220

https://doi.org/10.1161/01.cir.101.23.e215 (2000).

Figures

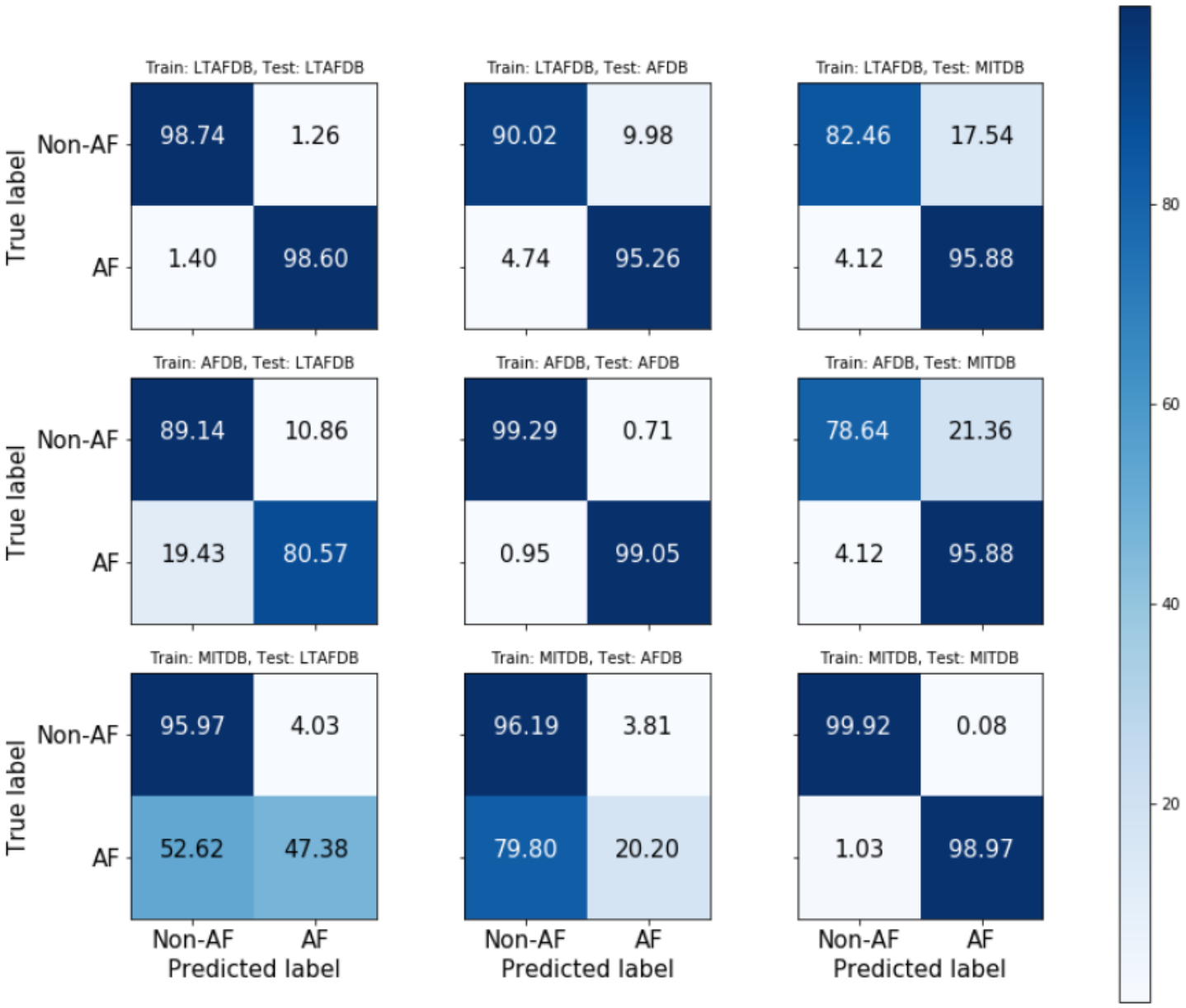

Figure 1

Confusion matrix of the Resnet 50 model for estimating data dependency. The confusion matrices in the first row are results of the trained models on LTAFDB. The second row shows results of the trained models on AFDB and the third row shows results of the trained models on MITDB. The confusion matrices in the first column are results evaluated by LTAFDB. Also, the second column show results evaluated by AFDB and the third column show results evaluated by MITDB.

