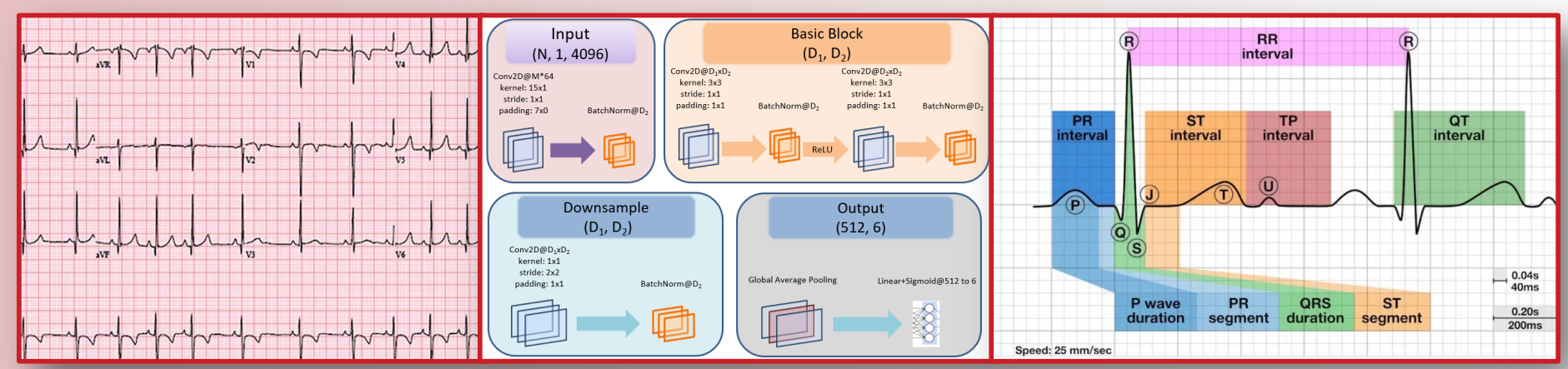
The Impact of ECG Channel Reduction on Multi-Label Cardiac Diagnosis



D. Alexandrov and J. Picone

Neural Engineering Data Consortium
Temple University

NEURAL ENGINEERING
DATA CONSORTIUM

Abstract

Research Question:

- What are the preprocessing requirements for electrocardiogram (ECG) data when using deep learning models, and how does channel reduction impact model performance and diagnostic accuracy?

Methodology:

- ResNet18 model on TNMG CODE Corpus
- Systematic comparison of 8 vs. 12 channel ECG data
- Ablation analysis to determine channel importance

Findings:

- Derived ECG channels showed minimal or detrimental effect
- Model performance consistently decreased with more channels
- Deep learning models are effective using raw ECG data

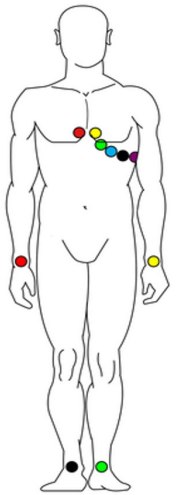
Impact:

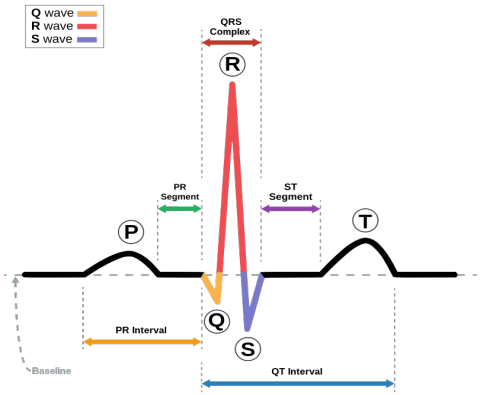
- Challenges traditional ECG preprocessing assumptions
- Reduces effort needed for machine learning (ML) system development
- Enables improved medical diagnostics

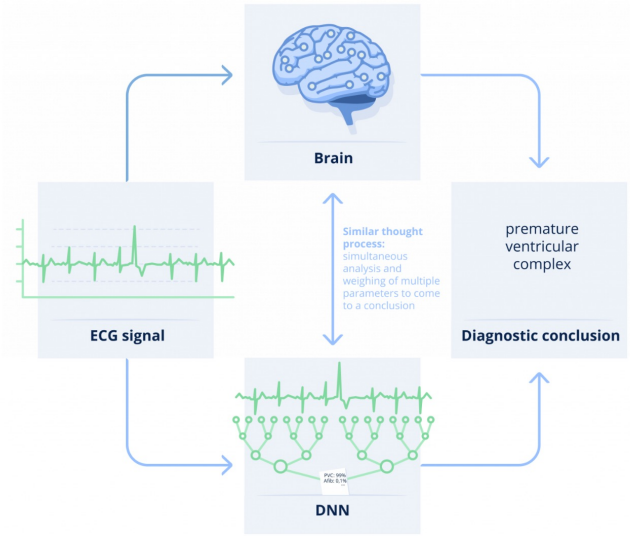
Introduction

Traditional ECG Processing:

