Current Natural Sciences & Engineering
Volume 2, Issue 3, 2025, 674-681

Classification and Analysis of 12-Lead Electrocardiograms

Revanth Reddy Pasula

Department of Computer Science, Wichita State University, Wichita, United States

Received date:01/07/2025, Acceptance date: 21/07/2025

DOI: ***http://doi.org/10.63015/2ai-2471.2.3***

**Corresponding Author: revanthreddy210799@gmail.com*

Abstract

This work investigates the classification of 12-lead electrocardiogram (ECGs) to detect abnormalities in the heart using three computational techniques. They are: (1) gradient-boosted ensembling following manual feature extraction, (2) deep learning with stacked autoencoders connected to the output of a multi-layer perceptron (MLP) classifier, and (3) a fusion model combining deep-learning and manually extracted features. An experiment is conducted using the PhysioNet/Computing in Cardiology Challenge 2020 database, addressing a multi-label classification task involving 27 heartbeat rhythm diagnoses. The best-performing model, which merges handcrafted features with autoencoder-derived features, achieves an average classification accuracy of 30.7% and a challenge metric score of 0.4366. The paper concludes by discussing potential improvements in multi-channel ECG classification methods.

Keywords: ECG Classification; 12-Lead ECG; Feature Extraction; Deep Learning; Autoencoders; Gradient Boosting

I. INTRODUCTION

Cardiac conditions still top the global causes of death at approximately 80% of deaths related to them, mainly due to heart attack and stroke. Twelve-lead electrocardiography (12-lead ECG) is the key to detecting cardiac pathology and assessing high-risk patients. An ECG captures the heart's electrical signals from electrodes positioned on the chest and limbs, producing waveforms corresponding to myocardial depolarization and repolarization. While computer-aided ECG analysis is widely adopted, current automated interpretation software sometimes fails to match the accuracy of specialist cardiologists, leading to missed or incorrect diagnoses. Technological advances have introduced a variety of ECG recording devices, ranging from portable single-lead designs to sophisticated clinical machines. Consumer-oriented devices like the a six-lead model, Apple's one-lead Apple Watch, and the three-lead Cardio Core wearable demonstrate the growing potential for personalized heart monitoring. However, in clinical settings, standard 12-lead systems produced by manufacturers such as General Electric and Philips remain the gold standard for comprehensive cardiac evaluation. This paper focuses on the traditional 12-lead ECG, which provides extensive coverage of cardiac electrical signals from various directions and is widely used in clinical practice. This study proposes a framework that integrates conventional signal processing with modern machine learning techniques for multi-label classification of 12-lead ECG data. It emphasizes three distinct modelling approaches for automated detection of various cardiac conditions from ECG signals. The study's hypotheses are framed based on these approaches:

- **Hypothesis 1:** Classic machine learning methods like gradient-boosted decision trees will perform similarly to, if not better than, deep learning methods (using autoencoders) in cumulative metrics such as the F-measure and general accuracy.
- **Hypothesis 2:** For tree-structured classifiers, systematic regularization of the input feature space and intentional feature selection will probably improve the challenge metric (a particular contest scoring criterion) better than simply augmenting the feature set with features synthesized by autoencoders.
- **Hypothesis 3:** Adding features extracted from a deep autoencoder to a decision-tree ensemble, along with manually engineered features, is expected to improve the overall classification accuracy of the model.

II. CONTRIBUTIONS

The main contributions of this research are summarized as follows:

- **Traditional feature-based classifier:** We created a structured approach to classifying 12-lead ECG signals with deep manual feature extraction followed by an ensemble of gradient-boosted trees. It was entered in the PhysioNet/CinC 2020 Challenge [11], where it attained a validation challenge score of 0.476 and a test (hidden set) score of –0.080, ranking 36 out of 41 valid submissions in the official ranking.
- **Deep learning autoencoder classifier:** We employed a deep learning approach using stacked autoencoders to obtain concise representations from segmented heartbeats and then a sequence model to predict full ECG records. Without access to official test data for this method, performance was assessed using Monte Carlo cross-validation within the public dataset (20 random 80/10/10 training-validation-testing splits). The model using autoencoder achieved an average challenge score of 0.248 on these test splits. While its overall accuracy was less than the feature-based model, the deep model yielded slightly better sensitivity to some conditions – for instance, incomplete right bundle branch block (IRBBB), left anterior fascicular block (LAnFB), prolonged PR interval, and right-axis deviation (RAD).
- **Hybrid feature–embedding ensemble:** We created a hybrid modelling approach that combines manually crafted features with feature learning from autoencoders to train an improved set of gradient-boosted tree classifiers. Compared to the purely manual method, the hybrid approach employs feature selection at the level of individual labels instead of one global ranking feature. The winning configuration, labelled “Top 1000 Features + Embeddings,” chose the top 1000 most significant features for each diagnostic label and produced a test-split challenge score of

0.4366 – well above the remaining configurations assessed within the study.

