Improved Generalization and Interpretability capabilities of a Deep Learning model for automatic diagnosis of 12-lead ECG

Syed Saqib Habeeb
`bm20btech11015@iith.ac.in`

Tanmay Goyal
`ai20btech11021@iith.ac.in`

Abstract

Globally, cardiovascular diseases (CVDs) are the leading cause of death, imposing significant health and economic strains. Electrocardiograms (ECGs) are essential for monitoring heart activity and detecting cardiac arrhythmias. These tools are simple, effective, and non-invasive, generally comprising 12 leads connected to electrodes on the body, with six attached to the limbs and six to the chest. There is an increasing need for experts who can reliably interpret these readings and predict abnormalities. This is where automating the task of performing diagnosis based on the ECGs comes into the picture. The authors of [6] developed a deep model that showed better performance than machine learning models which used expert extracted features. However, one of the limitations noted by the authors was the generalizability of the model. In our paper, we build upon this model by incorporating data from the PTB-XL dataset [5] and interpret the results using SHAP values and Grad-CAMs.

1. Introduction

Globally, cardiovascular diseases (CVDs) are the leading cause of death, imposing significant health and economic strains. Electrocardiograms (ECGs) are essential for monitoring heart activity and detecting cardiac arrhythmias. These tools are simple, effective, and non-invasive, generally comprising 12 leads connected to electrodes on the body, with six attached to the limbs and six to the chest. Despite the high reliability of ECGs, their interpretation can be complex, and even experienced cardiologists may face challenges that lead to less than optimal clinical outcomes.

Every year, about 300 million ECGs are performed across the globe, and the frequency of these tests is increasing. In areas, particularly low and middle-income countries, where there is a notable lack of experienced cardiologists, the reliance on computer-aided ECG interpretation is becoming more critical, pushing forward the advancement of automated ECG analysis technologies as a prominent research priority.

In our study, we expanded an existing model to a broader context by enhancing the training dataset. We mapped the 52 categories from the PTB-XL dataset to the 9 categories of the CPSC dataset, thus broadening the data's range and effectiveness. We train two different models: the first being the original model trained on the new data set and the second being a completely new model trained on both datasets from scratch. We find that both our models perform almost as well as the original model. To advance the model's interpretability, we employed SHAP values. Furthermore, we integrated Gradient-weighted Class Activation Mapping (Grad-CAM) techniques to boost both the performance and the explanatory capability of the model.

2. Problem Statement

The authors of [6] noted that their model was entirely trained on the CPSC dataset [2] which consists of data entirely collected from China hospitals. They wished to expand on the generalizability of the model. Towards this, we decided to incorporate data from the PTB-XL dataset. Once we ran preliminary tests on the PTB-XL dataset using the original model, we found that the authors were indeed right about their concerns regarding the model's generalizability. The results are shown in 5. Further, we wished to interpret this results and check if our model predictions' were in line with domain expertise.

3. Literature Review

Deep learning has been extensively applied in various medical imaging and diagnostic fields due to its ability to learn hierarchical representations and complex patterns from large datasets. Notably, convolutional neural networks and recurrent neural networks have shown particular promise in image and sequence data analysis. Focusing on ECG, deep learning models have been tailored to improve the diagnostic process of cardiovascular diseases. For instance, Hannun et al. (2019) [1] demonstrated that a deep CNN could outperform cardiologists in diagnosing atrial fibrillation from ECG data. Zhang et al. [6] developed a deep model which extracted features from the ECG to de-

tect cardiological abnormalities.
Recent innovations have introduced models incorporating attention mechanisms and transfer learning to enhance diagnostic accuracy and reduce the need for vast labeled datasets. Sajad et al. [3]implemented an attention-based deep learning framework that improves the interpretability of diagnostic decisions, a critical factor in clinical settings.

4. Methodology

4.1. Dataset

The PTB-XL dataset consists of 21,799 clinical 12-lead ECGs from 18,869 patients of 10 second length. The data consisted of 52 categories of various cardiovascular diseases which covered diagnostic, form, and rhythm statements.

