How does Augmentation affect Feature Space: A study using various augmentation methods in Distributed Learning

Nikil Sharan Prabahar Balasubramanian

A thesis/dissertation submitted in partial fulfillment

of the requirements for the degree of

MS in Computer Science

Department of Computer Science

Sagnik Dakshit, Ph.D., Committee Chair

Soules College of Business

The University of Texas at Tyler
Tyler, Texas

This is to certify that the Master's Thesis/Doctoral Dissertation of

NIKIL SHARAN PRABAHAR BALASUBRAMANIAN

has been approved for the thesis/dissertation requirement on
19th November 2024
for the Master's in Computer Science degree

Approvals:

Thesis/Dissertation Chair: Sagnik Dakshit, Ph.D.

Member: Yi Li, Ph.D.

Member: Arun Kulkarni, Ph.D.

Chair, Department of Computer Science: Leonard Brown, Ph.D.

Dean, Soules College of Business

Title: How does Augmentation affect Feature Space: A study using various augmentation methods in Distributed Learning

Author: Nikil Sharan Prabahar Balasubramanian

19th November 2024

Dissertation Chair: Dr. Sagnik Dakshit

Dissertation Committee Members: Dr. Arun Kulkarni
                                  Dr. Yi Li

Abstract

This thesis examines the impact of data augmentation techniques on model performance within a distributed learning framework, focusing on enhancing feature diversity and improving representation for under-represented classes. Data augmentation, commonly used to address data imbalance, significantly influences the feature space learned by deep learning models, with varied effects in distributed settings where data is split across nodes. Our study reveals that inconsistencies in feature learning across nodes reduce the benefits of local augmentation in capturing complex patterns, leading to suboptimal model performance. To address this, we propose a coherent augmentation approach that embeds consistent transformations in the central server, resulting in improved class distinction, and balanced learning across classes. Our findings underscore the potential of embedding augmentation to optimize distributed models, suggesting paths for refining augmentation parameters and enhancing feature space visualization in decentralized environments, with future applications across diverse data structures and distributed learning paradigms.

TABLE OF CONTENTS

Chapter 1

Introduction

The growth of deep learning, especially using convolutional neural networks (CNNs), has revolutionized medical diagnostics by enhancing pattern recognition capabilities, which are vital for accurate decision-making. However, in fields like healthcare, where data is often imbalanced and lacks sufficient diversity, these models face significant limitations. For example, certain medical conditions may be underrepresented in datasets, which can result in a model that performs well on common conditions but struggles with rarer, clinically significant ones. This imbalance can lead to bias and reduces the model's effectiveness in real-world scenarios, where diverse patient profiles and conditions are the norm.

Data augmentation has emerged as a key strategy to address these issues by creating synthetic data variations that enrich the training set. Through augmentation techniques, models can be trained on a more representative dataset that captures the range of variations seen in clinical data. This helps mitigate the effects of data imbalance and improves the model's robustness, enhancing its ability to generalize across diverse conditions.

In addition to data scarcity, the black-box nature of CNNs poses a challenge, as it limits the interpretability of the learned feature space, especially in applications requiring clinical transparency. Understanding how a model learns from augmented data and which features it

focuses on is essential for ensuring that the model's decisions align with clinical reasoning. To address this, techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) offer a means to visualize which regions of input data significantly impact the model's decisions, shedding light on the learned feature space. This study employs Grad-CAM to illustrate how different augmentation strategies influence feature learning, providing insight into the model's focus during multi-class and multi-label classification tasks.

For our investigation, we utilized two public ECG datasets, PTB-XL and MIT-BIH(Wagner et al. 2020; Plawiak 2017), to examine the effects of these augmentation methods across both single and multi-label classification scenarios. These datasets enable us to evaluate the impact of augmentation on handling diverse cardiac conditions while addressing the data imbalance challenges in model training.

Initial experiments were conducted on a centralized CNN model to observe the direct impact of these augmentations on feature learning. Using Grad-CAM, we could visualize how these augmentations enhanced model focus and robustness by highlighting relevant areas of the ECG signals that influence predictions within a custom CNN, applied to a modified version of the PTBXL dataset, a comprehensive ECG dataset widely used in cardiovascular research.

The need for this research stems from the challenges in developing effective deep learning models for healthcare, particularly in tasks like ECG classification. Despite the success of deep learning in other fields, healthcare data, especially time-series signals like ECG, often face

limitations in data availability and quality. Traditional augmentation methods and newer

generative models(Akhoondkazemi et al. 2023) can help create more diverse and robust datasets.

However, it's not always clear how different types of augmentation impact model learning and

the key features that the model picks up for clinical decision-making. Understanding these

effects is especially important for ECG data, where small changes in signal characteristics are

critical for accurate diagnosis. This study aims to fill this gap by examining how specific label-

invariant augmentation methods influence a model's ability to recognize important ECG patterns,

making it one of the first studies to do so.

To tackle the challenges of privacy and data accessibility, federated learning has emerged as a

promising solution. Rather than centralizing data in one location, federated learning enables

multiple devices or institutions to collaboratively train a model while retaining data locally. This

decentralized training process is critical in privacy-sensitive fields, like healthcare, where sharing

raw data can lead to confidentiality issues. However, federated learning introduces unique

obstacles in feature space alignment and model interpretability, as local models trained on

diverse data distributions may develop inconsistencies in feature representations. This

complexity can compromise the consistency of the global model's feature space and its

explainability, which becomes especially relevant when applied to ECG data, where subtle

patterns are crucial for clinical insights. In this study, we are not focusing on privacy or data

security in federated learning. Instead, we are looking closely at how using data augmentation

before training can improve the way features are learned across different local models. This

approach aims to make the combined global model more consistent and reliable for identifying

important ECG patterns.

In the next part of out research, we implement a federated learning setup using the Federated Averaging (FedAvg) algorithm, where model updates from local nodes are periodically aggregated to create a global model. The decentralized nature of federated learning complicates the explainability of the global model due to variations in the local data distributions, which affect how each local model learns and prioritizes features. For example, different devices or institutions may have ECG data from varied populations, potentially causing the local models to focus on different aspects of the ECG signals. These discrepancies pose a significant challenge to the interpretability of the global model, as local feature representations may conflict, leading to a fragmented feature space that reduces model transparency and alignment.

To better understand how data augmentations affect feature learning and explainability in this federated context, we employ a range of augmentation techniques, including jittering, scaling,

Flipping, Time/Magnitude warping and more advanced embedding augmentations like random walking and angular variance modifications based on class distribution. These visualizations provided insights into the potential of augmentation to enrich the model's feature space, making it more resilient and interpretable. However, when these augmentations were applied in a federated setup using FedAvg, the improvements in feature learning were less evident, indicating that the federated setup imposes unique constraints on how augmentation strategies impact the feature space.

To address this issue, we introduced embedding augmentations, where modifications are applied directly to the learned feature embeddings rather than to the raw input data. Unlike traditional augmentations that work at the input level, embedding augmentations offer a way to harmonize feature representations within the internal layers of the model. This approach is especially valuable in a federated learning environment, where aligning the feature space across diverse nodes is challenging. Embedding augmentations can effectively stabilize and unify the feature learning process, reducing the inconsistencies that arise in the decentralized setting. Grad-CAM visualizations of the augmented embeddings reveal a more consistent feature focus across nodes, suggesting that embedding-based augmentation can enhance the global model's interpretability by creating a more aligned feature space.

This thesis aims to advance the understanding of how different augmentation strategies influence the explainability and feature space alignment in federated learning models, specifically those trained using the FedAvg algorithm. By exploring the role of embedding augmentations, we highlight a pathway to improve both the robustness and transparency of models trained on decentralized data, addressing the unique challenges of federated learning. Our findings contribute to the growing field of explainable AI, particularly in applications where interpretability is critical, such as healthcare, where understanding a model's focus and feature prioritization can provide valuable insights into complex medical data.

Our initial investigation was centered on uncovering the effects of different label-invariant augmentation methods on feature learning within convolutional neural networks (CNNs) applied

to 1D time-series electrocardiogram (ECG) data. In deep learning, data augmentation is widely used to address data scarcity, improve generalization, and reduce overfitting. However, in healthcare applications, especially with ECG data, augmentation requires a nuanced approach: while enhancing model performance, transformations can unintentionally alter signal characteristics critical to clinical interpretation. Given the high stakes of healthcare applications, understanding how augmentations influence feature learning and model activation patterns is essential to ensuring model transparency and reliability.

In this experiment, we applied a series of augmentation techniques—including scaling, jitter, magnitude warping, and time warping—chosen based on their prevalence in ECG classification research. These methods aim to preserve the integrity of class labels while expanding data diversity. Using Grad-CAM (Gradient-weighted Class Activation Mapping) to analyze model responses, we were able to visually examine how each augmentation strategy impacted the CNN's focus areas on ECG signals. The resulting activation maps provided insights into how these augmentations could improve or degrade model performance by either highlighting relevant features or diverting focus to non-essential patterns.

Our findings revealed distinct variations in model behavior depending on the augmentation type. For example, scaling produced improvements in classification accuracy by enhancing the model's attention to key signal features, while jitter sometimes led to a decrease in performance by introducing noise that obscured clinically relevant patterns. These observations underscored the importance of informed augmentation selection, as certain transformations could

inadvertently alter features significant to healthcare outcomes. This experiment not only highlighted the value of visual explanations in understanding model internals but also established a foundation for developing augmentation strategies that enhance feature learning in ECG-based deep learning models. Through this analysis, we lay the groundwork for a more systematic approach to data augmentation in healthcare, one that aligns with the goal of explainable AI and prioritizes both model robustness and interpretability(Balasubramanian and Dakshit 2024).

Chapter 2

Augmentation in Time-Series Healthcare Signals

2.1 Overview of Time-Series Data in Healthcare

Time-series healthcare data, such as ECG and other physiological recordings, presents specific

challenges and needs in model training and application. The classification of ECG signals, as a

prime example of time-series data, often faces issues like class imbalance, where common

patterns, such as non-ectopic beats, vastly outnumber rarer types like supraventricular or

ventricular ectopic beats. This imbalance (Fan et al. 2022) skews model performance, making it

challenging to accurately identify rare but critical arrhythmias. To address this, various methods

have been proposed, such as resampling techniques, weighted models, and ensemble learning

approaches, which aim to improve sensitivity for minority classes and balance predictions across

categories(Xu, Zhang, and Xiao 2023).

Additionally, noise and variability in ECG signals, often arising from patient movement,

electrode placement, and equipment variations, introduce another layer of complexity.

Techniques like denoising autoencoders (DAEs) and feature extraction methods, such as

principal component analysis (PCA) and wavelet transformations, are employed to filter noise

while preserving clinically relevant features in the data. These methods help create stable

representations that improve model reliability across diverse patient conditions.

Generalization across varied patient populations remains a significant challenge, especially in distributed learning frameworks where privacy concerns limit data sharing between institutions. To address this, approaches like source-free domain adaptation (SFDA) have been explored, ensuring that models trained on one dataset can adapt to others without direct access to target data. Techniques such as synthetic data generation and pseudo-labeling strategies further enhance generalizability and support robust model performance across clinical settings(Yuan and Siyal 2023).

2.2 The Need for Augmentation in Healthcare Time-Series

Data augmentation is indispensable in healthcare time-series, where it addresses issues of data scarcity, class imbalance, and noise. By creating additional training data from existing samples, augmentation reduces overfitting and improves model generalization. In healthcare, this is especially valuable as it enables the model to handle patient and environmental variations effectively.

(Guo, Yang, and Sano 2023), in their paper "Empirical Study of Mix-based Data Augmentation Methods in Physiological Time Series Data," demonstrated that mix-based augmentations (like mixup, cutmix, and manifold mixup) provide consistent performance improvements across physiological datasets by introducing variability without extensive tuning. This approach highlights how augmentation can enhance robustness in healthcare time-series data, enabling models to generalize across diverse clinical scenarios.

2.3 Augmentation Techniques Applied in Time-Series Healthcare Signals

To address the challenges in healthcare time-series data, we implemented several augmentation techniques(Yang et al.) that introduce controlled variation. Each technique was chosen to enhance the model's ability to generalize to real-world data:

2.3.1 No Augmentation (Baseline)

The baseline condition, where no augmentation was applied, serves as a reference point to evaluate the effectiveness of each augmentation method. By comparing against the original, unaltered data, we can measure the degree of improvement that each augmentation method contributes to model performance.(Yang, Yu, and Sano 2022)
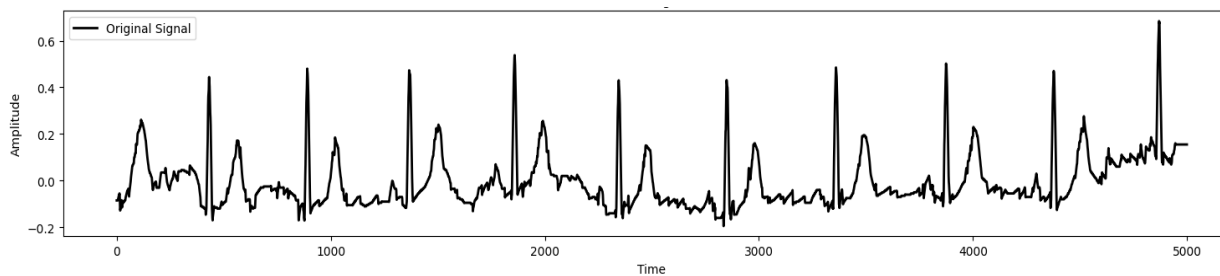
Figure 2.1- Base Signal

Formula: $x' = x$

2.3.2 Jittering

Jittering involves adding random noise to each data point in the time series, simulating small-scale variations typical in real-world signals. In healthcare data, slight fluctuations in sensor

readings are common due to environmental or physiological factors. By introducing controlled

noise, jittering encourages the model to ignore minor inconsistencies, improving its robustness to

data noise. In this study, we used a noise parameter, sigma, to control the magnitude of the jitter,

ensuring realistic variations that mirror true physiological conditions.(Yang, Yu, and Sano 2022)
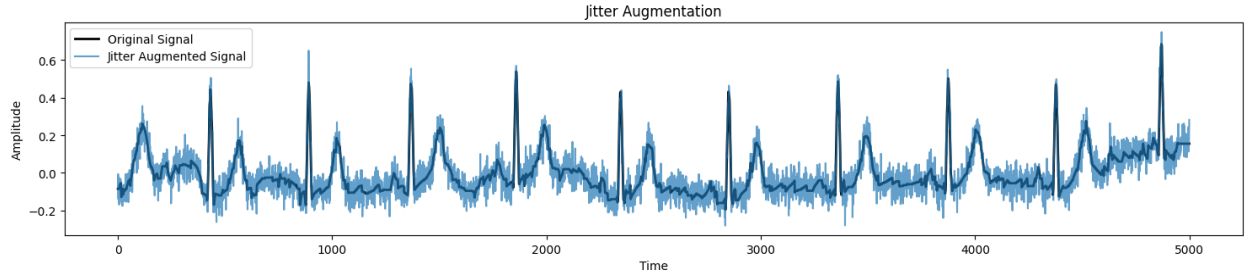
Figure 2.2 Jitter

Formula: $x' = x + \epsilon$, where $\epsilon \sim N(0,\sigma)$.

2.3.3 Scaling

Scaling adjusts the amplitude of each data point by a random factor, simulating the natural

variations in signal strength observed across different patients or sensor placements. This

approach is particularly useful for making the model adaptable to amplitude variations without

distorting the overall signal structure. While scaling enhances robustness, it also posed memory

load issues, which we addressed by applying scaling on a per-point basis within each data

segment.(Yang, Yu, and Sano 2022)
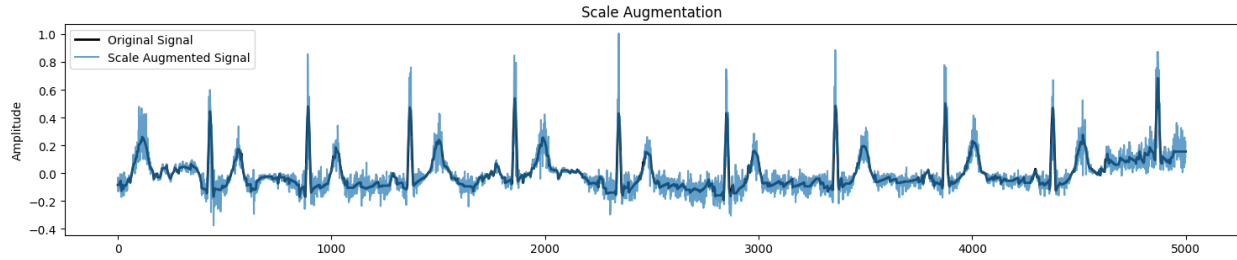
Figure 2.3 Scale

Formula: $x' = x + \epsilon$, where $\epsilon \sim N(0,\sigma)$.

2.3.4 Flipping

The flipping augmentation reverses the order of data points in the sequence. This is useful for time-series data where directionality may not impact interpretation, such as in cyclic or recurring signals. Flipping helps the model recognize patterns regardless of sequence direction, broadening the variety of patterns it can learn. We applied this technique specifically to 1D and 2D data with single-feature dimensions.(Yang, Yu, and Sano 2022)
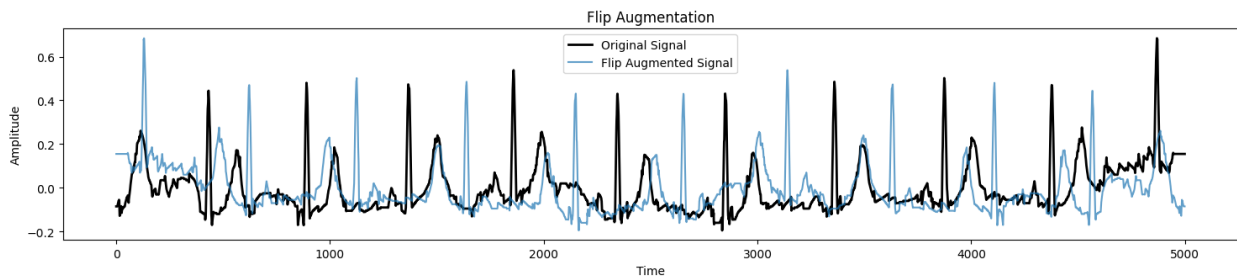
Figure 2.4 Flip

Formula: x'=x[end:0]

2.3.5 Magnitude Warping

Magnitude warping smoothly alters the amplitude of the time-series data using spline-based adjustments, simulating gradual shifts in signal intensity over time. This can reflect natural physiological variations, such as fluctuations in heart rate or respiratory patterns. By warping the signal amplitude, this technique helps the model learn from more diverse amplitude patterns, leading to improved generalization across different patient conditions.(Yang, Yu, and Sano 2022)
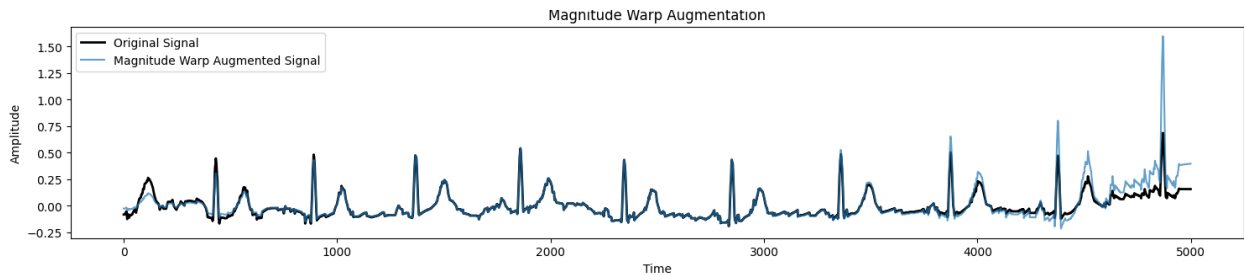
Figure 2.5 Magnitude warp

Formula: $x^{\wedge\prime} = x \cdot w(t)$, where $w(t)$ is a cubic spline interpolation.

