HIGH SCHOOL EDITION
Journal of Student Research

Volume 13 Issue 1 (2024)

An Interpretable Machine Learning Model for Heart Arrhythmia Classification

Karthik Vedula

Poolesville High School, USA

ABSTRACT

Machine learning (ML) has been a very effective tool for arrhythmia detection and classification using electrocardiograms (ECGs). However, in order for patients and healthcare professionals to trust the ML models, the models have to be interpretable to show how they arrived at a certain conclusion. We present an ML model that uses the Deep k-Nearest Neighbors framework in order to produce example-based explanations and uncertainty estimations. These examples are ECGs similar to the input derived from conducting a nearest neighbor search on encoded samples (which are values of a layer of the neural network after passing dataset samples through it). We introduce a new technique of using these example-based explanations in conjunction with saliency maps, and also use a neighbor-based uncertainty estimation technique. We show that the saliency maps provide good explanations, but the neighbor examples are needed to assess the credibility of those saliency maps. Our uncertainty estimations increase the accuracy of the model from 86% to 93% (when measured with coverage of 76%). Overall, our novel methods prove to be a promising solution in the field of interpretable ML for arrhythmia classification.

Introduction

About 1 in 20 people in the U.S. have an arrhythmia - also known as irregular heartbeat[1]. According to the CDC, by 2030, 12.1 million people in the U.S. will have atrial fibrillation, a very common arrhythmia which can increase the risk of stroke by fivefold[2][3]. Early detection and classification of these diseases is imperative for better treatment[4]. Electrocardiograms (ECGs) provide a non-invasive method of measuring the electrical activity of the heart, and can provide information regarding the presence and type of arrhythmia for a patient. However, the processing of these ECGs is a time-consuming process (can range from the day of procedure to several weeks[7][8]) as it requires professionally trained physicians to manually identify and classify irregular heartbeats - a task that can be automated[6]. Therefore, there is a need for automated ECG processing for the identification of heart arrhythmias.

For such automation, machine learning (ML) has been a very effective tool for arrhythmia identification[9]. Rajpurkar et. al have developed an ML model for arrhythmia identification through ECGs that provided accuracy similar to that of a cardiologist[10]. Other researchers have also developed ML models for this application. While these models have high accuracy, many do not have the ability for the doctor or patient to understand why the model came to a certain conclusion. This means that the model cannot have that much credibility as it cannot display reasoning, leading to the model becoming unusable in many of such critical situations[14][15]. In other words, in order for these ML models to be integrated into current healthcare, they have to be interpretable[16].

Interpretable ML for arrhythmia identification and classification is a relatively nascent field[17]. There has been research such as techniques applying LIME[18], saliency maps[19], and also feature engineering/selection[20] for heart arrhythmias. However, these techniques provide evidence from a prediction solely from the input ECGs (without comparing/constrasting to other ECGs), leading in some cases difficulty in interpretation of the maps as they might not display an easily identifiable reason, making it difficult to use such information as evidence for a prediction[23]. Therefore, there is a need for using example-based explanations, where samples similar to the input (preferably from the model's perspective) are shown to provide reasoning. We also use uncertainty estimation techniques to estimate the likelihood of an incorrect prediction.

Methods

We use the ML model architecture developed by Rajpurkar et. al, which has shown to provide (with training from their own large dataset) to be potentially better than a cardiologist[10]. We use the Physionet Challenge 2017 dataset (vs. proprietary dataset used by Rajpurkar et. al) with labels consisting of Normal Sinus Rhythm, Atrial Fibrillation, and Other arrhythmias. We then split our data into a *training* set, a *reference* set, and a *test* set. The reference dataset is for developing our interpretability methods, where we collect the values of a layer of the neural network after passing data from the *reference* dataset through it to get our encoded samples, as described by Papernot & McDaniel in their Deep k-Nearest Neighbors Neural Network[25]. Figure 1 shows the process to generate the encoded samples. Since the layer used had 256 dimensions, we applied Principal Component Analysis (PCA) to reduce the encoded samples' dimensions to 10.