- Cardiologists rely on preprocessed ECG data and derived channels
- Current methods combine raw signals through linear combinations
- Preprocessing steps can mask or eliminate subtle diagnostic patterns



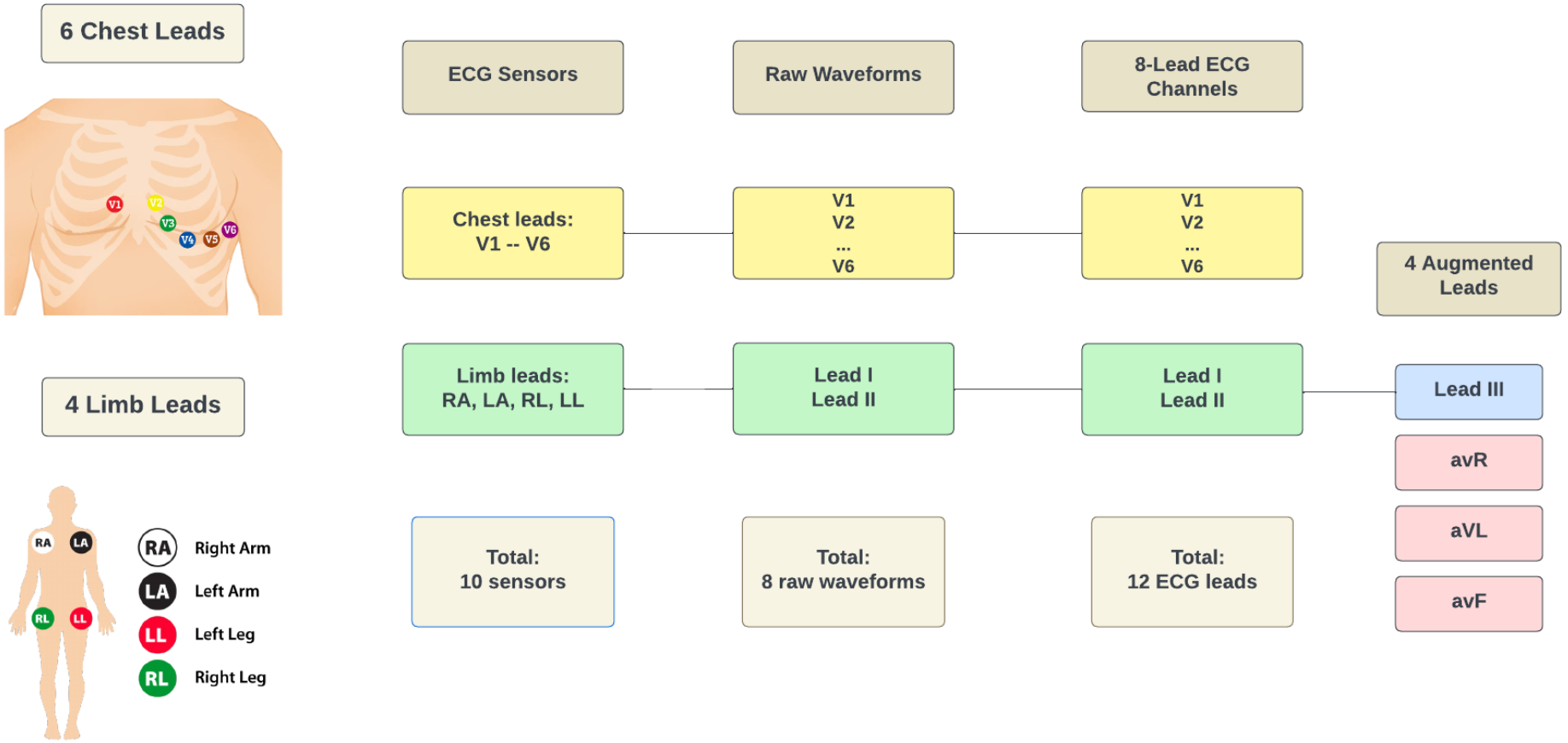




Deep Learning Advantages:

- Neural networks excel at extracting complex features autonomously
- Can process raw physiological signals without manual engineering
- Potential to identify patterns lost in traditional preprocessing
- Streamlines diagnostic pipeline while preserving signal integrity

ECG Lead Systems and Signal Processing



Standard 10-Lead ECG:

- 6 precordial leads (V1-V6) on chest
- 4 limb leads on arms and legs (RA, LA, LL, RL-ground)
- Records 8 raw waveforms ($V1–V6$, $DI = LA - RA$, $DII = LL - RA$)