2.3.6 Time Warping

Time warping distorts the time axis, adjusting the temporal positions of data points to simulate variable signal speed. This technique is useful in healthcare for mimicking irregular event durations, such as variations in heartbeat or breathing rate. By warping the time axis, we encourage the model to learn patterns that are resilient to timing fluctuations, a critical feature in real-world healthcare applications.(Yang, Yu, and Sano 2022)
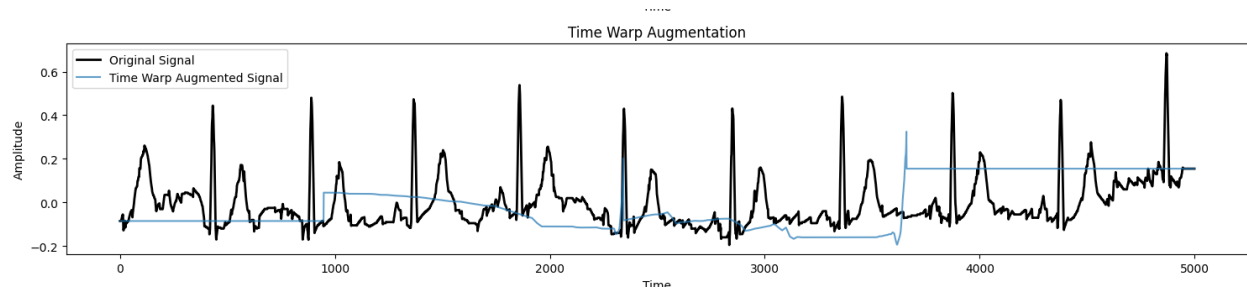
Figure 2.6 Time warp

Formula: t′ = S(t), where S is a spline function with added noise.

2.3.7 Window Slicing

Window slicing extracts a random segment from within the time series and pads it to fit the original length. This technique enables the model to train on partial observations, as is often the case in real-world data collection where some segments might be missing or incomplete. Window slicing provides the model with a variety of partial views, improving its adaptability and performance on incomplete data.(Yang, Yu, and Sano 2022)
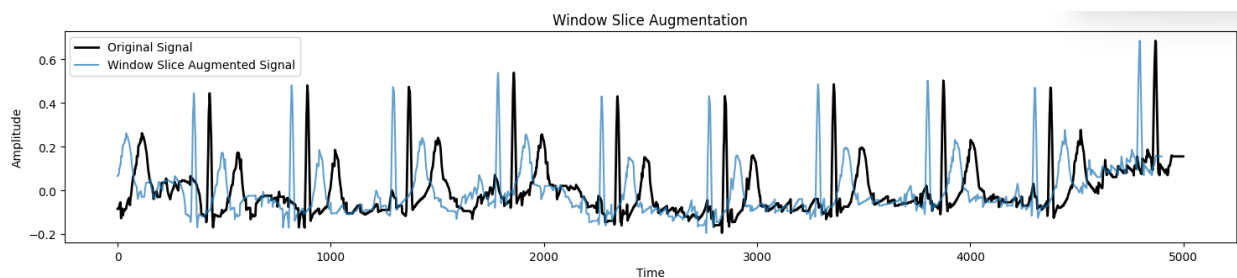
Figure 2.7 Window slice

Formula: x' = x[t0: t0 + L]

with padding,where t0 is a random start index and L is the window length.

2.3.8 Window Warping

Window warping involves selecting a specific window within the time series and repeating it, introducing a stretched effect in the sequence. This augmentation can simulate prolonged events, such as extended heartbeats or respiratory cycles. By altering the length of certain signal segments, window warping encourages the model to become more flexible with event duration, ensuring robustness across a range of signal lengths.(Yang, Yu, and Sano 2022)
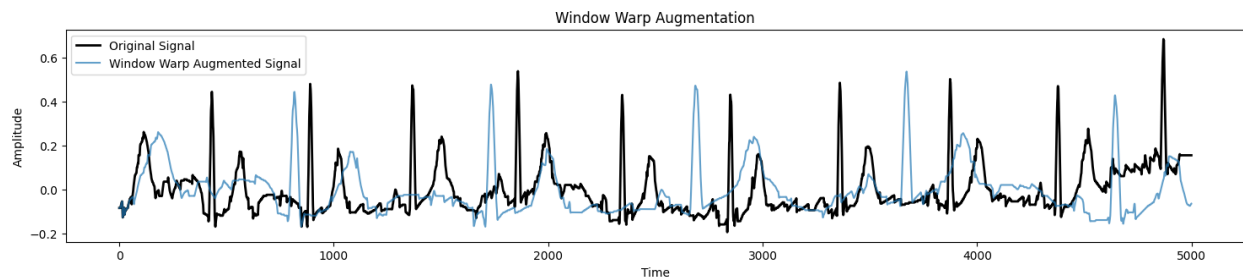
Figure 2.8 Window warp

Formula: $' = x[: t0] \cup x[t0: t0 + L] \cdot \beta \cup x[t0 + L:]$

where t0 is a random start index,L is the window length,and β is a stretching factor.

2.3.9 Advanced Techniques and Challenges in Augmenting Time-Series Healthcare Data

Traditional augmentations may require expert tuning, and sometimes compromise data integrity.

Advanced techniques like mix-based augmentations (e.g., mixup, cutmix, and manifold mixup)

create richer feature representations by combining multiple data samples, reducing dependence on domain-specific adjustments:

Mixup: Blends two signals, generating transitions that capture variability.(Guo, Yang, and Sano 2023)

Cutmix: Replaces a segment with another signal, enhancing diversity.(Guo, Yang, and Sano 2023)

Manifold Mixup: Blends features in hidden layers, capturing abstract patterns essential for high-dimensional data.(Guo, Yang, and Sano 2023)

These advanced techniques are shown to improve classification accuracy and generalization in physiological datasets, as reported by Authors Guo et al(Guo, Yang, and Sano 2023).

2.4 Datasets

To evaluate augmentation techniques for time-series healthcare data, we used two prominent ECG datasets: the MIT-BIH Arrhythmia dataset and the PTB-XL dataset(Wagner et al. 2020). These datasets provide a comprehensive foundation for testing data and embedding augmentation techniques in both single-label, multi-class, and multi-label classification scenarios, each representing real-world clinical challenges.

2.4.1 MIT-BIH Arrhythmia Dataset

The MIT-BIH Arrhythmia Dataset is a foundational resource in arrhythmia detection, widely utilized due to its comprehensive representation of different heart rhythms and associated abnormalities. Developed by the Massachusetts Institute of Technology (MIT) and Beth Israel

Hospital (BIH) in the 1980s, this dataset includes 48 half-hour ECG recordings from 47 individuals, annotated to reflect a wide array of heartbeat types and arrhythmias. It is frequently used to train and evaluate algorithms for arrhythmia detection, providing a valuable benchmark for studies aiming to improve diagnostic accuracy in clinical and remote monitoring applications(Apandi, Ikeura, and Hayakawa 2018)

In this study, we focused on the four most frequently observed arrhythmias within the dataset: Normal Sinus Rhythm, Atrial Fibrillation, Peri-Ventricular Contraction, and Left Bundle Branch Block, comprising 198, 95, 94, and 73 samples, respectively. This selection allowed us to analyze the dataset's potential in addressing multi-class classification challenges. These four arrhythmia types represent a range of conditions with varying prevalence, making the MIT-BIH dataset especially valuable for evaluating how data augmentation techniques impact model performance across both common and rare classes(Apandi, Ikeura, and Hayakawa 2018)

A significant challenge in working with the MIT-BIH dataset is its inherent class imbalance. Most ECG samples correspond to common rhythms like Normal Sinus Rhythm, while arrhythmias such as Left Bundle Branch Block and Peri-Ventricular Contractions are underrepresented. This imbalance can lead models to overfit on majority classes, thereby reducing their sensitivity to less frequent arrhythmias. Researchers have noted that without proper handling of this imbalance, models tend to perform poorly in detecting minority classes, which can have critical implications in clinical practice where detecting rare arrhythmias early is essential(Moody and Mark 2001).

To address this, we applied label-invariant augmentation techniques, which modify ECG signals without changing their fundamental labels. By introducing transformations such as jittering, scaling, and time-warping, we aimed to increase the dataset's diversity and balance the representation of each arrhythmia type. These augmentation strategies allow the model to learn from a broader range of variations within each class, reducing its dependence on majority classes and enhancing its sensitivity to underrepresented arrhythmias. Such approaches have been shown in other studies to improve the model's ability to detect subtle yet clinically significant features specific to each arrhythmia type, supporting better generalization across patient samples(Moody and Mark 2001).

Class 0 had 198 samples before augmentation and remained at 198 samples afterward. Class 1 had 95 samples initially, which increased to 198 post-augmentation. Class 2 started with 94 samples and was augmented to 198 samples, while Class 3 had 73 samples before augmentation, reaching 198 samples afterward. This augmentation process balanced the dataset, equalizing the sample count across classes and mitigating the initial class imbalance, providing the model with a more uniform representation of each category.

Thus, the MIT-BIH Arrhythmia Dataset, combined with strategic augmentation, offers a robust framework for testing and validating classification models in realistic, imbalanced conditions. By leveraging its diverse, annotated samples and applying augmentation, we aim to enhance our model's ability to generalize across arrhythmia classes, making it more reliable and adaptable for potential clinical applications.

2.4.2 PTB-XL Dataset

The PTB-XL(Wagner et al. 2020) dataset is a comprehensive collection that serves as a valuable resource for both multi-class and multi-label classification tasks, making it especially suitable for evaluating ECG analysis models. This dataset encompasses a range of cardiac conditions that are categorized under five primary diagnostic superclasses: normal ECG, conduction disturbances, myocardial infarction, hypertrophy, and ST/T changes. Each superclass contains thousands of samples, ensuring diverse representation and aiding in the training of models capable of distinguishing subtle variations across different heart conditions(Strodthoff et al. 2021)

In multi-label classification, the PTB-XL dataset provides a robust framework for models to handle conditions with overlapping labels. For example, conditions such as conduction disturbances (2,387 samples), myocardial infarction (4,919 samples), hypertrophy (815 samples), and ST-T changes (4,714 samples) represent specific ECG abnormalities that the model must learn to differentiate independently. This setup is instrumental in testing model specificity, where achieving high precision for distinct classes is essential in clinical diagnostics(Strodthoff et al. 2021)

Since this is a multi-label classification problem, achieving a perfectly balanced dataset with label-invariant augmentation is inherently challenging. In multi-label setups, each sample may belong to multiple classes simultaneously, meaning that augmenting data for one class can inadvertently increase sample counts for overlapping classes. This characteristic prevents strict class-level balancing, as adding new instances of underrepresented classes will often impact the sample distribution of co-occurring classes.

In this augmented PTB-XL dataset, while the final distribution improves class representation—resulting in 12,467 samples for Class 0, 10,847 for Class 1, 12,925 for Class 2, 8,952 for Class 3, and 13,906 for Class 4—it cannot reach an exact balance across classes. However, this balanced yet approximate distribution is beneficial in training the model to recognize distinct patterns while maintaining sensitivity to the multi-label nature of ECG abnormalities. The augmentation process thus supports a dataset that is both comprehensive and representative, essential for developing a model capable of nuanced, clinically relevant multi-class predictions.

The multi-label setup in PTB-XL, where ECG recordings may simultaneously exhibit multiple abnormalities, introduces an added layer of complexity. For instance, a single recording can manifest characteristics of both myocardial infarction and conduction disturbances. This structure reflects real-world clinical scenarios where patients often present with co-occurring cardiac conditions, making it critical for models to accurately identify overlapping features. Author (Śmigiel, Pałczyński, and Ledziński 2021) in their research utilized this dataset's multi-label aspect thus provides insights into the effectiveness of augmentation techniques in handling complex interactions between multiple abnormalities in ECG data. The richness of the PTB-XL dataset, combined with its application in multi-class and multi-label classification, makes it a benchmark for evaluating various augmentation strategies and machine learning models aimed at improving ECG classification accuracy in diverse clinical contexts(Śmigiel, Pałczyński, and Ledziński 2021)

2.5 Model Architecture
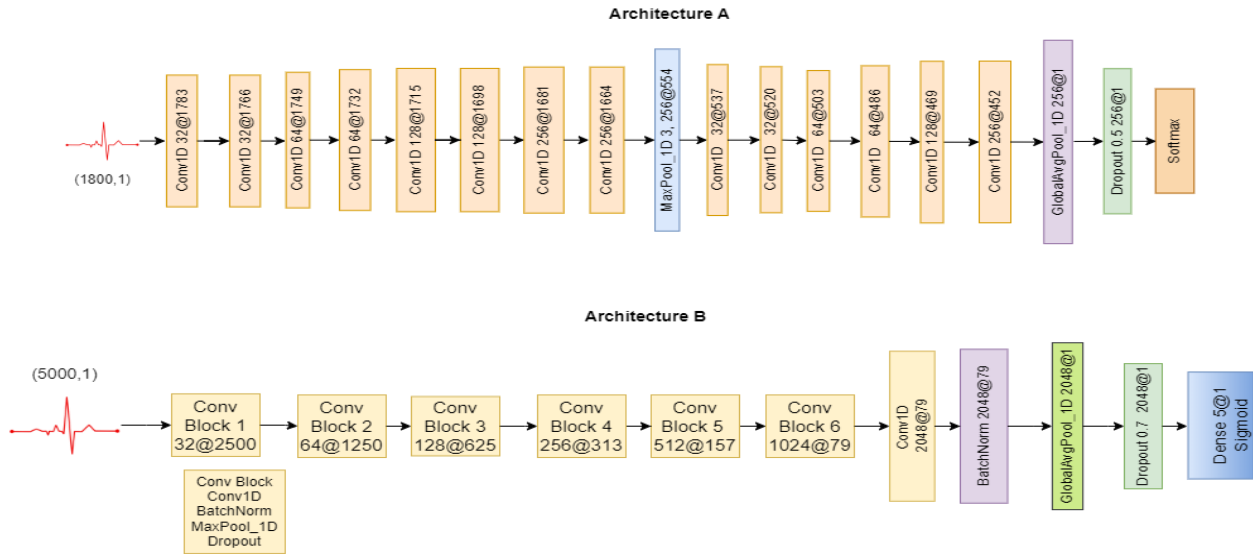
Figure 2.9 Deep Neural Network Architectures

To meet the specific demands of multi-class and multi-label ECG classification, we designed two distinct convolutional neural network (CNN) architectures—Architecture A and Architecture B—each tailored to excel in its designated classification task. These architectures, as illustrated in the figure, capture the intricate temporal and morphological patterns inherent in ECG signals, which are critical for accurate cardiac condition diagnosis.

2.5.1 Architecture A: Multi-Class Classification

Architecture A, illustrated in the upper section of the figure, is structured to handle multi-class classification tasks, where each ECG sample is associated with a single cardiac condition. This architecture is designed for efficiency and interpretability, providing a balanced approach to feature extraction without excessive computational complexity.

Input Layer: Architecture A accepts 1D ECG signals with an input shape of (1000,1), aligning with single-lead ECG formats.

2.5.1.1 Convolutional Blocks*:*

The architecture begins with a sequence of convolutional blocks that enhance feature depth progressively, a design choice found effective in capturing varied temporal patterns in ECG signals(Tang et al. 2020)The initial layers employ 1D convolutional filters with 32 and 64 units, designed to capture low-level ECG features. Each convolutional block is followed by max-pooling, which down samples the feature maps, conserving essential features while reducing spatial dimensions, a technique known to improve computational efficiency in ECG signal processing(Dakshit and Prabhakaran 2023)(Ahmed et al. 2023) Further convolutional layers with increased filter counts (128 and 256) extract more complex patterns, which are crucial for distinguishing between cardiac conditions. This setup is in line with approaches that emphasize hierarchical feature extraction through multiple convolutions and pooling operations mentioned by (Tang et al. 2020)

2.7.1.2 Intermediate Layers:

Following the initial convolutional blocks, two additional convolutional layers with 32 and 64 filters are applied to refine mid-level features, as depicted in the diagram. These layers are crucial for distinguishing between similar but distinct arrhythmias.

2.7.1.3 Final Convolutional Layers and Global Pooling:

Two final convolutional layers with 128 and 256 filters increase the model's capacity for extracting high-level features before the global average pooling layer reduces dimensionality. This reduction condenses the information into a compact representation, retaining only the most relevant features for classification.

2.5.1.4 Output Layer:

The output layer uses softmax activation, allowing the model to assign probabilities to each class. This configuration supports single-label predictions, making Architecture A suitable for multi-class classification tasks where each ECG sample has a unique class label.

2.5.2 Architecture B: Multi-Label Classification

Architecture B, illustrated in the lower section of the figure, is tailored for multi-label classification tasks, where an ECG sample may exhibit multiple cardiac conditions simultaneously. The architecture is deeper and incorporates more regularization, designed to manage the increased complexity associated with multi-label scenarios.

Input Layer: Architecture B processes 1D ECG signals with an input shape of (5000,1), capturing longer temporal sequences that are essential for multi-label tasks requiring comprehensive signal information.

2.5.2.1 Initial Convolutional Block:

The architecture begins with an extensive initial block comprising three convolutional layers with 32, 64, and 128 filters, each followed by batch normalization and dropout for regularization, as depicted in the diagram.

This block uses smaller kernel sizes to precisely capture subtle signal fluctuations, an essential feature for distinguishing multiple overlapping conditions.

2.5.2.2 Deep Convolutional Layers:

Following the initial block, the model includes deeper convolutional layers with significantly higher filter counts (256, 512, 1024, and 2048 filters) to capture complex, high-level features.

