End-to-end Deep Learning from Raw Sensor Data: Atrial Fibrillation Detection using Wearables

Igor Gotlibovych
Jawbone Health
London, UK
igor.gotlibovych@gmail.com

Stuart Crawford
Jawbone Health
San Francisco, California, USA
scrawford@jawbone.com

Dileep Goyal
Jawbone Health
San Francisco, California, USA
dgoyal@jawbone.com

Jiaqi Liu
Jawbone Health
San Francisco, California, USA
jliu@jawbone.com

Yaniv Kerem
Jawbone Health
San Francisco, California, USA
ykerem@jawbone.com

David Benaron
Jawbone Health
San Francisco, California, USA
dbenaron@jawbone.com

Defne Yilmaz
UCSF
San Francisco, California, USA
defne.yilmaz@ucsf.edu

Gregory Marcus
UCSF
San Francisco, California, USA
greg.marcus@ucsf.edu

Yihan (Jessie) Li*
Jawbone Health
London, UK
jessieli@jawbone.com

ABSTRACT

We present a convolutional-recurrent neural network architecture with long short-term memory for real-time processing and classification of digital sensor data. The network implicitly performs typical signal processing tasks such as filtering and peak detection, and learns time-resolved embeddings of the input signal.

We use a prototype multi-sensor wearable device to collect over 180 h of photoplethysmography (PPG) data sampled at 20 Hz, of which 36 h are during atrial fibrillation (AFib).

We use end-to-end learning to achieve state-of-the-art results in detecting AFib from raw PPG data. For classification labels output every 0.8 s, we demonstrate an area under ROC curve of 0.9999, with false positive and false negative rates both below 2×10^{-3}.

This constitutes a significant improvement on previous results utilising domain-specific feature engineering, such as heart rate extraction, and brings large-scale atrial fibrillation screenings within imminent reach.

CCS CONCEPTS

•Computing methodologies → Neural networks; •Applied computing → Consumer health; Health informatics;

KEYWORDS

atrial fibrillation, convolutional recurrent neural network, time series classification, wearable devices

*corresponding author

KDD'18 Deep Learning Day, London, UK

DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

1.1 Atrial fibrillation

Atrial fibrillation (AFib) is a condition characterised by an irregular and often rapid heartbeat due to abnormalities in the heart's electrical activity. It affects between 2–3% of the population [4], yet as many as 50% of cases remain undiagnosed for 5+ years [50]. It causes a range of complications, including stroke [42], yet can be managed successfully if diagnosed early [25]. Despite significant interest, AFib detection is complicated by several factors: First, it typically relies on an electrocardiogram (ECG) recorded in a hospital setting. Second, due to intermittent nature of the condition, patients may not exhibit any symptoms at the time of the recording, and may require prolonged monitoring. Finally, diagnosis is made by trained cardiologists, and screening efforts are thus difficult to scale to the larger population.

1.2 Wearable devices for AFib diagnostics

Photoplethysmography (PPG) has been proposed as a lower-cost alternative to ECG for the purpose of AFib detection. PPG heart rate monitors have already found wide-spread use in wearable consumer devices such as fitness trackers and smart watches. Unlike ECG, PPG measures changes in the intensity of light reflected by the user's skin due to varying volume and oxygenation of blood in the capillaries [2]. In a recent study [44], it was shown that heart rate readings from an Apple Watch could be useful in detecting AFib.

In the present study, we develop a neural-network-based algorithm to detect AFib from raw PPG signal. The sensor signal is provided by a wrist-worn prototype fitness tracker device, and sampled continuously at 20 Hz. By training a neural network to perform all stages of feature extraction and classification, we achieve performance far superior to what is possible from heart rate features alone.