Derived Channel Calculation:

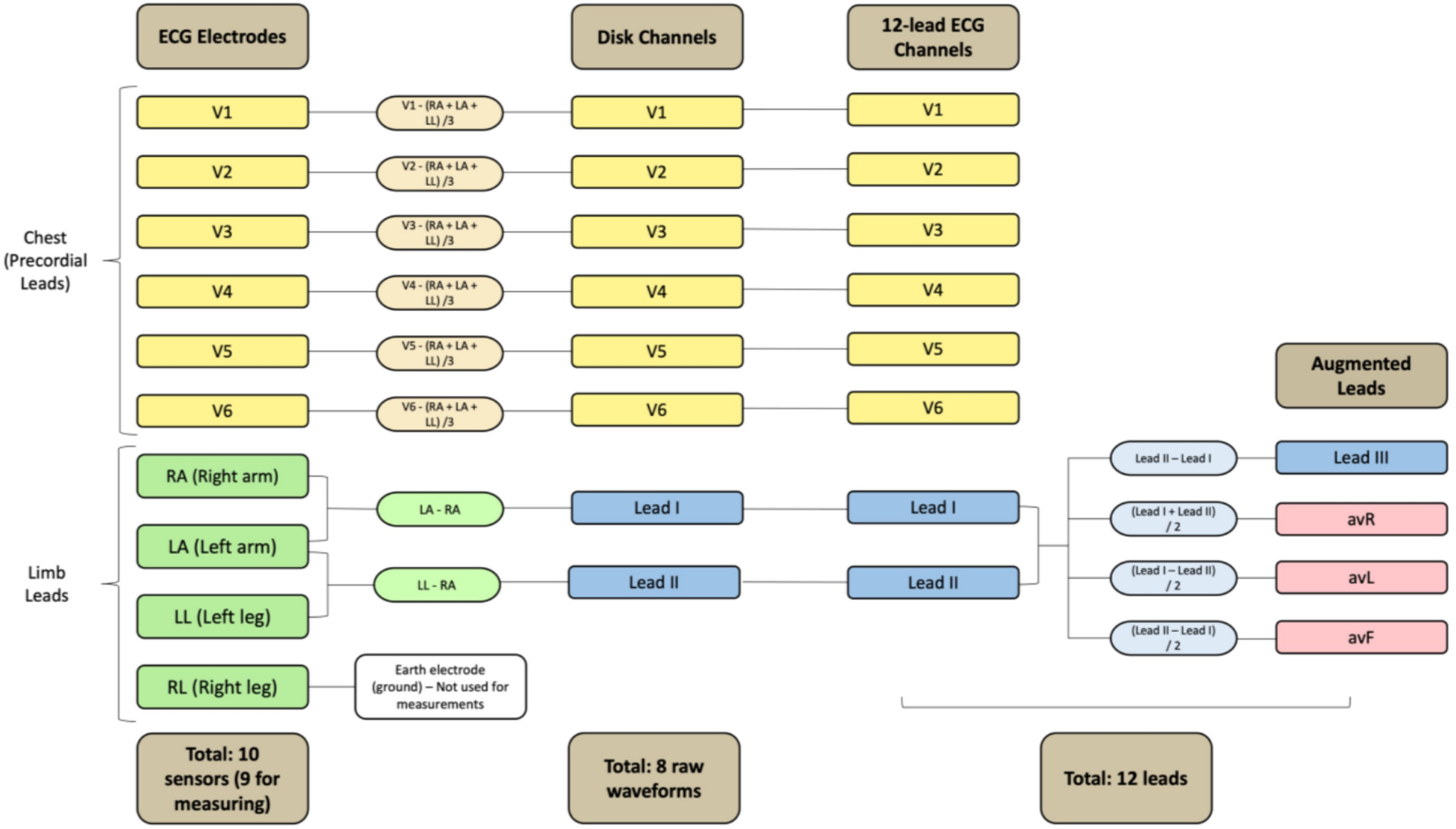
$$DIII = (DII - DI)$$

$$aVR = (DI + DII)/2$$

$$aVL = (DI - DII)/2$$

$$aVF = (DII - DI)/2$$

A 12-Channel ECG



The TNMG CODE Corpus

Dataset Overview:

- Telehealth Network of Minas Gerais (TNMG)
- Clinical Outcomes in Digital Electrocardiography (CODE) group
- Comprehensive ECG collection (2010-2016)
- Coverage: 811 counties in Minas Gerais, Brazil
- Total records: 6,716,317 annotated ECGs
- Patient population: 1,558,749 unique individuals

Evaluation Dataset (“Golden Dataset“):

- 827 carefully selected ECG recordings
- Annotation protocol:
  - Initial review by two independent cardiologists
  - Disagreements resolved by third specialist consultation
  - Consensus-based final annotations
- Labeled for six specific cardiac abnormalities
- Serves as high-quality benchmark for model evaluation

Annotations in TNMG

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

Dataset Imbalance

Distribution Analysis:

- Binary feature vectors track presence/absence of six conditions (1dAVb, RBBB, LBBB, SB, AF, ST).
- The vast majority of records in both training and evaluation sets are healthy cases.
- Single disease occurrences represent less than 10% of total dataset.
- Multiple disease combinations appear in very, very small fractions, yet this is what makes this problem challenging.
- The gold standard evaluation dataset particularly lacks representation of multiple disease cases.