III. METHODOLOGY

Deep Autoencoder + MLP Classification (Architecture Details): In our implementation, the stacked autoencoder comprised multiple fully connected layers to encode each heartbeat segment into a low-dimensional embedding. Each heartbeat (segmented via a standard R-peak detection algorithm with a fixed window length around each QRS complex) was resampled to a uniform length (approximately 500 samples) and fed into an encoder network with three dense layers of 256, 128, and 64 neurons (using ReLU activations). The encoder's bottleneck layer produced a **64-dimensional** latent vector representing the heartbeat. A symmetric decoder (64→128→256 neurons, ReLU activations, and a linear output layer) was trained to reconstruct the input waveform from this embedding. We trained the autoencoder on the training set heartbeats for up to 100 epochs using the Adam optimizer (learning rate ~0.001) with mean squared error as the loss, employing early stopping if reconstruction error on a validation subset did not improve for 10 consecutive epochs to prevent overfitting. After obtaining per-beat embeddings, a sequence model was used to aggregate these into a record-level representation. Specifically, we employed a one-layer LSTM with 128 hidden units: the sequence of heartbeat embeddings for an ECG record was fed into the LSTM, and the final hidden state (128-dimensional) was taken as the record-level embedding. (We also experimented with simple averaging of the heartbeat vectors as a pooling strategy, but the trainable LSTM encoder performed comparably and retained temporal information about beat sequence.) This record-level embedding was then input to a multi-layer perceptron classifier. The MLP classifier consisted of two dense hidden layers (128 and 64 neurons, ReLU activations) and an output layer of 27 sigmoid neurons (one per diagnosis) to produce multi-label predictions. We applied a dropout rate of 0.2 in the MLP to improve generalization, and optimized the classifier using binary cross-entropy loss (with Adam, learning rate 0.001). During supervised training of the MLP, we fine-tuned the encoder and LSTM weights (which were initially learned in the unsupervised phase) – we found that allowing fine-tuning improved validation performance slightly compared to keeping the encoder frozen. The autoencoder and classifier were trained for roughly 50 epochs (with early stopping on validation loss) in each cross-validation fold. This deep architecture, including its regularization (dropout and early stopping), was designed to balance model complexity with the risk of overfitting. The result was an end-to-end deep network that first compresses beats into latent features and then learns to classify entire ECG records from sequences of those features. However, as discussed later, this complex model did not outperform the simpler approaches.

IV. RESULTS

A Comparison of classification performance metrics on the test split for the XGBoost ensemble across different feature selection strategies is shown in Figure 1. The horizontal axis labels "A" through "J" correspond to the ten model configurations detailed in Table I (in order). Plotted values include the PhysioNet Challenge score (the primary metric) alongside secondary metrics such as overall accuracy and F1-score, all summarized over 20 cross-validation runs. Each colored marker, along with its error bar or box, shows the distribution (mean and variance) of a given metric for each configuration. The figure highlights the trade-offs in performance when using all features, the top-1000 features, or the top-100 features, with and without incorporating autoencoder embeddings.

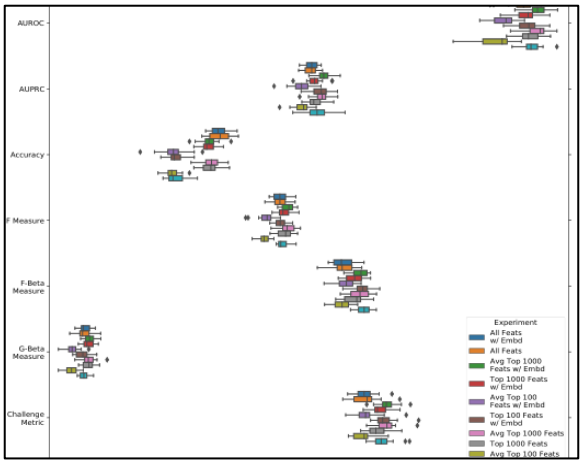

Fig 1. Comparison of classification performance metrics.

