INTERNSHIP REPORT

On

“Contrastive and Metric Learning on IMU-Based Gesture Recognition Dataset”

Submitted in partial fulfilment of the requirements for the award of

Bachelor of Technology

Submitted by-

Ritika Varshney

ROLL NO: -BT24EEE096

Department of Electrical and Electronics Engineering

Visvesvaraya National Institute of Technology, Nagpur

Internship carried out at

Indian Institute of Technology Indore

Department of Electrical Engineering

Under the guidance of

Dr. Ayush Tripathi

Assistant Professor

Department of Electrical Engineering

IIT Indore

Internship Duration: 1st of December 2025 – 31st of December 2025

Academic Year: 2025–2026

TABLE OF CONTENTS

CERTIFICATE

2) Abstract:

Inertial Measurement Unit (IMU) signals represent motion-related measurements obtained from sensors such as accelerometers and gyroscopes and are widely used in applications including gesture recognition, human activity recognition, and human–computer interaction. However, IMU signals are noisy, non-stationary, and user-dependent, which makes accurate gesture classification a challenging task. Deep learning models, particularly convolutional neural networks (CNNs), have shown strong performance in learning discriminative features directly from IMU time-series data.

This internship project focuses on IMU-based gesture recognition using a single fixed CNN architecture, with primary emphasis on studying the effect of different embedding-based loss functions. The CNN was evaluated under four experimental cases, including cross-entropy loss as a baseline and multiple embedding-oriented loss functions.

Model performance was evaluated using classification accuracy. In addition, the learned feature embeddings were visualized using class-wise color plots to analyze feature clustering and class separability. The results indicate that embedding-based loss functions improve the structure of the learned embedding space; however, overall classification performance remains limited due to the inherent variability and noise present in IMU signals.

3) Introduction

3.1 Background of the study

IMU time-series data characteristics, such as sensor noise, motion dynamics, and user-dependent movement patterns, significantly affect model generalization in real-world scenarios. CNNs are particularly effective for capturing local temporal patterns in IMU signals generated by accelerometer and gyroscope sensors. While many studies focus on improving model architectures, the influence of loss functions on learned feature representations has received relatively less attention.

This project focuses on analyzing IMU-based gesture recognition using a fixed CNN architecture, with the primary objective of studying how different embedding-based loss functions affect feature learning and classification performance.

3.2 Objectives of the Project

The specific objectives of the project are:

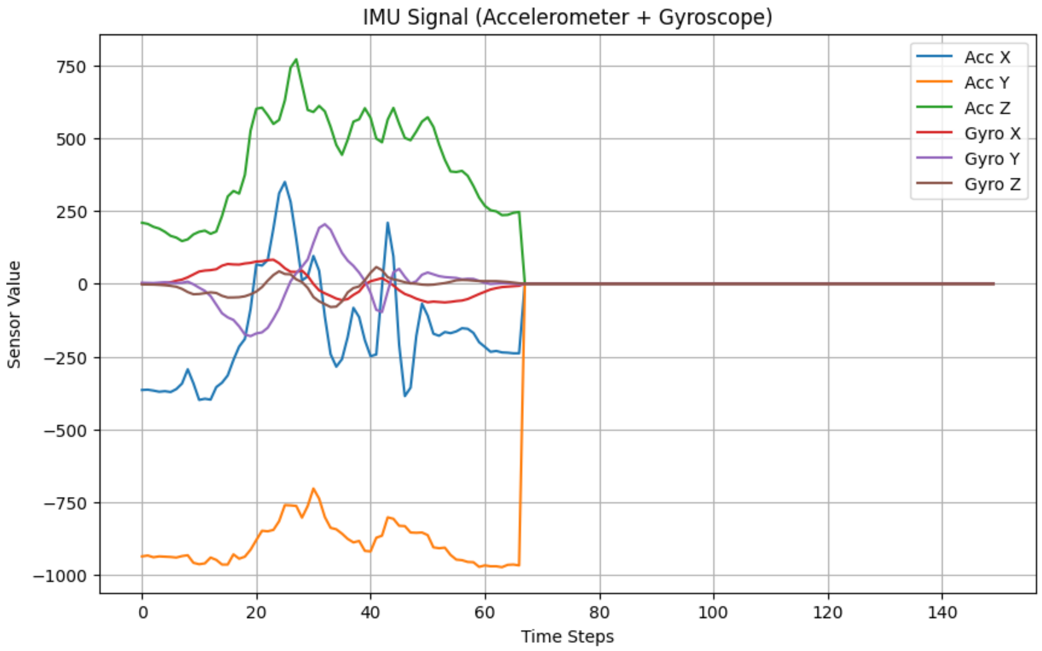
- To study the characteristics and challenges associated with multi-channel IMU time-series data
- To compare classification performance across all experimental cases under identical training conditions
- To visualize learned embeddings using class-wise color plots to study feature clustering.
- To analyze the relationship between embedding quality and classification accuracy.