Each layer is followed by batch normalization, max-pooling, and dropout layers, which stabilize learning and reduce overfitting. These layers enable the model to build a hierarchical understanding of the ECG signals, essential for recognizing multiple co-occurring cardiac conditions in a single recording.

2.5.2.3 Global Average Pooling Layer:

After the final convolutional layer, Architecture B incorporates a global average pooling layer to condense the learned features into a compact representation, making the data suitable for multi-label output. This layer averages each feature map channel, reducing the risk of overfitting associated with fully connected layers and preparing the model for simultaneous label prediction.

2.5.2.4 Output Layer with Sigmoid Activation:

The output layer employs sigmoid activation, allowing each output neuron to act independently. This configuration enables multi-label predictions, where the model can assign multiple labels to a single ECG sample, making it ideal for detecting multiple cardiac abnormalities within the same recording.

2.5.3 Summary of Architectural Design

Both architectures are optimized to meet the unique requirements of their respective classification tasks. Architecture A is designed for efficient, single-label classification, balancing interpretability with depth to distinguish between distinct cardiac conditions. In contrast, Architecture B uses a deeper and more regularized design, suitable for complex multi-label tasks where conditions may overlap within a single ECG signal. Together, these architectures provide a robust framework for evaluating augmentation techniques across varied classification scenarios, each structure optimized to maximize performance and generalization in ECG-based healthcare diagnostics.

2.6 Related Works

Data augmentation in healthcare time-series has been extensively studied:

GANs for Atrial Fibrillation Detection: GANs improve class representation for minority classes, enhancing model performance in ECG classification(Akhoondkazemi et al. 2023)

ECG Augment by (Nonaka and Seita 2020) Combines scaling, flipping, and noise addition, enhancing robustness for ECG classification without complex data generation.

These works highlight both traditional and advanced methods for improving classification performance, with GANs and mix-based augmentations providing promising results for realistic data synthesis and class balance.

2.7 Overfitting and Underfitting Challenges

One of the primary concerns in data augmentation is balancing between overfitting and underfitting:

Overfitting: Some augmentation methods can cause the model to overfit to augmented data patterns, especially when these patterns do not align with the real-world characteristics of ECG signals. For example, our study found that jittering, which introduces random noise to the data, often led to overfitting. The noise introduced by jittering may cause the model to focus on irrelevant features rather than clinically meaningful features, reducing its ability to generalize to new data. This result underscores the need for cautious use of noise-based augmentations in clinical datasets.

Underfitting: Conversely, underfitting can occur when augmentation methods are too simplistic or do not introduce enough variability, preventing the model from learning complex patterns in the data. Our study suggests that augmentations like flipping or permutation, while useful in increasing data variety, might not capture the depth of variability needed for ECG classification. These simpler augmentations could fail to expand the model's feature space adequately, resulting in limited performance improvements.

2.8 Summary

This chapter provides an in-depth overview of datasets, experimental setup, and augmentation techniques in healthcare time-series analysis. Using datasets like MIT-BIH and PTB-XL, alongside specialized architectures (Architecture A for multi-class and Architecture B for multi-label classification), we evaluated a range of augmentation methods. From traditional transformations like jittering to advanced techniques like mix-based augmentations, these methods prove effective in enhancing model robustness, accuracy, and generalization, addressing key challenges in healthcare time-series data. The related works underline the importance of tailored augmentation approaches to meet the complex demands of clinical applications.

Chapter 3:

How Augmentation affects Feature Space

3.1 Challenges due to Data Augmentation

Data augmentation is a powerful tool for increasing data variability, particularly valuable in

healthcare, where datasets are often limited in size and diversity. However, in fields like ECG

analysis, the application of augmentation presents specific challenges. While techniques like

jittering, scaling, and time warping can introduce useful variations, they also risk altering

clinically relevant features, which can reduce a model's interpretability and reliability. In

medical contexts, it is essential that augmentations do not introduce artifacts that could lead to

misdiagnoses or mask important clinical patterns. Thus, our study aims to rigorously evaluate

augmentation methods to determine which approaches improve model performance without

compromising clinical accuracy(Balasubramanian and Dakshit 2024).

3.2 Evaluating Augmentation Effects on Model Performance

Our study assesses the effectiveness of various augmentation techniques on ECG data using

multiclass and multilabel classification tasks. Performance metrics such as accuracy, precision,

recall, and AUC were used to provide a comprehensive view of how each augmentation method

impacts model learning and generalization. By applying these metrics, the study could observe

not only the gains in predictive performance but also how well the models handled variations

introduced through augmentation. The results highlight that while some augmentations, like

scaling and magnitude warping, contribute positively to the model's performance, others, such as jittering, may distort critical signal features, adversely affecting the classification accuracy.

3.3 Impact of Label-Invariant Augmentation Techniques on Multiclass and Multilabel ECG Classification

Our study conducted in-depth analysis across two well-known ECG datasets—MIT-BIH Arrhythmia (Dataset A) and PTB-XL (Dataset B)—to evaluate each augmentation method's effect on multiclass and multilabel classification:

Multiclass Classification on MIT-BIH: For the MIT-BIH dataset, which focuses on detecting different types of arrhythmias, the study applied a CNN model to examine how each augmentation affected predictive performance. Among the techniques tested, scaling yielded the most significant improvements, with a 3.13% increase in accuracy and comparable gains in precision and recall. This indicates that scaling introduces realistic variations in signal amplitude that help the model adapt to patient-specific differences in heart signal strength. Meanwhile, magnitude warping showed a 2.61% boost in accuracy, highlighting its value in enhancing class separability without distorting critical signal morphology(Balasubramanian and Dakshit 2024)

Multilabel Classification on PTB-XL: For the PTB-XL dataset, which includes a wide range of ECG diagnoses, the multilabel classification task posed unique challenges. Window slicing and jittering, which altered the signal structure, led to performance declines, with window slicing

resulting in a 3% drop in accuracy. This suggests that slicing could disrupt essential temporal

dependencies within ECG signals, negatively impacting multilabel classification where several

conditions are identified simultaneously. Conversely, time warping and window warping

produced modest performance gains, as these techniques helped the model adapt to slight

variations in the duration and structure of ECG waveforms.

3.4 Impact of Augmentation on Class Imbalance

Class imbalance is a common issue in ECG datasets, where certain conditions (e.g., rare

arrhythmias) are underrepresented. The study we conducted explored how augmentations could

address this imbalance by generating synthetic data for minority classes.

Scaling and Magnitude Warping: These methods proved particularly useful for balancing classes

in multiclass classification. For example, scaling allowed the model to learn from amplified

signals, which improved the identification of rarer arrhythmias without oversampling. This

adjustment led to an improvement in precision and recall across imbalanced classes, indicating

that these augmentations create a richer feature space for minority classes.

Limitations of Simple Augmentation: For multilabel classification, class imbalance posed

additional challenges. Time warping showed some success in improving recall for rare

conditions, with a modest 0.0204 increase. However, simpler augmentations like window slicing

failed to address imbalance effectively, sometimes omitting critical patterns necessary for identifying coexisting conditions in multilabel settings.

3.5 Clinical Relevance and Preservation of Signal Integrity

Maintaining the clinical relevance of ECG signals is paramount when applying augmentations. Our study emphasizes that augmentations must be carefully selected to avoid altering the signal in ways that could mislead the model.

Successful Methods: Scaling and magnitude warping maintained the integrity of ECG signals, enhancing the model's learning capacity without compromising clinical features. These augmentations preserved the underlying signal morphology, ensuring that the model learned relevant patterns rather than artificial distortions.

Challenging Augmentations: Jittering, while beneficial in other fields, added random noise that negatively impacted classification accuracy in ECG data. The study observed that such augmentations could obscure diagnostically important features, such as P-wave and QRS complex shapes, which are essential for accurate arrhythmia classification. The results underscore the need to prioritize augmentations that maintain signal integrity, particularly for sensitive medical data.(Balasubramanian and Dakshit 2024)

3.6 Limitations and Challenges in Real-World Applications

Despite t6he observed benefits, our study identifies practical limitations in applying these augmentation methods in real-world clinical environments:

Scalability Issues: Some augmentations, particularly time-based methods like time warping and magnitude warping, require substantial computational resources. For large-scale implementations, these methods may introduce latency and require high-performance computing capabilities, which can be costly and complex to integrate into clinical workflows.

Deployment Challenges: The need for careful validation and regulatory compliance limits the use of certain augmentations in clinical models. Augmented models must undergo rigorous testing to ensure they maintain diagnostic accuracy, particularly in applications with real-time requirements. The study suggests that further research into adaptive augmentation methods could help overcome these challenges by automatically adjusting the augmentation process based on real-time data characteristics.

3.7 Summary

Our study provides a thorough examination of augmentation methods for ECG classification, highlighting both effective techniques and challenges in implementation. Key takeaways include:

Effective Techniques: Scaling and magnitude warping proved most beneficial, improving classification accuracy without compromising clinical features. These methods showed consistent improvements in multiclass classification tasks and helped manage class imbalance in minority classes.

Areas for Improvement: Methods like jittering and window slicing had mixed results, often reducing accuracy or omitting crucial temporal information. The study recommends cautious application of such augmentations in healthcare data.

3.8 Results and Grad-CAM Analysis

The effectiveness of various augmentation techniques was empirically evaluated on two datasets, Dataset A (MIT-BIH Arrhythmia) and Dataset B (PTB-XL), across tasks of multiclass and multilabel classification. The models trained with each augmentation method were compared against a baseline model trained without augmentation, using metrics such as accuracy, precision, recall, and AUC to assess performance changes.

3.8.1 Multiclass Classification Results (Dataset A)

Scaling: Scaling yielded the highest improvements across metrics, with a 3.13% increase in accuracy, 2% in precision, and 3% in recall compared to the baseline. These improvements highlight the potential of scaling to enhance the model's ability to recognize variations in signal amplitude across classes.

Magnitude Warping: This method resulted in a 2.61% increase in accuracy. Magnitude warping was particularly effective in handling class imbalance by expanding the feature space, which helped the model learn distinguishing features for underrepresented classes.

Time Warping and Window Warping: Both methods contributed to notable improvements, with Time Warping showing a 2.61% improvement in accuracy, while Window Warping increased recall by 4%. These methods allowed the model to adapt to temporal variability within the ECG signals, which proved beneficial in recognizing arrhythmia patterns.

However, jitter augmentation led to decreased performance, suggesting that adding random noise to ECG signals may distort the clinical characteristics essential for accurate classification. This indicates that certain augmentations, while increasing robustness, may negatively impact model accuracy if they overly distort important signal features.

3.8.2 Multilabel Classification Results (Dataset B)

For multilabel classification of conditions such as conduction disturbance and myocardial infarction on the PTB-XL dataset, augmentation methods again proved beneficial, though results varied:

Window Slice: This augmentation method, which involved slicing the ECG signal into windows, showed the most significant drop in performance, with accuracy decreasing by approximately

3%. This result suggests that breaking down signals into smaller segments may omit crucial temporal information needed for accurate multilabel classification.

Magnitude Warp and Time Warp: Both augmentations achieved modest improvements in recall values, with Time Warp recording the highest gain at 0.0204. This supports the utility of warping techniques in preserving essential signal characteristics while introducing variations that improve the model's ability to generalize across different patient profiles.

Window Warping: Window Warping produced stable AUC values and performed well across metrics, indicating its effectiveness in handling multilabel classification. This technique's ability to stretch segments within signals appears to provide a beneficial diversity without losing critical temporal features.

Table 3.1: Model Accuracy Across Augmentation Methods for Multiclass and Multilabel Classification

	Experiment 1				Experiment 2			
Augmentation	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall	AUC
No-Augmentation	90.10%	0.94	0.89	0.9864	83.38%	0.6676	0.6891	0.8527
Flip	91.15%	0.92	0.91	0.9780	82.41%	0.6464	0.6791	0.8395
Jitter	89.06%	0.93	0.87	0.9798	82.97%	0.6644	0.6658	0.8486

Magnitude Warp	92.71%	0.94	0.93	0.9839	83.32%	0.6643	0.6945 0.8504
Scaling	93.23%	0.96	0.92	0.9784	82.02%	0.6332	0.6945 0.8405
Time Warp	92.71%	0.94	0.93	0.9694	82.24%	0.6345	0.7095 0.8433
Window Slice	91.15%	0.93	0.90	0.9676	80.38%	0.5968	0.7024 0.8298
Window Warp	92.19%	0.93	0.93	0.9868	83.13%	0.6599	0.6934 0.8504

Overall, the results emphasize the importance of selecting augmentation methods carefully based on the specific requirements of multiclass versus multilabel classification tasks. Certain augmentations, such as scaling and warping, show strong potential for enhancing model accuracy and robustness, while others, like jittering and window slicing, may introduce distortions that degrade performance.

3.8.3 Calculations for Experiment A

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 3.1 Accuracy

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

Equation 3.2 Precision

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

Equation 3.3 Recall

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Equation 3.4 F1-Score

3.8.3.1 No Augmentation

Confusion Matrix

$$\begin{bmatrix} 81 & 1 & 1 & 0 \\ 0 & 40 & 0 & 0 \\ 17 & 0 & 22 & 0 \\ 0 & 0 & 0 & 30 \end{bmatrix}$$

$$Accuracy = \frac{81 + 40 + 22 + 30}{192} = 0.9010$$

Class 0

$$\begin{bmatrix} 81 & 2 \\ 17 & 92 \end{bmatrix}$$

$$Precision = \frac{81}{81 + 17} = 0.826 \sim 0.83$$

$$Recall = \frac{81}{81 + 2} = 0.9759 \sim 0.98$$

$$F1\ Score = \frac{2 * 0.83 * 0.98}{0.83 + 0.98} = 0.898 \sim 0.90$$

Class 1

$$\begin{bmatrix} 40 & 2 \\ 0 & 150 \end{bmatrix}$$

$$Precision = \frac{40}{40 + 1} = 0.9756 \sim 0.98$$

$$Recall = \frac{40}{40 + 0} = 1.00$$

$$F1\ Score = \frac{2 * 1 * 0.98}{1 + 0.98} = 0.989 \sim 0.99$$

Class 2

$$\begin{bmatrix} 22 & 1 \\ 17 & 152 \end{bmatrix}$$

$$Precision = \frac{22}{22 + 1} = 0.956 \sim 0.96$$

$$Recall = \frac{22}{22 + 17} = 0.5641 \sim 0.56$$

$$F1\ Score = \frac{2 * 0.56 * 0.96}{0.56 + 0.96} = 0.707 \sim 0.71$$

Class 3

$$\begin{bmatrix} 30 & 0 \\ 0 & 162 \end{bmatrix}$$

$$Precision = \frac{30}{30 + 0} = 1.00$$

$$Recall = \frac{30}{30 + 0} = 1.00$$

$$F1\ Score = \frac{2 * 1 * 1}{1 + 1} = 1.00$$

$$Macro\ Avg\ Precision = \frac{0.83 + 0.98 + 0.96 + 1.00}{4} = 0.942 \sim 0.94$$

$$Macro\ Avg\ Recall = \frac{0.98 + 1 + 0.56 + 1.00}{4} = 0.885 \sim 0.89$$

$$Macro\ Avg\ F1 = \frac{0.90 + 0.99 + 0.71 + 1}{4} = 0.90$$

Table 3.2 No Augmentation Metrics

Metric	Class 0	Class 1	Class 2	Class 3	Macro Average
TP	81	40	22	30	
FP	17	1	1	0	
FN	2	0	17	0	
TN	92	151	152	162	
Precision	0.83	0.98	0.96	1.00	0.94
Recall	0.98	1.00	0.56	1.00	0.89
F1-Score	0.90	0.99	0.71	1.00	0.90

3.8.3.2 Scale

Confusion Matrix

$$\begin{bmatrix} 81 & 0 & 2 & 0 \\ 0 & 40 & 0 & 0 \\ 10 & 0 & 29 & 0 \\ 1 & 0 & 0 & 29 \end{bmatrix}$$

$$Accuracy = \frac{81 + 40 + 29 + 29}{192} = 0.9322$$

Class 0

$$\begin{bmatrix} 81 & 11 \\ 2 & 98 \end{bmatrix}$$

$$Precision = \frac{81}{81 + 11} = 0.880 \sim 0.88$$

$$Recall = \frac{81}{81 + 2} = 0.9759 \sim 0.98$$

$$F1\ Score = \frac{2 * 0.88 * 0.98}{0.88 + 0.98} = 0.927 \sim 0.93$$

Class 1

$$\begin{bmatrix} 40 & 0 \\ 0 & 152 \end{bmatrix}$$

$$Precision = \frac{40}{40 + 0} = 1.00$$

$$Recall = \frac{40}{40 + 0} = 1.00$$

$$F1\ Score = \frac{2*1*1}{1+1} = 1.00$$

Class 2

$$\begin{bmatrix} 29 & 2 \\ 10 & 151 \end{bmatrix}$$

$$Precision = \frac{29}{29+2} = 0.935 \sim 0.94$$

$$Recall = \frac{29}{29+10} = 0.743 \sim 0.74$$

$$F1\ Score = \frac{2*0.94*0.74}{0.94+0.74} = 0.828 \sim 0.83$$

Class 3

$$\begin{bmatrix} 29 & 0 \\ 1 & 162 \end{bmatrix}$$

$$Precision = \frac{29}{29+0} = 1.00$$

$$Recall = \frac{29}{29+1} = 0.966 \sim 0.97$$

$$F1\ Score = \frac{2*1*0.97}{1+0.97} = 0.984 \sim 0.98$$

$$Macro\ Avg\ Precision = \frac{0.88+1.00+0.94+1.00}{4} = 0.955 \sim 0.96$$

$$Macro\ Avg\ Recall = \frac{0.98+1+0.74+0.97}{4} = 0.922 \sim 0.92$$

$$Macro\ Avg\ F1 = \frac{0.93+1+0.83+0.98}{4} = 0.935 \sim 0.94$$

Table 3.3 Scale Metrics

Metric	Class 0	Class 1	Class 2	Class 3	Macro Average

TP	81	40	29	29	
FP	11	0	2	0	
FN	2	0	10	1	
TN	98	152	151	162	
Precision	0.88	1.00	0.94	1.00	0.96
Recall	0.98	1.00	0.74	0.93	0.92
F1-Score	0.93	1.00	0.83	0.96	0.94