The Output of the Wilcoxon signed-rank test analyzing the distributions of the Challenge metric for all pairs of model configs is shown in Figure 2. The entry in the matrix is the p-value for the null hypothesis the corresponding pairs of configs' performance is the same; darker hues represent lower p-values. The cells marked by the symbol (*) correspond to statistically significant differences at $\alpha = 0.001$. For example, configs with aggressive feature pruning by selecting Top 100 features have different configuration performance than some others ($p < 0.001$ in those rows), uncovering the impact of the feature selection approach. Smaller p-values in

general (dark blue cells in the heatmap) reflect configuration pairs where the performance had differed significantly, uncovering the modelling choice (label-specific selection of features, including embeddings, etc.) with the resultant impact on the Challenge score.

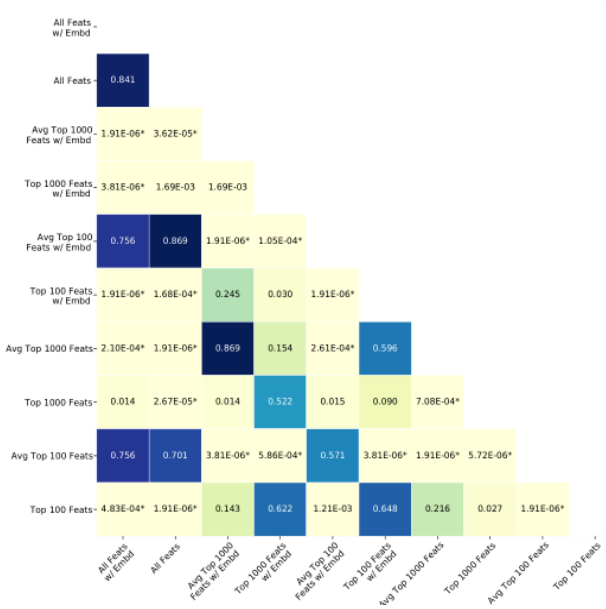

Figure 2. Wilcoxon signed-rank test distributions

V.DISCUSSION

The findings of our research emphasize a few key points, consistent with observations by other researchers. First, the inclusion of automatically learned deep features did not yield a performance gain, in line with comments by Bengio et al. that simply adding deep models to standard machine learning pipelines may not improve results. In our case, the gradient boosting ensemble achieved strong results with carefully selected time-domain and morphological features alone, and the added complexity of the autoencoder-derived features did not pay off in improved scoring. One likely reason is that the unsupervised autoencoder learned latent features that were not well aligned with the discriminative features needed for classification – the tree models could not effectively utilize the extra information when those deep features were essentially abstract combinations of raw signals. Furthermore, using deep features as input to a shallow classifier reduced interpretability of the system; it became difficult to trace which ECG lead or waveform characteristic contributed to a given autoencoder feature, obscuring the reasoning behind a particular prediction.

Several factors may explain why the autoencoder-based deep model underperformed the traditional feature-based model. Model depth and complexity: The deep autoencoder and LSTM introduced a large number of trainable parameters, increasing the risk of overfitting given the effective amount of labelled training data (43,000 records – substantial, but small relative to the complexity of a deep network). Training the autoencoder to reconstruct signals, while useful for unsupervised feature learning, does not guarantee that the learned features are optimal for distinguishing arrhythmias. The deep model might require even more data or more aggressive regularization to realize its potential, whereas the simpler XGBoost models could generalize well with the available data. Over-compression bottleneck: By compressing each heartbeat (hundreds of sample points) into a 64-dimensional code and then compressing an entire sequence of beats into a 128-dimensional record vector, the autoencoder may have discarded subtle but important information needed to differentiate certain diagnoses. This information bottleneck can hurt classification – for example, fine-grained timing differences or low-amplitude waveform nuances might be lost in the compression. Mismatch between learned vs. discriminative features: The autoencoder was optimized to minimize reconstruction error, not to maximize classification accuracy. Thus, it likely learned features capturing dominant morphological patterns (to faithfully rebuild signals) rather than the specific anomalies that signal different arrhythmias. Those latent features could be "orthogonal" to the features that best separate classes, making it hard for the MLP (or the hybrid model's trees) to translate them into better predictions. In short, the deep model's abstract features did not add significant new predictive signal beyond what the manually engineered features already provided. Consequently, we failed to support Hypothesis 3 – incorporating unsupervised deep features did not significantly enhance classification accuracy or the Challenge metric in this study.