Since we wished to retain the classes of our original model, we mapped each of these 52 categories to the 9 categories on which the original model was trained. The 9 original categories included normal sinus rhythm (SNR), atrial fibrillation (AF), first-degree atrioventricular block (IVAB), left bundle branch block (LBBB), right bundle branch block (RBBB), premature atrial contraction (PAC), premature ventricular contraction (PVC), ST-segment depression (STD), ST-segment elevation (STE).

After performing the mapping, we removed 11 classes from the PTB-XL dataset which were not remotely close to any of the 9 classes in the original model. We also dropped a few data points (258 in number) which were labelled by one or more of only these classes that were dropped. These 11 classes were: ischemic in inferior leads (ISCIN), ischemic in inferolateral leads (ISCIL), digitalis-effect (DIG), ischemic in lateral leads (ISCLA), right ventricular hypertrophy (RVH), long QT-interval (LNGQT), right atrial overload/enlargement (RAO/RAE), electrolytic disturbance or drug (former EDIS) (EL), ischemic in anterior leads (ISCAN), septal hypertrophy (SEHYP), and sinus tachycardia (STACH).

After the preprocessing, the class-wise distribution of the dataset is given in table 1

4.2. Architecture

We retain the same model architecture as the authors of the original paper. The model consists of 34 layers, which include 4 stacked residual blocks. Each residual block consists of two 1D convolutional layers, two batch normalization layers, 1 dropout layer, and two ReLU activation layers. The outputs from the residual blocks are pooled and these results are concatenated and sent to the output layer which used a sigmoid activation to make predictions. The architecture has been shown in figure 1.

Class	**Count (%)**	**Male (%)**	**Age**
SNR	14373 (66.72)	6921 (48.15)	58.88 (29.09)
AF	207 (0.96)	123 (59.42)	61.02 (29.23)
IAVB	1536 (7.13)	951 (61.91)	74.13 (41.18)
LBBB	2231 (10.36)	1343 (60.20)	77.51 (46.88)
RBBB	1735 (8.05)	1130 (65.13)	66.80 (38.05)
PAC	324 (1.50)	184 (56.79)	73.03 (38.64)
PVC	1027 (4.77)	609 (59.30)	73.54 (40.05)
STD	728 (3.38)	284 (39.01)	73.29 (41.47)
STE	5469 (25.39)	3407 (62.30)	71.44 (36.55)

Table 1. Description of the PTB-XL dataset after pre-processing

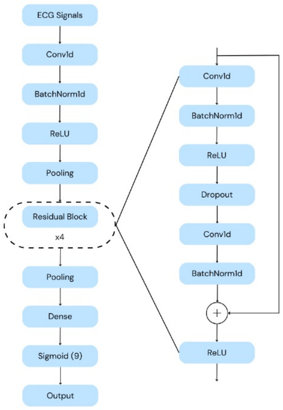


Figure 1. The architecture of the model

4.3. Training

We first used the original code given here to train the original model and reproduce the results of [6]. We found that a few of the metrics we obtained were slightly off than the metrics claimed in the original paper. We shall use the metrics we obtained in the rest of the paper.

We used the same hyperparameters as the authors of the original paper: learning rate of 0.0001, 30 epochs, and a batch size of 32. However, we trained two different models:

1. Trained the original model from [6] on the PTB-XL dataset. This model was then tested on a combination of both CPSC and PTB-XL data points. We shall refer to this model as ”Ours_1”

2. Created a model from scratch and trained it on a dataset comprising data from both CPSC and PTB-XL. This model was again tested on the combined dataset. We shall refer to this model as ”Ours_2”

4.4. Evaluation Metrics

We use the following metrics for evaluating our models:

1. Precision: For class i,

$$Precision_i = \frac{TP_i}{TP_i + FP_i}$$

2. Recall: For class i

$$Recall_i = \frac{TP_i}{TP_i + FN_i}$$

3. F1 score: For class i

$$F1_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}$$

4. AUC: Area under the ROC curve
5. Accuracy: For class i

$$Accuracy_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$$

4.5. SHAPs for interpretability

We use SHAP values to explain and interpret our results. Shapley Addditive explanations are a way to explain the output of the models by checking for a subset of features which best explains the output of the model. For feature i, we iterate over all possible subsets which do not include the feature i and check the output of the model with and without the feature i to find it's contribution.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S|| - 1)!}{|F|!}$$

$$\times [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]$$

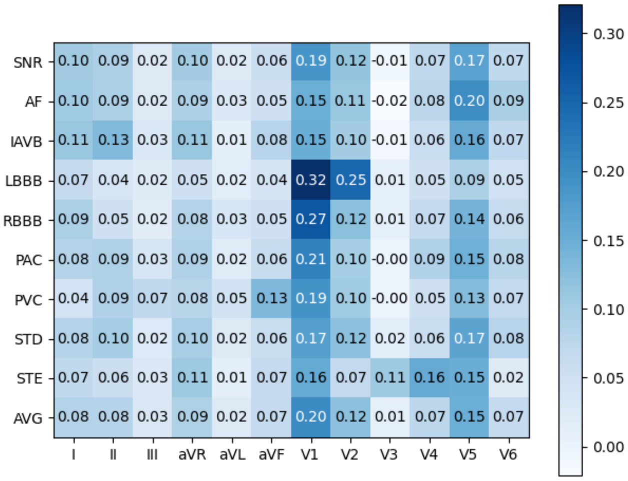
We first obtain the SHAP values for the original model on both datasets. Further, we then test both datasets on ”Our” model to check if the same patterns hold.

4.6. Grad-CAMs for interpretability

We also use Grad-CAMs to highlight the areas in the ECGs that get activated. Grad-CAM stand for gradient weighted Class Activation Mappings, which use the gradients flowing into a layer to produce a map highlighting the important reasons. Grad-CAM calculate the gradients of the output for a class before the softmax with respect to the feature map activations of a convolutional layer. In general, if α_k^c is the importance weight for the neuron representing class c and we wish to find the gradient with respect to feature activation maps A^k, then,

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

where $\frac{1}{Z}\sum\limits_i \sum\limits_j (.)$ computes the global average pooling of the gradients.



	I	II	III	aVR	aVL	aVF	V1	V2	V3	V4	V5	V6
SNR	0.10	0.09	0.02	0.10	0.02	0.06	0.19	0.12	-0.01	0.07	0.17	0.07
AF	0.10	0.09	0.02	0.09	0.03	0.05	0.15	0.11	-0.02	0.08	0.20	0.09
IAVB	0.11	0.13	0.03	0.11	0.01	0.08	0.15	0.10	-0.01	0.06	0.16	0.07
LBBB	0.07	0.04	0.02	0.05	0.02	0.04	0.32	0.25	0.01	0.05	0.09	0.05
RBBB	0.09	0.05	0.02	0.08	0.03	0.05	0.27	0.12	0.01	0.07	0.14	0.06
PAC	0.08	0.09	0.03	0.09	0.02	0.06	0.21	0.10	-0.00	0.09	0.15	0.08
PVC	0.04	0.09	0.07	0.08	0.05	0.13	0.19	0.10	-0.00	0.05	0.13	0.07
STD	0.08	0.10	0.02	0.10	0.02	0.06	0.17	0.12	0.02	0.06	0.17	0.08
STE	0.07	0.06	0.03	0.11	0.01	0.07	0.16	0.07	0.11	0.16	0.15	0.02
AVG	0.08	0.08	0.03	0.09	0.02	0.07	0.20	0.12	0.01	0.07	0.15	0.07

Figure 2. SHAP values obtained for the original model trained on the CPSC dataset

We compute the Grad-CAMs at the end of the 4th residual block and threshold the gradients at the 95^{th} percentile.

5. Results

We tested the original model from [6] on the PTB-XL dataset to test the generalizability of the model. The results are tabulated in table 2. Clearly, the metrics showcase that the model proposed in [6] was not generalizing well enough.

We also obtained the precision, recall, F1 score, AUC, and accuracy for each of the classes and compared them with our newly trained models, ”Ours_1” and ”Ours_2”. The results are tabulated in table 3

We find that ”Ours_2” model performs better than ”Ours_1” model in all aspects. From now on, we shall refer to ”Ours_2” as ”Our” model. We find that ”Our” model has comparable AUC and accuracy to the original model proposed in [6] even though the F1 score is slightly lower. We shall present the interpretability results only for this model.

