The Impact of ECG Channel Reduction on Multi-Label Cardiac Diagnosis

D. Alexandrov and J. Picone

Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA
{dmitry.alexandrov, picone}@temple.edu

***Abstract*— Electrocardiogram (ECG) recordings, which graphically represent the electrical activity of the heart over time, are essential for diagnosing a variety of heart diseases by identifying abnormal patterns in the signal. In this work, we investigate the need for preprocessing ECG data using a ResNet18 deep learning model for the TNMG CODE Corpus. While traditional ECG interpretation often relies on preprocessed data and channels derived from linear combinations of the raw signal, we hypothesize that these techniques may be unnecessary or even detrimental in modern deep learning approaches. We systematically demonstrate that the ResNet18 model performance consistently decreases with the increase in the number of channels. We also conduct an ablation analysis, which reveals that derived ECG channels have minimal or detrimental effect. This study demonstrates the effectiveness of deep learning models in processing ECG data, supporting the hypothesis that model-based features are no longer needed when there is sufficient training data available. This decreases the effort required to develop machine learning systems for new domains, contributing to potential improvements in medical diagnostics.**

I. Introduction

Electrocardiograms (ECGs) are a fundamental tool in cardiology, allowing physicians to diagnose a wide range of cardiac abnormalities. Traditionally, cardiologists utilize preprocessed ECG data and derived channels, which are linear combinations of the raw signal, to aid in their diagnosis. However, with the rapid advancements in machine learning, automated analysis of medical data is becoming a reality. Deep learning models have shown a remarkable ability for extracting relevant features from complex data without explicit feature engineering. *This suggests that these models may be able to effectively analyze raw ECG signals, potentially eliminating the need for preprocessing.* This new approach could simplify the analysis pipeline and preserve subtle patterns in the data that might be lost during traditional preprocessing steps.

The Residual Network architecture (ResNet18), introduced by He et al. [1], is a significant model in deep learning for image recognition tasks. Its main innovation is the use of residual blocks, which allow the network to learn residual functions with reference to layer inputs. This approach enables the training of much deeper networks by addressing the vanishing gradient problem. The ResNet18 model consists of 18 layers, including convolutional layers, batch normalization, ReLU activation functions, and skip connections that form the characteristic residual blocks. These skip connections allow the network to bypass one or more layers, providing a direct route for gradients to flow backwards through the network during training. While originally designed for image classification, ResNet's ability to capture hierarchical features makes it well-suited for complex pattern recognition tasks, including time series analysis, where identifying both local patterns and global trends is crucial. ResNet18 has been successfully applied to EEG analysis (reducing complexity and latency) [2], and image analysis [3], in additional to cardiology applications [4]. In this paper we do not argue that this is the best overall architecture. We simply use this approach as a well-established baseline.

Most clinical ECG recordings are collected with a 10-lead system. These channels are converted to eight signal channels as shown in Figure 1. Prior to the introduction of deep learning, these eight signal channels were converted to twelve leads using a well-known set of preprocessing techniques [5]. A typical system employs ten electrodes: six precordial leads (V1-V6) placed on the chest, and four limb leads places on the right arm (RA), left arm (LA), left leg (LL), and right leg (RL, used as a ground). From these, eight raw waveforms are recorded: the six precordial leads and two limb leads, DI and DII. DI is derived as the potential difference between LA and RA (LA – RA), while DII is the difference between LL and RA (LL – RA). The remaining four leads are derived from DI and DII as follows:

$$DIII = DII - DI \tag{1}$$

$$aVR = \frac{DI+DII}{2} \tag{2}$$

$$aVL = \frac{DI-DII}{2} \tag{3}$$

$$aVF = \frac{DII-DI}{2} \tag{4}$$

We hypothesize that extensive preprocessing of ECG data may not be beneficial for deep learning models and could potentially decrease their performance.