4) Literature Review:

4.1 IMU Signals

IMU signals are obtained from motion sensors such as accelerometers and gyroscopes, which measure linear acceleration and angular velocity along three orthogonal axes. These signals are widely used in gesture recognition, activity recognition, navigation, and wearable computing applications due to their low cost, portability, and ease of integration.

IMU signals are noisy, non-stationary, and highly sensitive to factors such as sensor placement, movement speed, and inter-subject variability, making robust gesture classification challenging.



4.2 IMU Signal Processing and Classification Techniques

Early approaches in IMU-based gesture recognition rely on manual signal processing and handcrafted feature extraction. Time-domain and frequency-domain features are extracted from accelerometer and gyroscope signals and used as inputs to traditional machine learning classifiers. Designing effective features requires domain expertise and extensive tuning, and such methods often struggle to handle noise and variability in real-world IMU recordings.

Techniques	**Features Used**	**Limitations**
Time-domain analysis	Mean absolute value (MAV), Root mean square (RMS)	Highly sensitive to noise and motion variability
Classical classifiers (SVM, k-NN)	Fixed handcrafted feature vectors	Poor generalization across users
Threshold-based methods	Simple amplitude-based thresholds	Not robust under dynamic motions

Limitations of Traditional EMG Signal Processing and Classification Techniques

4.3 Deep Learning Approaches for IMU Analysis

Deep learning models learn hierarchical feature representations directly from raw IMU time-series data, making them suitable for handling complex and noisy motion signals. CNNs are commonly applied in IMU-based gesture recognition because they can capture local temporal patterns associated with motion dynamics.

In most existing studies, deep learning is primarily used as a classification tool with emphasis on network architecture. The role of training objectives and loss functions in shaping learned representations is often not explicitly analyzed.

4.4 Representation Learning and Embedding-Based Loss Functions

Representation learning refers to the process of learning a mapping

$$f_\theta : \mathbb{R}^{T \times C} \rightarrow \mathbb{R}^d$$

that transforms raw input data into a lower-dimensional feature space, commonly referred to as an **embedding space**. The objective is to encode task-relevant information such that samples belonging to the same class are easily separable from those of different classes.

In IMU-based gesture recognition, raw accelerometer and gyroscope signals are high-dimensional, noisy, and exhibit strong inter-user variability. Representation learning addresses this challenge by projecting IMU signals into a structured embedding space where meaningful similarity relationships can be learned.

In this project, a fixed convolutional neural network (CNN) is used to learn embeddings from raw IMU time-series data. The network architecture is kept constant across all experiments, and representation learning behavior is controlled exclusively through the choice of loss function.

The different embedding loss functions used in this project are as follows :

1. **Supervised Contrastive Loss :** Supervised contrastive loss is a representation learning objective that operates directly on the learned embeddings and uses class labels to define similarity relationships between samples. For a given embedding, all samples from the same class are treated as positives, while samples from different classes are treated as negatives. The loss encourages higher similarity between embeddings of the same class and lower similarity between embeddings of different classes, typically using cosine similarity.