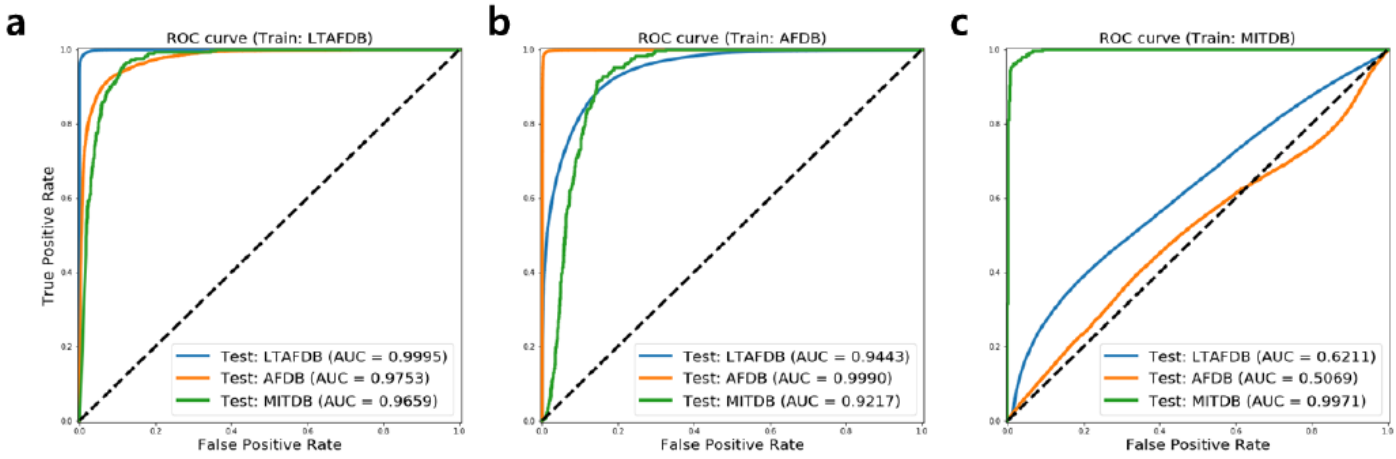

Figure 2

ROC curve of the Resnet 50 model for estimating data dependency. (a) The results of train model on LTAFDB. (b) The results of train model on AFDB. (c) The results of train model on MITDB.

