Ambulatory Atrial Fibrillation Monitoring Using Wearable Photoplethysmography with Deep Learning

Yichen Shen*
yichen.shen@samsung.com
Samsung Strategy and Innovation Center

Maxime Voisin*
maximev@stanford.edu
Department of Computer Science
Stanford University

Alireza Aliamiri
alireza.a@samsung.com
Samsung Strategy and Innovation Center

Anand Avati
Department of Computer Science
Stanford University

Awni Hannun
Department of Computer Science
Stanford University

Andrew Ng
Department of Computer Science
Stanford University

ABSTRACT

We develop an algorithm that accurately detects Atrial Fibrillation (AF) episodes from photoplethysmograms (PPG) recorded in ambulatory free-living conditions. We collect and annotate a dataset containing more than 4000 hours of PPG recorded from a wrist-worn device. Using a 50-layer convolutional neural network, we achieve a test AUC of 95% in presence of motion artifacts inherent to PPG signals. Such continuous and accurate detection of AF has the potential to transform consumer wearable devices into clinically useful medical monitoring tools.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Applied computing** → **Health informatics**.

KEYWORDS

Atrial fibrillation; convolutional neural network; deep learning; ambulatory; PPG

ACM Reference Format:
Yichen Shen, Maxime Voisin, Alireza Aliamiri, Anand Avati, Awni Hannun, and Andrew Ng. 2019. Ambulatory Atrial Fibrillation Monitoring Using Wearable Photoplethysmography with Deep Learning. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3292500.3330657

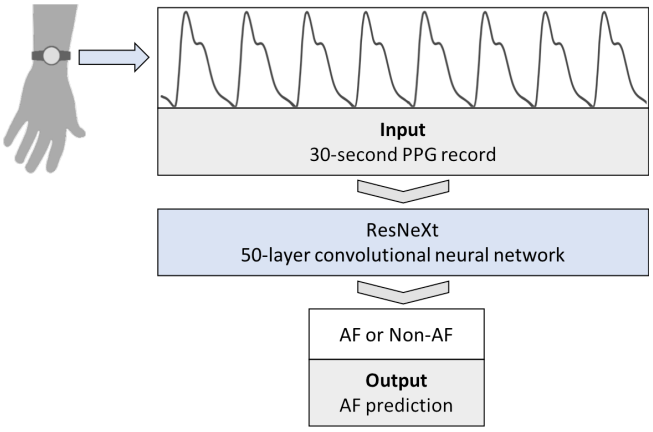


Figure 1: Our trained convolutional neural network correctly detects Atrial Fibrillation (AF) from other rhythms (Non-AF) on this PPG recorded with a wrist-wearable device

1 INTRODUCTION

Atrial fibrillation (AF) is the most common cardiac arrhythmia, affecting between 2.7 million and 6.1 million adults in the United States. This number is expected to double over the next 25 years [8]. AF is a risk factor for blood clots, cognitive impairment, heart failure and stroke [15, 16]. Sub-clinical or silent AF are mostly undetected. Silent AF results in a quarter of all ischemic strokes [13].

The diagnosis is usually performed by cardiologists by observing the electrical activity of the heart in an electrocardiogram (ECG) typically measured with a cardiac event recorder, a Holter monitor or a chest patch. Recently, automatic detection of AF using ECG signals has achieved expert-level performance by leveraging deep neural networks [11, 23].

Photoplethysmography (PPG) is an emerging technology that enables non-invasive heart rhythm measurement through optical sensing. A PPG sensor detects blood volume changes in the microvascular bed of tissue using a low intensity light. The optical mechanism PPG sensors use to measure blood volume change allows them to be placed in wearable devices like smartwatches.

Using PPG sensors to detect AF has several advantages over ECG sensors. A PPG sensor can measure continuously and does not require active participation from the user, unlike ECG event recorders which must be activated by the user at the onset of any symptoms. Because of this, PPG can more accurately quantify AF burden, which is the percentage of time a subject's heart rhythm is in AF. AF burden is a much more indicative risk factor for heart attacks than the binary presence or absence of AF [4, 9]. Another advantage of PPG sensors is that they are already embedded in

*Both authors contributed equally to this research.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330657

mainstream smartwatches which are deployed on mass scale. PPG-based monitoring via a smartwatch is seamless and can be activated over long periods of time with minimal discomfort compared to ECG-based monitors. Hence, continuous AF monitoring using PPG sensors in mainstream smartwatches has the opportunity to be a more convenient, cost-effective solution to systematic, proactive AF screening. This would help detect challenging AF cases such as paroxysmal and silent AF which are often not diagnosed by opportunistic, reactive ECG-based AF screening [7].