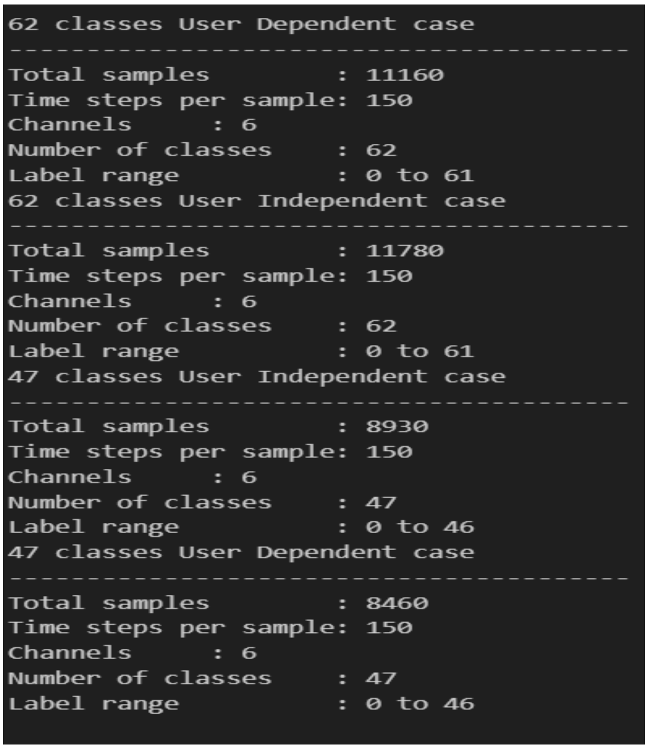
   From a geometric perspective, this enforces global clustering in the embedding space, resulting in compact class-wise clusters and improved inter-class separation. For IMU signals, where sensor noise, user-dependent motion patterns, and recording variations introduce large intra-class variability, supervised contrastive loss helps reduce this variability by pulling same-class embeddings closer together. Its effectiveness, however, depends on batch composition, as sufficient positive and negative samples must be present within each batch to provide a meaningful contrastive signal.

2. **Triplet Loss :** Triplet loss is a metric learning objective that enforces relative distance constraints between embeddings. It operates on triplets consisting of an anchor sample, a positive sample from the same class, and a negative sample from a different class. The loss encourages the distance between the anchor and positive embeddings to be smaller than the distance between the anchor and negative embeddings by at least a predefined margin. Geometrically, triplet loss enforces local ordering relationships in the embedding space rather than global clustering. This makes it effective for separating classes that are closely related, which is common in IMU-based gesture recognition, where different gestures often exhibit overlapping motion dynamics. However, triplet loss is sensitive to triplet selection, and its performance depends on the presence of informative anchor–positive–negative combinations during training.

3. **Margin Loss:** Margin-based embedding loss introduces an explicit margin between embeddings of different classes by penalizing samples that lie close to class boundaries. The objective is to enforce a minimum separation between positive and negative pairs in the embedding space, thereby strengthening class boundaries. From a geometric perspective, this loss increases inter-class separation and reduces ambiguity near decision regions. For IMU-based gesture classification, where sensor noise and similarity in motion patterns can cause class overlap, margin-based loss improves robustness by pushing embeddings away from boundary regions. However, excessive margin enforcement can over-constrain the embedding space, potentially limiting flexibility when natural overlap exists between certain gesture classes.

5) Methodology:

5.1 Dataset Description

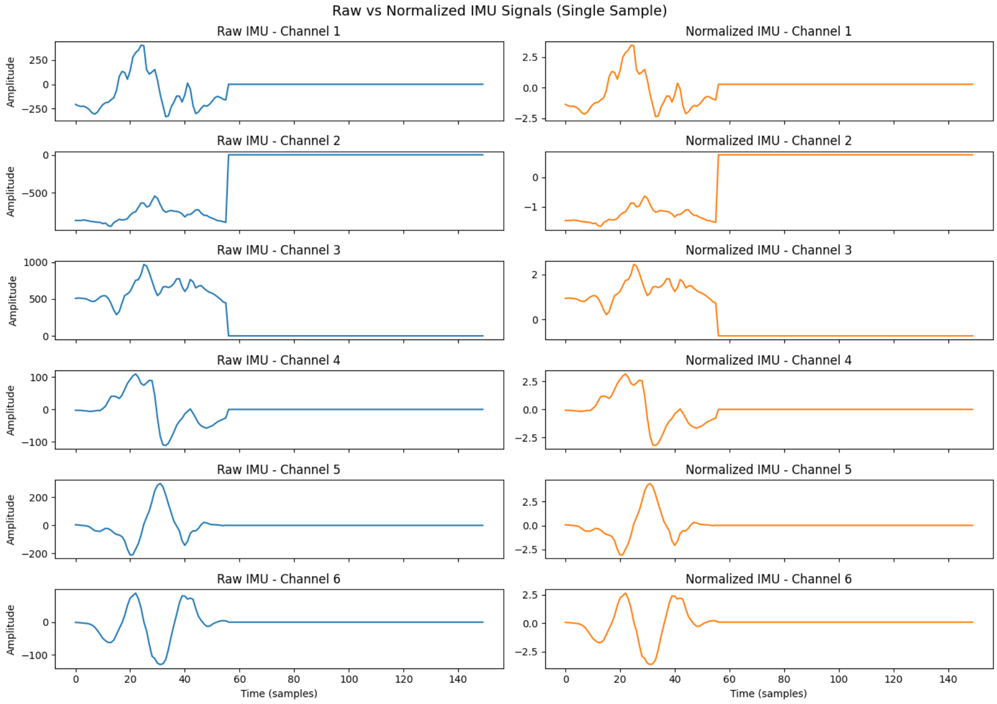


```
62 classes User Dependent case
----------------------------------------
Total samples        : 11160
Time steps per sample: 150
Channels     : 6
Number of classes    : 62
Label range          : 0 to 61
62 classes User Independent case
----------------------------------------
Total samples        : 11780
Time steps per sample: 150
Channels     : 6
Number of classes    : 62
Label range          : 0 to 61
47 classes User Independent case
----------------------------------------
Total samples        : 8930
Time steps per sample: 150
Channels     : 6
Number of classes    : 47
Label range          : 0 to 46
47 classes User Dependent case
----------------------------------------
Total samples        : 8460
Time steps per sample: 150
Channels     : 6
Number of classes    : 47
Label range          : 0 to 46
```

5.2 Data Preprocessing and Normalization

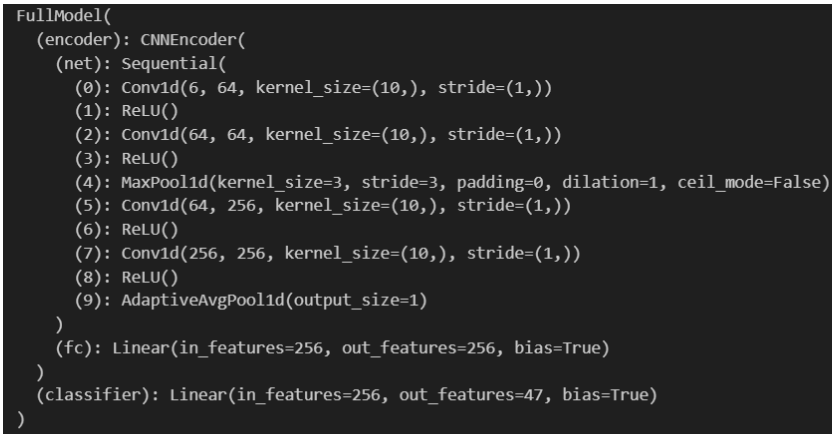
The preprocessing steps includes:

- Temporal downsampling of IMU signals to reduce sequence length and computational complexity.
- Channel-wise normalization was performed for each sample by subtracting the mean and dividing by the standard deviation of every sensor channel.
- Conversion of one-hot encoded labels into integer class indices suitable for loss computation.
- No aggressive filtering or handcrafted feature extraction was applied
- The normalized IMU signals were converted into tensor format and paired with their corresponding class labels for training and evaluation.