Table 1: Train and test data.

	source	subjects	rhythm	duration [h]
train	UCSF*	29	AFib‡	30
			NSR§	15
	internal†	13	NSR§	100
test	UCSF*	7	AFib‡	6
			NSR§	3
	internal†	4	NSR§	25

* patients undergoing cardioversion, awake
† volunteers with no known arrhythmias, asleep
‡ atrial fibrillation
§ normal sinus rhythm

2 DATA

The intermittent nature of AFib presents significant challenges to data collection. We collaborate with the University of California, San Francisco (UCSF) Division of Cardiology to record a range of signals as patients undergo cardioversion - a medical procedure that restores normal sinus rhythm (NSR) in patients with AFib through electric shocks. The procedure is performed under conscious sedation, limiting both the patient's discomfort and movement. Participants are of a diverse demographic, covering a range of ages (37–85 years), skin types (I–V on the Fitzpatrick scale [13]), races (77% white), and both genders (71% male). Cardiologist-reviewed ECGs are used to infer the ground truth labels before and after cardioversion. We exclude a minority of regions labelled by experts as other arrhythmias, and exclude one patient from the test set due to insufficient ECG data during recurring AFib episodes post-cardioversion. In addition, we record data from volunteers with no known arrhythmias during sleep outside the hospital setting. We assume that these internal recordings do not contain episodes of atrial fibrillation. We do not exclude recordings or parts thereof based on PPG signal quality, and allow for possibility of mislabelled regions in the training data due to insufficient ECG coverage. Table 1 summarizes the data used for algorithm development and testing.

Results presented here are based on approximately 180 h of data, of which 36 h are AFib. This is equivalent to approximately 10^7 raw samples, or 10^6 individual heartbeats. Using raw data maximises the information available for classification, and opens up numerous possibilities for generating augmented data, as discussed in Section 3.3.

3 CLASSIFYING RAW PPG SIGNALS

The bottom panel in Fig. 2 shows a typical PPG signal as the patient transitions from AFib to a normal sinus rhythm. Changes in the amplitude and periodicity of the signal are apparent, but presentation varies over time and between patients. By using a suitable heartbeat segmentation algorithm, it is possible to extract a range of features describing variability in periods and amplitudes, as well as morphology, of individual heartbeats. Insets of Fig. 2 (top panel) illustrate the value of this approach, yet choosing relevant features is a non-trivial task. Real-world issues such as signal discontinuities and noise from a range of sources further complicate

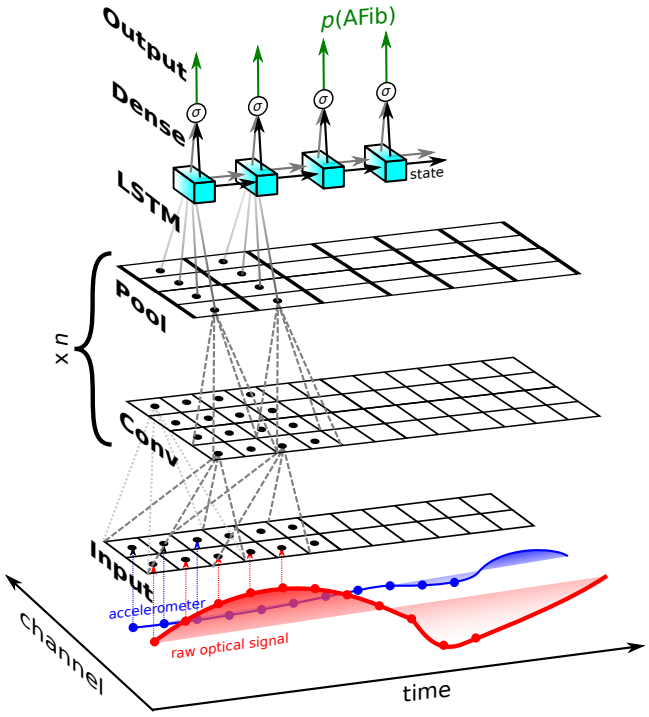

Figure 1: A convolutional-recurrent architecture for classification of raw time-series data. While the receptive field of each neuron in the convolutional (Conv) layers is well defined, the recurrent long short-term memory (LSTM) layer can learn variable-length correlations.

the classification problem. It is common practice to pre-process the signal, exclude noisy regions using a separate criterion, or introduce an additional label for such regions. Importantly, the information content in noisy signals per unit time may vary, which must be reflected in the classifier output.

3.1 Related work

A range of timeseries classification techniques have been proposed [47], with deep learning gaining increasing traction [3]. Recent work on classifying PPG signals can be broadly divided into time-domain heart rate approaches relying on heartbeat segmentation [33], and frequency-domain approaches generating features through Fourier or wavelet transforms [40]. Classification of ECG signals has received significant attention, with deep learning approaches employed almost exclusively in recent work [36, 39, 48, 49].