II. TNMG CODE Corpus

The TNMG CODE Corpus (TNMG) [5] represents a major advancement in the field of cardiology, and is the corpus we focus on in this study. TNMG is a dataset of ECG records collected by the Telehealth Network of Minas Gerais (TNMG) between 2010 and 2016 in 811 counties in the Brazilian state of Minas Gerais, organized by the Clinical Outcomes in Digital Electrocardiography (CODE) group. The dataset contains a total of 6,716,317

979-8-3503-8857-2/24/$31.00 ©2024 IEEE

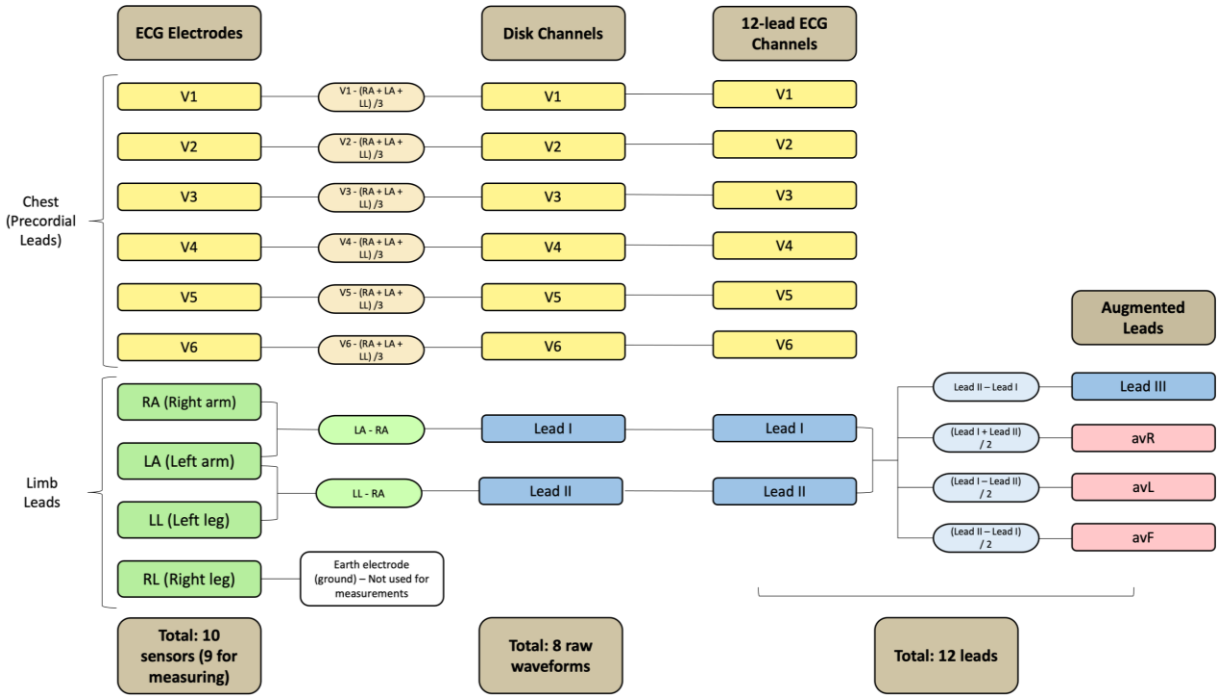

Figure 1. Conversion of an ECG collected with 10 leads to 8 and 12-channel waveforms [4]

annotated records from 1,558,749 patients.

The dataset includes a curated "golden dataset" of 827 ECG recordings, which serves as a high-quality evaluation set. These recordings were independently annotated by two cardiologists. In cases of disagreement, a third specialist reviewed the annotations to establish a consensus. The data set was labeled for six abnormalities as shown in Table 1.

In Table 2, we show a distribution in the number of records and percentage of feature vectors in both datasets, where presence or absence of each abnormality is marked as a binary vector in the same order as Table 1: 1dAVb, RBBB, LBBB, SB, AF, ST. It is clear that the majority of both datasets consists of healthy records. Tokens with a single disease occur in single digit percentages. An even smaller fraction of tokens with multiple diseases appears in the corpus. Of equal concern is that tokens with multiple diseases are not well represented in the evaluation dataset, known as the gold standard dataset.