Another important aspect is data quality. The public ECG dataset had several limitations that we did not fully address in preprocessing, and these likely affected all models' performance. There was evident label noise and inconsistency – for example, some records were clearly bradycardic (heart rate < 60 bpm) yet not labelled as such, and there were cases of low-voltage QRS complexes being labelled as atrial fibrillation or other rhythm abnormalities. We also observed instances where the distinction between atrial fibrillation and atrial flutter was inconsistently labelled. Such mislabels (or missing labels for certain conditions) introduce confusion during training: the classifiers might learn to predict "incorrect" patterns or ignore certain abnormalities because they are not reliably annotated. Additionally, the ECG signals

showed significant artifacts in some cases – e.g., baseline wander that led to unrealistic voltage shifts, or extremely low signal-to-noise ratios where true P/QRS/T waves were barely discernible. We did not perform advanced filtering or artifact removal beyond basic normalization, meaning the models had to cope with this noise. These dataset issues (noisy signals, missing or incorrect labels) likely prevented higher accuracy. Even an ideal algorithm would struggle if some arrhythmias are unlabelled or if noisy recordings are present with misleading labels. In future work, refining the dataset by removing or relabelling questionable records and reducing artifacts could lead to overall improvements in model performance. Ultimately, the limitations of the training data – including label noise, incomplete annotation of certain arrhythmias, and various ECG artifacts – constrained the accuracy achievable by both the shallow and deep learning approaches. We recognize that our choice to apply minimal preprocessing was a trade-off: it preserved data quantity and variability but came at the cost of introducing more noise. Addressing these data quality challenges will be essential to further enhance model performance. Lastly, we note that our choice of classifier and feature set also influences outcomes. We used gradient-boosted trees (XGBoost) for the feature-based models due to their robust handling of high-dimensional data and strong performance in many settings. It would be valuable to explore whether other classifiers (e.g., SVMs or random forests) using the same manual feature set could achieve similar results – perhaps the boosted trees had no special advantage beyond being well-tuned for this task. Moreover, our manual feature generation yielded thousands of features using general time-series libraries. While this broad approach helped initial performance, it likely included redundant or irrelevant features. A more targeted feature design using clinical expertise (focusing on known ECG markers for each condition) could produce a smaller, more interpretable feature set that rivals the larger set in accuracy. This could improve efficiency and transparency, as the model would rely on medically meaningful features.

VI.CLINICAL IMPLICATIONS

From the clinical point of view, the study shows the promise and the limitations of automated ECG classification. The algorithms were more accurate in some cardiac conditions than others. Notably, conditions with specific waveform changes were identified more reliably. For instance, bundle branch blocks and axis deviations – conditions with definite morphological changes in ECG – were some of the best-identified conditions. The deep-learning model had modestly higher sensitivity for conditions like incomplete right bundle branch block (IRBBB), left anterior fascicular block (LAnFB), prolonged PR interval, and right-axis deviation than the feature-based one. This makes sense, because these conditions affect specific intervals or waveforms (e.g., the QRS shape for IRBBB, the measurable interval of the PR for prolonged interval) that the algorithms – more specifically the autoencoder – were tuned to identify. When compared with other scenarios, the models underperformed in situations of mild, transient arrhythmia or otherwise noise-influenced situations. For example, separating atrial fibrillation from atrial flutter or other atrial 'arrhythmias proved challenging partly due to inaccurate categorizations in the datasets the models were trained on and partly because underlying features of AF (such as an irregularly irregular rhythm and an absence of P-waves) could easily be obscured or masked by noise or other atrial activity. Likewise, conditions of low-amplitude T-wave abnormalities or subtle ischemic changes were the most difficult to identify because they involve the detection of fine waveform variations neither captured adequately by our features nor by the autoencoder. Occasionally, the algorithms would make erroneous predictions – for instance, identifying a record to have a “T-wave abnormality” where the signal was noisy and where there were no visible T-waves, revealing likely false alarms caused by artifacts. Generally, high-amplitude or timing-based abnormalities (e.g., blocks and axis shifts) were more easily identified by the algorithms than were rhythm disorders or low-voltage changes hidden in noise. Balancing false alarms and missed events is important in evaluating clinical utility. Our top-performing model – a hybrid ensemble – tended to favor sensitivity for some diagnoses due to the weighting of the challenge metric. This caught more cases of severe arrhythmias but resulted in some false alarms. For instance, the model sometimes marked recordings as atrial fibrillation where irregularity resulted from motion artifacts. Such false alarms might result in unwarranted testing were it to be used in clinical practice. Misclassifications were also seen in recordings where baseline wander or noise was severe – the model outputted AF, atrial flutter, or “T-wave abnormality” where no actual arrhythmia was present. Left uncontrolled, these false alarms might lead to alarm fatigue in clinical environments. Conversely, the models at times missed arrhythmias detected by the cardiologist – e.g., infrequent premature beats or minor ST-segment shifts in ischemia suspicion. Incidentally, some records with overt bradycardia (severely slowed heart rate) were neither marked nor detected by the model,