Our work on classifying medical sensor signals benefits from the many advances made using convolutional and recurrent neural networks in the domains of audio labelling and synthesis [20, 38, 46], and image recognition [21, 28, 29, 37, 41].

3.2 Convolutional-recurrent architecture

To overcome the issues outlined above, we propose an end-to-end model mapping the inputs to a sequence of calibrated, instantaneous probabilities. The model is based on the convolutional-recurrent neural network architecture shown schematically in Fig. 1.

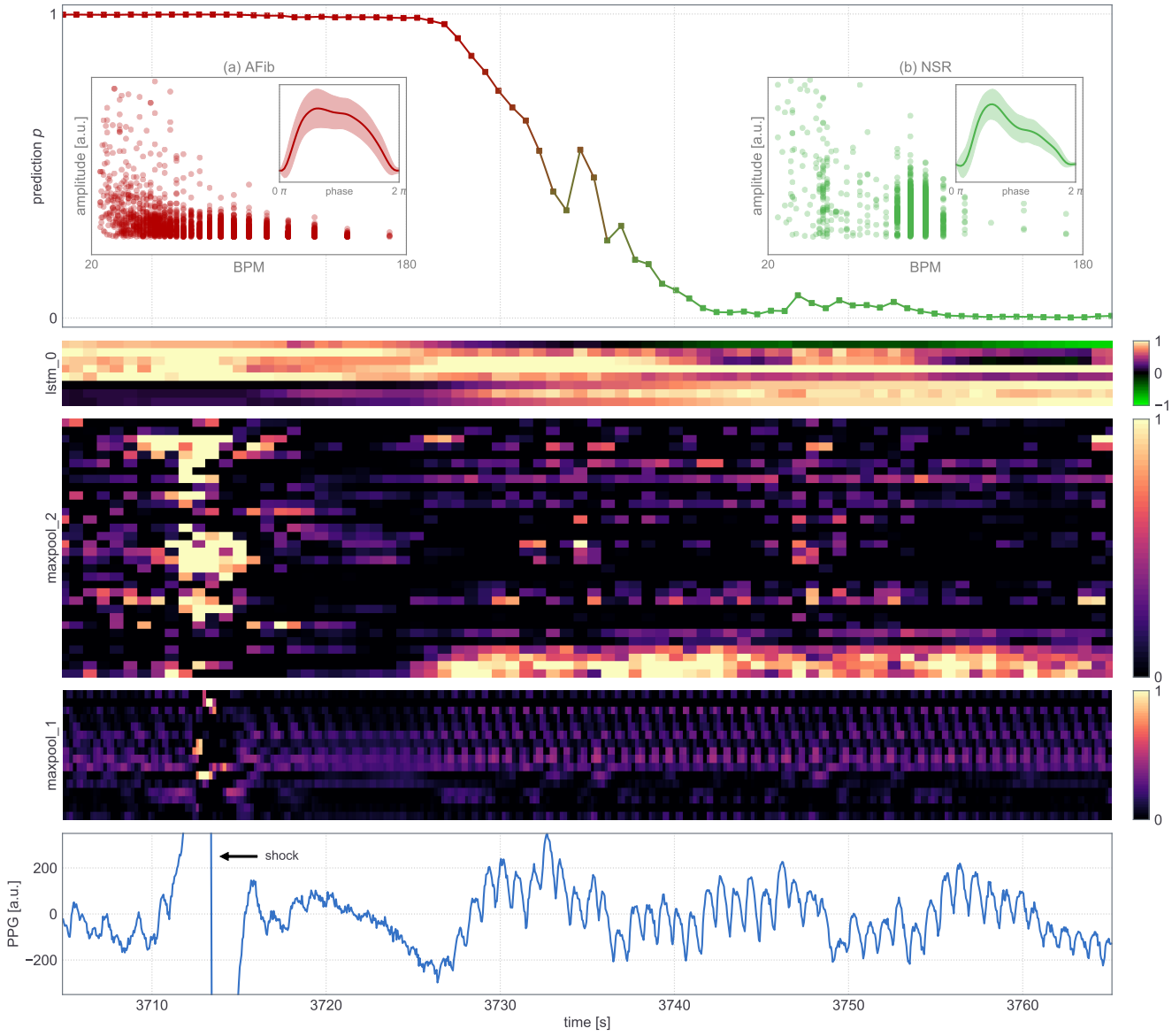

Figure 2: Real-time labelling of AFib vs NSR from raw PPG signal during cardioversion. Bottom to top: PPG signal; time-aligned activations in intermediate layers of the unrolled network; output probability p (see text). Insets show typical variations in heart rates (BPM), amplitudes, and PPG morphology for individual heartbeats during (a) AFib and (b) NSR. Details on visualizing individual activations are given in Appendix A.