Impact on Model Development:

- Severe imbalance creates training challenges (e.g. risk of model bias towards healthy cases)
- Harder to detect multiple concurrent conditions
- Limited examples of multiple diseases affects a model's ability to learn disease interactions

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

Relevant Prior Art

Foundational Study:

- Ribeiro et al. [5]:
  - Modified ResNet model trained on 2M+ ECG exams using a 12-lead ECG
  - Preprocessing steps:
    - 400 Hz resampling
    - Zero-padding to 4096 samples
    - Derived channel inclusion
    - Z-score normalization
  - Results outperformed cardiology residents
  - Did not investigate preprocessing impact

Recent Lead Reduction Studies:

- Pastika et al. [6]:
  - 8-lead ECGs for body mass index prediction
  - Demonstrated viability of reduced leads
- von Bachmann et al. [7]:
  - 8-lead approach for electrolyte prediction
  - Limited exploration of lead reduction rationale
  - Noted redundancy in derived channels

ResNet18 Model Implementation

Data Preprocessing:

- Time-series ECG data converted to 3D tensors
- Zero-padding to $4,096$ samples
- Z-score normalization for standardization
- Final tensor shape: $(N, 1, 4096)$
  - $N = number\ of\ ECG\ channels\ (8\ or\ 12)$
  - Singleton dimension for temporal processing

Model Modifications:

- Adapted ResNet18 for multi-label classification
- Modified first layer for 8/12 channel input
- Final layer outputs 6 condition probabilities
- Sigmoid activation for multi-label output

Training Parameters:

- Adam optimizer
- Learning rate: 0.001
- Binary Cross-Entropy loss:

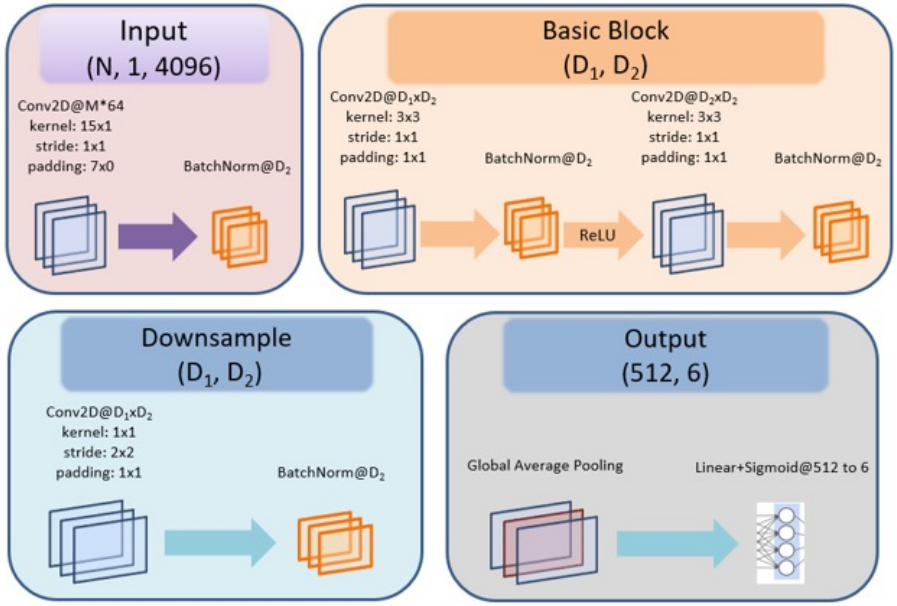
$$BCE(y, \hat{y}) = -\frac{1}{n}\sum_{i=0}^{n-1}[y_i\, log(\hat{y}_i) + (1 - y_i)\, log(1 - \hat{y}_i)]$$