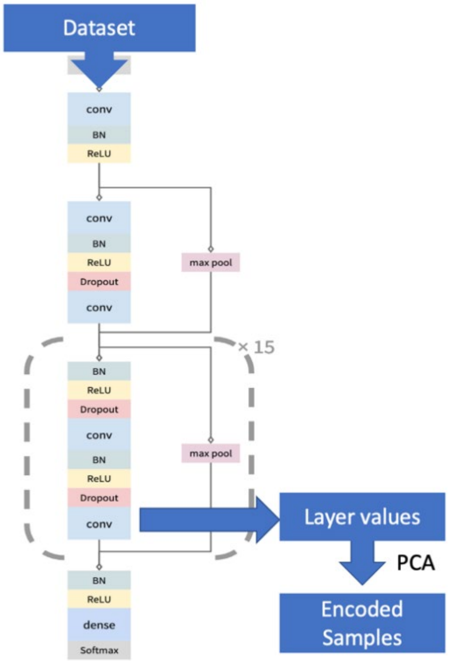

Figure 1. Process showing how encoded samples are generated. Adapted from Rajpurkar et al.

Example-Based Explanations Using Neighbors

For a specific input, we conduct a nearest neighbor search through the encoded samples to find similar ECG samples from the model's perspective (based on Euclidean distance). We also add saliency maps, which show features most significant to the prediction, and use NeuroKit2[26] for labeling of ECG features (such as P-waves) to aid in interpretation. This ensemble of example-based explanations, saliency maps, and feature labelling brings much more insight into the model's predictions.

Neighbor-Based Uncertainty Estimation

For each sample, we find the *n*th nearest neighbor of the input which was predicted incorrectly. Figure 2 shows an example of the neighbors for a given input sample shown in grey, where red represents an incorrect prediction and blue represents a correct one. Here, the 6th nearest neighbor (*n*=6) to the grey sample is predicted incorrectly. This implies that larger the *n*, the more correctly predicted neighbors, the higher the likelihood that the model predicts correctly (see Figure 3). We then use this data (correlation between *n* and model's correctness on the prediction) in order to approximate the likelihood of a model predicting incorrectly given an input. Our experimentation provides more exact values for these probabilities.

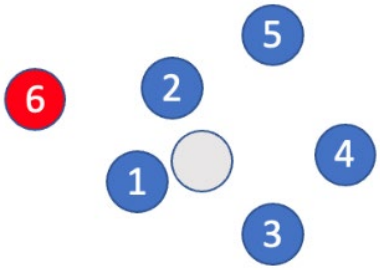

Figure 2. Nearest neighbor example. Blue indicates correctly predicted sample, red indicates incorrectly predicted sample.

Results

Figure 3 shows a stacked histogram of the distribution of model accuracy on samples with respect to confidence values derived from uncertainty estimation. Blue indicates correct predictions and red indicates incorrect ones. There is a clear positive correlation between *n*, which is the *n*th nearest neighbor that was predicted incorrectly, and the accuracy of the model. Our ML model achieved an accuracy of 86% (comparable to other models used with this dataset), which further increased to 93% after the use of uncertainty estimation methods. This accuracy was measured with coverage of 76%, (with 74% accuracy in non-confident predictions). It is important to note that the coverage-accuracy tradeoff can be adjusted.