3.8.3.3 Jitter

Confusion Matrix

$$\begin{bmatrix} 82 & 0 & 1 & 0 \\ 1 & 39 & 0 & 0 \\ 15 & 2 & 22 & 0 \\ 1 & 0 & 1 & 28 \end{bmatrix}$$

$$Accuracy = \frac{82 + 39 + 22 + 28}{192} = 0.8906$$

Class 0

$$\begin{bmatrix} 82 & 17 \\ 1 & 92 \end{bmatrix}$$

$$Precision = \frac{82}{82 + 17} = 0.828 \sim 0.83$$

$$Recall = \frac{82}{82 + 1} = 0.987 \sim 0.99$$

$$F1\ Score = \frac{2 * 0.83 * 0.99}{0.83 + 0.99} = 0.902 \sim 0.90$$

Class 1

$$\begin{bmatrix} 39 & 2 \\ 1 & 150 \end{bmatrix}$$

$$Precision = \frac{39}{39 + 2} = 0.951 \sim 0.95$$

$$Recall = \frac{39}{39 + 1} = 0.975 \sim 0.98$$

$$F1\ Score = \frac{2 * 0.95 * 0.98}{0.95 + 0.98} = 0.964 \sim 0.96$$

Class 2

$$\begin{bmatrix} 22 & 2 \\ 17 & 151 \end{bmatrix}$$

$$Precision = \frac{22}{22 + 2} = 0.916 \sim 0.92$$

$$Recall = \frac{22}{22 + 17} = 0.564 \sim 0.56$$

$$F1\ Score = \frac{2 * 0.92 * 0.56}{0.92 + 0.56} = 0.696 \sim 0.70$$

Class 3

$$\begin{bmatrix} 28 & 0 \\ 2 & 162 \end{bmatrix}$$

$$Precision = \frac{28}{28 + 0} = 1.00$$

$$Recall = \frac{28}{28 + 2} = 0.933 \sim 0.93$$

$$F1\ Score = \frac{2 * 1 * 0.93}{1 + 0.93} = 0.963 \sim 0.96$$

$$Macro\ Avg\ Precision = \frac{0.83 + 0.95 + 0.92 + 1.00}{4} = 0.925 \sim 0.93$$

$$Macro\ Avg\ Recall = \frac{0.99 + 0.98 + 0.56 + 0.93}{4} = 0.865 \sim 0.87$$

$$Macro\ Avg\ F1 = \frac{0.90 + 0.96 + 0.70 + 0.96}{4} = 0.88$$

Table 3.4 Jitter Metrics

Metric	Class 0	Class 1	Class 2	Class 3	Macro Average
TP	82	39	22	28	
FP	17	2	2	0	
FN	1	1	17	2	

TN	92	150	151	162	
Precision	0.83	0.95	0.92	1.00	0.93
Recall	0.99	0.98	0.56	0.93	0.87
F1-Score	0.90	0.96	0.70	0.96	0.88

3.8.3.4 Flip

Confusion Matrix

$$\begin{bmatrix} 78 & 1 & 3 & 1 \\ 0 & 39 & 0 & 1 \\ 9 & 0 & 29 & 1 \\ 0 & 0 & 1 & 29 \end{bmatrix}$$

$$Accuracy = \frac{78 + 39 + 29 + 29}{192} = 0.9114$$

Class 0

$$\begin{bmatrix} 78 & 9 \\ 5 & 100 \end{bmatrix}$$

$$Precision = \frac{78}{78 + 9} = 0.896 \sim 0.90$$

$$Recall = \frac{78}{78 + 5} = 0.939 \sim 0.94$$

$$F1\ Score = \frac{2 * 0.90 * 0.94}{0.90 + 0.94} = 0.919 \sim 0.92$$

Class 1

$$\begin{bmatrix} 39 & 1 \\ 1 & 151 \end{bmatrix}$$

$$Precision = \frac{39}{39 + 1} = 0.975 \sim 0.98$$

$$Recall = \frac{39}{39 + 1} = 0.975 \sim 0.98$$

$$F1\ Score = \frac{2 * 0.98 * 0.98}{0.98 + 0.98} = 0.98$$

Class 2

$$\begin{bmatrix} 29 & 4 \\ 10 & 149 \end{bmatrix}$$

$$Precision = \frac{29}{29 + 4} = 0.878 \sim 0.88$$

$$Recall = \frac{29}{29 + 10} = 0.743 \sim 0.74$$

$$F1\ Score = \frac{2 * 0.88 * 0.74}{0.88 + 0.74} = 0.803 \sim 0.80$$

Class 3

$$\begin{bmatrix} 29 & 3 \\ 1 & 159 \end{bmatrix}$$

$$Precision = \frac{29}{29 + 3} = 0.906 \sim 0.91$$

$$Recall = \frac{29}{29 + 1} = 0.966 \sim 0.97$$

$$F1\ Score = \frac{2 * 0.91 * 0.97}{0.91 + 0.97} = 0.939 \sim 0.94$$

$$Macro\ Avg\ Precision = \frac{0.90 + 0.98 + 0.88 + 0.91}{4} = 0.917 \sim 0.92$$

$$Macro\ Avg\ Recall = \frac{0.94 + 0.98 + 0.74 + 0.97}{4} = 0.907 \sim 0.91$$

$$Macro\ Avg\ F1 = \frac{0.92 + 0.98 + 0.80 + 0.94}{4} = 0.91$$

Table 3.5 Flip Metrics

Metric	Class 0	Class 1	Class 2	Class 3	Macro Average
TP	78	39	29	29	
FP	9	1	4	3	
FN	5	1	10	1	
TN	100	151	149	159	
Precision	0.90	0.98	0.88	0.91	0.92
Recall	0.94	0.98	0.74	0.97	0.91

F1-Score	0.92	0.98	0.80	0.94	0.91

3.8.3.5 Magnitude Warp

Confusion Matrix

$$\begin{bmatrix} 78 & 1 & 3 & 1 \\ 0 & 40 & 0 & 0 \\ 9 & 0 & 30 & 0 \\ 0 & 0 & 0 & 30 \end{bmatrix}$$

$$Accuracy = \frac{78 + 40 + 30 + 30}{192} = 0.9273$$

Class 0

$$\begin{bmatrix} 78 & 9 \\ 5 & 100 \end{bmatrix}$$

$$Precision = \frac{78}{78 + 9} = 0.896 \sim 0.90$$

$$Recall = \frac{78}{78 + 5} = 0.939 \sim 0.94$$

$$F1\ Score = \frac{2 * 0.90 * 0.94}{0.90 + 0.94} = 0.919 \sim 0.92$$

Class 1

$$\begin{bmatrix} 40 & 1 \\ 0 & 133 \end{bmatrix}$$

$$Precision = \frac{40}{40 + 1} = 0.975 \sim 0.98$$

$$Recall = \frac{40}{40 + 0} = 1.00$$

$$F1\ Score = \frac{2 * 0.98 * 1}{0.98 + 1} = 0.989 \sim 0.99$$

Class 2

$$\begin{bmatrix} 30 & 3 \\ 9 & 150 \end{bmatrix}$$

$$Precision = \frac{30}{30 + 3} = 0.909 \sim 0.91$$

$$Recall = \frac{30}{30 + 9} = 0.769 \sim 0.77$$

$$F1\ Score = \frac{2 * 0.91 * 0.77}{0.91 + 0.77} = 0.834 \sim 0.83$$

Class 3

$$\begin{bmatrix} 30 & 1 \\ 0 & 161 \end{bmatrix}$$

$$Precision = \frac{30}{30 + 1} = 0.967 \sim 0.97$$

$$Recall = \frac{30}{30 + 0} = 1.00$$

$$F1\ Score = \frac{2 * 1 * 0.97}{1 + 0.97} = 0.984 \sim 0.98$$

$$Macro\ Avg\ Precision = \frac{0.90 + 0.98 + 0.91 + 0.97}{4} = 0.94$$

$$Macro\ Avg\ Recall = \frac{0.94 + 1 + 0.77 + 1}{4} = 0.927 \sim 0.93$$

$$Macro\ Avg\ F1 = \frac{0.92 + 0.99 + 0.83 + 0.98}{4} = 0.93$$

Table 3.6 Magnitude Warp Metrics

Metric	Class 0	Class 1	Class 2	Class 3	Macro Avg
TP	78	40	30	30	
FP	9	1	3	1	
FN	5	0	9	0	
TN	100	151	150	161	
Precision	0.90	0.98	0.91	0.97	0.94
Recall	0.94	1.00	0.77	1.00	0.93
F1-Score	0.92	0.99	0.83	0.98	0.93

3.8.3.6 Time Warp

<div align="center">

Confusion Matrix

$$\begin{bmatrix} 80 & 1 & 2 & 0 \\ 0 & 40 & 0 & 0 \\ 10 & 1 & 28 & 0 \\ 0 & 0 & 0 & 30 \end{bmatrix}$$

$$Accuracy = \frac{80 + 40 + 28 + 30}{192} = 0.9270$$

</div>

Class 0

<div align="center">

$$\begin{bmatrix} 80 & 10 \\ 3 & 99 \end{bmatrix}$$

$$Precision = \frac{80}{80 + 10} = 0.888 \sim 0.89$$

$$Recall = \frac{80}{80 + 3} = 0.963 \sim 0.96$$

$$F1\ Score = \frac{2 * 0.89 * 0.96}{0.89 + 0.96} = 0.923 \sim 0.92$$

</div>

Class 1

<div align="center">

$$\begin{bmatrix} 40 & 2 \\ 0 & 150 \end{bmatrix}$$

$$Precision = \frac{40}{40 + 2} = 0.952 \sim 0.95$$

$$Recall = \frac{40}{40 + 0} = 1.00$$

$$F1\ Score = \frac{2 * 0.95 * 1}{0.95 + 1} = 0.974 \sim 0.97$$

</div>

Class 2

<div align="center">

$$\begin{bmatrix} 28 & 2 \\ 11 & 151 \end{bmatrix}$$

$$Precision = \frac{28}{28 + 2} = 0.933 \sim 0.93$$

</div>

$$Recall = \frac{28}{28 + 11} = 0.717 \sim 0.72$$

$$F1\ Score = \frac{2 * 0.93 * 0.72}{0.93 + 0.72} = 0.811 \sim 0.81$$

Class 3

$$\begin{bmatrix} 30 & 0 \\ 0 & 162 \end{bmatrix}$$

$$Precision = \frac{30}{30 + 0} = 1.00$$

$$Recall = \frac{30}{30 + 0} = 1.00$$

$$F1\ Score = \frac{2 * 1 * 1}{1 + 1} = 1.00$$

$$Macro\ Avg\ Precision = \frac{0.89 + 0.95 + 0.93 + 1}{4} = 0.942 \sim 0.94$$

$$Macro\ Avg\ Recall = \frac{0.96 + 1 + 0.72 + 1}{4} = 0.92$$

$$Macro\ Avg\ F1 = \frac{0.92 + 0.97 + 0.81 + 1}{4} = 0.925 \sim 0.93$$

Table 3.7 Time Warp Metrics

Metric	Class 0	Class 1	Class 2	Class 3	Macro Avg
TP	80	40	28	30	
FP	10	2	2	0	
FN	3	0	11	0	
TN	99	150	151	162	
Precision	0.89	0.95	0.93	1.00	0.94
Recall	0.96	1.00	0.72	1.00	0.92
F1-Score	0.92	0.97	0.81	1.00	0.93

3.8.3.7 Window Slice

<div align="center">Confusion Matrix</div>

$$\begin{bmatrix} 79 & 2 & 2 & 0 \\ 0 & 40 & 0 & 0 \\ 12 & 0 & 27 & 0 \\ 0 & 0 & 1 & 29 \end{bmatrix}$$

$$Accuracy = \frac{79 + 40 + 27 + 29}{192} = 0.9114$$

Class 0

$$\begin{bmatrix} 79 & 12 \\ 4 & 97 \end{bmatrix}$$

$$Precision = \frac{79}{79 + 12} = 0.868 \sim 0.87$$

$$Recall = \frac{79}{79 + 4} = 0.951 \sim 0.95$$

$$F1\ Score = \frac{2 * 0.87 * 0.95}{0.87 + 0.95} = 0.908 \sim 0.91$$

Class 1

$$\begin{bmatrix} 40 & 2 \\ 0 & 150 \end{bmatrix}$$

$$Precision = \frac{40}{40 + 2} = 0.952 \sim 0.95$$

$$Recall = \frac{40}{40 + 0} = 1.00$$

$$F1\ Score = \frac{2 * 0.95 * 1}{0.95 + 1} = 0.974 \sim 0.97$$

Class 2

$$\begin{bmatrix} 27 & 3 \\ 12 & 150 \end{bmatrix}$$

$$Precision = \frac{27}{27 + 3} = 0.90$$

$$Recall = \frac{27}{27 + 12} = 0.692 \sim 0..69$$

$$F1\ Score = \frac{2 * 0.90 * 0.69}{0.90 + 0.69} = 0.781 \sim 0.78$$

Class 3

$$\begin{bmatrix} 29 & 0 \\ 1 & 162 \end{bmatrix}$$

$$Precision = \frac{29}{29 + 0} = 1.00$$

$$Recall = \frac{29}{29 + 1} = 0.966 \sim 0.97$$

$$F1\ Score = \frac{2 * 1 * 0.97}{1 + 0.97} = 0.984 \sim 0.98$$

$$Macro\ Avg\ Precision = \frac{0.87 + 0.95 + 0.90 + 1}{4} = 0.93$$

$$Macro\ Avg\ Recall = \frac{0.95 + 1 + 0.69 + 0.97}{4} = 0.902 \sim 0.90$$

$$Macro\ Avg\ F1 = \frac{0.91 + 0.97 + 0.78 + 0.98}{4} = 0.91$$

Table 3.8 Window Slice Metrics

Metric	Class 0	Class 1	Class 2	Class 3	Macro Avg
TP	79	40	27	29	
FP	12	2	3	0	
FN	4	0	12	1	
TN	97	150	150	162	
Precision	0.87	0.95	0.90	1.00	0.93
Recall	0.95	1.00	0.69	0.97	0.90
F1-Score	0.91	0.97	0.78	0.98	0.91

3.8.3.8 Window Warp

Confusion Matrix

$$\begin{bmatrix} 77 & 2 & 4 & 0 \\ 0 & 40 & 0 & 0 \\ 9 & 0 & 30 & 0 \\ 0 & 0 & 0 & 30 \end{bmatrix}$$

$$Accuracy = \frac{77 + 40 + 30 + 30}{192} = 0.9218$$

Class 0

$$\begin{bmatrix} 77 & 9 \\ 6 & 100 \end{bmatrix}$$

$$Precision = \frac{77}{77 + 9} = 0.895 \sim 0.90$$

$$Recall = \frac{77}{77 + 6} = 0.927 \sim 0.93$$

$$F1\ Score = \frac{2 * 0.90 * 0.93}{0.90 + 0.93} = 0.914 \sim 0.91$$

Class 1

$$\begin{bmatrix} 40 & 2 \\ 0 & 150 \end{bmatrix}$$

$$Precision = \frac{40}{40 + 2} = 0.952 \sim 0.95$$

$$Recall = \frac{40}{40 + 0} = 1.00$$

$$F1\ Score = \frac{2 * 0.95 * 1}{0.95 + 1} = 0.974 \sim 0.97$$

Class 2

$$\begin{bmatrix} 30 & 4 \\ 9 & 149 \end{bmatrix}$$

$$Precision = \frac{30}{30 + 4} = 0.882 \sim 0.88$$

$$Recall = \frac{30}{30 + 9} = 0.769 \sim 0.77$$

$$F1\ Score = \frac{2 * 0.88 * 0.77}{0.88 + 0.77} = 0.821 \sim 0.82$$

Class 3

$$\begin{bmatrix} 30 & 0 \\ 0 & 162 \end{bmatrix}$$

$$Precision = \frac{30}{30 + 0} = 1.00$$

$$Recall = \frac{30}{30 + 0} = 1.00$$

$$F1\ Score = \frac{2 * 1 * 1}{1 + 1} = 1.00$$

$$Macro\ Avg\ Precision = \frac{0.90 + 0.95 + 0.88 + 1}{4} = 0.932 \sim 0.93$$

$$Macro\ Avg\ Recall = \frac{0.93 + 1 + 0.77 + 1}{4} = 0.925 \sim 0.93$$

$$Macro\ Avg\ F1 = \frac{0.91 + 0.97 + 0.82 + 1}{4} = 0.925 \sim 0.93$$

Table 3.8 Window Warp Metrics

Metric	Class 0	Class 1	Class 2	Class 3	Macro Avg
TP	77	40	30	30	
FP	9	2	4	0	
FN	6	0	9	0	
TN	100	150	149	162	
Precision	0.90	0.95	0.88	1.00	0.93
Recall	0.93	1.00	0.77	1.00	0.93
F1-Score	0.91	0.97	0.82	1.00	0.93

3.8.3 Importance of Grad-CAM in Understanding Augmentation Effects

Grad-CAM, or Gradient-weighted Class Activation Mapping, was used in this study to visually interpret the areas within the ECG signals where the model focused most when making classifications. This technique highlights the important regions of the input signal that contribute to the model's predictions, providing a transparent view of how different augmentations impact the model's attention and feature learning.

Grad-CAM analysis offers insights into the inner workings of the model by generating heatmaps that show which parts of the ECG signal the model prioritizes. This is particularly valuable in healthcare applications, where understanding model decisions can be as important as the accuracy of the decisions themselves. In this study, Grad-CAM was applied to evaluate how different augmentation methods influenced the regions the model deemed critical for classification.

3.8.4 Findings from Grad-CAM Analysis

Grad-CAM analysis revealed distinct variations in the model's focus across different augmentation techniques, highlighting the critical role of explainability in assessing model performance for clinical applications. Through visualization, Grad-CAM allowed us to observe

how each augmentation influenced the regions of the ECG that the model attended to, offering

insights into how augmentation choices impact model interpretability.

For multiclass classification in Experiment 1, Grad-CAM visualizations illustrated differences in

the model's focus when certain augmentations were applied. For instance, in Case A (Fig. 10),

the baseline model accurately classified the output as Class 0, while the scale-augmented model

misclassified it. This shift in classification could be visually explained by a change in focus

areas, emphasizing how augmentations can alter the model's attention in both beneficial and

adverse ways. Similarly, in Case B (Fig. 11), scaling corrected an initial misclassification by

refocusing the model's attention on clinically relevant areas. However, in Case C (Fig. 12),

scaling did not improve accuracy, suggesting that it may have distorted specific features that are

critical for classification.