Table 1. Annotations present in TNMG

Label	Description
1dAVb	First-degree atrioventricular block: A delay in the conduction of electrical impulses from the atria to the ventricles.
RBBB	Right bundle branch block: A condition where the right side of the heart's electrical conduction system is impaired.
LBBB	Left bundle branch block: A condition where the left side of the heart's electrical conduction system is impaired.
SB	Sinus bradycardia: A slower-than-normal heart rhythm, defined as a heart rate below 60 beats per minute in adults.
AF	Atrial fibrillation: An irregular heart rhythm, characterized by chaotic electrical activity.
ST	Sinus tachycardia: A higher-than-normal heart rhythm, defined as a heart rate above 100 beats per minute in adults.

The imbalance in this data has a profound impact on our ability to train high performance models, as we will show in Section IV. However, it is not our intention here to focus on techniques to deal with imbalance. Instead, we are following the process described in [5] so that our experiments can be directly compared.

III. Application of Deep Learning

A previous study by Ribeiro et al. demonstrated the effectiveness of deep neural networks for automatic classification of 12-lead ECGs [5]. Similar to our work, they developed a ResNet-18 model trained on over 2 million ECG exams from the TNMG database. Their model was able to detect six types of ECG abnormalities with high accuracy, outperforming cardiology residents. For preprocessing, they resampled all ECGs to 400 Hz, zero-padded signals to 4096 samples per lead, used the derived channels, and applied z-score normalization. While their approach showed promising results, the impact of this preprocessing on model performance was not thoroughly investigated.

Table 2. Distribution of classes in TNMG CODE

Feature Vector	Train		Gold (Eval)	
	#	%	#	%
000000	6,014,462	89.55000	681	82.34583
010000	145,208	2.16202	28	3.38573
000001	131,820	1.96268	35	4.23216
000010	100,865	1.50179	11	1.33011
000100	94,500	1.40702	15	1.81378
001000	86,487	1.28771	20	2.41838
100000	75,924	1.13044	25	3.02297
010010	11,910	0.17733	1	0.12092
110000	11,168	0.16628	0	0.00000
101000	7,580	0.11286	3	0.36276
001010	7,019	0.10451	0	0.00000
010100	5,713	0.08506	0	0.00000
100100	4,215	0.06276	0	0.00000
010001	3,408	0.05074	1	0.12092
001001	3,066	0.04565	0	0.00000
000011	2,860	0.04258	0	0.00000
100010	1,871	0.02786	1	0.12092
001100	1,625	0.02419	1	0.12092
011000	1,621	0.02414	4	0.48368
110100	1,165	0.01735	0	0.00000
100001	560	0.00834	1	0.12092
110010	515	0.00767	0	0.00000
000110	506	0.00753	0	0.00000
101100	331	0.00493	0	0.00000
011100	329	0.00490	0	0.00000
010011	292	0.00435	0	0.00000
101010	248	0.00369	0	0.00000
111000	220	0.00328	0	0.00000
011010	189	0.00281	0	0.00000
001011	143	0.00213	0	0.00000
100110	88	0.00131	0	0.00000
010110	82	0.00122	0	0.00000
110110	64	0.00095	0	0.00000
111100	56	0.00083	0	0.00000
011001	45	0.00067	0	0.00000
110001	43	0.00064	0	0.00000
100011	36	0.00054	0	0.00000
111010	19	0.00028	0	0.00000
101001	16	0.00024	0	0.00000
001110	14	0.00021	0	0.00000
000101	10	0.00015	0	0.00000
011011	5	0.00007	0	0.00000
010101	5	0.00007	0	0.00000
011110	5	0.00007	0	0.00000
111011	4	0.00006	0	0.00000
111110	3	0.00004	0	0.00000
110011	1	0.00001	0	0.00000
101110	1	0.00001	0	0.00000

Newer studies by Pastika et al. [6] and von Bachmann et al. [7] have also adopted reduced lead configurations, utilizing 8-lead ECGs in their deep learning models for body mass index and electrolyte prediction, respectively. However, these studies did not extensively discuss the rationale for lead reduction. The choice of using raw leads stems from the fact that they are linear combinations of raw leads, making them redundant.