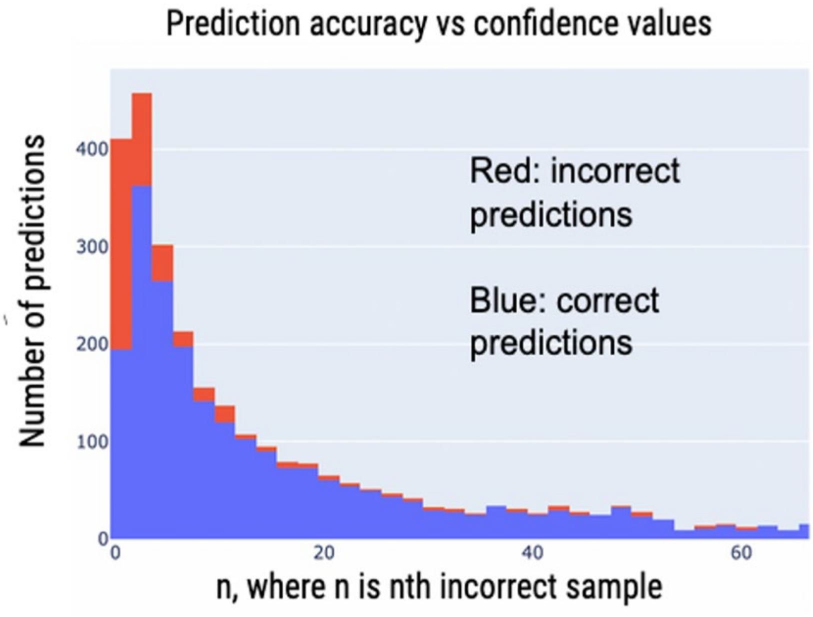

Figure 3. Distribution of model accuracy on samples with respect to confidence values derived from uncertainty estimation.

Figures 4 to 7 show results that are discussed in the next section.

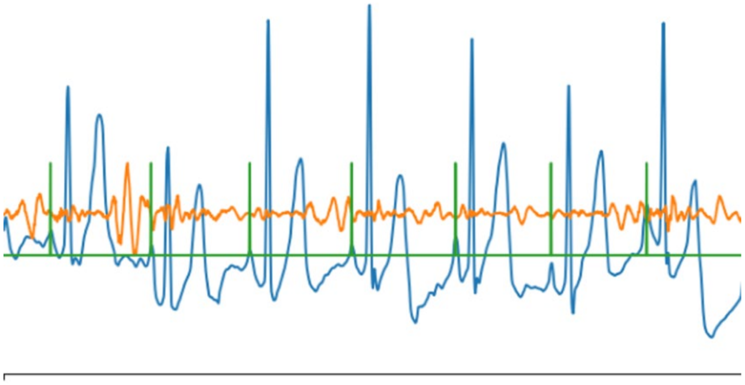

Figure 4. Example of an input signal (blue) with saliency maps (orange) and p-wave labelling (green)