We also obtained the population level SHAP values for all the leads and the classes.

We know that lead II is positioned from the negative pole of the right arm to the positive pole of the left leg. This positioning makes it a good vantage point for both atrial and ventricular depolarization paths. This can be also be observed in the SHAP metrics obtained shown in 5. Higher comparative SHAP values for lead II can be explained as follows :

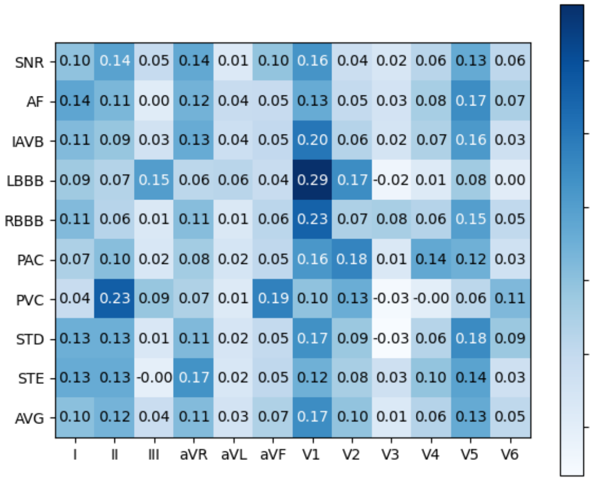
1. For Atrial Fibrillation (AF), this lead shows a moderately high SHAP value due to its alignment with

	Precision		Recall		F1		AUC		Accuracy	
	CPSC	**PTB_XL**	**CPSC**	**PTB_XL**	**CPSC**	**PTB_XL**	**CPSC**	**PTB_XL**	**CPSC**	**PTB_XL**
SNR	0.786	0.863	0.906	0.387	0.842	0.534	0.972	0.713	0.958	0.550
AF	0.938	0.066	0.931	0.522	0.934	0.117	0.989	0.721	0.975	0.925
IAVB	0.894	0.259	0.894	0.393	0.894	0.312	0.988	0.728	0.974	0.877
LBBB	0.935	0.502	1.000	0.427	0.967	0.461	1.000	0.805	0.997	0.897
RBBB	0.913	0.554	0.966	0.813	0.939	0.659	0.993	0.943	0.968	0.932
PAC	0.771	0.035	0.857	0.068	0.812	0.046	0.972	0.203	0.964	0.957
PVC	0.922	0.760	0.787	0.537	0.849	0.629	0.973	0.947	0.969	0.970
STD	0.782	0.078	0.756	0.387	0.768	0.130	0.938	0.635	0.940	0.824
STE	0.474	0.330	0.500	0.157	0.486	0.213	0.911	0.544	0.972	0.705
AVG	0.824	0.383	0.844	0.410	0.832	0.345	0.971	0.693	0.969	0.848

Table 2. Testing the original model on the CPSC and PTB-XL datasets to test for generalizability