PPG-based AF detection has received traction over the past few years. Early attempts leveraged hand-crafted features about inter-beat intervals in PPGs [3, 18, 20–22, 26, 27], while recent approaches trained deep neural networks on PPGs to detect AF [1, 10, 24]. However, PPGs used in these studies were collected in controlled environments often inside a hospital, or were only a few minutes long. Hence, it is unclear how the results of these studies would degrade in ambulatory and free-living conditions. A few attempts were made to detect AF from PPG in ambulatory free-living conditions for prolonged periods of time. These approaches either obtained moderate performance [28], or deleted a significant portion of PPG segments – e.g at least 33% of PPGs [2]. This is largely due to the presence of noise and motion artifacts which corrupt the PPG. As a result, previous attempts have not been able to accurately identify AF episodes in PPG collected in an ambulatory free-living setting for a prolonged period of time.

In this work, we present the first model to continuously and accurately detect AF episodes in PPG collected in an ambulatory free-living setting. The model achieves an AUC of 95% on the test set. Furthermore we do not discard any PPG segment and show robustness to motion artifacts. To achieve these results, we train a 50-layer convolutional neural network to detect AF on more than 4000 hours of PPG signals collected from 81 patients. Our work can be used for challenging downstream tasks like measuring AF burden in ambulatory conditions.

2 MODEL

2.1 Problem Formulation

Our goal is to detect AF episodes in a continuous PPG signal collected from free-living subjects. We extract consecutive 30-second records from the full PPG recording. For each 30-second record x, our model outputs a binary score $y \in \{0, 1\}$ indicating respectively the absence or presence of AF. We optimize the binary cross-entropy objective function

$$\mathcal{L} = -\sum_{i=1}^{N} y^{(i)} \log p(y = 1 \mid x^{(i)}) + (1 - y^{(i)}) \log p(y = 0 \mid x^{(i)}),$$

where i is the index of the PPG record (there are N records in total) and $p(y = l \mid x^{(i)})$ is the probability that the network assigns to label l given the input record $x^{(i)}$.

2.2 Model Architecture and Training

The AF prediction network is a 1D convolutional neural network (CNN). The input to the network is a 30-second PPG record sampled at 20 Hz. For each record we apply a Finite Impulse Response (FIR) low-pass filter with a cutoff frequency of 5Hz. We also scale the input record to have a mean of zero and a unit variance. The output of the network is a binary label indicating the absence or presence of AF in the input record.

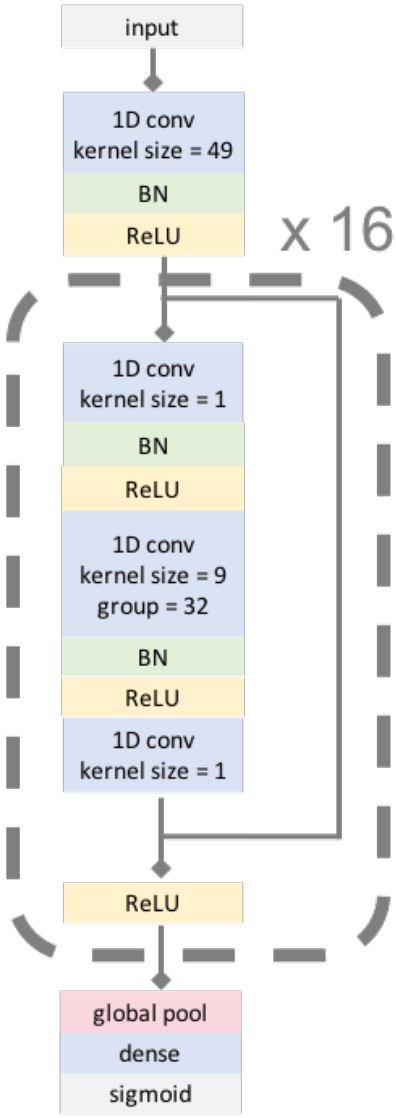


Figure 2: The architecture of the network. The network consists of 49 layers of convolution followed by a global pooling layer, a fully-connected layer and a sigmoid.

The high-level architecture of the network is shown in Figure 2. The CNN consists of 49 layers of 1D convolutions. This is followed by a global average pooling layer, a dense layer and a sigmoid layer to produce an output between 0 and 1.

The network consists of an initial 1D convolution layer with kernel size of 49 followed by 16 ResNeXt bottleneck blocks [29] sharing the same topology. The blocks rely on grouped convolutions which yield higher representation power than other state-of-the-art convolutional networks with the same number of parameters. Each block consists of 3 convolutional layers. First, a convolutional bottleneck

layer with kernel size of 1 reduces the number of feature maps. It is followed by a grouped convolution with a kernel size of 9 which provides more expressive power in each block. Finally, a convolutional layer with a kernel size of 1 restores the original number of feature maps. These 16 blocks are grouped into 4 stages containing 3,4,6 and 3 blocks respectively. The spatial map is downsampled at the grouped convolution layer of the first block of each stage. The ResNeXt architecture was initially designed for 2D data. We adapt it for our 1D data. To ensure that each block has roughly the same computational complexity in terms of FLOPs, we downsample the spatial map using stride-4 convolutions – rather than stride-2 convolutions in the 2D architecture – at the grouped convolution layer of the first block of each stage. We also remove the initial pooling layer to avoid downsampling the input record by too much. Finally, we fine-tune the cardinality – number of groups in the grouped convolution – and the bottleneck width of each group. Our best performing network has a cardinality of 32 and a bottleneck width of $4d$, where d starts out as 1 and is incremented at each stage of the network.

