

June 11, 2025

JY Prompt:

I am working on training a deep learning model that will correctly classify ECG signals output from a prototype ECG device I am building. I have very small output data from this prototype ECG device. But I have 100,000 patient data from hospital 12-signal ECG machines that are publicly available, such as PTB-XL and SPH.

Let's say I want to use these public 12-lead ECG dataset, to create training and test data for a ResNet + SE model to classify ten different types of heart conditions. But instead of hospital-style 12-signal measurements with conductive gel, the data comes from my prototype that has leads embedded into an elastic shirt. It's likely that there will be some displacement of each lead in the shirt from where the lead would be if placed using each patient's fiducials. Also, there will be added noise due to body & shirt movement. How would you augment the public and test data to approximate the displacements and noise from the prototype. Would there be a much larger augmented dataset (using the public dataset as the basis)?

Claude 4:

This is a fascinating domain adaptation problem where you need to bridge the gap between high-quality hospital ECG data and your prototype wearable device. Here's a comprehensive approach to data augmentation that should help your model generalize to your shirt-based ECG system:

## Lead Displacement Augmentation

### Spatial Transformations:

- Apply small rotations ( $\pm 5-15^\circ$ ) to individual leads to simulate electrode misplacement
- Use time-domain stretching/compression ( $\pm 5-10\%$ ) to simulate different chest geometries
- Add inter-lead timing delays ( $\pm 10-50\text{ms}$ ) to simulate asynchronous sampling from displaced electrodes
- Apply amplitude scaling per lead ( $0.8-1.2\times$ ) to account for varying contact quality

### Anatomical Variation Simulation:

- Create lead-specific transformation matrices based on typical displacement patterns in shirt-based systems
- Use morphological transformations to simulate how signals change with electrode position relative to heart anatomy
- Apply different scaling factors for precordial vs limb leads since shirt-based systems primarily affect chest leads

# Noise and Artifact Augmentation

## Movement Artifacts:

- Add low-frequency drift (0.05-2 Hz) with varying amplitudes to simulate baseline wander from breathing and movement
- Inject periodic artifacts (walking frequency ~1-2 Hz) with harmonics
- Apply sudden amplitude shifts to simulate contact interruptions
- Add muscle artifact noise (20-200 Hz) with realistic EMG characteristics

## Contact Quality Variations:

- Simulate impedance changes by adding multiplicative noise
- Apply random dropout events (brief signal loss) to individual leads
- Add 50/60 Hz powerline interference with varying intensities
- Introduce gaussian noise with signal-dependent variance

# Systematic Augmentation Strategy

## Multi-Stage Augmentation Pipeline:

1. **Base Augmentation** (2-3x dataset size): Apply standard ECG augmentations (time shifts, amplitude scaling, noise addition)
2. **Displacement Augmentation** (5-10x): Add lead-specific spatial transformations and timing variations
3. **Wearable-Specific Augmentation** (10-20x): Layer on movement artifacts, contact variations, and textile-specific noise patterns

## Realistic Noise Modeling:

- Record actual noise characteristics from your prototype when not measuring ECG
- Create noise profiles from different activities (sitting, walking, arm movement)
- Use these profiles to generate synthetic noise that matches your device's characteristics

# Advanced Techniques

## Generative Approaches:

- Train a conditional GAN to generate "shirt-ECG" versions of hospital ECG data
- Use cycle-consistent adversarial networks to learn the mapping between hospital and wearable ECG domains
- Consider diffusion models for more controlled noise generation

## Physical Modeling:

- Use forward modeling based on ECG generation principles to simulate how lead displacement affects signal morphology
- Apply inverse problem techniques to understand how your specific lead configuration samples the cardiac electrical field

## Validation Strategy

### Progressive Validation:

1. Start with heavily augmented data and gradually reduce augmentation intensity
2. Use cross-validation with different augmentation strategies
3. Test on real data from your prototype as it becomes available
4. Implement uncertainty quantification to identify when the model encounters out-of-distribution samples

**Recommended Dataset Size:** Given your 100K hospital ECG samples, I'd suggest creating a 500K-1M sample augmented dataset. This 5-10x expansion should provide sufficient diversity while maintaining computational feasibility.

The key is to make your augmentation as realistic as possible by understanding the specific failure modes and noise characteristics of your shirt-based system. Start with conservative augmentation parameters and tune them based on validation performance with your actual prototype data.

Would you like me to elaborate on any of these augmentation techniques or help you implement specific components of this pipeline?

JY Prompt:

This is a very good list. Thanks! I would like sample python codes to implement the first 3 augmentation strategies, (1) Spatial Transformations, (2) Anatomical Variation Simulation, and (3) Noise and Artifact Augmentation. Please put each item into a separate python file. I will use Numpy and PyTorch as my main libraries. Please add explanation of code at the end of each file or each function. Please provide examples of usage using sample data.

Claude 4:

I'll create three separate Python files implementing the augmentation strategies you requested. Each will include comprehensive code, explanations, and usage examples.

Files created:

- spatial\_transformations.py
- anatomical\_variations\_v0.py
- anatomical\_variations.py
- noise\_and\_artifacts.py