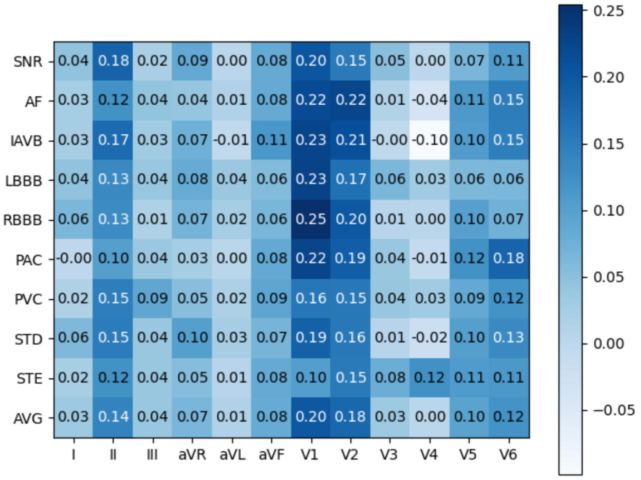
	Precision			Recall			F1			AUC			Accuracy		
	Original Model	**Ours_1 Model**	**Ours_2 Model**	**Original Model**	**Ours_1 Model**	**Ours_2 Model**	**Original Model**	**Ours_1 Model**	**Ours_2 Model**	**Original Model**	**Ours_1 Model**	**Ours_2 Model**	**Original Model**	**Ours_1 Model**	**Ours_2 Model**
SNR	0.786	0.756	0.805	0.906	0.894	0.925	0.842	0.819	0.861	0.972	0.881	0.929	0.958	0.787	0.839
AF	0.938	0.500	0.880	0.931	0.508	0.780	0.934	0.504	0.827	0.989	0.927	0.983	0.975	0.954	0.985
IAVB	0.894	0.636	0.600	0.894	0.502	0.579	0.894	0.561	0.590	0.988	0.916	0.892	0.974	0.936	0.934
LBBB	0.935	0.734	0.777	1.000	0.817	0.850	0.967	0.773	0.812	1.000	0.978	0.984	0.997	0.958	0.966
RBBB	0.913	0.864	0.920	0.966	0.874	0.837	0.939	0.869	0.876	0.993	0.989	0.992	0.968	0.967	0.970
PAC	0.771	0.862	0.833	0.857	0.258	0.722	0.812	0.397	0.773	0.972	0.826	0.968	0.964	0.973	0.986
PVC	0.922	0.735	0.709	0.787	0.925	0.908	0.849	0.819	0.796	0.973	0.985	0.989	0.969	0.975	0.971
STD	0.782	0.310	0.790	0.756	0.427	0.482	0.768	0.359	0.598	0.938	0.855	0.941	0.940	0.912	0.963
STE	0.474	0.638	0.723	0.500	0.689	0.773	0.486	0.663	0.747	0.911	0.903	0.941	0.972	0.859	0.894
AVG	0.824	0.670	0.782	0.844	0.655	0.762	0.832	0.640	0.765	0.971	0.918	0.958	0.969	0.924	0.945

Table 3. Metrics for our new models compared to the metrics for the original model



	I	II	III	aVR	aVL	aVF	V1	V2	V3	V4	V5	V6
SNR	0.10	0.14	0.05	0.14	0.01	0.10	0.16	0.04	0.02	0.06	0.13	0.06
AF	0.14	0.11	0.00	0.12	0.04	0.05	0.13	0.05	0.03	0.08	0.17	0.07
IAVB	0.11	0.09	0.03	0.13	0.04	0.05	0.20	0.06	0.02	0.07	0.16	0.03
LBBB	0.09	0.07	0.15	0.06	0.06	0.04	0.29	0.17	-0.02	0.01	0.08	0.00
RBBB	0.11	0.06	0.01	0.11	0.01	0.06	0.23	0.07	0.08	0.06	0.15	0.05
PAC	0.07	0.10	0.02	0.08	0.02	0.05	0.16	0.18	0.01	0.14	0.12	0.03
PVC	0.04	0.23	0.09	0.07	0.01	0.19	0.10	0.13	-0.03	-0.00	0.06	0.11
STD	0.13	0.13	0.01	0.11	0.02	0.05	0.17	0.09	-0.03	0.06	0.18	0.09
STE	0.13	0.13	-0.00	0.17	0.02	0.05	0.12	0.08	0.03	0.10	0.14	0.03
AVG	0.10	0.12	0.04	0.11	0.03	0.07	0.17	0.10	0.01	0.06	0.13	0.05

Figure 3. SHAP values obtained for the original model trained on the PTB-XL dataset