In order to make the optimization of such a network tractable, we employ shortcut connections in a similar manner to those found in ResNeXt architectures [29]. The shortcut connections between neural network layers optimize training by allowing information to propagate well in very deep neural networks.

Batch normalization [14] and a rectified linear activation are applied after each convolutional layer. We train the network from scratch and use He initialization for the weights of the convolutional layers [12]. We use the Adam optimizer [17] with the default parameters and minibatches of size 64. We save the best model as evaluated on the validation set during the optimization process.

2.3 Baseline Model

Inter-Beat Interval (IBI) is a feature commonly used for PPG-based AF detection. We implement the IBI algorithm described in [18]. For each 30-second PPG record, we identify beats in the PPG signal and compute the IBIs. The feature-based baseline algorithm then predicts AF by putting a threshold on the IBI variation measured in terms of Root Mean Square of the Successive Differences (RMSSD).

3 DATA

We used two datasets to train the model: the *clinician-annotated* and the *NSR* datasets. The *clinician-annotated* dataset consists of 402 continuous PPG recordings collected from 29 free-living subjects. Each continuous PPG recording is 8 hours long on average. We simultaneously collected a reference ECG for rhythm annotation using an ECG patch. Out of the 29 subjects, 13 have persistent AF throughout their recordings, 2 have persistent normal sinus rhythm, and the remaining 14 display rhythms that change over time – including 8 arrhythmias other than AF and normal sinus rhythm. The *NSR* dataset consists of 341 continuous PPG recordings collected from 53 healthy free-living subjects who self-reported as not having any symptoms of an arrhythmia. Each continuous PPG recording is 3 hours long on average. In summary, the two datasets in aggregate contain 743 continuous PPG recordings, each having a time span of a few hours.

Dataset	train+val subjects	train+val records	test subjects	test records
Clinician Annotated	19	238345	10	147968
NSR	32	76374	20	47879

Table 1: Two datasets were collected from free-living ambulatory subjects. The PPG in the clinician-annotated dataset are fully annotated by clinicians using a reference ECG. The NSR dataset was collected from healthy subjects who report themselves as not having any arrhythmia. We give the total number of subjects and records for both the training and test sets.

All PPGs are recorded using a Samsung wrist-wearable device with a sampling frequency of 20 Hz. The device also records tri-axial acceleration, which is used in Section 4.2 to evaluate the model's robustness to motion artifacts. In the clinician-annotated dataset, the reference ECG is collected from a single-lead, continuous monitoring patch with a sampling frequency of 500Hz. Each ECG is fully annotated by an ECG technician. The expert technician highlights segments of the continuous signal and marks them as corresponding to one of 10 rhythm classes: 8 heart arrhythmias, normal sinus rhythm and noise. All rhythms were labeled from their corresponding onset to offset, resulting in a full segmentation of the ECG. The noise label is assigned when it is impossible to identify the underlying rhythm from the ECG.

We break down the 743 continuous PPG recordings into 510,566 PPG records of 30 seconds. Each 30-second PPG record has one binary label which indicates if the 10-class rhythm segmentation of the corresponding ECG record contains AF. The binary label serves as the ground truth for training and evaluating models. PPG records whose corresponding ECG record is labeled as noise are discarded, since the ground truth rhythm is not known. They represent 1% of the data.

The PPG records are split into a training, validation and test set. We ensure that there is no subject overlap between these sets. We also ensure that each set has almost balanced AF and non-AF records and that the proportion of subjects with respectively persistent AF, persistent normal sinus rhythm and multiple rhythms is similar across each set. The training, validation and test set contain respectively 42, 10 and 30 subjects in total, representing 50%, 12% and 38% of the PPG records, as detailed in Table 1.

The test set contains 147,968 records from the clinician-annotated dataset (10 test subjects) and 47,879 records from the NSR dataset (20 test subjects). Test subjects with persistent AF, persistent normal sinus rhythm and multiple rhythms represent respectively 45%, 33% and 22% of the test records. 50.2% of the test records are labeled AF.