5.3 Model Architecture

A one-dimensional CNN was used as the core model for IMU signal analysis. Each input sample consists of six channels corresponding to three-axis accelerometer and three-axis gyroscope signals recorded over a fixed time window. The network comprises four convolutional layers that extract temporal motion features, followed by non-linear activations. The output is a fixed-length embedding vector used for classification or optimized using embedding-based loss functions. The architecture was kept fixed across all experiments.



```
FullModel(
  (encoder): CNNEncoder(
    (net): Sequential(
      (0): Conv1d(6, 64, kernel_size=(10,), stride=(1,))
      (1): ReLU()
      (2): Conv1d(64, 64, kernel_size=(10,), stride=(1,))
      (3): ReLU()
      (4): MaxPool1d(kernel_size=3, stride=3, padding=0, dilation=1, ceil_mode=False)
      (5): Conv1d(64, 256, kernel_size=(10,), stride=(1,))
      (6): ReLU()
      (7): Conv1d(256, 256, kernel_size=(10,), stride=(1,))
      (8): ReLU()
      (9): AdaptiveAvgPool1d(output_size=1)
    )
    (fc): Linear(in_features=256, out_features=256, bias=True)
  )
  (classifier): Linear(in_features=256, out_features=47, bias=True)
)
```

6) Training Strategy:

The following training strategies were used:

- The model was trained using the Adam optimizer with a fixed learning rate.
- Mini-batch training was used to efficiently train the model on the IMU dataset.
- An early stopping strategy based on validation accuracy was employed to prevent overfitting.
- For each user-independent data split, the model was trained and evaluated independently.

- Final performance was obtained by aggregating results across all data splits to ensure stability and reproducibility.

7) Experimental Work and Observations:

The IMU dataset was divided into multiple user-independent train–test splits. Four experimental cases were evaluated by varying only the loss function. Supervised contrastive loss produced structured embeddings but required diverse batch composition. Triplet loss was sensitive to sample selection, while margin-based loss improved class separation but required careful tuning.

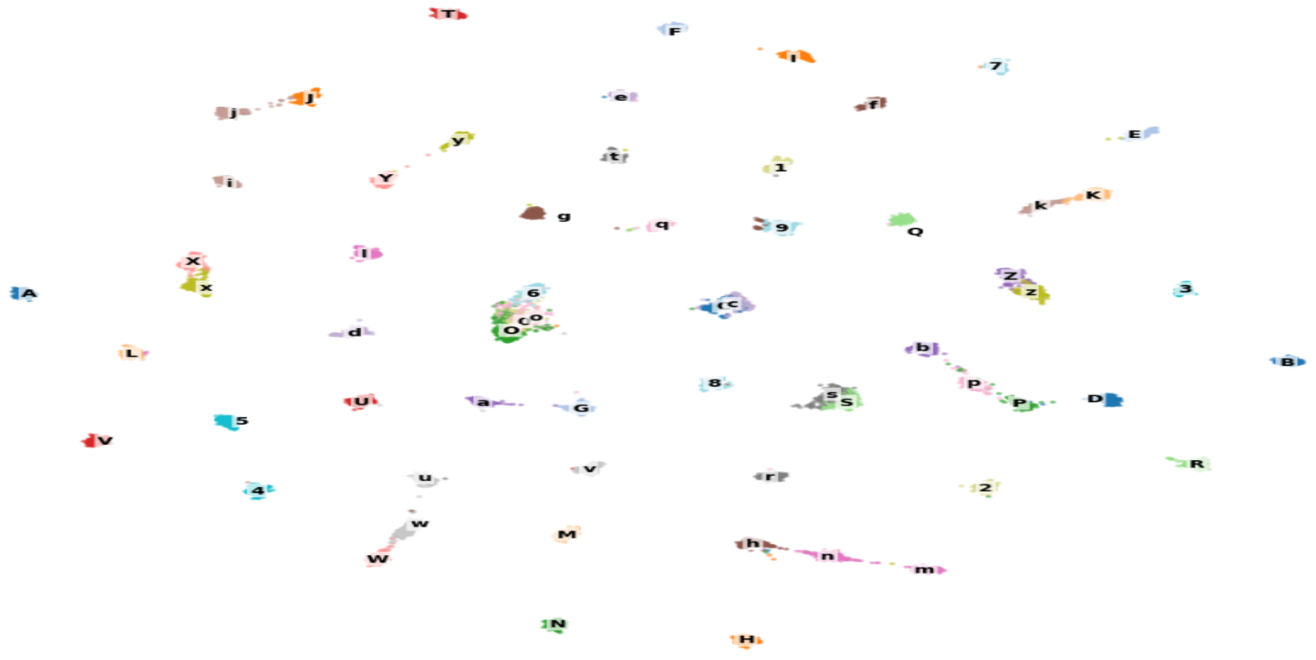
8) Results and Discussions:

Sr. no	Experimental Cases	Cross Entropy Loss	Supervised Contrastive Loss	Triplet Loss	Margin Loss
1	User Independent (62 classes)	59.36	62.13	L2: 58.47 Cosine:59.13 NTXent:59.53	59.20
2	User Dependent (62 classes)	82.93	87.02	L2: 86.43 Cosine: 85.91 NTXent:87.06	85.45
3	User Dependent (47 classes)	86.68	90.42	L2: 68.00 Cosine: 67.12 NTXent:68.91	89.04
4	User Independent (47 classes)	66.23	69.51	L2:90.95 Cosine:90.38 NTXent: 90.5	66.9

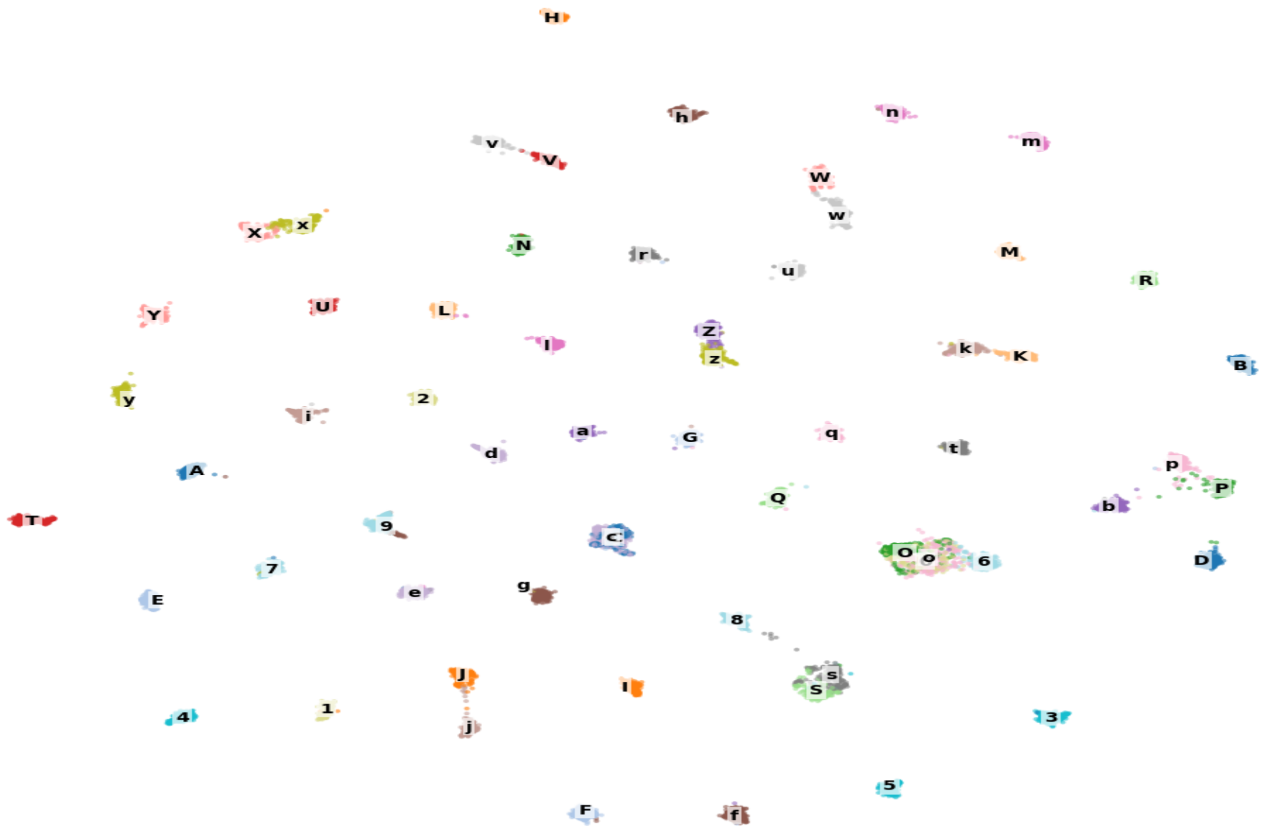
The above table represents the accuracy of CNN model with different loss functions in four experimental cases.