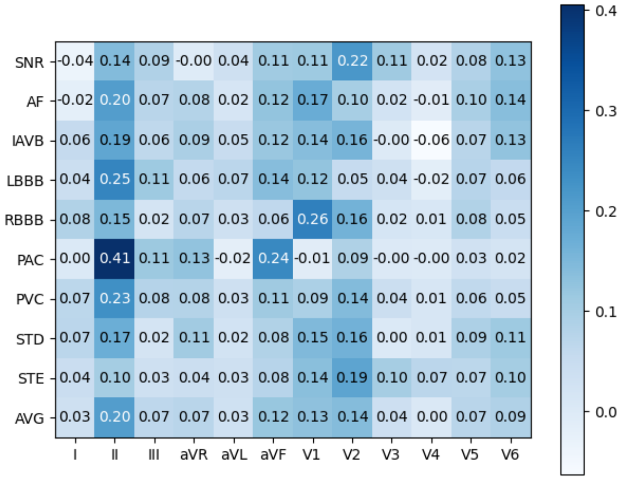


	I	II	III	aVR	aVL	aVF	V1	V2	V3	V4	V5	V6
SNR	0.04	0.18	0.02	0.09	0.00	0.08	0.20	0.15	0.05	0.00	0.07	0.11
AF	0.03	0.12	0.04	0.04	0.01	0.08	0.22	0.22	0.01	-0.04	0.11	0.15
IAVB	0.03	0.17	0.03	0.07	-0.01	0.11	0.23	0.21	-0.00	-0.10	0.10	0.15
LBBB	0.04	0.13	0.04	0.08	0.04	0.06	0.23	0.17	0.06	0.03	0.06	0.06
RBBB	0.06	0.13	0.01	0.07	0.02	0.06	0.25	0.20	0.01	0.00	0.10	0.07
PAC	-0.00	0.10	0.04	0.03	0.00	0.08	0.22	0.19	0.04	-0.01	0.12	0.18
PVC	0.02	0.15	0.09	0.05	0.02	0.09	0.16	0.15	0.04	0.03	0.09	0.12
STD	0.06	0.15	0.04	0.10	0.03	0.07	0.19	0.16	0.01	-0.02	0.10	0.13
STE	0.02	0.12	0.04	0.05	0.01	0.08	0.10	0.15	0.08	0.12	0.11	0.11
AVG	0.03	0.14	0.04	0.07	0.01	0.08	0.20	0.18	0.03	0.00	0.10	0.12

Figure 4. SHAP values obtained for the Ours_2 model for the CPSC dataset

the axis of atrial depolarization, making it sensitive to AF's irregular atrial activity characteristic.

2. Left and Right Bundle Branch Blocks (LBBB and RBBB) have a moderate SHAP value for lead II, implying its contribution to detecting ventricular depolarization delays. This lead can also show broadened QRS complexes typically seen in bundle branch blocks.



	I	II	III	aVR	aVL	aVF	V1	V2	V3	V4	V5	V6
SNR	-0.04	0.14	0.09	-0.00	0.04	0.11	0.11	0.22	0.11	0.02	0.08	0.13
AF	-0.02	0.20	0.07	0.08	0.02	0.12	0.17	0.10	0.02	-0.01	0.10	0.14
IAVB	0.06	0.19	0.06	0.09	0.05	0.12	0.14	0.16	-0.00	-0.06	0.07	0.13
LBBB	0.04	0.25	0.11	0.06	0.07	0.14	0.12	0.05	0.04	-0.02	0.07	0.06
RBBB	0.08	0.15	0.02	0.07	0.03	0.06	0.26	0.16	0.02	0.01	0.08	0.05
PAC	0.00	0.41	0.11	0.13	-0.02	0.24	-0.01	0.09	-0.00	-0.00	0.03	0.02
PVC	0.07	0.23	0.08	0.08	0.03	0.11	0.09	0.14	0.04	0.01	0.06	0.05
STD	0.07	0.17	0.02	0.11	0.02	0.08	0.15	0.16	0.00	0.01	0.09	0.11
STE	0.04	0.10	0.03	0.04	0.03	0.08	0.14	0.19	0.10	0.07	0.07	0.10
AVG	0.03	0.20	0.07	0.07	0.03	0.12	0.13	0.14	0.04	0.00	0.07	0.09

Figure 5. SHAP values obtained for the Ours_2 model for the PTB-XL dataset

3. For PAC and PVC, lead II has a high SHAP value, especially for PAC, likely because it can effectively show premature beats originating from the atria (for PAC) or the ventricles (for PVC), which can appear as early P waves or aberrant QRS complexes, respectively.