4 RESULTS

Models are compared based on their AUC, area under the Receiver Operating Characteristic (ROC) curve, which is independent of the prevalence of AF in the data. Each point on the ROC curve represents a sensitivity-specificity pair corresponding to a particular

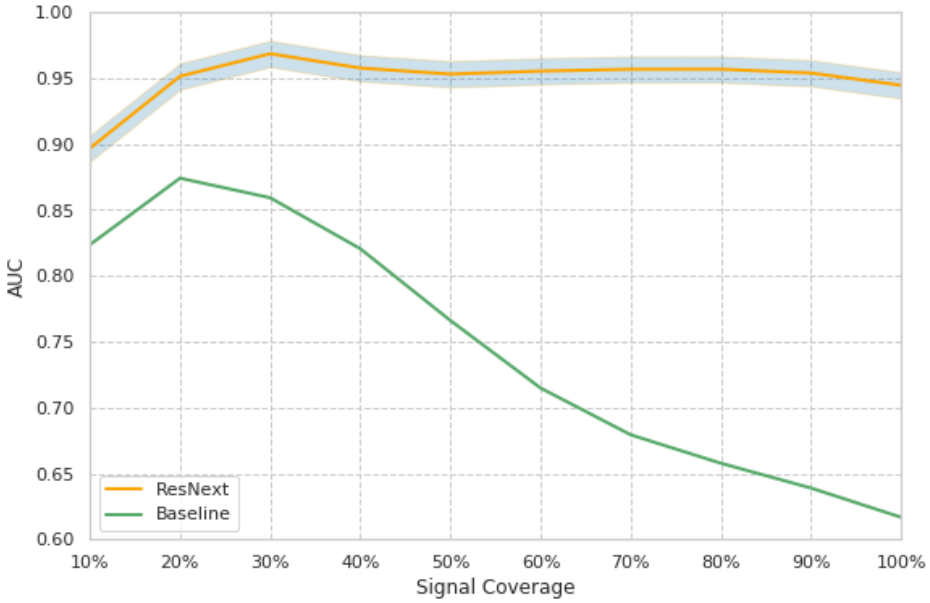


Figure 3: The AUC-coverage curves (see Section 4.1) of our deep learning model and of a baseline feature-based model. The AUC of the deep learning model does not degrade as predictions are done on test records with increasingly high motion intensity. This suggests that our deep learning model is robust to motion artifacts inherent to the ambulatory free-living setting. The performance of the deep learning model is averaged over 3 random seeds.

decision threshold. A model with high AUC enables practitioners to choose the sensitivity-specificity trade-off that best suits their use case. Models are compared based on AUC computed on the test set.

4.1 Impact of motion artifacts

In ambulatory free-living conditions, motion artifacts are expected to corrupt the PPG and degrade the accuracy of AF predictions. We evaluate the robustness of our model to such motion artifacts. To do so, a motion intensity score is assigned to each PPG record. This score is calculated as the standard deviation of the amplitude of the tri-axial acceleration. PPG records in the test set are ordered by increasing motion intensity. We then evaluate the model AUC on the subset of test records whose motion intensity is in the lowest c-th percentile. By sweeping c in $\{10, 20, ..100\}$, an *AUC-coverage curve* is created. Each point (c, p) on the curve indicates that the model has an AUC of p on the test records whose motion intensity is in the lowest c-th percentile. Note that the AUC reported for a coverage $c = 100\%$ is the AUC on the full test set. A model robust to motion artifacts is expected to exhibit a flat AUC-coverage curve.

4.2 Analysis

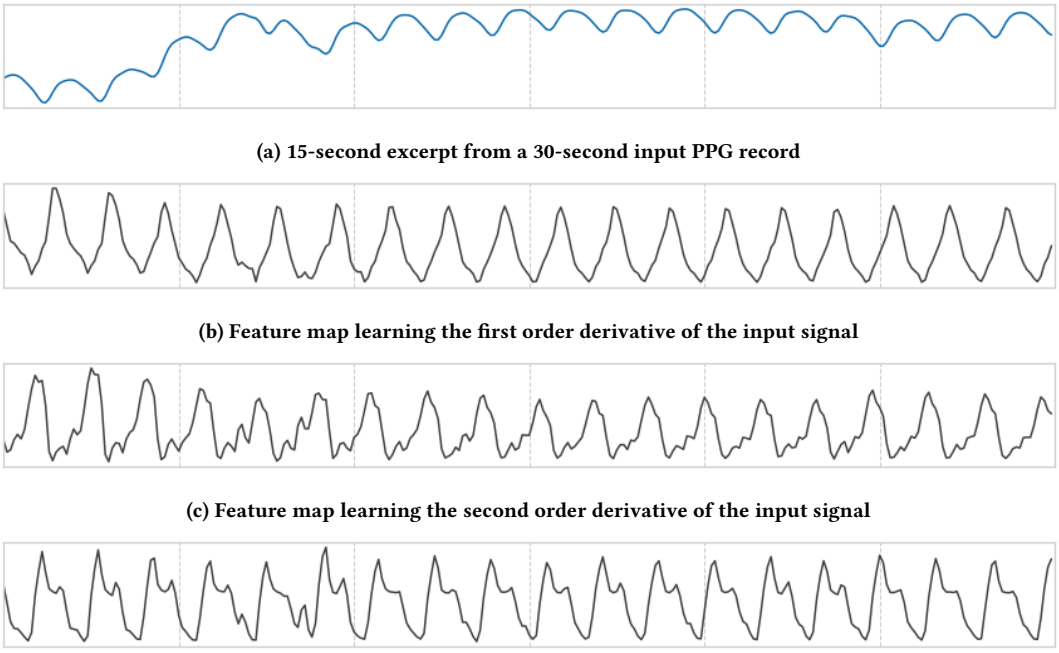
The deep learning model obtains an AUC of 94.8% on the test set. The network largely outperforms the feature-based baseline. Interestingly, the performance of the deep learning model does not degrade when predicting on test records with higher motion intensity, unlike the baseline (see Figures 3 and 8). This suggests that our model is robust to motion artifacts typically encountered in free-living conditions.