Following Ribeiro et al. [5], we utilize a ResNet18 architecture adapted for multi-label classification. To create input tensors for our model, we transform the time-series ECG data into multi-channel 3D tensors. All signals are zero-padded to 4096 samples and undergo z-score normalization to standardize the data. The normalized signals are then reshaped into tensors of shape (N, 1, 4096), where N represents the number of ECG channels. In this representation, each channel corresponds to a separate ECG lead, and the temporal samples span the width of the tensor. Although the tensor has three dimensions, the singleton height dimension (1) facilitates the processing of temporal data similarly to how CNNs handle spatial information in images.

A typical block in our architecture is shown in Figure 2. The overall architecture is illustrated in Figure 3. The first convolutional layer of our model was modified to take either eight or twelve channels as input. The final layer was adapted to output probabilities for each of the six cardiac conditions by utilizing a sigmoid activation function. We employ the Adam optimization algorithm [8] with a learning rate of 0.001. Due to the multi-label nature of our task, we use Binary Cross-Entropy loss [9] as the objective function:

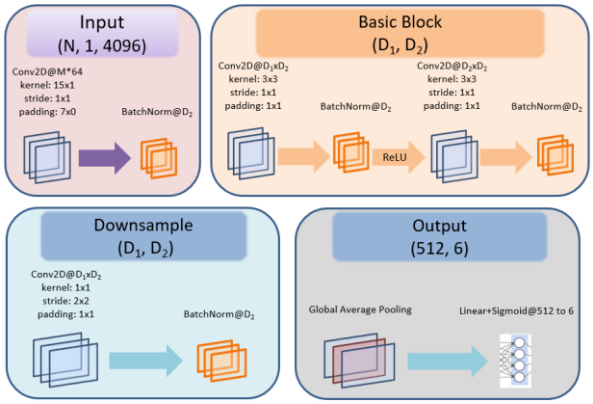

Figure 2. A typical block in the ResNet-18 architecture

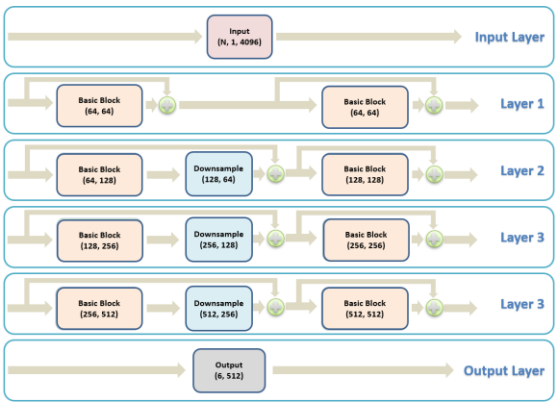

Figure 3. The composite ResNet18 architecture uses four internal layers similar to those shown in Figure 2, in addition to input and output layers.

$$BCE(y,\hat{y}) = -\frac{1}{n}\sum[y_i\, log(\hat{y}_i) + (1-y_i)\, log(1-\hat{y}_i)] \quad (5)$$

where n is the number of classes, y_i is the true label, and $\hat{y}_i$ is the predicted probability for class i.

In all experiments, the training process iterated over 10 epochs with a batch size of 32. We monitored training and validation losses, along with accuracy, micro-averaged F1score, and macro-averaged F1 score to assess the model performance. This approach allows us to evaluate the effectiveness of our model in processing both raw and minimally preprocessed ECG signals.