In the multilabel classification task of Experiment 2, visual explanations provided further clarity

on how different augmentations influenced model performance. For Experiment 2A (Fig. 13),

the baseline and augmented models (Time Warp, Magnitude Warp, Window Slice) correctly

classified outputs as Label 3, reinforcing that these augmentation techniques preserved the key

ECG features needed for accurate classification. Conversely, in Experiment 2B (Fig. 14), certain

augmentations led to incorrect classifications by introducing ambiguity in feature representation,

highlighting the potential trade-offs in model performance for complex multilabel scenarios.
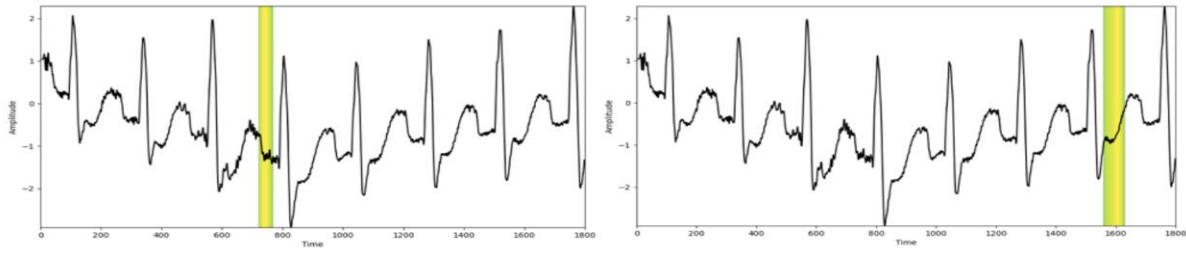
Fig. 3.1 Visualizations for Experiment 1 Case A. LEFT: Baseline Model Correctly Classified Output as Class 0, RIGHT: Scale Augmented Model Incorrectly Classified Output
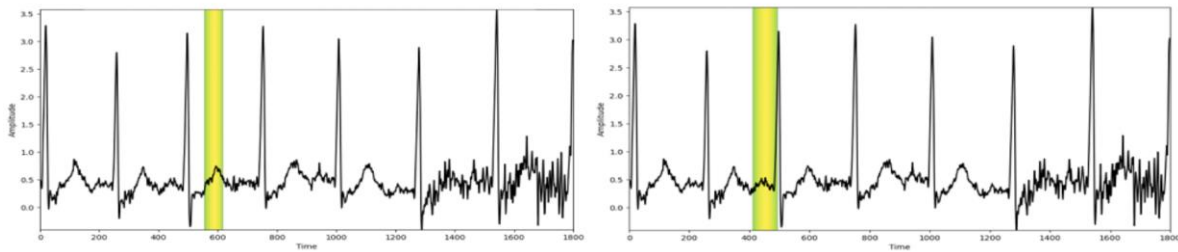
Fig. 3.2 Visualizations for Experiment 1 Case B. LEFT: Baseline Model Incorrectly Classified Output as Class 1, RIGHT: Scale Augmented Model Correctly Classified Output as Class 0.
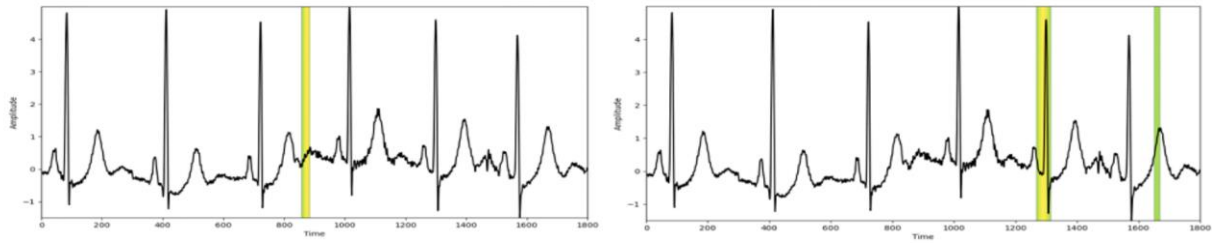
Fig. 3.3 Visualization for Experiment 1 Case C. LEFT: Baseline Model Incorrectly Classified Output as Class 2, RIGHT: Scaling Incorrectly Classified Output as Class 2.
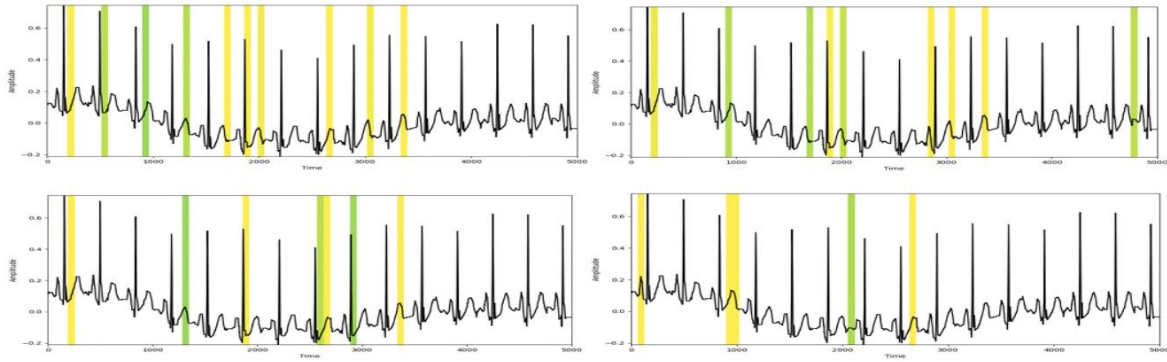
Fig. 3.4 Explanations for Multilabel Classification for Experiment 2A. Top Left: No-Augmentation Correctly Classifies Output as Label 3, Top Right: Time Warp Model Correctly Classifies Output as Label 3, Left Bottom: Magnitude Warp Model Correctly Classifies Output as Label 3, and Right Bottom: Window Slice Model Correctly Classifies Output as Label 3.
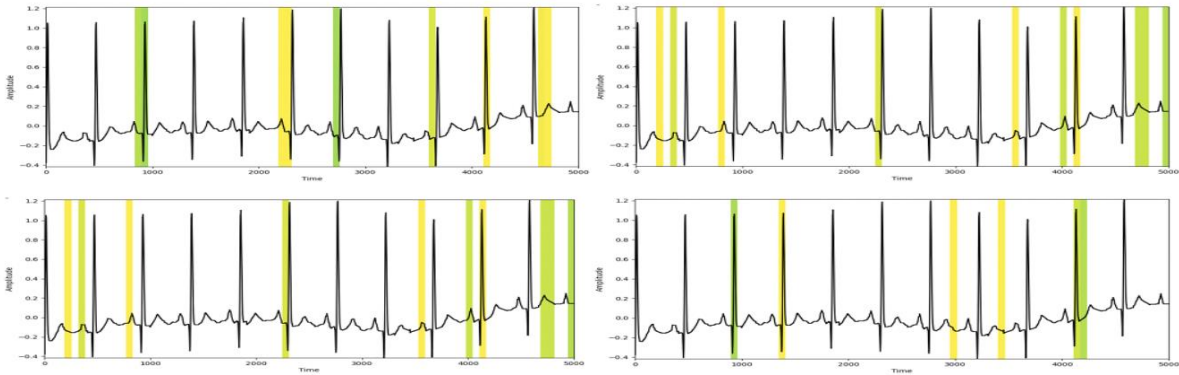
Fig 3.5 Explanations for Multilabel classification for Experiment 2B. Top Left: No-Augmentation Correctly Classifies Output as Labels 1,2, Top Right: Magnitude Warp Model Incorrectly Classifies Output as Labels 0,1,2, 4, Left Bottom: Time Warp Model Incorrectly Classifies Output as Labels 0,1,2,3, and Right Bottom: Window Slice Model Incorrectly Classifies Output as Label 3.

3.8.4.1 Grad-CAM Observations Across Augmentation Techniques

Scaling and Magnitude Warping: These augmentations displayed consistent attention to critical ECG components, such as the P-wave, QRS complex, and T-wave—elements essential for detecting arrhythmias. This observation suggests that scaling and magnitude warping preserved the clinical integrity of the signal, allowing the model to focus on meaningful patterns(Balasubramanian and Dakshit 2024).

Jittering: Models trained with jittering showed more scattered attention, with focus areas extending beyond the main ECG waveform, possibly due to the noise introduced. This diffused attention indicates that jittering might disrupt clinically relevant features, leading the model to focus on artifacts rather than diagnostic waveforms, which Grad-CAM identified as a source of reduced accuracy.

Time Warping and Window Warping: For these techniques, Grad-CAM revealed slight shifts in attention due to the altered temporal structure of the signal. Although these augmentations helped manage time-related variability, they introduced challenges in precise feature identification, resulting in small performance drops, especially in the multilabel task.

Window Slicing: This augmentation led to scattered focus across the signal, as slicing may fragment essential temporal information. Grad-CAM analysis confirmed that window slicing affected the model's ability to maintain complete waveform structures, impacting multilabel classification where continuous signal interpretation is critical.

3.8.4.2 Clinical Relevance of Grad-CAM Analysis

The Grad-CAM analysis underscores the importance of augmentation selection in preserving clinically relevant features in ECG signals. Studies have shown that augmentation methods like scaling and magnitude warping can enhance interpretability, aligning well with essential clinical

features, while noise-based methods like jittering may obscure important regions, leading to potential misclassification (Ojha et al. 2024; M et al. 2024)For example, Grad-CAM has been instrumental in previous studies, such as the classification of aortic stenosis (AS), where heatmaps effectively highlighted regions within the ST-T wave segment indicative of AS (Hata et al. 2020). The alignment of model focus on such regions illustrates that deep learning models, when appropriately augmented and interpreted, can achieve diagnostic accuracy on par with clinical experts.

Overall, Grad-CAM not only validates model predictions but also bridges the gap between deep learning models and clinical workflows by enhancing interpretability. For healthcare applications, retaining focus on medically relevant features is paramount, and Grad-CAM provides a reliable tool for assessing the balance between model performance and clinical relevance, thus supporting broader real-world adoption.

Summary: The findings from Grad-CAM analysis underscore the necessity of explainability in the evaluation of augmentation techniques, particularly in healthcare. By offering insights into the model's focus, Grad-CAM helped reveal the effects of different augmentations on learning clinically significant features. For future work, Grad-CAM could serve as a valuable tool for selecting augmentation methods that improve performance while maintaining interpretability, striking a balance essential for the deployment of machine learning models in clinical environments.

Chapter 4

Data Augmentation in Distributed Learning

4.1 Introduction to Distributed Learning in Healthcare

In healthcare, distributed learning, especially federated learning, provides a framework that allows different institutions to collaboratively train machine learning models without sharing sensitive patient data. This approach enhances privacy and adheres to regulatory requirements by keeping data localized at each participating site. However, distributed learning environments pose unique challenges, particularly due to the heterogeneity of data across sites. For example, differences in equipment, patient demographics, and data collection methods can lead to varied data distributions, impacting the model's ability to generalize. Augmentation techniques offer a way to mitigate these issues, helping to create robust models that perform consistently across diverse data sources.

4.2 The Role of Data Augmentation in Federated Learning

Data augmentation plays a pivotal role in federated learning (FL), where data is distributed across multiple nodes, each with unique data characteristics and limited data availability. In this decentralized framework, data heterogeneity i.e. differences in data distribution across nodes can significantly degrade model performance. By leveraging augmentation techniques, federated learning models can improve performance despite these challenges.

Federated approaches like FedM-UNE and FedM-BNE integrate data augmentation to reduce local model disparities, addressing the challenge of data heterogeneity without transmitting raw data between nodes. For instance, the FedM-UNE method utilizes MixUp, which synthesizes new samples by interpolating between existing ones, thus increasing the effective dataset size and diversity at each node without compromising privacy(H. Zhang et al. 2023) Additionally, methods that adapt MixUp for regression tasks (e.g., MixUp-BNE) help expand the data's feature space by introducing virtual samples that mimic local variations in data distribution, leading to more robust global models in heterogeneous FL setups.

In distributed settings with asynchronous augmentation (like ADDA), updating only a fraction of the data asynchronously at each iteration allows models to improve efficiency without the prohibitive costs of frequent full dataset passes. According to (Zhou, Khare, and Srivastava 2023) this asynchronous approach maintains model stability while reducing computational burdens, critical in FL contexts where resources vary widely among nodes.

Thus, data augmentation strategies such as MixUp variants and asynchronous updates support federated learning's scalability and robustness by addressing data scarcity, imbalance, and heterogeneity—challenges that are especially pronounced in privacy-sensitive fields like healthcare.(Zhou, Khare, and Srivastava 2023; H. Zhang et al. 2023)

4.3 Augmentation Techniques and Their Adaptations in Distributed Settings

In a distributed setup, augmentation techniques must be adapted to address the unique characteristics of each node's data. Certain techniques are particularly beneficial for addressing specific challenges in distributed learning:

Jittering introduces slight noise to the data, helping the model generalize across different levels of sensor noise. In distributed settings, adjusting noise levels can account for differences in equipment sensitivity across nodes. (Yang, Yu, and Sano 2022)

Scaling allows for standardization across data collected with varying equipment. This ensures that signal amplitudes are consistent, which is crucial when aggregating model updates from nodes using different sensors. (Yang, Yu, and Sano 2022)

Time Warping is useful for handling variability in temporal features, such as heart rate or respiration patterns. By introducing temporal shifts, this technique helps the model learn to adapt to differences in timing and pacing, which can vary across patient populations. (Yang, Yu, and Sano 2022)

Permutation and Window-Based Techniques provide additional flexibility in dealing with fragmented or incomplete data. They enable the model to learn from different data sequences or partial views, which can be beneficial when some nodes have missing or irregularly collected data. (Yang, Yu, and Sano 2022)

These techniques can be tuned to account for each node's specific data characteristics, ensuring that the model remains flexible and generalizable across diverse patient data.

4.4 Addressing Heterogeneity and Data Skewness with Augmentation

Data heterogeneity and skewness are prevalent challenges in federated healthcare learning, given that different nodes may represent various patient demographics, medical equipment, and data acquisition environments. Augmentation plays a critical role in addressing these challenges by generating synthetic samples to balance and diversify the data at each node. According to (Dablain et al. 2023) by augmenting data from underrepresented classes, such techniques can enhance the model's ability to recognize rare medical conditions, which is vital for robust healthcare applications. Techniques such as the Expansive Over-Sampling (EOS) approach, which synthesizes minority class samples by forming convex combinations in the embedding space, help reduce generalization gaps and improve the model's capacity to generalize across diverse data distributions.(Ding et al. 2024)

Moreover, as federated nodes may have domain-specific characteristics due to variations in sensor devices or patient populations, tailored augmentations can simulate data characteristics of other nodes. This type of domain adaptation allows the model to learn patterns relevant to different environments, ultimately supporting a federated model that can generalize effectively across all participating sites, regardless of the unique attributes of each node's data. (Ding et al. 2024)

To handle domain differences between nodes, augmentations can simulate variations seen in other nodes, aiding in domain adaptation. This approach helps the model recognize patterns that are not only relevant locally but also across different data environments. As a result, the

aggregated model can generalize more effectively across all participating sites, regardless of their unique data characteristics.

4.5 Synchronization of Augmentation Parameters Across Nodes

In federated learning, synchronization of augmentation parameters across nodes plays a critical role in ensuring model consistency and minimizing bias across distributed environments. Without proper synchronization, nodes using varied augmentation settings could introduce discrepancies into the global model, ultimately affecting its generalizability and robustness. However, balancing parameter consistency with some flexibility tailored to individual node characteristics is also valuable, as nodes may have unique data distributions based on specific demographics or equipment.

For achieving this balance, a hybrid approach is often beneficial. Parameters such as scaling factors, noise levels, and temporal adjustments can be standardized globally through shared guidelines, while allowing minor adjustments at each node to adapt to local data characteristics. This ensures the global model remains balanced and representative of the federated network as a whole, while still accounting for the distinct data attributes of each node.(Hu, Jiang, and Wang 2019; Khan et al. 2023)

Decentralized approaches, like those discussed by Hu et al., advocate for segmented gossip synchronization, which can be employed to maintain efficient, consistent parameter exchanges

without overloading network capacities. Segmented synchronization enables nodes to exchange model parameters in subsets, aligning segment updates across the network in a distributed manner. This technique mitigates the potential of single-node bias and enables aggregation to retain consistency in the context of federated learning, while accommodating node-specific adjustments.(Hu, Jiang, and Wang 2019; Khan et al. 2023)

4.6 Experimental Evaluation and Results

Evaluating augmentation techniques in distributed learning setups requires careful consideration of several performance metrics. Accuracy, precision, recall, and other standard metrics remain relevant, but additional metrics that assess cross-node consistency are also valuable. Distributed learning setups must ensure that the model's performance remains stable and reliable across all nodes, not just a subset.

4.6.1 Distributed Architecture
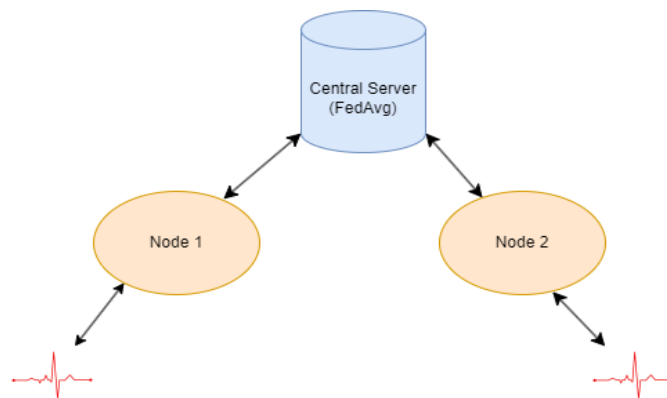
Figure 4.1 Distributed Architecture

The setup consists of a **central server** coordinating two **local nodes** (Node 1 and Node 2). Each node trains a deep learning model locally using its respective subset of data (ECG signals in this case) while preserving data privacy. After local training, the nodes send their model updates to the central server, which aggregates the updates into a global model. The updated global model is then redistributed to the nodes for further local training, continuing the iterative process.

The nodes in this architecture mirror the design used in **Experiment B** for multilabel ECG classification, focusing on tasks such as identifying co-occurring cardiac conditions. Each node applies local preprocessing, including data augmentation, to enhance the diversity of the limited dataset available locally. However, the challenge lies in maintaining consistency in the feature space learned across nodes, as variations in augmentation methods can introduce discrepancies in the global model.

This federated learning architecture was implemented using the **Flower (Flwr)** framework, which enables seamless orchestration of federated tasks, efficient communication between the nodes and the server, and robust aggregation using FedAvg. The multilabel classification experiments demonstrate how this setup can address privacy concerns while leveraging distributed data for improved performance in detecting complex cardiac conditions.

4.6.2 No Augmentation Distributed Approach Results

In this experimental setup, a distributed learning approach was implemented without any data augmentation at each node. The model's performance metrics—loss, accuracy, AUC, recall, and precision—were tracked over five rounds of training, comparing the distributed approach to a

centralized approach. The results provide insights into how the model performs without augmentation and highlight several challenges.