Often the errors made by the deep learning model are understandable. First, although the model generally shows robustness against motion artifacts, it is still misled when too much noise corrupts the PPG. Second, we note in Figure 3 that performance decreases on PPG records whose motion intensity is in the bottom 20-th percentile. This is explained by the fact that low motion intensity does not necessarily correspond to clear PPG signal. For example, records collected from improperly worn wearable devices may have low motion intensity while not containing any relevant PPG morphology to predict AF. Finally, the AUC drops to 85.5% on test patients with mixed AF and non-AF rhythms (partial AF). These records may be challenging to classify since they sometimes exhibit multiple arrhythmia other than AF as well as normal sinus rhythm. Also, only a handful of patients with partial AF are in the training set, so the model has relatively few such patients to learn from. These patients also have noisier labels than persistent patients since the boundaries between different rhythms are fuzzy.

5 MODEL INTERPRETATION

5.1 Visualization of the learned low-level feature maps

To understand which discriminative features the model uses to predict AF, we randomly select a test PPG record and visualize the feature maps obtained at the end of the first stage of the deep neural network. As shown in Figure 4, the model learns to remove the low-frequency baseline wander. Some feature maps, shown in (d), identify systolic and diastolic peaks in the PPG. Having irregular beats is a symptom of AF. Other feature maps, shown in (b) and



(a) 15-second excerpt from a 30-second input PPG record

(b) Feature map learning the first order derivative of the input signal

(c) Feature map learning the second order derivative of the input signal

(d) Feature map identifying systolic and diastolic peaks

Figure 4: Examples of feature maps learned by the model at the end of the first stage of the deep learning model, using (a) as input

(c), seem to approximate the first and second order derivatives of the PPG signal. Previous work [6] suggests that first and second order derivatives of a PPG signal contain information about the cardiovascular system such as hypertension and arterial stiffness, and other work [5] suggests that the latter is an important predictor of AF in hypertensive patients. The features learned by our deep neural network therefore seem to be consistent with previous works in the medical field.

5.2 Visualization of salient regions

To further interpret the predictions of our deep learning model, we use the saliency mapping technique [25]. Saliency maps indicate which time steps influence most the prediction by computing the gradient of the binary cross-entropy loss with respect to each input time step. Formally, each time step I_k in the input record $I = [I_0..I_N]$ has a saliency score S_k:

$$S_k = |\frac{\partial L}{\partial I_k}|$$

where L is the binary cross-entropy loss of the AF detection network for input record I.

Figure 5 provides the saliency mapping of two PPG samples. Colored regions are those with high saliency scores. They contribute most to the prediction. The network seems to focus on specific substructures in the PPG morphology such as systolic and diastolic peaks as well as slopes to the left of the systolic peaks.

5.3 Visualization of high-level feature space

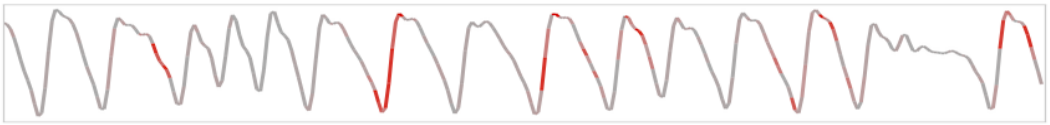
We visualize the high-level representations learned by the model from 8000 randomly selected test records, using the t-SNE method [19]. The representations learned by the last convolutional layer of the network are mapped to a 2D space. The mapping is such that the joint probability of records close to each other in the high-dimensional representation space is similar to their joint probability in the 2D space.