Our input is a sequence of samples x_{t_i}, recorded at times t_i. The corresponding sampling frequency is $f_x = (t_{i+1} - ti)^{-1}$. We seek to predict the sequence of probabilities

$$p_{\tau_j} = P(\text{AFib at } \tau_j | x_{t_i \leq \tau_j})$$

Notice that our approach allows for the output p to depend on all previous values of x_{t_i}. Convolutional layers [30] with ReLU non-linearities [17, 26] extract multiple new features each layer, based on a receptive field of fixed length. [1] Convolution kernels can be seen as digital signal filters, and remove the need for hand-engineered signal processing operations. Max-pooling [23] is commonly used in deep convolutional neural networks, and in the context of signal processing it can be interpreted as a down-sampling operation.[2] A variable receptive field of each output is achieved by applying a

[1] the receptive field could be expanded significantly, e.g. using dilated convolutions as in [46]
[2] another way to down-sample the signal is through strided convolutions [12]

long short-term memory (LSTM) recurrent layer [15, 22],[3] followed by a single dense layer with sigmoid activation for the final output p.

The convolutional-recurrent architecture has further practical advantages: the sequence lengths used for training or prediction are flexible, and a real-time implementation is possible on a range of platforms.

The output frequency $f_p = (\tau_{j+1} - \tau_j)^{-1}$ is constrained to the divisors of f_x. The overall down-sampling ratio we use is $f_p/f_s = 1/16$, i.e. a new label is output every 0.8 s for an input signal sampled at 20 Hz.

Our implementation uses proven open-source libraries [1, 10, 35]. The model hyperparameters are chosen through cross-validation. [4] We find that our model is robust over a wide range of hyperparameters, with overfitting largely controlled by data augmentation at training time, as described in the following section.

3.3 Model training

We seek to minimize the binary cross-entropy loss function, summed over all outputs. The loss function is adjusted for class imbalance [19].

Our model contains ca. 10000 trainable parameters, and we follow best practices to improve convergence, reduce training time, and control over-fitting. These include weights initialization [16, 43], batch normalization between layers [24], dropout in the LSTM layer [14], and the choice of optimizer [27].

We train our network on mini-batches of fixed-length subsequences of the training data. The LSTM state is initialized at random for each example, and example length is chosen to allow the learning of long-range dependencies. Each epoch, we perform random augmentation of the training batches. Data augmentation has become a standard technique for training neural networks for image classification [9], audio tasks [11], and other timeseries applications [18, 45]. Using raw data allows us to identify domain-specific heuristics for data augmentation, and thus account for e.g. variations in user skin tone and varying light conditions. We randomly offset selected examples within the raw training signals (random cropping), and apply scaling, additive shifts and random Gaussian noise with random amplitudes per example. Random augmentation proves crucial to obtaining a model with superior performance on real-world signals.

To monitor convergence, we use a validation set of non-overlapping, unaugmented subsequences, reducing the learning rate every time the validation loss stops decreasing, as seen in Figure 3.

We generally achieve better performance on unaltered validation data compared to randomly augmented training batches. Similarly, we find that the performance of the trained model on the test set is unaffected by the presence of noisy recordings in the training set, and is robust to the presence of some mislabelled training examples. This is especially important given the limitations of our dataset explained in Section 2.