Given that our problem is a multi-label classification task, we employ micro and macro F1 scores [10] as a key metric for evaluating model performance. These metrics provide a comprehensive assessment of the overall model's ability to identify several cardiac conditions at once. The micro F1 score calculates metrics by counting the true positives, false negatives and false positives across all classes. It is computed as the harmonic mean of precision and recall:

$$Micro\ F1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

where TP, FP, and FN represent true positives, false positives, and false negatives, respectively.

In contrast, the macro F1 score calculates the F1 score for each class independently and then averages these scores. Micro F1 tends to give more weight to frequent classes, while macro F1 gives equal weight to all classes, regardless of their frequency in the dataset.

IV. Experimentation

We conducted eight experiments to systematically evaluate the impact of ECG channel reduction and dataset size on deep learning performance. We used four different training dataset sizes: 2K (2,000), 20K (20,000), 200K (200,000), and 2,000K (2,000,000) records, each tested with both 8-channel and 12-channel ECG configurations. For the 12-channel sets, we applied minimal preprocessing consisting of resampling to 400 Hz and 2x scaling of the signal amplitude. To address class imbalance, we chose the distribution for each training set that balances the frequency of occurrence of the class labels to the extent possible. However, as the dataset size increased, we had to include a higher proportion of healthy records due to their prevalence in the corpus.

In each experiment, we trained a separate ResNet18 model and evaluated its performance on the golden test set. We also utilized a fixed development set of 5,000 records in all experiments to monitor training process and prevent overfitting. The development set was balanced to represent a variety of combinations of cardiac conditions: approximately 4,000 examples were evenly split between single-condition cases and healthy records, 750 evenly split examples with two conditions, 249 examples containing three conditions, and one rare example with four concurrent conditions. The results for each experiment are shown in Table 3.

Our results reveal a consistent pattern across all dataset sizes: the models trained on 8-channel ECG data outperformed those trained on 12-channel data. The performance difference was more pronounced in the smaller datasets and gradually diminished as the training dataset size increased.

We observed a significant decline in performance in both 8- and 12-channel models in experiments with 2,000K records. We attribute this decrease to the inherent class imbalance in the larger dataset. As we expanded to a much higher number of records, the proportion of healthy ECG examples increased significantly. Although it reflects the prevalence of these records in the general population, this imbalance led to a bias in the model's predictions, favoring the majority class at the expense of less common combinations of cardiac conditions.

Another observation is that models trained on 2,000K records performed poorly on the balanced development set but showed a noticeably higher performance on the evaluation set. This discrepancy is likely caused by the higher proportion of healthy records in the evaluation set, which more closely mirrors the distribution in the training data. These observations are an example of the importance of considering dataset composition and carefully balancing class distributions within datasets.

To assess the stability and reproducibility of our findings, we conducted several experiments to estimate the variance of the F1 scores on 8-channel data. For each of the three dataset sizes (2K, 20K and 200K) we performed five independent training runs. Each run utilized a different random seed for data shuffling and model initialization. The results are shown in Table 4. As

Table 3. Micro F1 scores as a function of the training set size

Train Size	No. Chans	Train	Dev	Eval
2K	8	0.8810	0.7024	0.5029
2K	12	0.8690	0.7050	0.2127
20K	8	0.8870	0.8288	0.7022
20K	12	0.8812	0.8366	0.5509
200K	8	0.9310	0.8461	0.8421
200K	12	0.9286	0.8545	0.7956
2,000K	8	0.8809	0.7787	0.8649
2,000K	12	0.8787	0.7708	0.8522