Figure 6 provides the t-SNE visualization of the learned representations. In (a), we color records based on their ground-truth label. We observe that the cluster of AF records is mostly separable from the cluster of non-AF records. In (b), we color the same records based on their motion intensity – records with higher motion intensity have darker colors. We observe a continuous progression of motion intensity in the learned feature space. Records with high motion intensities are clustered in the bottom left quadrant of (b), whereas records with low motion intensities are in the top right quadrant of (b). By comparing (a) and (b), it appears that the cluster of PPG records with highest motion intensity lies close to the decision boundary and corresponds to the most difficult records to predict AF on. This confirms that motion artifacts are a major challenge in PPG-based AF prediction in the ambulatory free-living setup.

6 CONCLUSION

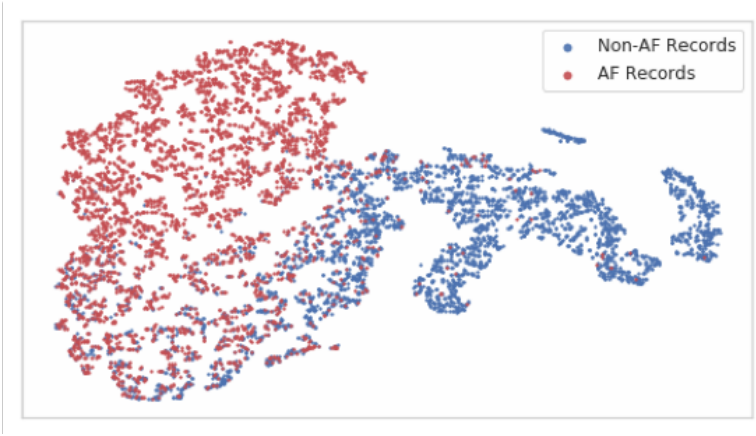
We develop a model which can accurately detect AF from continuous PPG records collected in the ambulatory free-living setting.

(a) Record with normal sinus rhythm. Predicted probability of AF = 0.0001

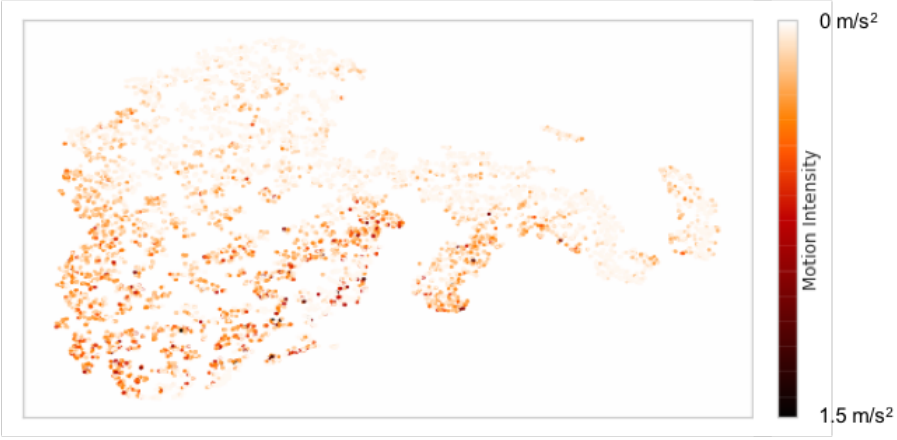


(b) Record with AF. Predicted probability of AF = 0.993

Figure 5: Saliency maps of 15-second excerpts from two 30-second PPG records. Colored regions indicate salient regions which impact the model predictions. The model seems to focus on systolic and diastolic peaks as well as slopes to the left of the systolic peaks.



(a)



(b)

Figure 6: (a) t-SNE plot with the label of each record and (b) t-SNE plot with the motion intensity of each record. AF and non-AF PPG records with low motion intensity are separable. PPG records with high motion intensity, mostly found in the bottom left quadrant, are more challenging to separate.

Key to our approach is a large annotated dataset and a deep convolutional neural network. For future work, we will consider diagnosis and monitoring of other types of arrhythmia as well as optimization of network size and latency for deployment on the wearable devices. With the prevalence of inexpensive wearable devices, high-accuracy continuous arrhythmia monitoring from PPG can not only lower the risk of undiagnosed AF in general but also save precious time and resources from expert clinicians and cardiologists in resource intensive tasks like measuring AF burden. Furthermore, we hope that this technology can eventually provide accurate diagnostic information in places with constrained access to cardiologists and other medical resources.

REFERENCES

[1] Alireza Aliamiri and Yichen Shen. 2018. Deep learning based atrial fibrillation detection using wearable photoplethysmography sensor. In *Biomedical & Health Informatics (BHI), 2018 IEEE EMBS International Conference on*. IEEE, 442–445.

[2] Alberto G Bonomi, Fons Schipper, Linda M Eerikäinen, Jenny Margarito, Ronald M Aarts, Saeed Babaeizadeh, Helma M de Morree, and Lukas Dekker. 2016. Atrial fibrillation detection using photo-plethysmography and acceleration data at the wrist. In *Computing in Cardiology Conference (CinC), 2016*. IEEE, 277–280.