presumably because bradycardia was sporadically tagged in the training material. Omitting such important events (false negatives) is especially troublesome in medicine because it would result in the patient's clinical status being undertreated. Though we did not report sensitivity for life-threatening arrhythmias per se because the setup is multi-label, the modest total sensitivity suggests some clinically significant events would routinely be omitted by the model in its present incarnation. It's also important to interpret the metrics of the performance in context. Our top model obtained around 30.7% overall accuracy in the test-split, by which we mean the complete list of multi-label diagnoses was correct in approximately one-third of cases. At first, 30% accuracy might appear poor in comparison to the average single-label task. All the same, in the multi-label classification task of 27 possible diagnoses, this is not directly comparable to 90% accuracy in e.g. the two-class task. The random guess or the frivolous classifier would obtain way below 30%, hence the model is undoubtedly extracting signal from the information. Nonetheless, from the clinical point of view, 30% accuracy (as well as the Challenge score of ~0.4366) is way from being enough for one's own diagnostic usage. In practice, it would mean the algorithm's output set of diagnoses for an ECG would be correct in the full set simply less than one-third of the time – quite insufficient for clinical decision-making. Doctors cannot tolerate missing 70% of the diagnoses or tolerating constant false alarms in the everyday workflow. At the current performance level, the model is best thought of as a decision support tool instead of an independent diagnostic system. For instance, model performance could pre-screen or flag some ECGs; even at 30% accuracy level, the model may mark ECGs as potentially abnormal for further consideration or provide a suggested list of conditions for clinicians to consider. This may highlight cases that may go otherwise unnoticed and provide a "second set of eyes". False alarms would need to be low though; too many false positive alerts is a recipe for clinicians to lose trust in the system. In our these results, the precision for some conditions were low and associated with many false positives. This illustrates the need for further refinements so these alerts are more specific. Overall, we have demonstrated proof of principle that the current model achieves accuracy and error rates for multi-label ECG classification with traditional and deep features combined, but the system is not yet clinically useful. There is much room for improvement, especially where it comes to improving sensitivity for critical arrhythmias and reducing false positives before such a model could meaningfully decrease either missed events or false alarms in the context of cardiac monitoring. Improvements such as more complete and reliable data, adding additional leads or patient data, or using more sophisticated architectures (i.e. transformer or attention models appropriate for the 12-lead ECG), may be the key to attaining the accuracy needed for clinically meaningful use.

VII.CONCLUSION

This study presented and compared three different methods for multi-label classification of 12-lead ECG records. As a starting point, we applied a methodology using conventional signal processing and extensive feature extraction with a shallow gradient-boosted trees ensemble. Second, we built a deep "beat-to-sequence" autoencoder model to autonomously learn features from raw ECG signals and used its embeddings within an MLP classifier. Lastly, we experimented with a hybrid approach, where deep autoencoder features were integrated with manually extracted features in an ensemble of gradient-boosted trees (with label-specific feature selection). The experimental evaluation addressed the hypotheses from the introduction. We confirmed Hypothesis 1: the classic feature-based ensemble performed better than the purely deep learning approach in terms of the Challenge metric and F-measure, supporting our expectation that a thoughtfully designed shallow model can rival or beat a deep neural network in this setting. We partially confirmed Hypothesis 2: prioritizing regularization of the feature inputs – through pruning and selecting the most informative features – was more beneficial than simply adding more features from the autoencoder without selection. In other words, judicious feature selection improved the Challenge score more than the naive inclusion of additional deep features. We did not find support for Hypothesis 3: combining deep autoencoder-derived features with the handcrafted feature set did not produce a statistically significant increase in classification performance. Despite the intuitive appeal of enriching the feature space with unsupervised learned features, our best results were achieved by the hybrid model with label wise top-1000 feature selection of autoencoder embeddings – and even that was on par with, not significantly above, the purely manual feature model. This winning configuration attained an average Challenge score of 0.4366 and an overall accuracy of ~30.7% on our test splits. These figures, while modest in absolute terms, were the highest in our comparisons. They highlight that combining traditional ECG features with modern machine learning can yield competitive results, but also that the deep features did