Table 4. Micro F1 scores as a function of the random seed

Data	Train	Dev	Eval
2K 8 Channels (1)	0.8898	0.7235	0.4596
2K 8 Channels (2)	0.8747	0.7047	0.4569
2K 8 Channels (3)	0.8766	0.7074	0.5111
2K 8 Channels (4)	0.8892	0.7163	0.4426
2K 8 Channels (5)	0.8720	0.6876	0.5251
StDev	0.0084	0.01356	0.03655
20K 8 Channels (1)	0.8887	0.8187	0.7335
20K 8 Channels (2)	0.8852	0.8132	0.7214
20K 8 Channels (3)	0.8880	0.8229	0.6897
20K 8 Channels (4)	0.8884	0.8241	0.6729
20K 8 Channels (5)	0.8882	0.8243	0.6905
StDev	0.0014	0.0047	0.0249
200K 8 Channels (1)	0.9312	0.8534	0.8251
200K 8 Channels (2)	0.9298	0.8523	0.8278
200K 8 Channels (3)	0.9307	0.8510	0.7831
200K 8 Channels (4)	0.9318	0.8451	0.8278
200K 8 Channels (5)	0.9298	0.8481	0.7752
StDev	0.0009	0.0034	0.0263

expected, there is a significant reduction in the standard deviation as the training set size increases.

This translates to an improvement in the statistical significance of these scores. For a sample size of 2K at 95% confidence, a difference in the F1 score of 0.0174 is statistically significant on the training data. For sample sizes of 20K and 200K, differences greater than 0.0054 and 0.0017, respectively, are significant. Hence, we see that the differences due to randomization in Table 4 are statistically significant, underscoring how sensitive these deep learning systems are to randomization (which makes reproducibility a challenge).

Nevertheless, the key point here is that performance for the 8-channel system is not statistically different from the 12-channel system, indicating that the deep learning system is able to implement whatever signal processing is necessary to extract meaningful information.

V. Ablation Analysis and Multicollinearity

As a further investigation of the impact of additional ECG channels on model performance, we conducted an ablation analysis [11] to determine channel importance. We systematically randomized each channel's data and measured the resulting change in the model's micro F1 score. The ablation process was performed for 250 randomized iterations for each channel to ensure consistent results. The importance of each channel was quantified as the average decrease in micro F1 score when that channel was randomized.

Figure 4 presents the results of three experiments using 2K, 20K, 200K and 2,000K records. The x-axis represents the 12 ECG channels, while the y-axis represents the change in micro F1 score. A positive score indicates a feature is important, while a negative score indicates a feature is redundant. Our analysis demonstrated that the precordial leads (V1-V6) tend to have the higher importance scores, while the derived