[3] Pak-Hei Chan, Chun-Ka Wong, Yukkee C Poh, Louise Pun, Wangie Wan-Chiu Leung, Yu-Fai Wong, Michelle Man-Ying Wong, Ming-Zher Poh, Daniel Wai-Sing Chu, and Chung-Wah Siu. 2016. Diagnostic performance of a smartphone-based photoplethysmographic application for atrial fibrillation screening in a primary care setting. *Journal of the American Heart Association* 5, 7 (2016), e003428.

[4] Lin Y Chen, Mina K Chung, Larry A Allen, Michael Ezekowitz, Karen L Furie, Pamela McCabe, Peter A Noseworthy, Marco V Perez, and Mintu P Turakhia. 2018. Atrial fibrillation burden: moving beyond atrial fibrillation as a binary entity: a scientific statement from the American Heart Association. *Circulation* 137, 20 (2018), e623–e644.

[5] Antoine Cremer, Marion Lainé, Georgios Papaioannou, Sunthareth Yeim, and Philippe Gosse. 2015. Increased arterial stiffness is an independent predictor of atrial fibrillation in hypertensive patients. *Journal of hypertension* 33, 10 (2015), 2150–2155.

[6] Mohamed Elgendi. 2012. On the analysis of fingertip photoplethysmogram signals. *Current cardiology reviews* 8, 1 (2012), 14–25.

[7] Ben Freedman, John Camm, Hugh Calkins, Jeffrey S Healey, Mårten Rosenqvist, Jiguang Wang, Christine M Albert, Craig S Anderson, Sotiris Antoniou, Emelia J Benjamin, et al. 2017. Screening for atrial fibrillation: a report of the AF-SCREEN International Collaboration. *Circulation* 135, 19 (2017), 1851–1867.

[8] Alan S Go, Dariush Mozaffarian, Véronique L Roger, Emelia J Benjamin, Jarett D Berry, Michael J Blaha, Shifan Dai, Earl S Ford, Caroline S Fox, Sheila Franco, et al. 2013. Heart disease and stroke statistics-2014 update: a report from the American Heart Association. *Circulation* (2013), 01–cir.

[9] Alan S Go, Kristi Reynolds, Jingrong Yang, Nigel Gupta, Judith Lenane, Sue Hee Sung, Teresa N Harrison, Taylor I Liu, and Matthew D Solomon. 2018. Association of Burden of Atrial Fibrillation With Risk of Ischemic Stroke in Adults With Paroxysmal Atrial Fibrillation: The KP-RHYTHM Study. *JAMA cardiology* (2018).

[10] Igor Gotlibovych, Stuart Crawford, Dileep Goyal, Jiaqi Liu, Yaniv Kerem, David Benaron, Defne Yilmaz, Gregory Marcus, and Yihan Li. 2018. End-to-end Deep Learning from Raw Sensor Data: Atrial Fibrillation Detection using Wearables. *arXiv preprint arXiv:1807.10707* (2018).

[11] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine* 25, 1 (2019), 65.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[13] Jeff S Healey, Stuart J Connolly, Michael R Gold, Carsten W Israel, Isabelle C Van Gelder, Alessandro Capucci, CP Lau, Eric Fain, Sean Yang, Christophe Bailleul, et al. 2012. Subclinical atrial fibrillation and the risk of stroke. *New England Journal of Medicine* 366, 2 (2012), 120–129.

[14] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[15] Craig T January, L Samuel Wann, Joseph S Alpert, Hugh Calkins, Joaquin E Cigarroa, Jamie B Conti, Patrick T Ellinor, Michael D Ezekowitz, Michael E Field, Katherine T Murray, et al. 2014. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. *Journal of the American College of Cardiology* 64, 21 (2014), e1–e76.

[16] Shadi Kalantarian, Theodore A Stern, Moussa Mansour, and Jeremy N Ruskin. 2013. Cognitive impairment associated with atrial fibrillation: a meta-analysis. *Annals of internal medicine* 158, 5_Part_1 (2013), 338–346.

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Jinseok Lee, Bersain A Reyes, David D McManus, Oscar Mathias, and Ki H Chon. 2012. Atrial fibrillation detection using a smart phone. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 1177–1180.

[19] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[20] Shamim Nemati, Mohammad M Ghassemi, Vaidehi Ambai, Nino Isakadze, Oleksiy Levantsevych, Amit Shah, and Gari D Clifford. 2016. Monitoring and detecting atrial fibrillation using wearable technology. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 3394–3397.

[21] Ming-Zher Poh, Yukkee Cheung Poh, Pak-Hei Chan, Chun-Ka Wong, Louise Pun, Wangie Wan-Chiu Leung, Yu-Fai Wong, Michelle Man-Ying Wong, Daniel Wai-Sing Chu, and Chung-Wah Siu. 2018. Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms. *Heart* (2018), heartjnl–2018.