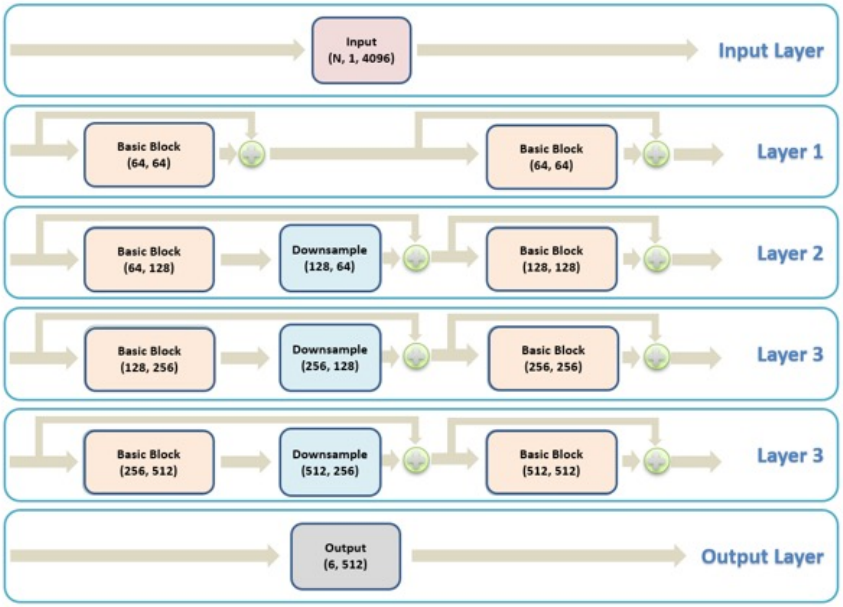
where:

$n = number\ of\ classes$

$y_i = true\ label$

$\hat{y}_i = predicted\ probability$





Training Process and Performance Metrics

Training Parameters:

- **Epochs: 10, batch size: 32**
- **Monitored: training/validation losses, accuracy, F1 scores**
- **Evaluated both raw and minimally preprocessed**

Performance Metrics:

- **Multi-label classification requires comprehensive metrics**
- **Primary metrics: Micro and Macro F1 scores**

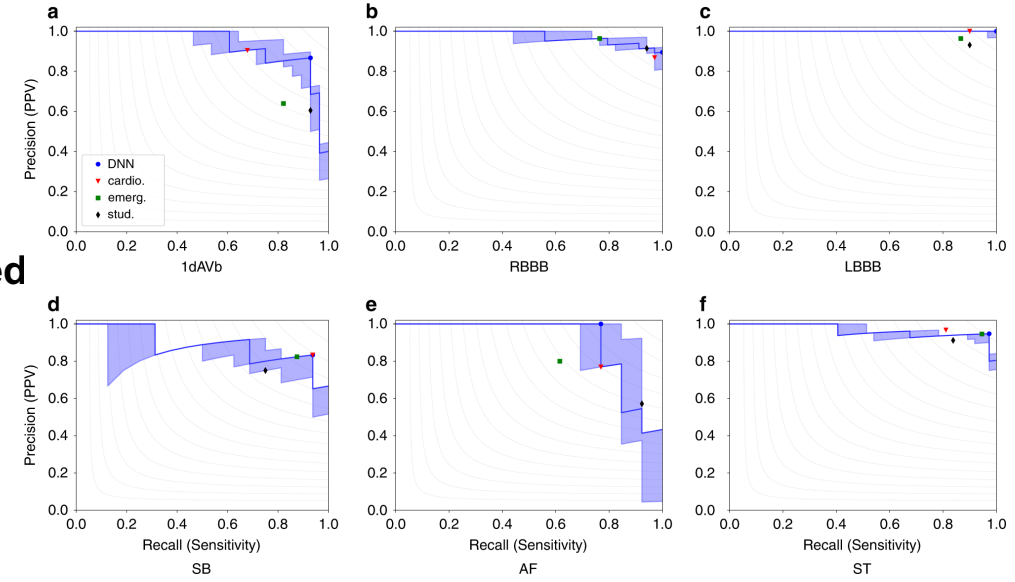
Micro F1 Score:

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

$$Precision = \frac{TP}{TP+FP},$$

$$Recall = \frac{TP}{TP+FN}.$$

- **Counts true positives (TP), false positives (FP), false negatives (FN)**
- **Weights frequent classes more heavily**



	DNN			cardio.				emerg.				stud.			
	meas.	noise	unexplain.	meas.	noise	concep.	atte.	meas.	noise	concep.	atte.	meas.	noise	concep.	atte.
1dAVb	3	2	1	8	3			15	3			13	3	3	
RBBB	3		1	4		2		1		8		3		2	
LBBB				1	1	1			1	4			2	3	
SB	4			4				4			1	5		2	1
AF		2	1		4	2			2	5			3	7	
ST	2		1	2	1		5	1	1	1	1	1	2	1	5

Macro F1 Score:

- **Calculates F1 independently for each class**
- **Gives equal weight regardless of class frequency**
- **Averages scores across classes**
- **Better for imbalanced datasets**

Experimental Design

Dataset Configurations:

- Four training set sizes: 2K, 20K, 200K, 2,000K
- Each tested with 8 and 12 channels
- 12-channel preprocessing: 400 Hz resampling, 2x amplitude scaling