[3] in theory, LSTM state will depend on all previous x_{t_i}, though practical limitations exist [6]

[4] like the train-test split, all cross-validation splits are by subject to obtain an unbiased estimate of model performance

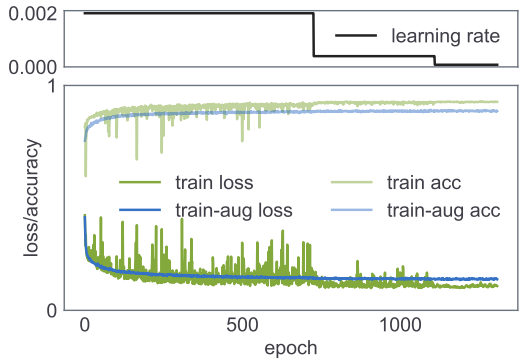

Figure 3: Learning curves (bottom) and learning rate annealing (top) with random augmentation

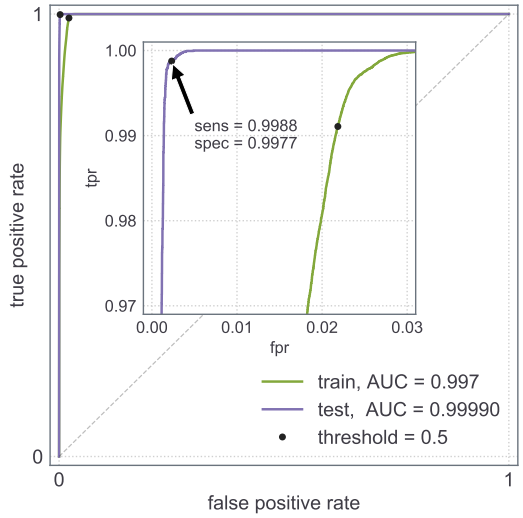

Figure 4: ROC of the trained model. Highlighted point corresponds to a probability threshold of 0.5.

4 RESULTS

4.1 Classifier performance

We evaluate performance of our model on the test set of recordings, as summarised in Table 1. We use raw sequence labels at 1.25 Hz. Figure 4 shows the receiver operating characteristic (ROC) curve for our probability predictions, for both train and test data. On the test set, we achieve AFib vs NSR classification with a specificity and sensitivity of 0.998 and 0.999, respectively, at a probability threshold of 0.5. This corresponds to a false positive rate of 2×10^{-3} and a false negative rate of 1×10^{-3}. The probability output is well calibrated, with a Brier score [7] of 0.002. As noted above, we have chosen to not exclude recordings with low signal quality, nor have we excluded a recording with suspected heart rhythm changes inbetween ECG spot checks from the training data - the ability to train a highly accurate classifier despite the likely presence of

mis-labelled data is important given the nature of physiological signals.

In large-scale screening applications, we expect a low false positive rate to be of key importance: not only is the fraction of individuals with AFib small, they are expected to exhibit AFib for a fraction of the time, with episodes varying in duration and frequency. [5] Considering the recordings from (presumed) healthy individuals during sleep only, we observe a false positive rate of 0.0016 at the same probability threshold.

4.2 Learned signal filtering

While the meaning of individual network weights is difficult to interpret, we can identify one specific task our network learns through training: that of signal filtering. The first convolutional layer can be seen as a bank of finite impulse response (FIR) filters, and we find that they adapt to perform high-pass filtering, with DC attenuations ranging from −37 dB to −64 dB. Thus, our approach removes the need for signal pre-processing, and the attenuation is consistent with the range of DC amplitudes seen in training.

4.3 Neuron function

Visualisation and interpretation of the function of individual neurons in convolutional [34] and recurrent [8] neural networks is an area of active research. Figure 2 shows time-resolved activations after two intermediate max-pooling layers, as well as the LSTM hidden state, time-aligned with the input signal. We can see how a number of neurons appear to specialize in tasks such as detecting peaks in layer `maxpool_1`, tracking persistent heart rhythm in layer `maxpool_2`, and finally encoding presence of AFib and/or NSR in the LSTM layer. It is interesting to note the time offset between transitions in individual LSTM hidden state values, and also the robust behaviour in the presence of input signal discontinuities and variable signal-to-noise ratios.

4.4 Heart rhythm embeddings

The hidden state of the LSTM layer can be interpreted as a time-dependent latent-space embedding of the underlying heart rhythm. We visualize 2D projections of these vectors for a range of patients in Figure 5. While our network learns a global decision boundary (shown for $p = 0.5$), we can see that the heart rhythm embeddings for both AFib and NSR vary between patients. We propose that standard unsupervised clustering techniques [32], applied to heart rhythm embeddings produced by our network, can further improve classification accuracy. More importantly, we envision being able to detect heart rhythm anomalies in individual subjects as outliers in the latent space, and extending our approach to other heart rhythm anomalies in the future.