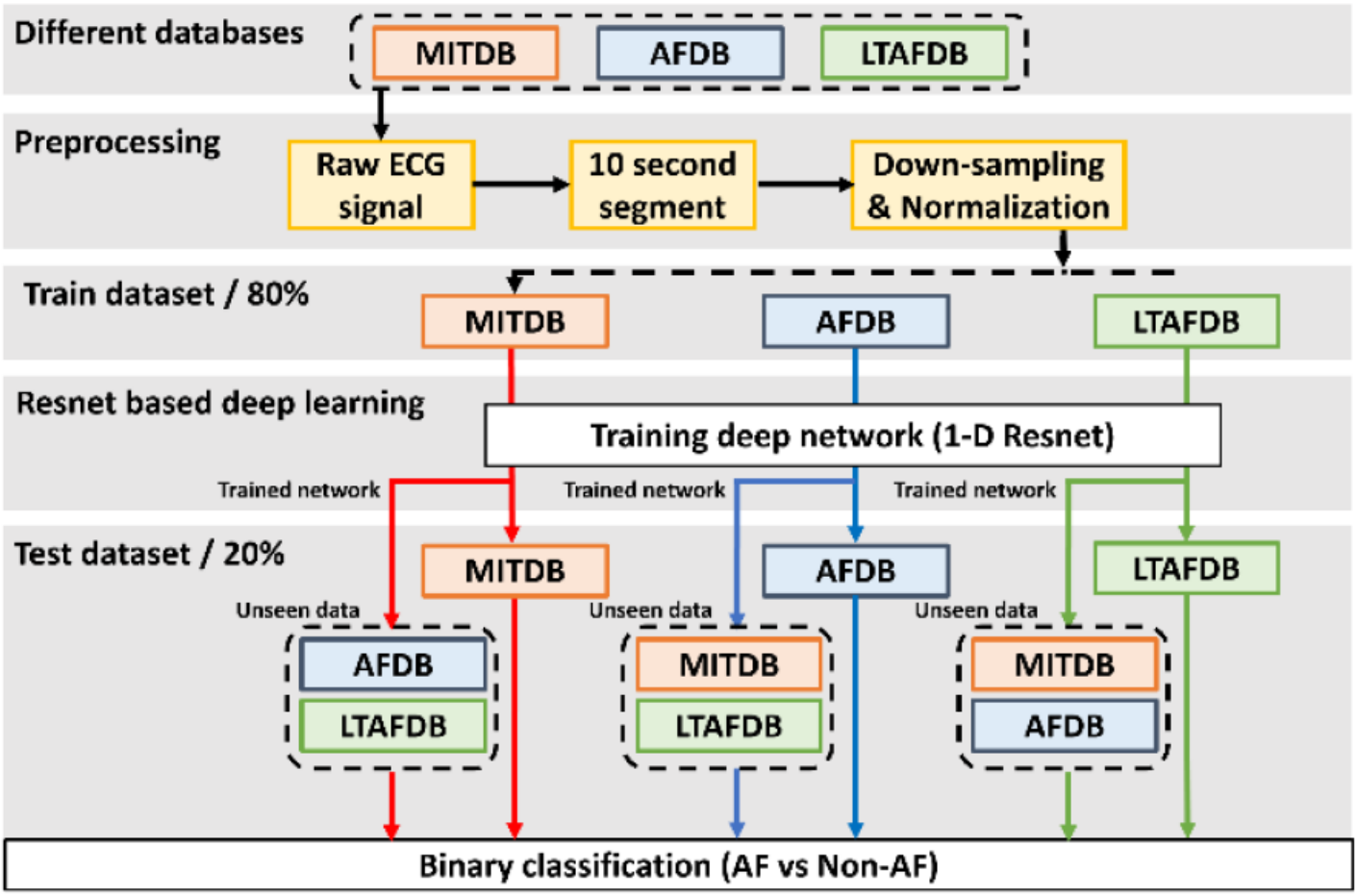

Figure 3

Overview of our study

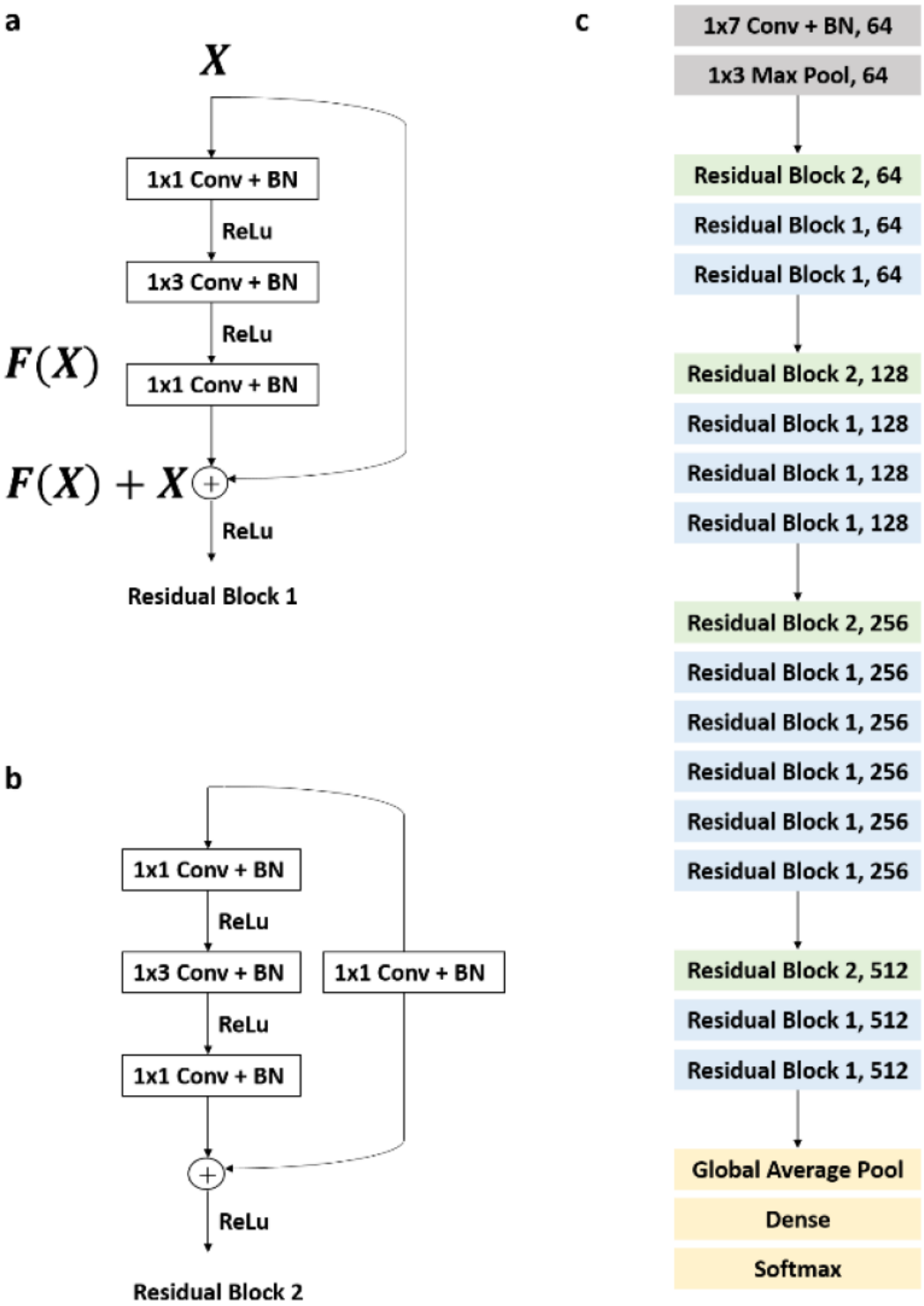

Figure 4

Schematic diagram of residual block. (a) Residual block when previous layer and present layer are same dimensions. (b) Residual block when previous layer and present layer are different dimensions. (c) Resnet 50 architecture.