not offer a breakthrough improvement given our approach. In conclusion, our work illustrates both the potential and the challenges of multi-label 12-lead ECG classification: with careful feature design and model tuning, a relatively interpretable model (boosted trees on engineered features) can perform on par with a deep learning model, and a fusion of the two can work if feature selection is employed. However, the lack of a clear performance boost from the autoencoder features suggests that future deep learning efforts need to capture information complementary to what traditional features provide. We believe that incorporating more advanced deep architectures (e.g., 12-lead convolutional or transformer networks) and improving data quality will be important steps forward. The metrics achieved here set a baseline, but are not yet at a level for clinical adoption – bridging that gap will require both algorithmic innovations and perhaps new forms of model validation focusing on clinical relevance (e.g., reducing critical arrhythmia misses and alarm fatigue).

VIII. FUTURE WORK

There are multiple valuable avenues for further research to develop this work. One main focus is improving and augmenting the dataset. As we observed, the current training dataset is impacted by label uncertainty and other issues. Taking measures to clean the dataset - e.g., fixing mislabelled records, excluding excessively noisy ECG records, if any, ensuring labellers adhere to a well-defined set of labelling criteria - would likely improve model performance significantly. Furthermore, adding more data, particularly for under-represented arrhythmias, and/or utilizing data augmentation techniques, may similarly improve the generalization of deep learning models to previously unseen cases.

Another avenue that would be most valuable to explore would be assessing other lead arrangements and modalities. For instance, the dataset from the PhysioNet/CinC 2020 Challenge used 12-lead ECGs, but the 2021 Challenge was based on 2-lead recordings. The exploration of the features-based and hybrid models with minimizing number of leads would provide information on the robustness of the models, and potentially lead to modifications to the models, such as features that are most relevant to specific leads. On the other hand, the incorporation of some additional sources of complementary information, such as demographics of patients or symptoms, could provide additional context to the model - e.g., some arrhythmias are seen more commonly with older patients, or patients who have certain risk factors.

On the modelling side, cutting-edge deep learning techniques for time series hold strong potential. Transformer-based architectures, in particular, have shown great success at capturing long-range dependencies in sequential data. Recent studies, such as work by Natarajan et al., have demonstrated that "wide and deep" transformer models can process raw 12-lead ECG waveforms alongside derived features to achieve state-of-the-art arrhythmia classification. Extending these transformer approaches to our multi-label task – perhaps combined with the expert features we developed – is a natural next step. Such models might uncover subtle waveform patterns or lead interactions that our autoencoder or manually engineered features missed.

To conclude, expanding the classification paradigm to a larger set of ECG findings would improve the clinical utility of the model. Our work, like the referenced challenge, was limited to 27 diagnoses but real-life ECG interpretation requires the consideration of a....future work could try to train a more comprehensive multi-label model that consisted of additional arrhythmias and ECG abnormalities (e.g., more subtle ST/T changes, patterns of hypertrophy, etc.). While this would create new challenges (e.g., larger number of classes, extreme imbalance in the data), any success in this domain would propel us closer to achieving the concept of an AI generalist assistant for ECG interpretation. In conclusion, the follow-up steps are performed in parallel: improvements in data, experimenting with more advanced deep learning architectures (while maintaining interpretability), and incorporating more diagnostic categories into the model's training - all in the interest of establishing some degree of reliability and clinical meaningfulness of an ECG classifier.

Conflict of Interest

There is no conflict to declare.

Acknowledgement

The author would like to acknowledge the Department of Computer Science at Wichita State University for its support and resources that contributed to the successful completion of this research.

REFERENCES

[1] S. S. Virani, A. Alonso, E. J. Benjamin, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, F. N. Delling, L. Djousse, M. S. Elkind, J. F. Ferguson, M. Fornage, S. S. Khan, B. M. Kissela, K. L. Knutson, T. W.