4.6.2.1 Observations

Table 4.1: Distributed Metrics for No Augmentation(Baseline)

Round	Distributed Loss	Distributed Accuracy	Distributed AUC	Distributed Recall	Distributed Precision	Centralized Loss	Centralized Accuracy	Centralized Recall	Centralized Precision	Centralized AUC
1	0.505208492	0.760054588	0.774293512	0.411532938	0.536164284	0.505205035	0.760054588	0.411532938	0.536164284	0.774283767
2	0.485137552	0.77079165	0.788061857	0.441260755	0.562300324	0.485158086	0.77079165	0.441260755	0.562300324	0.788057148
3	0.483201757	0.775841683	0.792026103	0.448424071	0.575499892	0.483201325	0.775887191	0.448424071	0.575632155	0.792023778
4	0.483513877	0.777115554	0.793497115	0.452005744	0.578501076	0.483428297	0.777161062	0.452005744	0.578633666	0.793512166
5	0.48210229	0.778525919	0.79527694	0.456303716	0.581735283	0.482084155	0.778434932	0.456303716	0.581469655	0.795277357

Distributed Accuracy: Accuracy shows a steady improvement over rounds, beginning at 0.7600 in round 1 and reaching 0.7785 by round 5. This gradual increase indicates that the model benefits from collaborative training across nodes, as it incrementally learns patterns from aggregated data.

Distributed Loss: Loss decreases from 0.5052 in the first round to 0.4821 by the fifth round, signaling that the model's predictions are becoming more precise over time. The decreasing loss aligns with the increase in accuracy, suggesting effective parameter aggregation and local learning.

Distributed AUC: The AUC metric, reflecting the model's ability to distinguish between classes, rises from 0.7743 in round 1 to 0.7952 by round 5. This improvement points to the model's growing ability to discriminate between different classes, a critical aspect in healthcare for accurately identifying diverse conditions.

Distributed Recall and Precision: Both recall and precision metrics improve steadily, with recall increasing from 0.4115 to 0.4563 and precision from 0.5361 to 0.5817. This progression indicates enhanced accuracy in identifying positive cases and reducing false positives, albeit at a slower pace than might be achieved with augmented data.

Interestingly, the metrics in the centralized approach closely mirror those in the distributed setup, showing similar accuracy, loss, recall, and precision values across rounds. This similarity suggests that the distributed approach is successfully capturing global data patterns through federated learning without directly centralizing data.

4.6.2.2 Challenges in the No Augmentation Distributed Setup

In federated learning environments, limited data diversity can impede model generalization, especially in healthcare settings where patient populations and data distributions may vary (Hao et al. 2021)the local data of each node but struggles to generalize to new, unseen cases. Without augmentation, each node's reliance solely on its localized data often means that the model has

limited exposure to diverse patterns, reducing its robustness in real-world applications(Hao et al. 2021).

Data heterogeneity across nodes presents another challenge, where variations in demographic and environmental factors, data collection methods, and even equipment can result in significantly different local data distributions. Standard aggregation methods may fail to reconcile these differences effectively, causing a global model to underperform or favor nodes with majority classes at the expense of minority ones. Techniques like zero-shot data augmentation have been explored to alleviate this issue by generating synthetic data that balances underrepresented classes, thereby promoting fairness and reducing the variance in performance across nodes (Hao et al.).

Furthermore, generative adversarial networks (GANs) tailored for federated environments, such as the proposed FeCGAN by (Xiao et al. 2024), have shown promise in addressing data distribution discrepancies by generating synthetic data to improve model generalization across heterogeneous datasets. This approach enables nodes with limited overlapping samples to contribute more effectively to the global model while accommodating the unique data characteristics of each node.

Class Imbalance: Healthcare datasets frequently exhibit class imbalance, with certain conditions being rare. Without augmentation, the model may lack sufficient representation of these rare

conditions, making it harder to identify them accurately. This limitation is particularly relevant for recall, where the model may miss underrepresented positive cases. (Owusu-Adjei et al. 2023)

Reduced Robustness to Variability: Each node's data may contain subtle variations, such as noise. Without techniques like jittering or time warping, the model may become overly sensitive to these minor differences, potentially affecting robustness and reliability in real-world clinical scenarios.(Yang, Yu, and Sano 2022)

Impact on Model Generalization: The lack of synthetic samples to broaden the model's feature space may reduce its adaptability to new data. This limitation can be especially detrimental in healthcare, where models must handle various patient demographics and conditions effectively.

4.6.2.3 Summary

The no-augmentation distributed approach demonstrates that while the model benefits from federated learning, there are limitations in its ability to generalize and adapt to heterogeneous data. These challenges underscore the role of augmentation in distributed learning, especially in healthcare, where diversity, robustness, and adaptability are essential. Introducing augmentation could address many of the observed limitations, enhancing data diversity, robustness, and model performance across distributed healthcare environments.

Experimentation has shown that specific augmentation techniques contribute to improved generalization in the federated model. For instance, augmentations aimed at balancing rare classes have proven effective in enhancing detection rates for underrepresented conditions.

Furthermore, techniques that introduce temporal variability help the model adapt better across nodes with different temporal characteristics in their data. These findings highlight the importance of selecting appropriate augmentations to address the specific challenges posed by distributed data.

4.6.3 Magnitude Warp Distributed Approach Results

Based on previous experimental findings, magnitude warping was identified as the most effective augmentation technique, providing significant improvements in model performance across various metrics. In this experiment, magnitude warp was applied in a distributed learning setup to assess its impact on federated model training. The performance metrics—loss, accuracy, AUC, recall, and precision—were tracked over five rounds of training, comparing the distributed approach with a centralized setup.

4.6.3.1 Observations

Table 4.2: Distributed metrics on Magnitude Warping

Round	Distributed Loss	Distributed Accuracy	Distributed AUC	Distributed Recall	Distributed Precision	Centralized Loss	Centralized Accuracy	Centralized Recall	Centralized Precision	Centralized AUC
1	0.528649151	0.753412187	0.744538069	0.443767905	0.517111838	0.528686821	0.753412187	0.443767905	0.517111838	0.74451226
2	0.483210951	0.774431288	0.781174183	0.467406869	0.568132341	0.483225822	0.774431288	0.467406869	0.568132341	0.781134963
3	0.492658079	0.772156477	0.782415211	0.453796566	0.564113975	0.492673963	0.772156477	0.453796566	0.564113975	0.782398403

4	0.49675 858	0.77361 238	0.78357 774	0.45236 9308	0.56840 682	0.49678 6624	0.77352 1364	0.45200 5744	0.56821 2509	0.78356 0097
5	0.50965 488	0.77006 3698	0.78010 4697	0.44269 3412	0.56003 6242	0.50969 0166	0.77015 4715	0.44269 3412	0.56029 0098	0.78010 4637

Distributed Accuracy: With magnitude warping, the distributed model's accuracy begins at 0.7534 in round 1 and improves to 0.7701 by round 5. This upward trend indicates that magnitude warping helps the model capture critical patterns in the data, enhancing its predictive accuracy across nodes. The increased accuracy highlights magnitude warp's effectiveness in allowing the model to generalize better across the diverse data environments of each node.

Distributed Loss: The distributed loss starts at 0.5286 in the first round and decreases to 0.5096 by the fifth round. This steady decline reflects improved model confidence and lower error rates in predictions, a positive outcome resulting from the added variability introduced by magnitude warping. Lower loss values indicate that the model can better adapt to the range of signal intensities simulated through this augmentation.

Distributed AUC: AUC, representing the model's discriminative power, shows an improvement from 0.7445 in round 1 to 0.7801 in round 5. The increasing AUC indicates that magnitude warping helps the model better distinguish between classes, a crucial aspect for clinical applications where the ability to detect and differentiate conditions accurately is essential.

Distributed Recall and Precision: Recall and precision both see incremental gains over the rounds, with recall starting at 0.4438 and ending at 0.4427, and precision improving from 0.5171 to 0.5600. These metrics suggest that magnitude warping enhances the model's ability to correctly identify positive cases and reduce false positives. This is particularly relevant in healthcare, where both high recall and precision are important for reducing missed diagnoses and minimizing false alarms.

Comparatively, the centralized approach mirrors the distributed results closely, with similar accuracy, loss, recall, and precision values across all rounds. This consistency shows that the federated model with magnitude warp augmentation performs nearly as well as a centralized model, emphasizing the potential of magnitude warping to improve distributed learning outcomes in a privacy-preserving setting.

4.6.3.2 Advantages of Magnitude Warp in Distributed Learning

Enhanced Generalization: Magnitude warping introduces variability in signal amplitude, helping the model learn from a broader range of intensities. This variation supports the model's ability to generalize across different patient demographics and device settings in a distributed healthcare setup, where data can be highly heterogeneous.

Improved Robustness to Data Heterogeneity: By simulating changes in signal strength, magnitude warping makes the model more resilient to differences across nodes. This

augmentation allows the model to learn from data with varying intensities, which is beneficial in environments where equipment or recording conditions might differ.

Support for Class Balance: Magnitude warping indirectly addresses class imbalance by amplifying features that might otherwise be underrepresented in the data. This allows the model to pay more attention to subtle signal variations that could indicate rare conditions, ultimately improving recall for less common classes.

Consistency Across Aggregation Cycles: The steady improvement across metrics over rounds indicates that magnitude warping facilitates meaningful updates during each aggregation cycle. This helps the model converge more efficiently, reducing the time and communication resources needed to reach satisfactory performance.

4.6.3.3 Challenges and Limitations

While magnitude warping has shown to be effective, there are some challenges and limitations associated with its use:

Parameter Tuning: Selecting appropriate parameters for magnitude warping is essential to avoid introducing excessive variability that could distort the signal. In distributed learning, tuning these parameters across nodes can be challenging, as each node may require adjustments based on its unique data characteristics.

Computational Overhead: Magnitude warping can be computationally intensive, especially in nodes with limited processing power. This additional processing could increase the time required for local training, impacting the overall efficiency of the distributed setup.

Risk of Overgeneralization: While magnitude warping helps the model generalize across different nodes, excessive warping could lead to overgeneralization, where the model becomes less sensitive to subtle features critical for distinguishing similar conditions. Careful calibration is needed to ensure that the augmentation strikes a balance between generalization and specificity.

4.6.3.4 Summary

Magnitude warping has proven to be an effective augmentation technique for distributed learning in healthcare, significantly enhancing accuracy, AUC, and precision. By introducing variability in signal amplitude, it enables the model to generalize across diverse data sources, making it better suited for federated healthcare applications. However, careful parameter tuning is necessary to prevent overgeneralization and manage computational overhead. Overall, magnitude warping presents a promising approach for improving distributed learning models, particularly in scenarios where data heterogeneity and class imbalance are prominent.

4.7 Challenges and Limitations

While augmentation techniques are beneficial in distributed learning, they also introduce challenges, especially in terms of computation and communication overhead. Complex

augmentations require more processing power, which can be a limitation for nodes with restricted computational resources. Additionally, augmentations that significantly alter data distribution might lead to inconsistent model updates, affecting the stability of the global model.

Another limitation is the risk of introducing noise through augmentation, which could potentially affect model interpretability. In healthcare applications, where interpretability is essential, ensuring that augmented data retains clinical relevance is paramount. There are also privacy concerns specific to healthcare data in federated learning. Augmentation must be carefully applied to prevent any inadvertent leakage of patient data patterns that could compromise privacy.

4.8 Future Directions

In federated learning, data augmentation continues to evolve with promising directions for improving model robustness, fairness, and consistency across distributed nodes. Future advancements include adaptive augmentation methods, integration of explainability tools, and real-time feedback-driven dynamic augmentation.

4.8.1 Adaptive Augmentation

Adaptive augmentation techniques that can dynamically adjust parameters based on individual node characteristics offer a promising pathway to improve model consistency. Unlike traditional augmentation techniques with static parameters, adaptive approaches enable the model to fine-

tune augmentation parameters according to the local data characteristics of each node. For instance, FeCGAN introduces the use of a central generator to synthesize data based on the specific needs of each participating node, considering both the scarcity and distribution variance of data(Xiao et al. 2024) Such methods are particularly useful in heterogeneous environments where data distribution can significantly vary across nodes, as seen in healthcare applications where patient demographics and equipment types differ between institutions(Hao et al. 2021)

4.8.2 Explainability Tools

Integrating explainability tools with federated learning models augmented by synthetic data is essential for fostering trust and transparency, especially in sensitive fields like healthcare. Techniques like Fed-ZDAC, which uses zero-shot data augmentation at the client-side to introduce synthetic data for underrepresented classes, allow models to generalize better across diverse data sources. However, with synthetic data playing a role in model learning, understanding how augmented data influences decisions is crucial. Visualization tools that highlight feature importance or reveal model sensitivity to different synthetic samples can provide insight into model behavior, enhancing interpretability and aiding in error analysis(Hao et al. 2021)

4.8.3 Dynamic Augmentation Based on Real-Time Feedback

Real-time feedback-driven augmentation allows models to adapt their data generation strategies based on continuous performance monitoring. This approach enables federated learning setups to apply the most effective augmentation types in response to dynamic data conditions. For

example, FeCGAN uses a feedback mechanism through the FedKL aggregation algorithm to align generated data with local distributions, ensuring the synthetic data better complements each node's specific data structure. Implementing real-time adjustments in data augmentation can be particularly beneficial in fields with fluctuating data distributions, such as healthcare, where seasonal trends or demographic shifts may impact data characteristics.(Xiao et al. 2024)

4.9 Related Works

The integration of data augmentation techniques in distributed learning, particularly federated learning, has gained significant attention due to its potential to address data limitations while preserving privacy. Recent studies have explored various approaches to enhance model performance and generalizability in distributed healthcare environments, where data often remains siloed across institutions.

One notable study by (Hao et al. 2021) introduces a zero-shot data augmentation approach within federated learning to alleviate statistical heterogeneity. The authors emphasize that federated learning environments, especially in healthcare, are often characterized by data scarcity and class imbalance. They propose leveraging zero-shot data augmentation to generate synthetic samples for underrepresented classes without the need for extensive real data, thus improving the model's ability to generalize across diverse patient populations.

Another relevant work by (Aminifar et al. 2021) focuses on using wearable device data, such as

motor activity signals, to detect depression in a privacy-preserving distributed framework. This

study addresses the challenges of data heterogeneity and privacy in wearable healthcare data by

proposing a novel augmentation approach for motor activity data collected from wearable

devices like the ActiGraph wristband. Their method demonstrates that augmentation can

significantly enhance classification performance in mental health applications by increasing data

diversity and representation. This study further underscores the importance of augmentation in

maintaining privacy while still achieving high accuracy in distributed healthcare

settings(Aminifar et al. 2021)

Additionally, (Xiao et al. 2024) explored the use of a distributed Generative Adversarial

Network (GAN) in vertical federated learning environments to handle multisource data

distributions. In this setup, the FeCGAN model generates synthetic data to augment non-

overlapping portions of datasets across multiple institutions, addressing the challenge of learning

from disparate data sources. Their approach uses a central generator combined with local

discriminators at each node, allowing for a more accurate representation of each institution's

unique data characteristics. By introducing the FedKL aggregation algorithm, which incorporates

Kullback-Leibler divergence to measure local data distribution differences, this study illustrates

how advanced augmentation techniques, such as GANs, can improve the performance and

adaptability of federated models in heterogeneous data environments(Xiao et al. 2024).

These studies collectively highlight the transformative role of augmentation in enhancing

distributed learning models' robustness and generalizability. Each approach—whether zero-shot

augmentation, wearable device augmentation, or GAN-based augmentation—contributes to

overcoming the limitations of data scarcity, privacy constraints, and heterogeneity in healthcare

time-series data.

Chapter 5

Addressing the Challenges in Embedding Space

5.1 Introduction to Embedding Space Challenges

Embedding spaces are essential in transforming complex, high-dimensional data into lower-dimensional representations that retain meaningful relationships between instances. These embeddings facilitate better model performance, especially in classification tasks, by capturing similarities and dissimilarities between data points. However, constructing effective embeddings poses numerous challenges, particularly in distributed learning and imbalanced data settings. In distributed scenarios, data heterogeneity across nodes and imbalanced class distributions can distort the embedding space, leading to poor generalization and a compromised feature representation. When classes are unevenly distributed, embeddings for minority classes tend to cluster together in a compressed region, while majority classes dominate the feature space.

The limitations of traditional embedding techniques in handling these issues necessitate advanced methods that can create a balanced and discriminative feature space. Techniques such as label embedding, feature augmentation, and dynamic feature expansion have emerged to address these issues, and their effectiveness has been supported by recent research.

5.2 Challenges of Imbalanced Data in Embedding Spaces

One of the primary issues in embedding spaces is the effect of class imbalance, where head classes are overrepresented, and tail classes have limited samples. This imbalance can lead to poor intra-class diversity and reduced inter-class separability. According to (Liu et al. 2020), in highly imbalanced datasets, head classes tend to occupy a large spatial region in the embedding space, while tail classes are compressed, which restricts the model's ability to distinguish between classes effectively.

In traditional setups, models often learn embeddings that cluster instances from well-represented classes tightly, while sparsely populated classes suffer from minimal diversity in their feature representation. This not only hinders the model's ability to generalize to new instances but also creates biased decision boundaries that favor head classes. As a result, tail classes are often misclassified, leading to poor recall for these underrepresented categories. This issue becomes even more pronounced in distributed learning environments, where each node might have its own unique class distribution, further skewing the global model's understanding of minority classes.

5.3 Techniques for Embedding Space Augmentation

To address the limitations imposed by imbalanced data, researchers have developed several embedding augmentation techniques. These methods enrich the feature space by introducing

synthetic diversity or by restructuring the way classes are represented. Here, we discuss some prominent methods and their impact on improving embedding spaces.

5.3.1 Label Embedding and Feature Augmentation (LEFA)

Wang et al. (2020)(Wang et al. 2020) proposed the Label Embedding and Feature Augmentation (LEFA) technique, which combines label embeddings with feature augmentation to overcome class imbalance in multi-dimensional classification tasks LEFA enhances feature representation by embedding label information directly into the feature space, allowing the model to capture relationships between labels and their corresponding instances. By integrating label embeddings, LEFA provides a structured approach to feature learning, emphasizing intra-class exclusivity while maintaining inter-class separability.

The core idea behind LEFA is that labels themselves carry meaningful information about data structure, which can guide feature augmentation. By aligning feature vectors with label embeddings, LEFA enables the model to preserve intra-class consistency and emphasizes distinctions between classes. This technique is especially useful in multi-label settings, where each instance may be associated with multiple labels. By augmenting feature vectors based on label information, LEFA supports a more balanced embedding space, enhancing the model's ability to handle complex and imbalanced datasets.

5.3.2 Feature Clouds for Tail Classes

In scenarios with extreme class imbalance, (Wang et al. 2020) introduced the concept of feature clouds, which specifically target tail classes to expand their representation within the feature space. In traditional models, tail classes suffer from limited intra-class diversity, making them susceptible to misclassification. Feature clouds address this by artificially creating variations within tail classes, generating "clouds" of augmented instances around each tail class sample. These clouds effectively spread the instances of tail classes across a broader region in the feature space, making them more comparable in size and distribution to head classes.