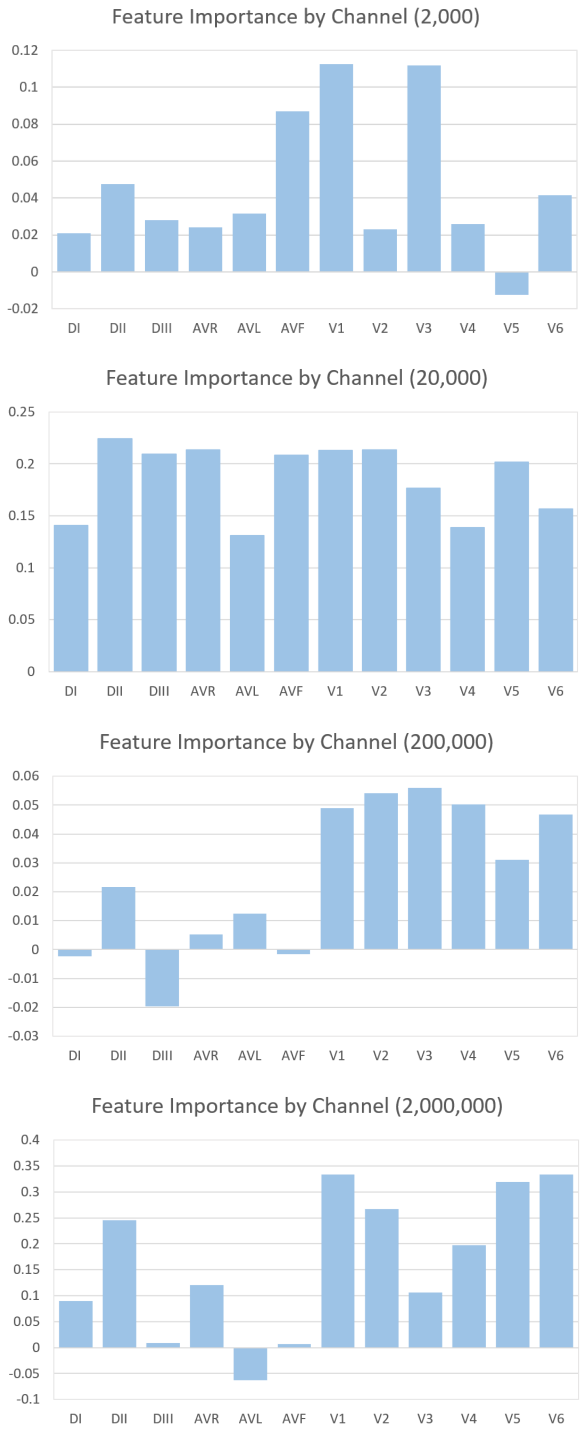

Figure 4. Results of the ablation analysis

channels (DIII, aVR, aVL, and aVF) showed lower scores and sometimes negative values, indicating their redundancy. Channels DI and DII showed mixed results, either being as important as some precordial leads or relatively insignificant.

Following our ablation analysis, it is important to consider the issue of multicollinearity in our ECG data. Multicollinearity [12] occurs when there is an approximately linear relationship between two or more independent variables in a regression model. While it is typically discussed in the context of regression models, it can also affect classification models, including our deep learning model for ECG classification.

To further validate our findings regarding channel redundancy, we employed Variance Inflation Factor (VIF) analysis on a balanced subset of 20,000 ECG recordings. VIF quantifies the severity of multicollinearity in regression analysis by measuring how much the variance of a coefficient estimate is increased due to collinearity with other variables. Our analysis revealed substantial redundancy among limb leads and their derived channels, with DI and DII showing high VIF scores of 11.64 and 22.38 respectively. The derived channels (DIII, aVR, aVL, and aVF) similarly exhibited high VIF scores ranging from 12.69 to 21.48, indicating severe multicollinearity. In contrast, all precordial leads (V1-V6) demonstrated low VIF scores between 2.60 and 3.63, well below the standard multicollinearity threshold of 5.0. These results provide strong statistical evidence that limb leads and their derivatives contain redundant information, while precordial leads contribute unique and independent signal characteristics.

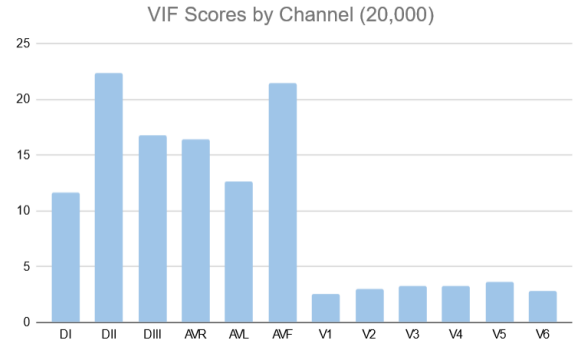

Figure 5. VIF scores

In our study, by using only the eight independent leads and omitting the derived leads, we performed a form of variable selection that addresses the issue of multicollinearity in ECG data. Our findings in model training, ablation analysis, and VIF scores support the hypothesis, indicating that derived ECG channels introduce multicollinearity and do not provide any additional predictive power.

VI. Conclusions and Future Work

Our study provides evidence supporting our hypothesis that extensive preprocessing of ECG data may not be beneficial for deep learning in cardiac diagnosis. Across various dataset sizes, models trained on raw signal outperformed those using derived channels and minor preprocessing. The ablation analysis further revealed that derived channels have little or slightly negative impact on model performance.

These findings highlight the capability of deep learning algorithms to extract meaningful patterns from complex physiological data without relying on handcrafted features. This suggests a potential for simpler, more direct approach for data input that may yield more accurate and robust models.