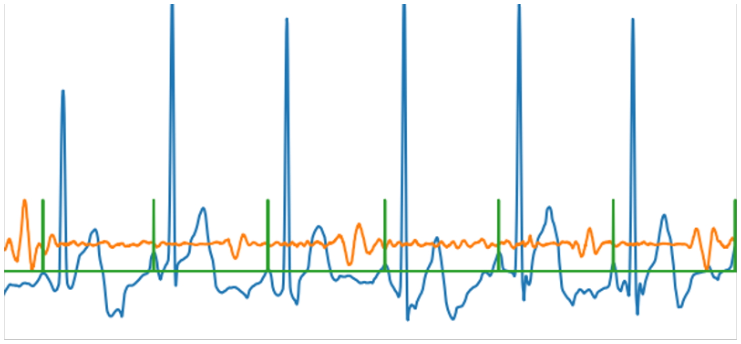

Figure 5: Example of a neighbor to the input signal from Figure 4

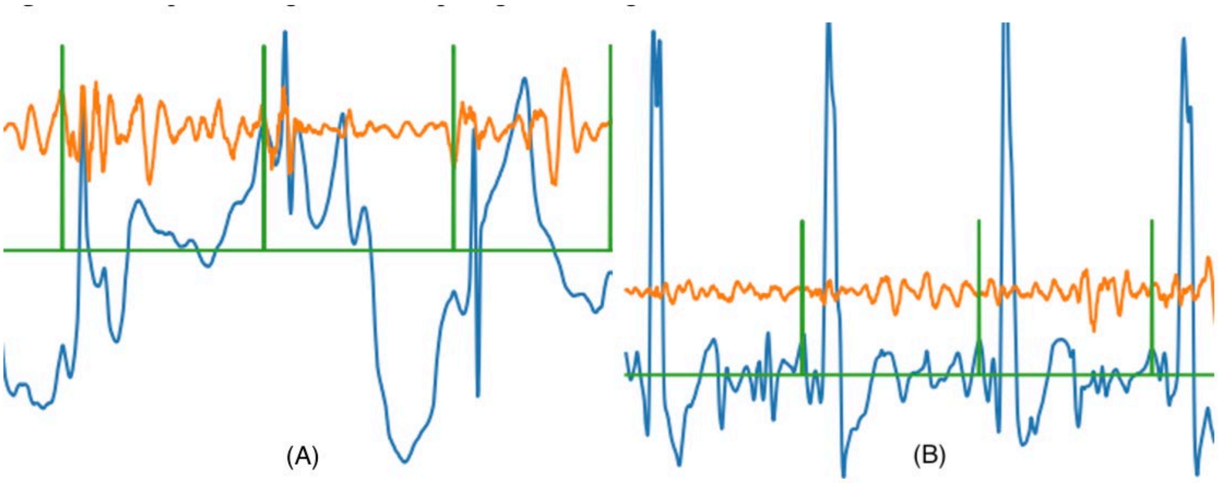

Figure 6. Misclassified neighbor comparison. (A) poorly recorded ECG; (B) a neighbor. The model associated the two samples by the noise of one sample with the noise of another.

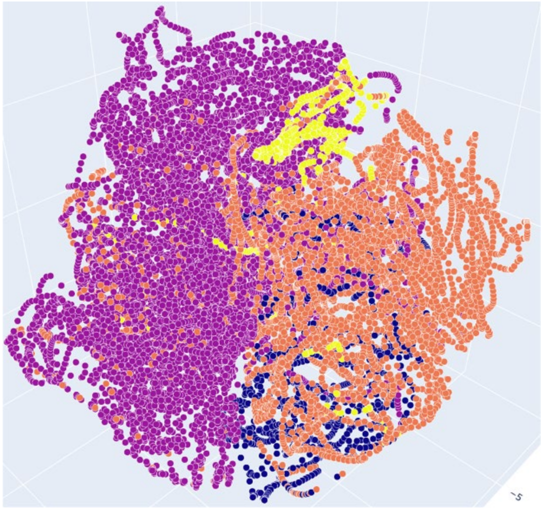

Figure 7. 3D t-SNE plot of the encoded samples (as derived in process shown in Figure 1). Color signifies the label of the data point. Points that seem to form a chain are ECG beats of the same patient.

Discussion

As shown in Figures 4 & 5, our nearest neighbor search was able to find similar samples to be used as explanations. For both the input in Figure 4 and neighbor sample in Figure 5, we display a plot which contains the signal itself, the model's saliency map for that sample, and labelled P-waves from NeuroKit2. The saliency map indicates that the model emphasized the region between (and including) the T-wave and the P-wave (Figure 5). This makes sense as atrial fibrillation has no P-wave and instead can comprise fibrillatory waves in the middle. This region can also be important for categorizing an input sample as "other". In this case, since both the input and its neighbor show presence of a clear P-wave among other reasons, the input is likely normal sinus rhythm (as predicted).

Along with using neighbors to infer on accuracy of our model predictions, we also utilize neighbors to infer on the validity of our saliency maps. The similarity between Figures 4 & 5 shows that the saliency map does indeed indicate legitimate reasoning (with both saliency maps highlighting before the P-waves). Figure 6 shows how the model likely mistook the noise in the input for noise between T and the next P-waves of a neighboring sample. This highlights the fact that without neighbors, it would be difficult to see whether the model's reasoning was backed with good evidence or not. Overall, saliency maps provide explanations and examples provide assessment of credibility to those explanations. Therefore, our technique of using saliency maps in conjunction with neighbor-based examples is advantageous.