[22] Shih-Ming Shan, Sung-Chun Tang, Pei-Wen Huang, Yu-Min Lin, Wei-Han Huang, Dar-Ming Lai, and An-Yeu Andy Wu. 2016. Reliable PPG-based algorithm in atrial fibrillation detection. In *Biomedical Circuits and Systems Conference (BioCAS), 2016 IEEE*. IEEE, 340–343.

[23] Supreeth P Shashikumar, Amit J Shah, Gari D Clifford, and Shamim Nemati. 2018. Detection of Paroxysmal Atrial Fibrillation using Attention-based Bidirectional Recurrent Neural Networks. *arXiv preprint arXiv:1805.09133* (2018).

[24] Supreeth Prajwal Shashikumar, Amit J Shah, Qiao Li, Gari D Clifford, and Shamim Nemati. 2017. A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology. In *Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on*. IEEE, 141–144.

[25] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[26] Dainius Stankevičius, Andrius Petrėnas, Andrius Sološenko, Mantas Grigutis, Tomas Januškevičius, Laurynas Rimševičius, and Vaidotas Marozas. 2016. Photoplethysmography-based system for atrial fibrillation detection during hemodialysis. In *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*. Springer, 79–82.

[27] Sung-Chun Tang, Pei-Wen Huang, Chi-Sheng Hung, Shih-Ming Shan, Yen-Hung Lin, Jiann-Shing Shieh, Dar-Ming Lai, An-Yeu Wu, and Jiann-Shing Jeng. 2017. Identification of atrial fibrillation by quantitative analyses of fingertip photoplethysmogram. *Scientific reports* 7 (2017), 45644.

[28] Geoffrey H Tison, José M Sanchez, Brandon Ballinger, Avesh Singh, Jeffrey E Olgin, Mark J Pletcher, Eric Vittinghoff, Emily S Lee, Shannon M Fan, Rachel A Gladstone, et al. 2018. Passive detection of atrial fibrillation using a commercially available smartwatch. *JAMA cardiology* 3, 5 (2018), 409–416.

[29] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 5987–5995.

A DISTRIBUTION OF MOTION INTENSITY SCORES

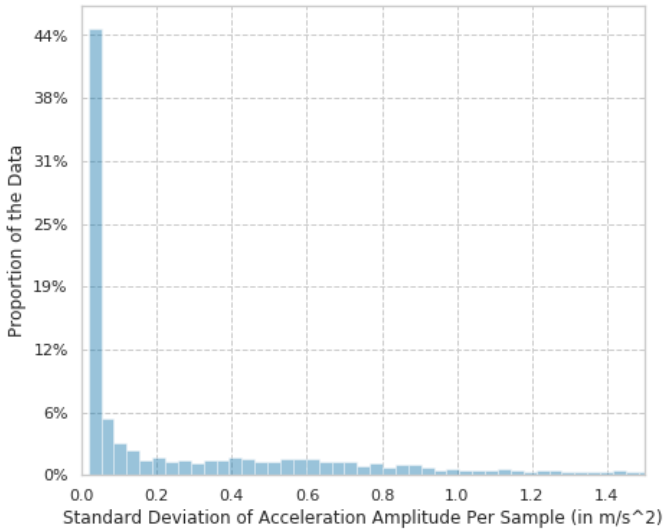


Figure 7: Distribution of motion intensity scores of the 30-second PPG records. Motion intensity is calculated as the standard deviation of the amplitude of the tri-axial acceleration.

B ROC CURVES OF THE DEEP LEARNING AND BASELINE MODELS

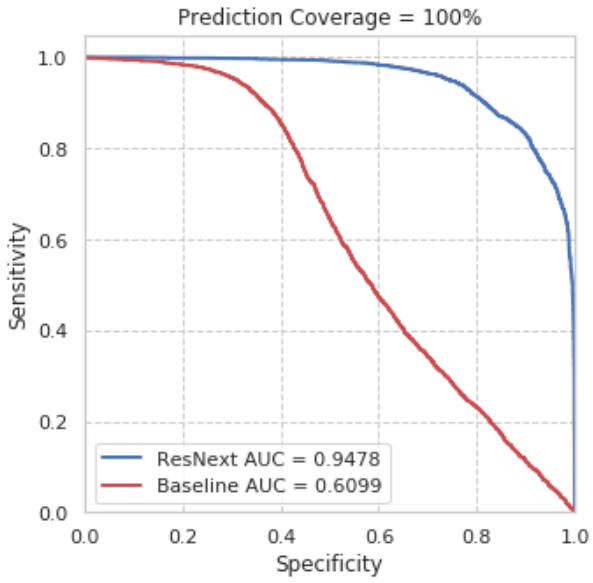


Figure 8: ROC curves (specificity vs sensitivity) of the deep learning model and a baseline model deployed on the full test set.