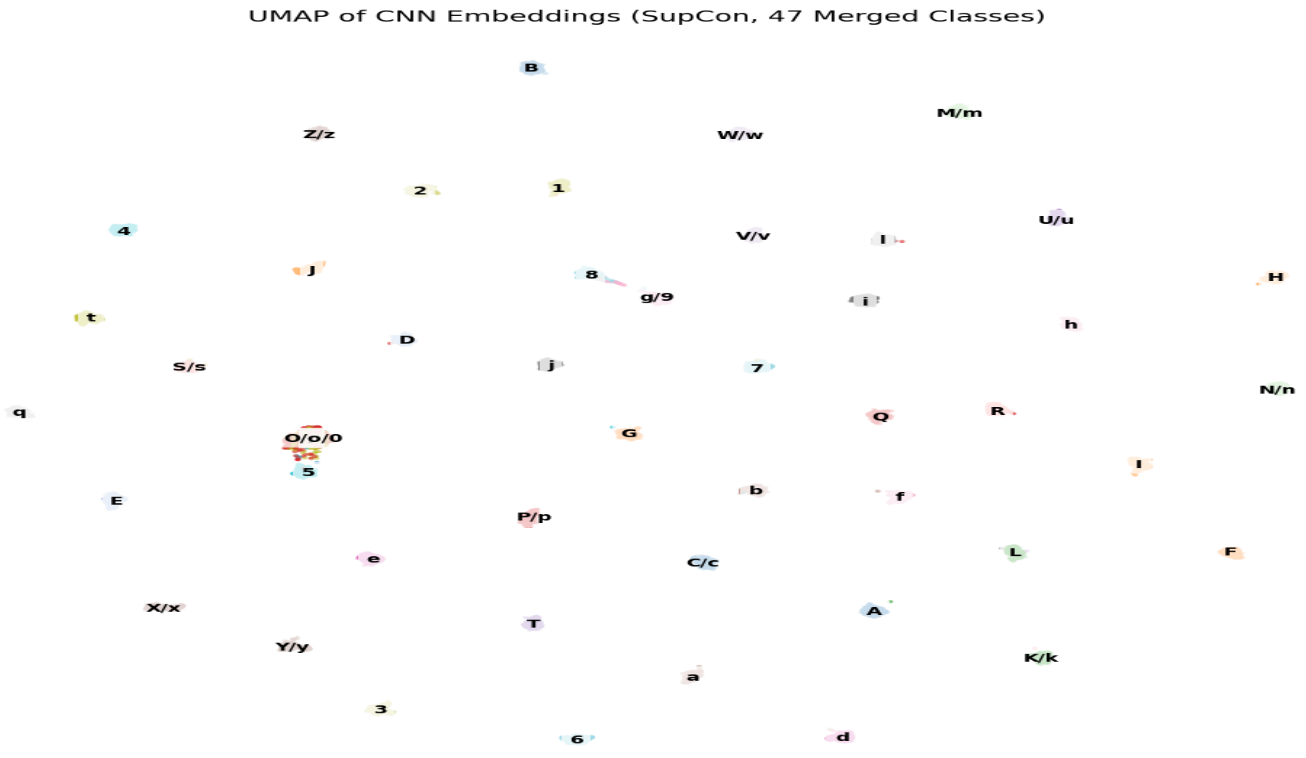
To further analyze the learned representations, embedding visualizations were generated using class-wise color plots. Since supervised contrastive loss achieved the highest classification accuracy, embedding analysis was focused on this best-performing setting.