It is important to note that our nearest neighbor search often gave dissimilar samples even with high model confidence. We use t-SNE (t-Distributed Stochastic Neighbor Embedding), a dimensionality reduction technique on encoded samples to find pitfalls of our technique (Figure 7). We find that there are regions where samples of different labels or regions with low number of samples have the most dissimilar neighbors. This makes sense as there is not much data to support the interpretability methods in this region. This can be amended with the use of large datasets (as shown by Rajpurkar et al.) instead, which we will leave for future work.

Conclusion

We introduced a machine learning model which utilizes both example-based explanations (using neighbor-based analysis) in conjunction with saliency maps for interpretability in the application of heart arrhythmia classification. We also used this neighbor-based analysis to develop a novel uncertainty estimation technique to find confidence values of a model's prediction. We showed the benefits of using both saliency maps and example-based explanations in ensemble to give explanations with credibility assessment, which other methods might not provide. We hope to continue this research in the future in the form of utilizing larger datasets, in order to provide better and more diverse samples to use for example-based explanations and boost model performance. We also point to using CAM or GradCAM with the neighbor analysis as techniques to explore as well.

Limitations

The dataset used did have enough datapoints for the ML model to be able to train with effective accuracy, but it did not have enough diversity as compared with other proprietary datasets – such as the dataset used by Rajpurkar et. al. Our Deep-kNN based approach, because of its reliance on datapoints neighboring each other, leverages highly diverse, abundant data in order to provide uncertainty estimations and interpretability. We point to applying our proposed approach to these larger datasets as future work.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

[1] Desai DS, Hajouli S. Arrhythmias. [Updated 2022 Jun 11]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK558923/

[2] Atrial Fibrillation. (2022, October). Centers for Disease Control and Prevention (CDC). Retrieved January 27, 2023, from https://www.cdc.gov/heartdisease/atrial_fibrillation.htm

[3] Odutayo A., Wong C.X., Hsiao A.J., Hopewell S., Altman U.G., A Emdin C. Atrial fibrillation and risks of cardiovascular disease, renal disease, and death: Systematic review and meta-analysis. BMJ. 2016;354:i4482. doi: 10.1136/bmj.i4482.

[4] Ahmed N, Zhu Y. Early Detection of Atrial Fibrillation Based on ECG Signals. Bioengineering (Basel). 2020 Feb 13;7(1):16. doi: 10.3390/bioengineering7010016. PMID: 32069949; PMCID: PMC7148541.

[5] Holter monitor. (n.d.). Mayo Clinic. Retrieved January 27, 2023, from https://www.mayoclinic.org/tests-procedures/holter-monitor/about/pac-20385039

[6] S. Osowski, L. T. Hoai and T. Markiewicz, "Support vector machine-based expert system for reliable heartbeat recognition," in IEEE Transactions on Biomedical Engineering, vol. 51, no. 4, pp. 582-589, April 2004, doi: 10.1109/TBME.2004.824138.

[7] ECG (electrocardiogram). (2022, January 28). Cancer Research UK. Retrieved January 27, 2023, from https://www.cancerresearchuk.org/about-cancer/tests-and-scans/ecg

[8] Electrocardiogram (ECG or EKG). (n.d.). Seattle Children's. Retrieved January 27, 2023, from https://www.seattlechildrens.org/clinics/heart/what-to-expect/electrocardiogram/

[9] Trayanova, N. A., Popescu, D. M., & Shade, J. K. (2021). Machine learning in arrhythmia and electrophysi- ology. Circulation Research, 128(4). doi:10.1161/CIRCRESAHA.120.317872