5 CONCLUSIONS

In this article, we have demonstrated how applying best practices from domains such as image classification and natural language processing to the hitherto under-explored application area of real-time sensor data classification yields state-of-the-art results in PPG-based diagnostics of atrial fibrillation. We show that digital signal pre-processing can be learned by a suitably chosen neural network architecture, in a way that easily generalises to a multi-sensor, multi-channel setting. By interpreting intermediate outputs of a pre-trained neural network as latent-space embeddings of the physiological signal, we can further personalize diagnostics through unsupervised learning.

One aspect that could affect real-world performance of the model is the minimum duration of an isolated AFib episode that we are able to detect. Our experiments with synthetic data [6] show that minimum to be between 20–200 s, with a strong variation between patients, and dependent on signal-to-noise ratios. We believe that using synthetic data at training time may improve this further -

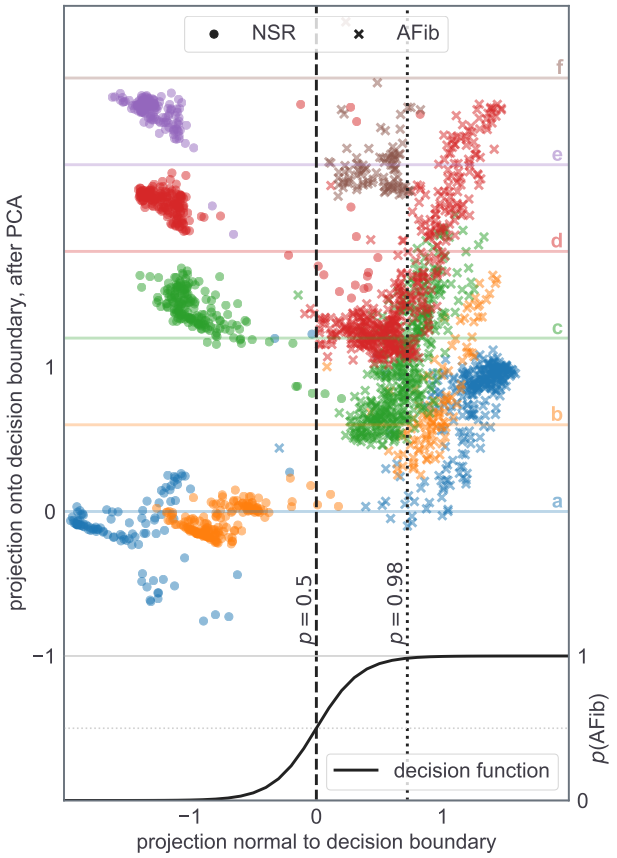

Figure 5: Heart rhythm embeddings, shown as 2D projections of LSTM hidden state vectors. Each point corresponds to one temporal step in the output sequence. For visual clarity, one in every 15 time steps is shown. Marker shape indicates the ground truth label; colours and letters correspond to unique patients. Projections were obtained as described in Appendix B.

[5] the total fraction of time spent in AFib by a given individual is known as the *AFib burden*; we are not aware of a study describing the distribution of burdens nor episode lengths

[6] obtained by splicing regions with different heart rhythms

concurrently, our ongoing data collection and labelling efforts focus on capturing a variety of real-world episodes.

While we have made every effort to train a robust and generalizable model, we have only accessed performance on data collected either in a hospital setting or during sleep. It remains to be seen how other factors such as motion and differing demographics affect the results. At the same time, we are confident that our approach will be applicable to new and larger datasets.

Three main issues have thus far precluded large-scale preventive diagnostics of AFib: the cost and availability of ECG monitoring devices, the episodic nature of the condition, and the need for expert review. By combining low-cost wearable sensors with deep learning algorithms, we pave the way to real time detection of atrial fibrillation in millions of users.

ACKNOWLEDGMENTS

We are grateful to Vasilis Kontis and David Grimes for their constructive comments on the manuscript.