In addition to lead II, moderately high SHAP values can also be observed for lead aVF for PAC, and lead v1 for RBBB suggesting that in addition to lead II, these leads also play an important role in addition II in detection of the respective abnormalities.

An example where the Grad-CAMs help in detecting the features in an ECG wave is shown in figure 6. Here, for the correct prediction of LBBB, the Grad-CAMs detect the following features :

1. In leads V1 and V2, the Grad-CAMs detect deeper S waves and also delayed intrinsicoid deflection (slower than usual time to peak of the R wave) which are both indicative of LBBB.

2. The presence of notched - M shaped - R wave in lead V5, which is a strong indicactor of LBBB

3. Grad-CAMs detect broad R wave in leads I and aVL indicating abnormal depolarization of the lateral wall of the left ventricle. This is a charecteristic of LBBB.

Another example can be seen in figure 7. Here the Grad-CAMs observes the following features which helps in predicting the correct label.

1. Leads I, aVL emphasize on the S wave.

2. Lead V1 highlights a widened R' wave.

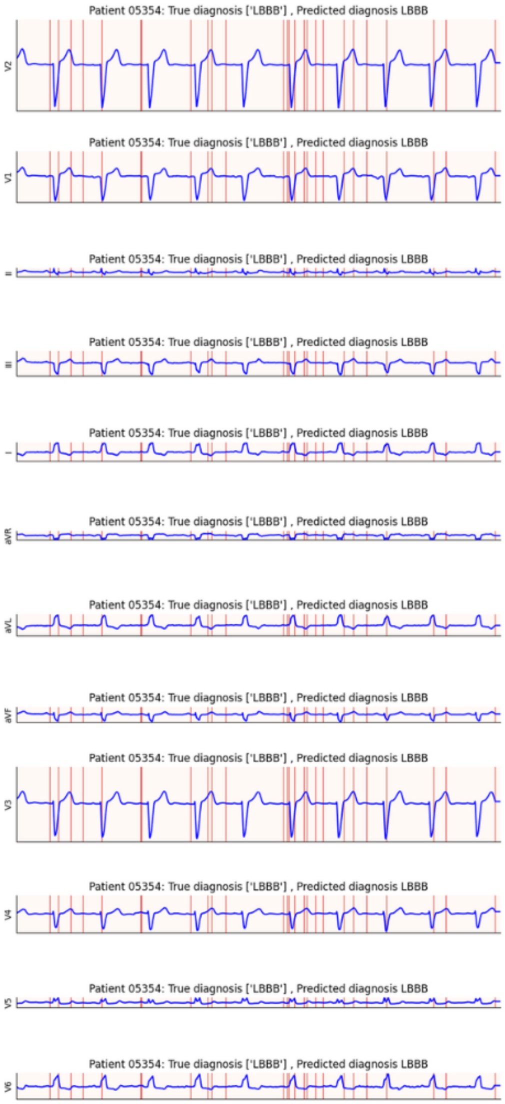


Figure 6. Grad-CAMs obtained for the correct prediction of LBBB

3. Leads V5, V6 concentrate on the delayed S wave.

The link to our code can be found here.

6. Conclusion

While our model has been trained on diverse data and should be able to solve the issue of generalizability, a few of the issues noted by the authors in [6] still persist. For instance, the response to adversarial instances, as well as, the combination of leads that should be used still remain open questions. Also, interpretability using Grad-CAMs does not always result in the correct areas being highlighted, as we shall depict in 7.