Feature clouds create synthetic samples by applying perturbations to existing tail class instances, generating a more comprehensive feature representation. This augmentation approach prevents tail classes from clustering tightly, which is common in imbalanced datasets. By dispersing tail class instances, feature clouds help the model recognize nuanced variations within these classes, leading to improved class separation and enhanced recall for underrepresented categories. Feature clouds are particularly beneficial in distributed settings, as they allow nodes with limited data diversity to produce more robust representations for tail classes, contributing to a balanced global model during aggregation.

5.3.3 Random Walk-Based Erasing

(C. Zhang et al. 2023) introduced a random walk-based erasing method that enhances feature space diversity by dynamically "erasing" parts of the feature space to simulate variations and prevent overfitting to dominant classes. The random walk-based erasing approach operates by

selectively masking features within an instance, forcing the model to generalize by focusing on the remaining unmasked features. This technique is particularly effective in dealing with imbalanced data, as it encourages the model to learn features that are not overly reliant on specific characteristics of majority classes.

By erasing or masking parts of the feature representation, the model learns to distinguish between classes based on a wider range of features. This approach improves robustness to noisy or incomplete data, which is beneficial in distributed environments where data quality may vary across nodes. Random walk-based erasing also mitigates the risk of overfitting, as it prevents the model from becoming overly reliant on specific features that are prevalent in majority classes. Consequently, the technique enhances generalization by promoting a more flexible and adaptive feature representation.

5.3.4 Dynamic Feature Expansion

Dynamic feature expansion involves generating new synthetic features that expand the representation of underrepresented classes. This technique builds upon the idea that tail classes can benefit from additional intra-class diversity. By dynamically creating new features based on the characteristics of existing samples, dynamic feature expansion enriches the feature space, allowing tail classes to occupy a more substantial region in the embedding space. This expanded representation reduces the likelihood of overlap between tail and head classes, thereby improving classification accuracy for minority classes.

In distributed learning, dynamic feature expansion can be applied at each node, enabling each local model to create richer feature representations. These enhanced local embeddings are then aggregated to form a balanced global model. Dynamic feature expansion helps counteract the effects of imbalanced data distributions across nodes, leading to improved consistency in the federated model.

5.4 Embedding Augmentation in Distributed Learning

Embedding augmentation techniques play a vital role in distributed learning, where data heterogeneity and imbalances can severely impact model performance. By enriching the feature space through methods like LEFA, feature clouds, and random walk-based erasing, embedding augmentation addresses the variability across nodes and enhances the robustness of the global model. In distributed setups, each node processes data that might differ in distribution due to factors such as regional demographics or data collection protocols. Embedding augmentation techniques provide a means to create a more uniform representation across nodes, even if their underlying data characteristics vary significantly.

For instance, nodes in urban areas may have more data on common conditions, while rural nodes may have limited data, particularly for rare conditions. Embedding augmentation enables each node to produce feature representations that are comprehensive, even for minority classes, contributing to a balanced and discriminative feature space when aggregated at the server. Techniques like random walk-based erasing are particularly useful in distributed learning, as

they introduce controlled randomness, allowing models to generalize better and avoid overfitting to local patterns.

5.4.1 Addressing Node-Specific Variability

Embedding augmentation methods can be tailored to account for node-specific characteristics, allowing each node to adapt its embedding space based on local data properties. For example, nodes with significant class imbalance can prioritize feature clouds to enhance the diversity of tail classes, while nodes with a more balanced dataset might use random walk-based erasing to improve generalization. By applying augmentation techniques dynamically based on the node's specific data profile, distributed learning setups can maintain consistency in the embedding space while still addressing local nuances.

5.4.2 Enhancing Global Consistency

Embedding augmentation techniques not only improve local model performance but also enhance the consistency of the aggregated global model. Techniques like LEFA ensure that label information is embedded consistently across nodes, supporting a more harmonized feature representation. During aggregation, these augmented embeddings contribute to a balanced and robust global feature space, enabling the model to perform reliably across diverse data environments. This is crucial in distributed learning, where the objective is to develop a model that generalizes effectively despite variability across participating nodes.

5.4.3 Reducing Communication Overhead

Another advantage of embedding augmentation in distributed learning is the potential reduction in communication overhead. By improving local embedding representations, nodes can produce embeddings that are more representative, reducing the need for frequent updates or extensive communication between nodes and the central server. For example, dynamic feature expansion can help a node generate diverse representations from a limited dataset, decreasing the frequency of data transmission while still contributing meaningfully to the global model. This is especially beneficial in healthcare, where data privacy and bandwidth constraints are often limiting factors in federated learning setups.

5.5 Modified Random Walk-Based Embedding (RWE) Augmentation

In this experiment, a modification was applied to the original random walk-based embedding (RWE) augmentation technique, which traditionally erases portions of the embedding space to enhance generalization. Instead of erasing data, the modified RWE augmentation in this experiment involves adding synthetic data points into the embedding space, increasing robustness and enhancing the model's capacity to generalize across classes. This modified approach not only mitigates the limitations of traditional erasing but also introduces controlled diversity within each class, which is crucial in distributed learning settings where nodes might exhibit varying data characteristics.

5.5.1 Results and Observations

The results of this modified RWE augmentation, shown across multiple rounds, illustrate a

notable improvement in various metrics such as recall, precision, and AUC. Here are the primary

observations:

Table 5.1: Distributed and Centralized Accuracy/Loss per Round for Random Walk Embedding

Augmentation

Round	Distributed Loss	Distributed Accuracy	Centralized Loss	Centralized Accuracy	Recall	Precision	AUC
1.0	0.48953	0.49249	0.69264	0.48817	0.27865	0.17726	0.51916
2.0	0.43953	0.5778	0.48951	0.77307	0.50537	0.55903	0.76399
3.0	0.42196	0.59054	0.43954	0.798	0.61533	0.59986	0.82551
4.0	0.414	0.59327	0.42197	0.80783	0.60029	0.62725	0.83588
5.0	0.40707	0.596	0.414	0.81438	0.60172	0.64417	0.84191
6.0	0.40135	0.60328	0.40706	0.8192	0.60638	0.65595	0.84724
7.0	0.39791	0.60828	0.40134	0.82329	0.60638	0.66759	0.85132
8.0	0.39395	0.60828	0.39792	0.82712	0.60279	0.68027	0.85327
9.0	0.39187	0.61192	0.39395	0.82957	0.60279	0.68778	0.85592
10.0	0.39146	0.61283	0.39187	0.83221	0.59814	0.69816	0.85719

11.0	0.38869	0.6151	0.39146	0.83276	0.58883	0.70437	0.85758
12.0	0.38832	0.61374	0.38869	0.8354	0.59814	0.70853	0.85931
13.0	0.38682	0.61692	0.38832	0.8354	0.58703	0.71416	0.86
14.0	0.38648	0.61465	0.38683	0.83631	0.58632	0.71767	0.86106
15.0	0.38697	0.61374	0.38648	0.83703	0.58883	0.71885	0.86133

This table shows the progression of distributed and centralized accuracy and loss across 15 training rounds using Random Walk Embedding Augmentation. The centralized model consistently achieves higher accuracy and lower loss per round, illustrating the advantages of centralized access, while the distributed model shows steady improvements as well.

Distributed Loss and Accuracy: The distributed model's loss progressively decreases over the rounds, with a final loss of approximately 0.38648 by the 14th round. Accuracy follows an upward trend, reaching 0.61465 for the distributed setup by round 14. This steady improvement in distributed accuracy and reduction in loss indicates that the modified RWE augmentation effectively supports learning in a federated setup, enabling the model to capture nuanced patterns more consistently across nodes.

Centralized Performance Stability: The centralized accuracy and loss stabilize as well, with the final centralized accuracy and AUC converging to 0.83703 and 0.86133, respectively, by round 15. This convergence signifies that the augmented distributed learning setup is able to achieve

performance levels comparable to a centralized approach, demonstrating the efficacy of the modified RWE augmentation in creating a balanced and generalizable feature space across nodes.

Recall and Precision Gains: Recall improves from an initial 0.27865 to over 0.58883 by round 15, and precision increases from 0.17726 to 0.71885. This indicates that the modified RWE technique, by adding synthetic data points instead of erasing features, enhances the model's capacity to correctly identify positive cases and reduce false positives. This is particularly relevant in healthcare applications, where both high recall and precision are essential for accurate diagnosis and minimizing missed cases.

5.5.2 Advantages of Data Addition over Data Erasure in RWE

The choice to add synthetic data instead of erasing portions by (C. Zhang et al. 2023) of the feature space addresses several limitations associated with traditional RWE augmentation:

Enhanced Class Representation: By augmenting each class with additional synthetic samples, the modified RWE helps to expand the representation of underrepresented classes. This is especially advantageous in distributed learning setups where individual nodes may have data scarcity for certain classes. Adding synthetic data improves intra-class diversity, making each class more distinguishable within the feature space.

Reduced Risk of Overgeneralization: Erasing data in traditional RWE can lead to overgeneralization, as the model might lose critical details necessary for accurate class separation. By adding data instead, the modified RWE ensures that the model retains important features while still benefiting from a wider range of synthetic samples, reducing the risk of overgeneralization.

Increased Robustness to Heterogeneous Data: In distributed learning environments, nodes often have heterogeneous data due to differences in demographics, equipment, or data collection protocols. The modified RWE's data addition approach strengthens the model's ability to generalize across diverse data sources by introducing a controlled variety of synthetic samples, which helps the model learn invariant features that are consistent across nodes.

Improved Learning for Tail Classes: By enriching the feature space with synthetic samples, the modified RWE also indirectly addresses class imbalance issues. Tail classes, which may be underrepresented, benefit from this augmentation as it increases their presence in the embedding space, allowing the model to better capture subtle variations within these classes and improve classification performance for minority categories.

Consistency Across Aggregation Cycles: The steady improvement across rounds shows that adding data through modified RWE allows each aggregation cycle to incorporate meaningful updates, enhancing convergence. The synthetic data enriches local representations, allowing the federated model to converge faster with consistent performance improvements over each round.

5.5.3 Embedding Space and Feature Distribution Insights

The modified RWE augmentation enriches the embedding space by filling in gaps and smoothing the distribution of features across classes. By adding synthetic points within each class, the technique creates a more evenly distributed feature space, minimizing the risks of sparsity or clustering that can lead to biased representations. This augmented embedding space supports better class separability, as it ensures that even minority classes have a comprehensive feature distribution, which improves overall model interpretability and performance.

Furthermore, the addition-based approach reduces noise sensitivity, as the model no longer needs to adjust to missing or "erased" features. Instead, it learns to navigate a more stable feature landscape where each class is well-represented, leading to improved robustness and reliability.

5.5.4 Practical Implications for Distributed Learning

The modified RWE technique's approach to adding synthetic data rather than erasing it provides several practical advantages in distributed learning:

Data Privacy: In federated learning, data privacy is a top concern, as raw data remains localized at each node. By generating synthetic data locally, each node can enhance its feature representation without compromising data privacy or needing external augmentation resources.

This local augmentation aligns well with federated learning principles, as nodes retain full control over the augmentation process.

Reduced Communication Overhead: With enriched embeddings generated at each node, there is less need for frequent parameter updates or extensive communication between nodes and the central server. This efficiency can reduce the communication costs of distributed learning, making the system more scalable and sustainable in real-world applications.

Adaptability Across Nodes: Each node can adapt the augmentation to its specific data distribution, making this technique highly flexible. For example, nodes with fewer data samples can generate more synthetic points to match the diversity levels of other nodes, promoting consistency in the aggregated model.

5.6 Angular Variance in Tail Class Embedding Augmentation: Modified LEFA Without Feature Clouds

In this experiment, a modified version of the Label Embedding and Feature Augmentation (LEFA) technique was applied, specifically focusing on addressing tail class representation without incorporating feature clouds. This choice was based on certain challenges highlighted in the original LEFA paper, where the implementation of feature clouds presented several practical limitations in distributed learning setups. Instead, the augmentation method here relies on angular variance to enhance tail class representations, as shown in the results.

5.6.1 Limitations of Feature Clouds in Capturing Intra-Class Exclusiveness

One of the key challenges with feature clouds, as identified in the LEFA study by (Wang et al. 2020), is their inability to effectively capture the intra-class exclusiveness required for robust feature representation, especially in multi-dimensional classification tasks. Intra-class exclusiveness refers to the distinct and unique characteristics within each class that allow the model to differentiate between labels or sub-categories within the same class. Failing to preserve this exclusiveness can lead to "degenerated performance," where the model's ability to accurately classify and separate instances within the same class deteriorates.

The primary reason for this limitation is that many existing feature augmentation approaches, including feature clouds, rely on relatively simple base classifiers, such as Support Vector Machines (SVM) and Naive Bayes. These classifiers are often insufficiently powerful to manage the complex label correlations found in high-dimensional and multi-label datasets. Since these base classifiers have limited capacity to learn the nuanced relationships and dependencies between labels, they struggle to maintain intra-class exclusiveness, especially when augmentations introduce synthetic variations that blur these fine-grained distinctions.

In multi-dimensional classification tasks, where instances can have multiple labels and complex relationships, the lack of intra-class exclusiveness can severely impact performance. For example, in healthcare applications, a model may need to distinguish between overlapping conditions or symptoms within the same class (e.g., different types of heart disease), each with

its own distinct patterns. When feature clouds are used without sufficient intra-class

exclusiveness, synthetic samples generated for these conditions may overlap, leading to reduced

separability and increasing the likelihood of misclassification. This can be particularly

problematic in applications where accurate differentiation within a class is essential for reliable

diagnoses and treatment recommendations.

5.6.2 Degenerated Performance Due to Label Correlation Complexity

The complexity of label correlations in multi-dimensional tasks presents a significant challenge

for simple base classifiers when handling augmented data. Augmentation techniques like feature

clouds often rely on random perturbations or distributions around existing samples, which may

inadvertently introduce similarities between labels within the same class. Without a robust

mechanism to preserve exclusiveness, these augmented samples can converge towards the same

feature space, thereby diminishing the model's discriminative power.(Wang et al. 2020)

For example, if an augmentation method generates synthetic samples around a heart disease class

without considering the specific sub-categories (e.g., ischemic heart disease vs. arrhythmia), the

model might learn an overly generalized representation that fails to capture the distinct features

of each sub-condition. This leads to what is described as "degenerated performance," where the

model's overall accuracy may appear stable, but its ability to correctly identify specific

conditions deteriorates. In a distributed learning setup, where data heterogeneity across nodes

already poses challenges, this degradation in performance can become even more pronounced, as

each node may contribute slightly different variations that further obscure intra-class exclusiveness.

5.6.3 The Role of Classifier Complexity in Effective Augmentation

The effectiveness of embedding augmentation techniques like feature clouds is often constrained by the choice of classifiers used in learning these embeddings. Simple classifiers such as SVM and Naive Bayes, though effective for straightforward binary or linear problems, lack the representational power required to handle intricate label relationships. Multi-dimensional and multi-label classification problems demand more sophisticated models that can learn hierarchical and complex patterns within each class. When augmentation relies on simple classifiers, it inadvertently introduces noise and blurs distinctions within the feature space, making it challenging to capture the exclusive attributes of each label within a class.(Wang et al. 2020)

Advanced classifiers, such as deep neural networks or ensemble methods, are generally more capable of handling complex label correlations. However, integrating these into augmentation processes like feature clouds would require substantial computational resources, making it challenging in distributed environments. Thus, the choice of classifier plays a crucial role in determining the success of augmentation techniques in preserving intra-class exclusiveness.(Wang et al. 2020)

5.6.4 Addressing the Intra-Class Exclusiveness Challenge Through Angular Variance

Given the limitations of feature clouds in maintaining intra-class exclusiveness(Liu et al. 2020), this experiment adopts angular variance as an alternative strategy to enhance tail class representation. Angular variance allows tail classes to spread out within the embedding space without relying on synthetic samples that could overlap with other labels or classes. By focusing on the angular distribution of instances within each class, this approach preserves the unique characteristics of each label, ensuring that each sub-category within a class retains its distinctiveness.

In contrast to feature clouds, which rely on synthetic sample generation(Liu et al. 2020), angular variance adjusts the spatial arrangement of existing instances in a way that respects intra-class distinctions. This method minimizes the risk of overlapping features, which can occur when using simpler classifiers that fail to capture complex label correlations. By avoiding synthetic samples, angular variance provides a clearer representation of each class, supporting better generalization and improving the model's ability to accurately distinguish between labels within the same class.

5.6.5 Practical Implications of Avoiding Feature Clouds in Distributed Learning

The decision to avoid feature clouds in favor of angular variance has practical benefits for distributed learning setups:

Enhanced Consistency Across Nodes: In a distributed system, where each node may have different data characteristics, maintaining intra-class exclusiveness is crucial for achieving a balanced global model. By employing angular variance rather than feature clouds, each node can independently adjust the spread of tail classes without introducing synthetic variations that may lead to inconsistencies during aggregation.

Reduced Computational Burden: Generating feature clouds can be resource-intensive, particularly in environments with limited computational capabilities. Angular variance, on the other hand, leverages existing data points to achieve class spread, making it more suitable for resource-constrained nodes and reducing the overall computational load in a distributed setup.

Improved Model Interpretability: Without synthetic samples, the model's decision boundaries and feature representations remain closer to real data, enhancing interpretability. This is particularly important in healthcare and other sensitive domains, where understanding the model's decision-making process is critical. Angular variance ensures that each class maintains its distinct characteristics, supporting clearer and more interpretable predictions.

Consistency in Performance: By avoiding synthetic augmentation methods that may inadvertently introduce overlapping features, angular variance promotes consistent performance across nodes. This approach minimizes the risk of degenerated performance, where intra-class distinctions are lost, and ensures that tail classes retain their unique representation across rounds of distributed learning.

In summary, the modified LEFA approach using angular variance rather than feature clouds addresses the limitations associated with intra-class exclusiveness. By adopting a method that preserves the distinctiveness of each label within a class, this approach overcomes the primary disadvantage of feature clouds—namely, the tendency to degrade performance due to insufficient classifier complexity and overlapping embeddings. Angular variance offers a more consistent, interpretable, and computationally efficient solution for distributed learning, making it well-suited for environments that require precise and reliable classification across a variety of classes.

5.6.6 Observations

The application of angular variance as an augmentation strategy in place of feature clouds produced a range of notable improvements across multiple metrics over the 15 training rounds. The following key observations highlight the effectiveness of angular variance in enhancing tail class representation while preserving intra-class exclusiveness.