A VISUALISING NEURONS

In Figure 2, we show activations of intermediate-layer neurons over time. We aim to show groups of neurons that learn similar functions. ReLU, and therefore max-pooling activations, are in the range $[0, \infty)$, while the hidden state values of an LSTM are in the range $[-1, 1]$. To better visualise the function of our network, we order individual channels i in each layer l by similarity of the activation timeseries $a_{it}^{(l)}$, where t denotes the time index. We use the optimal leaf ordering [5] obtained through hierarchical agglomerative clustering [31] with a suitable pairwise distance function. We find that the distance metric $d_{ij}^{(l)} = 1 - \left|\mathrm{corr}\left(a_{it}^{(l)}, a_{jt}^{(l)}\right)\right|$, computed over all times, yields good results. For the LSTM state, we invert the sign for channels with predominantly negative values (this is equivalent to flipping the sign of some weights to yield an equivalent network).

B VISUALISING VECTOR EMBEDDINGS

In Figure 5, we visualise multi-dimensional vector embeddings by projecting them onto 2D. This is done in a way that preserves the decision boundary, as defined by $\mathbf{x} \cdot \mathbf{w} + \mathbf{b} = 0$ for embeddings $\mathbf{x} \in \mathbb{R}^n$, and parameters of the simple linear classifier $\mathbf{w}, \mathbf{b} \in \mathbb{R}^n$. The corresponding logistic regression decision function is given by $p(\mathbf{x}) = \sigma(\mathbf{x} \cdot \mathbf{w} + \mathbf{b})$, with sigmoid activation $\sigma(a) = (1 + e^{-a})^{-1}$. $\mathbf{w}$ and $\mathbf{b}$ are learned by the output layer of the network during training.

To obtain 2D projections $\mathbf{y} = (y_0, y_1)$, we write $y_0 = \mathbf{x} \cdot \hat{\mathbf{w}} + \hat{\mathbf{b}}$ and $y_1 = \mathrm{PCA}_0\left(\mathbf{x} - \hat{\mathbf{w}} y_0\right)$. We use the notation $\hat{\mathbf{w}} = \mathbf{w}/|\mathbf{w}|$, $\hat{\mathbf{b}} = \mathbf{b}/|\mathbf{w}|$ for normalized vectors, and $\mathrm{PCA}_n(\mathbf{v})$ denotes the nth principal component of $\mathbf{v}$.

REFERENCES

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Software available from tensorflow.org.

[2] J. Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3):R1–39, Mar. 2007.

[3] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh. The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660, May 2017.

[4] J. Ball, M. J. Carrington, J. J. V. McMurray, and S. Stewart. Atrial fibrillation: Profile and burden of an evolving epidemic in the 21st century. *International Journal of Cardiology*, 167(5):1807–1824, Sept. 2013.

[5] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17(suppl_1):S22–S29, June 2001.

[6] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, Mar. 1994.

[7] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, Jan. 1950.

[8] S. Carter, D. Ha, I. Johnson, and C. Olah. Experiments in Handwriting with a Neural Network. *Distill*, 1(12):e4, Dec. 2016.

[9] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the Devil in the Details: Delving Deep into Convolutional Nets. *arXiv:1405.3531 [cs]*, May 2014.

[10] F. Chollet and others. Keras. 2015.

[11] X. Cui, V. Goel, and B. Kingsbury. Data Augmentation for deep neural network acoustic modeling. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5582–5586, May 2014.

[12] V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning. *arXiv:1603.07285 [cs, stat]*, Mar. 2016.

[13] T. B. Fitzpatrick. The Validity and Practicality of Sun-Reactive Skin Types I Through VI. *Archives of Dermatology*, 124(6):869, June 1988.

[14] Y. Gal and Z. Ghahramani. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 1027–1035, USA, 2016. Curran Associates Inc.

[15] F. A. Gers, J. A. Schmidhuber, and F. A. Cummins. Learning to Forget: Continual Prediction with LSTM. *Neural Comput.*, 12(10):2451–2471, Oct. 2000.

[16] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, Mar. 2010.

[17] X. Glorot, A. Bordes, and Y. Bengio. Deep Sparse Rectifier Neural Networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, June 2011.

[18] A. L. Guennec, S. Malinowski, and R. Tavenard. Data Augmentation for Time Series Classification using Convolutional Neural Networks. Sept. 2016.

[19] Haibo He and E. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, Sept. 2009.

[20] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. Deep Speech: Scaling up end-to-end speech recognition. *arXiv:1412.5567 [cs]*, Dec. 2014.