Kwan, D. T. Lackland, T. T. Lewis, J. H. Lichtman, C. T. Longenecker, M. S. Loop, P. L. Lutsey, S. S. Martin, K. Matsushita, A. E. Moran, M. E. Mussolino, A. M. Perak, W. D. Rosamond, G. A. Roth, U. K. Sampson, G. M. Satou, E. B. Schroeder, S. H. Shah, C. M. Shay, N. L. Spartano, A. Stokes, D. L. Tirschwell, L. B. VanWagner, and C. W. Tsao, "Heart disease and stroke statistics update: A report from the American Heart Association," *Circulation*, vol. 141, no. 9, pp. e139–e596, 2020. https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000000757

[2] H. Smulyan, "The computerized ECG: friend and foe," *The American Journal of Medicine*, vol. 132, no. 2, pp. 153–160, 2019. http://www.sciencedirect.com/science/article/pii/S0002934318308532

[3] R. O. Bonow, D. L. Mann, D. P. Zipes, and P. Libby, *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*. Elsevier Health Sciences, 2011.

[4] C. Breen, G. Kelly, and W. Kernohan, "ECG interpretation skill acquisition: A review of learning, teaching and assessment," *Journal of Electrocardiology*, 2019. http://www.sciencedirect.com/science/article/pii/S0022073618306411

[5] AliveCor. AliveCor KardiaMobile & KardiaMobile 6L. https://www.alivecor.com/ (Accessed November 4, 2020). https://www.alivecor.com/

[6] Apple. Apple Watch. https://www.apple.com/ca/watch/ (Accessed November 4, 2020). https://www.apple.com/ca/watch/

[7] QardioMD. QardioMD: Wireless ECG Monitoring with QardioCore. https://www.getqardio.com/en/qardiomd-ecg/ (Accessed November 4, 2020). [Online]. Available: https://www.getqardio.com/en/qardiomd-ecg/

[8] General Electric Healthcare – Diagnostic ECG. https://www.gehealthcare.com/products/diagnostic-ecg (Accessed November 4, 2020). https://www.gehealthcare.com/products/diagnostic-ecg

[9] Koninklijke Philips – Diagnostic ECG. https://www.usa.philips.com/healthcare/solutions/diagnostic-ecg/diagnostic-ecg (Accessed November 4, 2020). https://www.usa.philips.com/healthcare/solutions/diagnostic-ecg/diagnostic-ecg

[10] Kligfield Paul, Gettes Leonard S., Bailey James J., Childers Rory, Deal Barbara J., Hancock E. William, van Herpen Gerard, Kors Jan A., Macfarlane Peter, Mirvis David M., Pahlm Olle, Rautaharju Pentti, and Wagner Galen S., "Recommendations for the Standardization and Interpretation of the Electrocardiogram," *Journal of the American College of Cardiology*, vol. 49, no. 10, pp. 1109–1127, Mar. 2007, publisher: American College of Cardiology Foundation. https://www.jacc.org/doi/full/10.1016/j.jacc.2007.01.024

[11] E. A. Perez Alday, A. Gu, A. Shah, C. Robichaux, A.-K. I. Wong, C. Liu, F. Liu, A. B. Rad, A. Elola, S. Seyedi, Q. Li, A. Sharma, G. D. Clifford, and M. A. Reyna, "Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020," *Physiological Measurement*, 2020, In Press.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794.

[13] R. K. Vinayak and R. Gilad-Bachrach, "DART: Dropouts meet Multiple Additive Regression Trees," in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, Feb. 2015, pp. 489–497. http://proceedings.mlr.press/v38/korlakaivinayak15.html

[14] Y. Bengio. Deep learning challenges. CS-Can 2020. https://cscan-infocan.ca/feature-on-homepage/watch-deep-learning-challenges-with-yoshua-bengio/ (Accessed Nov 3, 2020). https://cscan-infocan.ca/feature-on-homepage/watch-deep-learning-challenges-with-yoshua-bengio/

[15] A. Natarajan, Y. Chang, S. Mariani, A. Rahman, G. Boverman, S. Vij, and J. Rubin, "A Wide and Deep Transformer Neural Network for 12-Lead ECG Classification," in *2020 Computing in Cardiology (CinC) Challenge*, 2020, pp. 1–4. https://raw.githubusercontent.com/physionetchallenges/physionetchallenges.github.io/master/2020/papers/107.pdf