Table 5.2: Distributed and Centralized Accuracy/Loss per Round for Angular Variance Embedding Augmentation

Round	Distributed Loss	Distributed Accuracy	Centralized Loss	Centralized Accuracy	Recall	Precision	AUC
1.0	0.43153	0.57598	0.51222	0.76069	0.41905	0.53719	0.76572
2.0	0.42111	0.5869	0.43153	0.80737	0.53832	0.64479	0.82361

3.0	0.41246	0.59349	0.4211	0.81301	0.54907	0.65822	0.83434
4.0	0.41124	0.596	0.41246	0.81947	0.56447	0.67235	0.842
5.0	0.40407	0.59918	0.41125	0.82056	0.5616	0.67703	0.84397
6.0	0.4015	0.60191	0.40407	0.82402	0.56519	0.68668	0.84929
7.0	0.39649	0.60692	0.4015	0.82848	0.56554	0.70147	0.85177
8.0	0.39886	0.60191	0.39649	0.83021	0.56769	0.70633	0.85422
9.0	0.39295	0.60328	0.39886	0.82957	0.55946	0.70839	0.85431
10.0	0.39238	0.60555	0.39295	0.83376	0.56554	0.72002	0.85755
11.0	0.38861	0.60464	0.39238	0.83412	0.5659	0.72113	0.85834
12.0	0.38765	0.60601	0.38861	0.83594	0.57128	0.72467	0.86032
13.0	0.38739	0.60646	0.38765	0.83594	0.56877	0.72611	0.86077
14.0	0.38574	0.61101	0.38739	0.83567	0.56734	0.72594	0.86123
15.0	0.38573	0.61101	0.38574	0.83776	0.56948	0.73238	0.86193

This table presents the distributed and centralized accuracy and loss values for each training round with Angular Variance Embedding Augmentation. Similar to the Random Walk method, centralized accuracy and loss outperform distributed metrics, though both setups demonstrate significant gains in accuracy and reduction in loss over the training rounds.

5.6.6.1 Steady Improvement in Key Metrics

Throughout the 15 rounds, there was a consistent upward trend in key performance metrics such as recall, precision, and AUC. Recall improved from **0.41905** in the first round to **0.56948** by the 15th round, while precision rose significantly from **0.53719** to **0.73238**. This indicates that the angular variance approach not only enhanced the model's ability to capture tail classes but also improved its precision, meaning the model was better at minimizing false positives. The substantial increase in AUC, from **0.76572** in the first round to **0.86193** by the 15th round, reflects the model's enhanced capacity to discriminate between classes as a result of angular variance, especially for underrepresented classes.

5.6.6.2 Enhanced Tail Class Representation and Class Separability

Angular variance successfully dispersed the embeddings of tail class instances, reducing the compression typically observed in minority classes. By allowing tail classes to occupy a larger and more distinct region in the feature space, angular variance helped improve class separability without introducing overlapping features. This improved separation was reflected in the increasing AUC and precision scores, suggesting that the model became more capable of correctly identifying and differentiating minority class instances over time. This effect is particularly valuable in healthcare applications, where accurately distinguishing between conditions within the same category (e.g., different types of heart disease) can be critical for patient outcomes.

5.6.6.3 Efficiency and Scalability in Distributed Learning

The results indicate that angular variance offers an efficient and scalable solution for distributed

learning setups. In federated environments where each node may have computational limitations,

feature clouds can be resource-intensive, as they require generating synthetic data points around

tail classes. In contrast, angular variance disperses existing data points within the feature space

without the need for synthetic sample generation. This led to a gradual decrease in distributed

loss, dropping from **0.43153** in the first round to **0.38573** by the last round, highlighting the

method's efficiency and suitability for distributed settings. The model's ability to perform well

without synthetic augmentation reduces the burden on individual nodes, making it easier to

implement across heterogeneous environments.

5.6.6.4 Stability Between Centralized and Distributed Metrics

The similarity between centralized and distributed metrics suggests that angular variance

provides a consistent augmentation effect across nodes, maintaining stable performance in both

localized and aggregated learning contexts. For example, the distributed and centralized AUC

metrics closely aligned, both reaching around **0.86193** in the final rounds. This stability is

essential for distributed systems, as it ensures that each node contributes consistently to the

global model without introducing discrepancies that could hinder aggregation. Such consistency

enhances the reliability of the federated model, making it more robust when deployed across

diverse data environments.

5.6.6.5 Improved Robustness to Data Imbalance and Limited Samples

Angular variance effectively mitigates issues associated with data imbalance, especially for tail

classes that lack sufficient samples. By dispersing instances within the embedding space, the

method allows each tail class to develop a broader and more comprehensive representation. This,

in turn, helps prevent the minority classes from being overshadowed by majority classes during

model training, leading to higher recall and precision for underrepresented groups. The gradual

improvement in recall, precision, and AUC over rounds indicates that angular variance helped

the model better capture the unique characteristics of minority classes, addressing data imbalance

without requiring extensive additional data.

5.6.6.6 Increased Interpretability of Feature Space

Unlike feature clouds, which generate synthetic samples that could obscure the model's

interpretability, angular variance adjusts the spatial distribution of real data points within the

feature space. This approach preserves the inherent structure of the original data, allowing for

clearer and more interpretable decision boundaries. The steady rise in precision and recall

reflects the model's capacity to make reliable distinctions based on real data variations, rather

than synthetic data that may not fully capture the complexities of each class. This increased

interpretability is particularly valuable in sensitive applications, such as healthcare, where

understanding model decisions is critical for clinical validation and trust.

5.6.6.7 Adaptability Across Diverse Node Characteristics

The angular variance method proved adaptable across nodes with differing data distributions, as

seen in the distributed metrics that closely tracked centralized performance. By adjusting the

angular distribution of instances, the technique accommodated variations in node data

characteristics, helping to harmonize the global model's feature space. This adaptability is

particularly advantageous in federated learning settings, where data heterogeneity can vary

widely between nodes. With angular variance, each node can effectively represent minority

classes without introducing synthetic data that could destabilize the global model.

Chapter 6

Conclusion

This study has investigated the impact of various data augmentation methods on model

performance within a distributed learning framework. In distributed learning, data resides on

multiple nodes, each with access to only a subset of the complete dataset. This setup introduces

challenges, particularly for ensuring the model learns a comprehensive and diverse

representation across all classes. Our findings indicate that traditional data augmentation

methods—such as adding noise or scaling—yielded limited improvements when applied alone.

These methods introduce minor variations in the input data but lack the transformative capacity

to enable the model to capture complex, diverse patterns across distributed nodes.

The most significant advancements were observed when embedding augmentation techniques,

specifically Random Walk and Angular Variance, were combined with traditional data

augmentation. Unlike traditional methods that make slight alterations to the input data,

embedding augmentation operates within the model's feature space, creating synthetic

representations that expand the diversity of learned features. This approach allows the model to

generate new internal representations, enriching the feature space and facilitating improved

generalization across nodes.

Embedding augmentation proved particularly advantageous for underrepresented, or "tail,"

classes, which are often sparsely represented in distributed datasets. This challenge is amplified

in distributed learning setups, where individual nodes may have skewed class distributions. By

directly expanding the feature space for tail classes, embedding augmentation enhances the

representation of these classes, ensuring a more balanced learning process. This approach

mitigates the risks of underfitting for minority classes, resulting in a model that can effectively

distinguish between diverse class distributions.

The combined approach of embedding and traditional augmentation provided several measurable

benefits:

6.1 Enhanced Accuracy and Accelerated Convergence

 Models utilizing both types of augmentation exhibited faster learning rates, reaching stable and

high-performance levels sooner than those relying solely on traditional methods. This

improvement was consistent in both distributed and centralized settings, though centralized

models achieved slightly higher accuracy due to their access to the entire dataset.

6.2 Improved Class Separability

Embedding augmentation techniques contributed to clearer class boundaries, which reduced

misclassifications and enhanced the model's decision-making accuracy. By promoting better-

defined distinctions within the feature space, embedding augmentation enabled the model to

handle complex class relationships more effectively.

6.3 Practical Recommendations

Based on these findings, I propose several practical recommendations for applying these

augmentation strategies to enhance distributed learning models:

6.3.1 Tailored Fine-Tuning for Distributed Setups

Fine-tuning the parameters of embedding augmentation techniques can optimize model

performance in distributed environments, aligning distributed models more closely with

centralized results. This may involve adjusting the intensity of augmentation to better suit each

node's unique data distribution.

6.3.2 Integrating Multiple Augmentation Types

Combining input-level and embedding-level augmentation creates a more adaptable model

capable of handling diverse data conditions. This integration enhances generalization by

providing multiple forms of variation, which are particularly useful in heterogeneous data

environments.

6.3.4 Visualization of Feature Space Dynamics

Visualizing changes in the feature space induced by both data and embedding augmentations can

offer deeper insights into how class separability and feature diversity evolve over training. This

analysis could inform the refinement of augmentation techniques for specific tasks, aiding in the development of more targeted approaches.

In conclusion, embedding-based augmentation, especially when combined with traditional data augmentation, presents a promising solution for improving model performance in distributed learning contexts. By enriching feature space diversity and enhancing model robustness, embedding augmentation enables distributed models to achieve reliable, high-quality performance even with limited data access at each node. Future research could explore the adaptation of these techniques for different data types and their application to other distributed learning paradigms. This combined augmentation strategy offers a practical path forward, enabling the creation of models that are both resilient and accurate within complex, decentralized learning environments.

Bibliography

Ahmed, Adel A., Waleed Ali, Talal A. A. Abdullah, and Sharaf J. Malebary. 2023. "Classifying Cardiac Arrhythmia from ECG Signal Using 1D CNN Deep Learning Model." *Mathematics* 11 (3): 562. https://doi.org/10.3390/math11030562.

Akhoondkazemi, Amirhossein, Arash Vashagh, Sayed Jalal Zahabi, and Davood Shafie. 2023. "Atrial Fibrillation (AF) Detection Using Deep Learning with GAN-Based Data Augmentation." In *2023 31st International Conference on Electrical Engineering (ICEE)*, 955–59. https://doi.org/10.1109/ICEE59167.2023.10334666.

Aminifar, Amin, Fazle Rabbi, Violet Ka I Pun, and Yngve Lamo. 2021. "Monitoring Motor Activity Data for Detecting Patients' Depression Using Data Augmentation and Privacy-Preserving Distributed Learning." In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2163–69. https://doi.org/10.1109/EMBC46164.2021.9630592.

Apandi, Ziti Fariha Mohd, Ryojun Ikeura, and Soichiro Hayakawa. 2018. "Arrhythmia Detection Using MIT-BIH Dataset: A Review." In *2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA)*, 1–5. https://doi.org/10.1109/ICASSDA.2018.8477620.

Balasubramanian, Nikil Sharan Prabahar, and Sagnik Dakshit. 2024. "Explanations of Augmentation Methods for Deep Learning ECG Classification." In *Artificial Intelligence in Medicine*, edited by Joseph Finkelstein, Robert Moskovitch, and Enea Parimbelli, 14845:277–87. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-66535-6_30.

Dablain, Damien A., Colin Bellinger, Bartosz Krawczyk, and Nitesh V. Chawla. 2023. "Efficient Augmentation for Imbalanced Deep Learning." In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 1433–46. https://doi.org/10.1109/ICDE55515.2023.00114.

Dakshit, Sagnik, and Balakrishnan Prabhakaran. 2023. "CVAE-Based Generator for Variable Length Synthetic ECG." In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, 235–44. Houston, TX, USA: IEEE. https://doi.org/10.1109/ICHI57859.2023.00040.

Ding, Hongwei, Nana Huang, Yaoxin Wu, and Xiaohui Cui. 2024. "LEGAN: Addressing Intraclass Imbalance in GAN-Based Medical Image Augmentation for Improved Imbalanced Data Classification." *IEEE Transactions on Instrumentation and Measurement* 73:1–14. https://doi.org/10.1109/TIM.2024.3396853.

Fan, Wei, Yujuan Si, Weiyi Yang, and Meiqi Sun. 2022. "Imbalanced ECG Data Classification Using a Novel Model Based on Active Training Subset Selection and Modified Broad Learning System." *Measurement* 198 (July):111412. https://doi.org/10.1016/j.measurement.2022.111412.

Guo, Peikun, Huiyuan Yang, and Akane Sano. 2023. "Empirical Study of Mix-Based Data Augmentation Methods in Physiological Time Series Data." arXiv. http://arxiv.org/abs/2309.09970.

Hao, Weituo, Mostafa El-Khamy, Jungwon Lee, Jianyi Zhang, Kevin J Liang, Changyou Chen, and Lawrence Carin. 2021. "Towards Fair Federated Learning with Zero-Shot Data Augmentation." In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3305–14. Nashville, TN, USA: IEEE. https://doi.org/10.1109/CVPRW53098.2021.00369.

Hata, Erika, Chanjin Seo, Masafumi Nakayama, Kiyotaka Iwasaki, Takaaki Ohkawauchi, and Jun Ohya. 2020. "Classification of Aortic Stenosis Using ECG by Deep Learning and Its Analysis Using Grad-CAM." In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1548–51. https://doi.org/10.1109/EMBC44109.2020.9175151.

Hu, Chenghao, Jingyan Jiang, and Zhi Wang. 2019. "Decentralized Federated Learning: A Segmented Gossip Approach." arXiv. http://arxiv.org/abs/1908.07782.

Khan, Qazi Waqas, Anam Nawaz Khan, Atif Rizwan, Rashid Ahmad, Salabat Khan, and Do-Hyeun Kim. 2023. "Decentralized Machine Learning Training: A Survey on Synchronization, Consolidation, and Topologies." *IEEE Access* 11:68031–50. https://doi.org/10.1109/ACCESS.2023.3284976.

Liu, Jialun, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. 2020. "Deep Representation Learning on Long-Tailed Data: A Learnable Embedding Augmentation Perspective." In , 2970–79. https://openaccess.thecvf.com/content_CVPR_2020/html/Liu_Deep_Representation_Learning_on_Long-Tailed_Data_A_Learnable_Embedding_Augmentation_CVPR_2020_paper.html.

M, Mohamed Musthafa, Mahesh T. R, Vinoth Kumar V, and Suresh Guluwadi. 2024. "Enhancing Brain Tumor Detection in MRI Images through Explainable AI Using Grad-CAM with Resnet 50." *BMC Medical Imaging* 24 (1): 107. https://doi.org/10.1186/s12880-024-01292-7.

Moody, G.B., and R.G. Mark. 2001. "The Impact of the MIT-BIH Arrhythmia Database." *IEEE Engineering in Medicine and Biology Magazine* 20 (3): 45–50. https://doi.org/10.1109/51.932724.

Nonaka, Naoki, and Jun Seita. 2020. "Data Augmentation for Electrocardiogram Classification with Deep Neural Network." arXiv. http://arxiv.org/abs/2009.04398.

Ojha, Jaya, Hårek Haugerud, Anis Yazidi, and Pedro G. Lind. 2024. "Exploring Interpretable AI Methods for ECG Data Classification." In *The Fifth Workshop on Intelligent Cross-Data Analysis and Retrieval*, 11–18. Phuket Thailand: ACM. https://doi.org/10.1145/3643488.3660294.

Owusu-Adjei, Michael, James Ben Hayfron-Acquah, Twum Frimpong, and Gaddafi Abdul-Salaam. 2023. "Imbalanced Class Distribution and Performance Evaluation Metrics: A Systematic Review of Prediction Accuracy for Determining Model Performance in Healthcare Systems." *PLOS Digital Health* 2 (11): e0000290. https://doi.org/10.1371/journal.pdig.0000290.

Plawiak, Pawel. 2017. "ECG Signals (1000 Fragments)" 3 (November). https://doi.org/10.17632/7dybx7wyfn.3.

Śmigiel, Sandra, Krzysztof Pałczyński, and Damian Ledziński. 2021. "ECG Signal Classification Using Deep Learning Techniques Based on the PTB-XL Dataset." *Entropy* 23 (9): 1121. https://doi.org/10.3390/e23091121.

Strodthoff, Nils, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. 2021. "Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL." *IEEE Journal of Biomedical and Health Informatics* 25 (5): 1519–28. https://doi.org/10.1109/JBHI.2020.3022989.

Tang, Wensi, Guodong Long, Lu Liu, Tianyi Zhou, Jing Jiang, and Michael Blumenstein. 2020. *Rethinking 1D-CNN for Time Series Classification: A Stronger Baseline*. https://doi.org/10.48550/arXiv.2002.10061.

Wagner, Patrick, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. 2020. "PTB-XL, a Large Publicly Available Electrocardiography Dataset." *Scientific Data* 7 (1): 154. https://doi.org/10.1038/s41597-020-0495-6.

Wang, Haobo, Chen Chen, Weiwei Liu, Ke Chen, Tianlei Hu, and Gang Chen. 2020. "Incorporating Label Embedding and Feature Augmentation for Multi-Dimensional Classification." *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (04): 6178–85. https://doi.org/10.1609/aaai.v34i04.6083.

Xiao, Yunpeng, Xufeng Li, Tun Li, Rong Wang, Yucai Pang, and Guoyin Wang. 2024. "A Distributed Generative Adversarial Network for Data Augmentation under Vertical Federated Learning." *IEEE Transactions on Big Data*, 1–12. https://doi.org/10.1109/TBDATA.2024.3375150.

Xu, Yuefan, Sen Zhang, and Wendong Xiao. 2023. "Inter-Patient ECG Classification with Intra-Class Coherence Based Weighted Kernel Extreme Learning Machine." *Expert Systems with Applications* 227 (October):120095. https://doi.org/10.1016/j.eswa.2023.120095.

Yang, Huiyuan, Han Yu, and Akane Sano. 2022. "Empirical Evaluation of Data Augmentations for Biobehavioral Time Series Data with Deep Learning." arXiv. http://arxiv.org/abs/2210.06701.

Yuan, Liqiang, and Mohammed Yakoob Siyal. 2023. "Target-Oriented Augmentation Privacy-Protection Domain Adaptation for Imbalanced ECG Beat Classification." *Biomedical Signal Processing and Control* 86 (September):105308. https://doi.org/10.1016/j.bspc.2023.105308.

Zhang, Chao, Weifeng Zhong, Changfeng Li, and Haipeng Deng. 2023. "Random Walk-Based Erasing Data Augmentation for Deep Learning." *Signal, Image and Video Processing* 17 (5): 2447–54. https://doi.org/10.1007/s11760-022-02461-3.

Zhang, Hao, Qingying Hou, Tingting Wu, Siyao Cheng, and Jie Liu. 2023. "Data-Augmentation-Based Federated Learning." *IEEE Internet of Things Journal* 10 (24): 22530–41. https://doi.org/10.1109/JIOT.2023.3303889.

Zhou, Jiayuan, Kshitij Khare, and Sanvesh Srivastava. 2023. "Asynchronous and Distributed Data Augmentation for Massive Data Settings." *Journal of Computational and Graphical Statistics* 32 (3): 895–907. https://doi.org/10.1080/10618600.2022.2130928.