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[22] S. Hochreiter and J. Schmidhuber. Long Short-term Memory. *Neural computation*, 9:1735–80, Dec. 1997.

[23] D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, D. Scherer, A. Müller, and S. Behnke. Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition. In K. Diamantaras, W. Duch, and L. S. Iliadis, editors, *Artificial Neural Networks – ICANN 2010*, volume 6354, pages 92–101. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[24] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]*, Feb. 2015.

[25] C. T. January, L. S. Wann, J. S. Alpert, H. Calkins, J. E. Cigarroa, J. C. Cleveland, J. B. Conti, P. T. Ellinor, M. D. Ezekowitz, M. E. Field, K. T. Murray, R. L. Sacco, W. G. Stevenson, P. J. Tchou, C. M. Tracy, and C. W. Yancy. 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. *Circulation*, 130(23):e199–e267, Dec. 2014.

[26] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pages 2146–2153, Sept. 2009.

[27] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, Dec. 2014.

[28] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, Dec. 1989.

[29] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov. 1998.
[30] Y. LeCun, K. Kavukcuoglu, and C. Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 253–256, May 2010.
[31] D. Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv:1109.2378 [cs, stat]*, Sept. 2011.
[32] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.
[33] S. Nemati, M. M. Ghassemi, V. Ambai, N. Isakadze, O. Levantsevych, A. Shah, and G. D. Clifford. Monitoring and detecting atrial fibrillation using wearable technology. *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2016:3394–3397, Aug. 2016.
[34] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev. The Building Blocks of Interpretability. *Distill*, 3(3):e10, Mar. 2018.
[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
[36] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. *arXiv:1707.01836 [cs]*, July 2017.
[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec. 2015.
[38] H. Sak, A. Senior, K. Rao, and F. Beaufays. Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition. *arXiv:1507.06947 [cs, stat]*, July 2015.
[39] S. P. Shashikumar, A. J. Shah, G. D. Clifford, and S. Nemati. Detection of Paroxysmal Atrial Fibrillation using Attention-based Bidirectional Recurrent Neural Networks. *arXiv:1805.09133 [cs, q-bio]*, May 2018.
[40] S. P. Shashikumar, A. J. Shah, Q. Li, G. D. Clifford, and S. Nemati. A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology. In *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 141–144, Feb. 2017.
[41] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training Very Deep Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2377–2385, Cambridge, MA, USA, 2015. MIT Press.
[42] S. Stewart, C. L. Hart, D. J. Hole, and J. J. V. McMurray. A population-based study of the long-term risks associated with atrial fibrillation: 20-year follow-up of the Renfrew/Paisley study. *The American Journal of Medicine*, 113(5):359–364, Oct. 2002.
[43] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147, Feb. 2013.
[44] G. H. Tison, J. M. Sanchez, B. Ballinger, A. Singh, J. E. Olgin, M. J. Pletcher, E. Vittinghoff, E. S. Lee, S. M. Fan, R. A. Gladstone, C. Mikell, N. Sohoni, J. Hsieh, and G. M. Marcus. Passive Detection of Atrial Fibrillation Using a Commercially Available Smartwatch. *JAMA cardiology*, 3(5):409–416, May 2018.
[45] T. T. Um, F. M. J. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulić. Data Augmentation of Wearable Sensor Data for Parkinson's Disease Monitoring using Convolutional Neural Networks. *arXiv:1706.00527 [cs]*, pages 216–220, 2017.
[46] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv:1609.03499 [cs]*, Sept. 2016.
[47] Z. Wang, W. Yan, and T. Oates. Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline. *arXiv:1611.06455 [cs, stat]*, Nov. 2016.
[48] Y. Xia, N. Wulan, K. Wang, and H. Zhang. Detecting atrial fibrillation by deep convolutional neural networks. *Computers in Biology and Medicine*, 93:84–92, Feb. 2018.
[49] M. Zihlmann, D. Perekrestenko, and M. Tschannen. Convolutional Recurrent Neural Networks for Electrocardiogram Classification. *arXiv:1710.06122 [cs]*, Oct. 2017.
[50] M. Zoni-Berisso, F. Lercari, T. Carazza, and S. Domenicucci. Epidemiology of atrial fibrillation: European perspective. *Clinical Epidemiology*, 6:213–220, June 2014.