Development Set Structure:

- Fixed 5,000 records for all experiments
- Balanced distribution:
  - 4,000 records: single conditions and healthy cases
  - 750 records: two conditions
  - 249 records: three conditions
  - 1 record: four conditions

Evaluation Method:

- Separate ResNet18 model for each experiment
- Performance tested on golden dataset
- Development set monitored for overfitting
- Class balance maintained where possible
- Larger datasets required more healthy records

Results

Channel Performance:

- 8-channel models consistently outperformed 12-channel models
- Performance gap largest in smaller datasets
- Difference decreased with larger training sets

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

Dataset Size Impact (2,000K Records):

- Significant performance decline in both configurations
- Root cause: Class imbalance issues
- Higher proportion of healthy records skewed predictions
- Model bias toward majority class
- Reduced accuracy for rare condition combinations

Development vs. Evaluation Performance:

- 2,000K models:
  - Poor performance on balanced development set
  - Better performance on evaluation set
  - Explained by similar healthy record distribution in training and evaluation
- Highlights critical importance of dataset composition
- Demonstrates impact of class balance on model reliability

Variance Experiments

Variance Experiments:

- Tested 8-channel data with 2K, 20K, 200K datasets
- 5 independent training runs per size
- Different random seeds for each run
- Evaluated data shuffling and model initialization impact

Statistical Significance (95% Confidence):

- 2K dataset: F1 difference > 0.0174 significant
- 20K dataset: F1 difference > 0.0054 significant
- 200K dataset: F1 difference > 0.0017 significant
- Standard deviation decreases with larger datasets

Key Implications:

- Deep learning systems are highly sensitive to randomization
- Reproducibility remains a significant challenge
- 8-channel performance statistically equivalent to 12-channel
- Suggests deep learning can replace traditional signal processing
- Model learns necessary feature extraction independently

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

Ablation Analysis

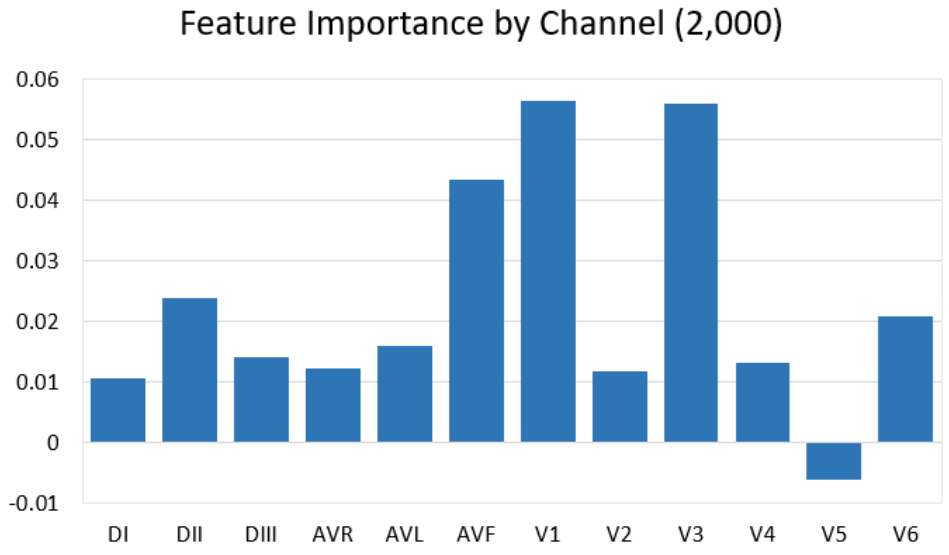
Ablation Process:

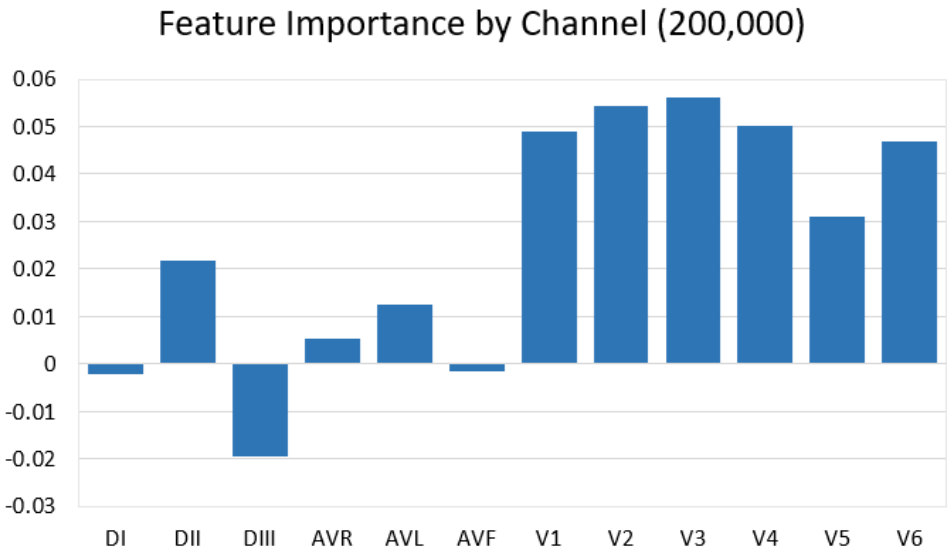
- Systematically evaluates each channel's importance
- Tests model performance when channels are “damaged”
- Reveals which features are truly essential