In future studies, we plan to further refine our approach by conducting additional experiments with 8-channel ECG data. We will focus on comparing the performance of models trained on raw 8-channel signals against those trained on extensively preprocessed 8-channel data, excluding the derived channels entirely and isolating the effects of preprocessing on the primary ECG leads.

Acknowledgments

We would like to express our gratitude to Ribeiro et al. [5] for providing the TNMG CODE dataset used in this study, which was instrumental in our analysis.

This material is based upon work supported in part by the by the Temple University Office of the Vice President for Research, and by the Temple University College of Science and Technology Research Experience for Undergraduates program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of Temple University.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, USA, 2016, pp. 770–778. doi: *10.1109/CVPR.2016.90*.

[2] V. Khalkhali, N. Shawki, V. Shah, M. Golmohammadi, I. Obeid, and J. Picone, “Low Latency Real-Time Seizure Detection Using Transfer Deep Learning,” in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, I. Obeid, I. Selesnick, and J. Picone, Eds., Philadelphia, Pennsylvania, USA: IEEE, 2021, pp. 1–7. doi: *10.1109/SPMB52430.2021.9672285*.

[3] M. Bagritsevich, D. Hackel, I. Obeid, and J. Picone, “Annotation of the Fox Chase Cancer Center Digital Pathology Corpus,” in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, Philadelphia, Pennsylvania, USA: IEEE, Dec. 2024, pp. 1–4. [In Publication]. url:

https://isip.piconepress.com/publications/unpublished/conferences/2024/ieee_spmb/dpath/.

[4] A. Paroya, L. Dewees, and A. Chau, "Cardiovascular Disease Detection Using 12-Lead Electrocariodgram (ECG) Machine Learning," in *Senior Design I, College of Engineering, Temple University*, Philadelphia, Pennsylvania, USA, 2023, pp. 1–17. url: *www.isip.piconepress.com/publications/presentations_misc/2023/senior_design/ml/*.

[5] A. H. Ribeiro *et al.*, "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature Communications*, vol. 11, no. 1, p. 1760, Apr. 2020, doi: *10.1038/s41467-020-15432-4*.

[6] L. Pastika *et al.*, "Artificial intelligence-enhanced electrocardiography derived body mass index as a predictor of future cardiometabolic disease," *npj Digital Medicine*, vol. 7, no. 1, p. 167, Jun. 2024, doi: *10.1038/s41746-024-01170-0*.

[7] P. von Bachmann *et al.*, "Evaluating regression and probabilistic methods for ECG-based electrolyte prediction," *Scientific Reports*, vol. 14, no. 1, p. 15273, Jul. 2024, doi: *10.1038/s41598-024-65223-w*.

[8] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," in *Proceedings of the International Conference on Learning Representations*, San Diego, California, USA, 2015, pp. 1–15. [Online]. url: *https://arxiv.org/abs/1412.6980*.

[9] A. Mao, M. Mohri, and Y. Zhong, "Cross-Entropy Loss Functions: Theoretical Analysis and Applications," in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., in *Proceedings of Machine Learning Research*, vol. 202. PMLR, Jul. 2023, pp. 23803–23828. [Online]. url: *https://proceedings.mlr.press/v202/mao23b.html*.

[10] I. C. Mogotsi, "Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze: Introduction to information retrieval," *Information Retrieval*, vol. 13, no. 2, pp. 192–195, Apr. 2010, doi: *10.1007/s10791-009-9115-y*.

[11] R. Meyes, M. Lu, C. W. de Puiseau, and T. Meisen, "Ablation Studies in Artificial Neural Networks." 2019. url: *https://arxiv.org/abs/1901.08644*.

[12] M. H. Graham, "Confronting Multicollinearity in Ecological Multiple Regression," *Ecology*, vol. 84, no. 11, pp. 2809–2815, Nov. 2003, doi: *10.1890/02-3114*.