[10] Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., & Ng, A. Y. (2017). Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. doi:10.48550/ARXIV.1707.01836

[11] Alfaras, M., Soriano, M., & Ortín, S. (07 2019). A Fast Machine Learning Model for ECG-Based Heartbeat Classification and Arrhythmia Detection. Frontiers in Physics, 7. doi:10.3389/fphy.2019.00103

[12] Rezaei, M. J., Woodward, J. R., Ramírez, J., & Munroe, P. (2021). A Novel Two-Stage Heart Arrhythmia Ensemble Classifier. Computers, 10(5). doi:10.3390/computers10050060

[13] Isin, A., & Ozdalili, S. (2017). Cardiac arrhythmia detection using deep learning. Procedia Computer Science, 120, 268–275. doi:10.1016/j.procs.2017.11.238

[14] Kolyshkina, I., & Simoff, S. (05 2021). Interpretability of Machine Learning Solutions in Public Healthcare: The CRISP-ML Approach. Frontiers in Big Data, 4, 660206. doi:10.3389/fdata.2021.660206

[15] Abdullah, T. A. A., Zahid, M. S. M., & Ali, W. (2021). A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions. Symmetry, 13(12). doi:10.3390/sym13122439
6

[16] Ahmad, M. A., Eckert, C., & Teredesai, A. (2018, August). Interpretable machine learning in healthcare. In Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics (pp. 559-560). doi: 10.1109/ICHI.2018.00095

[17] Gohel, P., Singh, P., & Mohanty, M. (2021). Explainable AI: current status and future directions. doi:10.48550/ARXIV.2107.07045

[18] M. Bodini, M. W. Rivolta and R. Sassi, "Interpretability Analysis of Machine Learning Algorithms in the Detection of ST-Elevation Myocardial Infarction," 2020 Computing in Cardiology, Rimini, Italy, 2020, pp. 1-4, doi: 10.22489/CinC.2020.403.

[19] Bridge, J., Fu, L., Xue, Y., Lip, G. Y. H., & Zheng, Y. (2022). Artificial intelligence to detect abnormal heart rhythm from scanned electrocardiogram tracings. Journal of Arrhythmia, 38(3). doi: 10.1002/joa3.12707

[20] Sager, S., Bernhardt, F., Kehrle, F., Merkert, M., Potschka, A., Meder, B., ... Scholz, E. (12 2021). Expert-enhanced machine learning for cardiac arrhythmia classification. PLOS ONE, 16(12), 1–22. doi:10.1371/journal.pone.0261571

[21] Jo, Y.-Y., Kwon, J.-M., Jeon, K.-H., Cho, Y.-H., Shin, J.-H., Lee, Y.-J., . . . Oh, B.-H. (2021). Detection and classification of arrhythmia using an explainable deep learning model. Journal of Electrocardiology, 67, 124–132. doi:10.1016/j.jelectrocard.2021.06.006

[22] N. Strodthoff, P. Wagner, T. Schaeffter and W. Samek, "Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 5, pp. 1519-1528, May 2021, doi: 10.1109/JBHI.2020.3022989.

[23] Ayano, Y. M., Schwenker, F., Dufera, B. D., & Debelee, T. G. (2023). Interpretable Machine Learn- ing Techniques in ECG-Based Heart Disease Classification: A Systematic Review. Diagnostics, 13(1). doi:10.3390/diagnostics13010111

[24] Cai, C. J., Jongejan, J., & Holbrook, J. S. (2019). The Effects of Example-Based Explanations in a Machine Learning Interface. doi:10.1145/3301275.3302289

[25] Papernot, N., & McDaniel, P. (2018). Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. arXiv preprint doi: 10.48550/arXiv.1803.04765

[26] Makowski, D., Pham, T., Lau, Z.J. et al. NeuroKit2: A Python toolbox for neurophysiological signal processing. Behav Res 53, 1689–1696 (2021). https://doi.org/10.3758/s13428-020-01516-y