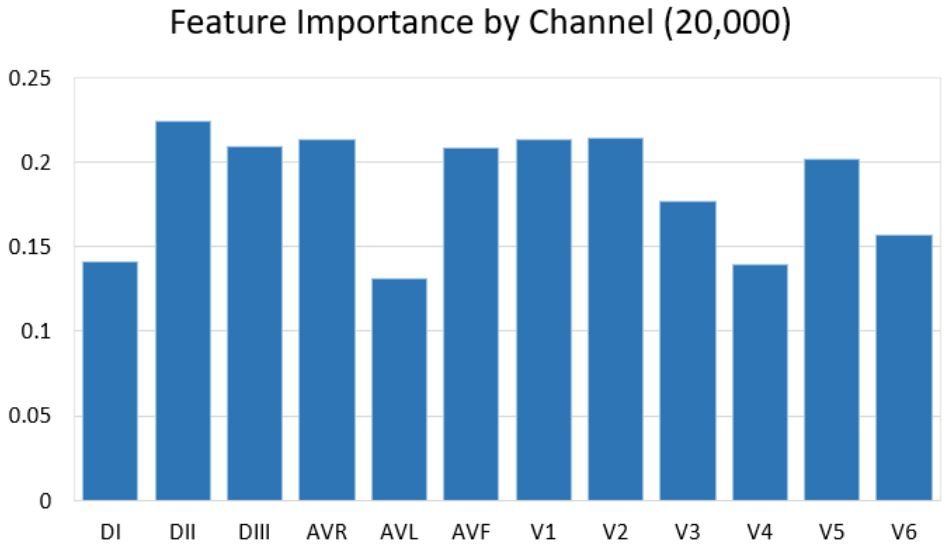
Methodology:

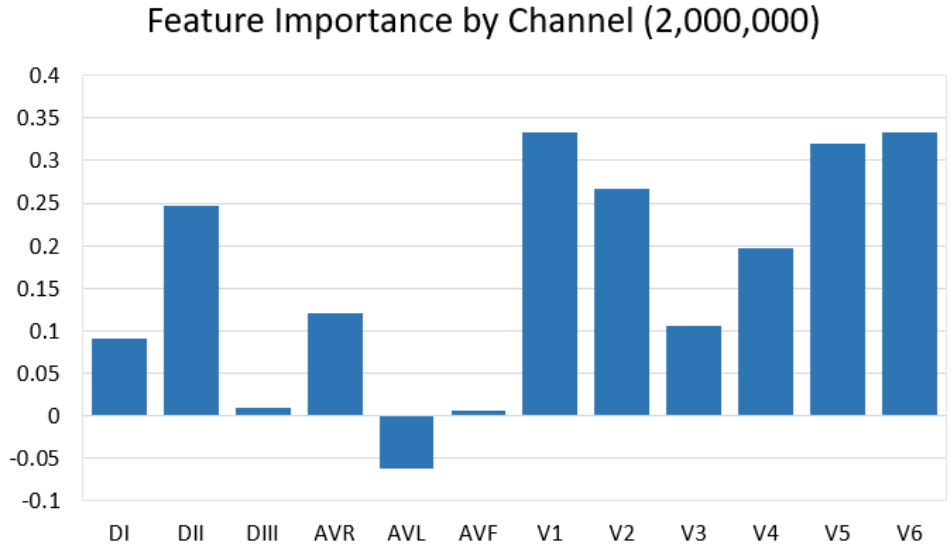
1. Start with fully trained model
2. For each channel:
    - Randomly scramble channel's data
    - Keep other channels intact
    - Measure performance drop
    - Repeat 250 times for reliability and average the results
3. Calculate importance:
    - Higher performance drop = More important channel
    - Negative impact = Channel potentially harmful
    - Near-zero impact = Redundant channel

Ablation Analysis Results









Variance Inflation Factor (VIF) Analysis

Variance Inflation Factor (VIF) Analysis:

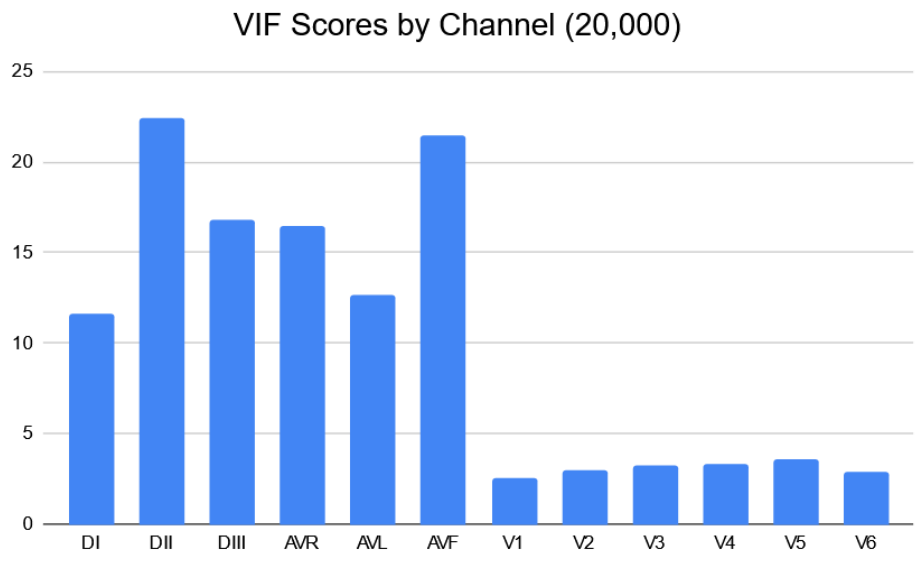
- Conducted on 20,000 balanced ECG recordings
- Measures how much variables overlap/correlate
- Higher VIF = More redundant information
- Standard threshold: VIF > 5.0 indicates high redundancy

How VIF Works:

- Regresses each variable against all others
- Measures increase in variance due to correlations
- $VIF = 1/(1 - R^2)$ where R^2 is from regression
- Example: VIF of 4 means 4x more variance due to correlation

Results:

- Limb Leads (DI, DII):
  - Very high VIF: 11.64 - 22.38
  - Shows severe redundancy
- Derived Channels (DIII, aVR, aVL, aVF):
  - High VIF: 12.69 - 21.48
  - Confirms redundant information
- Precordial Leads (V1-V6):
  - Low VIF: 2.60 - 3.63 (well below 5.0 threshold)
  - Indicates unique, independent signals



Conclusions and Future Work

Key Findings:

- ECG preprocessing may be unnecessary for deep learning
- Raw signal models outperform preprocessed ones
- Derived channels show minimal or negative impact
- Deep learning can extract features without manual engineering
- Simpler approach yields more robust results

Research Impact:

- Challenges traditional ECG analysis assumptions
- Demonstrates deep learning's pattern extraction capability
- Suggests streamlined diagnostic pipeline
- Potential for more accurate cardiac diagnosis
- Reduces complexity in model development

Future Work:

- Additional 8-channel experiments planned
- Compare raw vs. preprocessed 8-channel signals
- Isolate preprocessing effects on primary leads
- Evaluate impact without derived channels
- Focus on raw signal processing optimization

Acknowledgements

We would like to express our gratitude to Ribeiro et al. 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.

Selected References

1. K. He et al., “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
2. V. Khalkhali et al., “Low Latency Real-Time Seizure Detection Using Transfer Deep Learning,” in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2021.
3. M. Bagritsevich et al., “Annotation of the Fox Chase Cancer Center Digital Pathology Corpus,” in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 2024.
4. A. Paroya et al., “Cardiovascular Disease Detection Using 12-Lead Electrocariodgram (ECG) Machine Learning,” in *Senior Design I, College of Engineering, Temple University*, 2023.
5. A. H. Ribeiro et al., “Automatic diagnosis of the 12-lead ECG using a deep neural network,” *Nature Communications*, 2020.
6. L. Pastika et al., “Artificial intelligence-enhanced electrocardiography derived body mass index,” *npj Digital Medicine*, 2024.
7. P. von Bachmann et al., “Evaluating regression and probabilistic methods for ECG-based electrolyte prediction,” *Scientific Reports*, 2024.
8. D. P. Kingma and J. L. Ba, “Adam: A Method for Stochastic Optimization,” *International Conference on Learning Representations*, 2015.
9. A. Mao et al., “Cross-Entropy Loss Functions: Theoretical Analysis and Applications,” *International Conference on Machine Learning*, 2023.
10. I. C. Mogotsi, “Introduction to information retrieval,” *Information Retrieval*, 2010.
11. R. Meyes et al., “Ablation Studies in Artificial Neural Networks,” 2019.
12. M. H. Graham, “Confronting Multicollinearity in Ecological Multiple Regression,” *Ecology*, 2003.