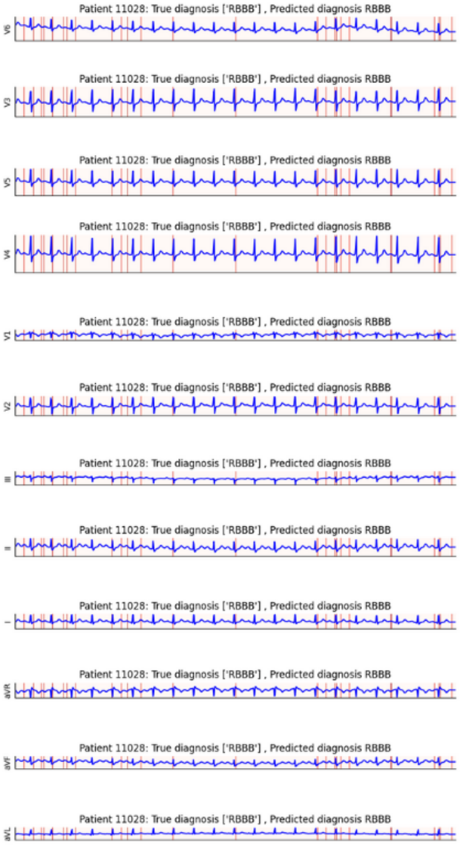


Figure 7. Grad-CAMs obtained for the correct prediction of RBBB

References

[1] Masoumeh Hagpanahi Geoffrey H. Tison Codie Bourn Mintu P. Turakhia Andrew Y. NG Awni Y. Hannun, Pranav Rajpurkar. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25:65–69, 2019. 1

[2] F F Liu, C Y Liu, L N Zhao, X Y Zhang, X L Wu, X Y Xu, Y L Liu, C Y Ma, S S Wei, Z Q He, J Q Li, and N Y Kwee. An open access database for evaluating the algorithms of ECG rhythm and morphology abnormal detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018. 1

[3] Sajad Mousavi, Fatemeh Afghah, Abolfazl Razi, and U. Rajendra Acharya. Ecgnet: Learning where to attend for detection of atrial fibrillation with deep visual attention. In *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 1–4, 2019. 2

[4] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct. 2019.

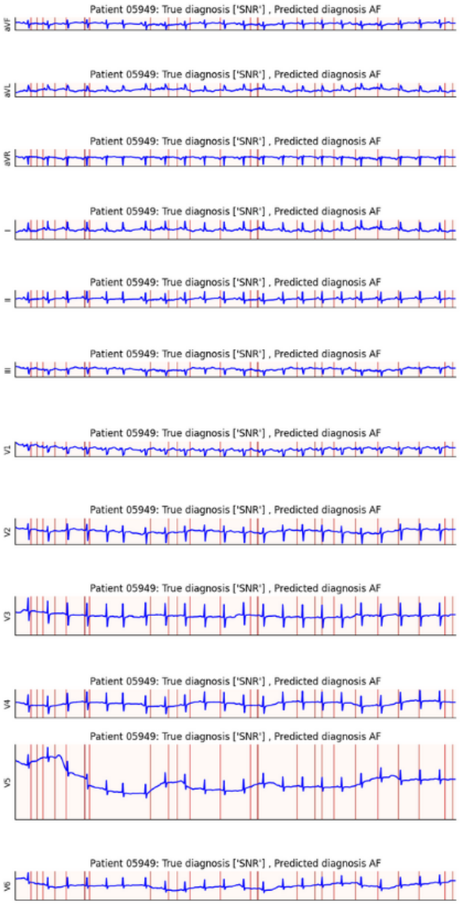


Figure 8. Grad-CAMs obtained for the incorrect prediction of SNR

[5] Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl, 2020. 1

[6] Dongdong Zhang, Xiaohui Yuan, and Ping Zhang. Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram, 2020. 1, 2, 3, 5

7. Appendix

The Grad-CAMs approach does not always give the right prediction. We can see this clearly in figure 8. Here the model incorrectly predicts SNR as AF. Here, the Grad-CAMs observe the following in the ECG:

1. Leads I, III, aVF, V3-V6 show a heavy emphasis on the QRS complex and T-waves.
2. Leads II, aVL focus on P-waves.
3. No significant observation in the lead aVR and V1.

4. Lead V2's focus is misplaced on the QRS complex rather than on baseline irregularities and P-wave morphology.

Some possible reasons for this wrong prediction are

1. Overemphasis on Ventricular Activity: The model appears to misinterpret the significance of ventricular activity (QRS and T-waves) over atrial activity.
2. P-wave Misinterpretation: The model's incorrect focus on P-waves in some leads suggests it might be recognizing atypical P-wave shapes as indicative of AF, possibly due to anomalies in the training examples.