UMAP of CNN Embeddings (SupCon, 62 Classes),User Independent

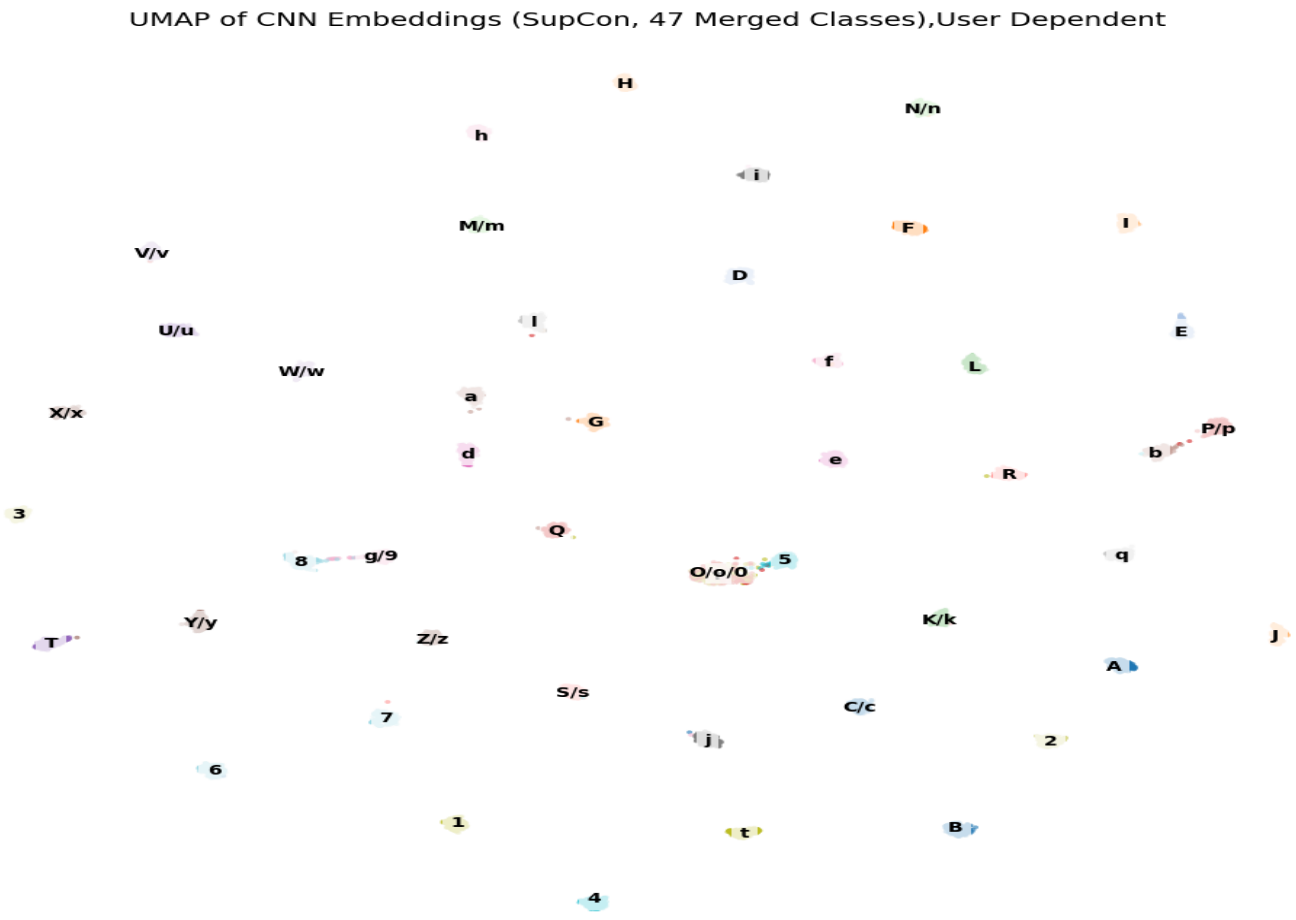


UMAP of CNN Embeddings (SupCon, 62 Classes,User Dependent)





UMAP of CNN Embeddings (SupCon, 47 Merged Classes)



UMAP of CNN Embeddings (SupCon, 47 Merged Classes),User Dependent

9) Challenges faced during the work:

- One of the primary challenges was the high level of noise and variability present in IMU signals. Factors such as sensor placement, variation in motion execution speed, and inter-subject movement differences resulted in significant variation across samples. This made it difficult to achieve consistently high classification accuracy, even with deep learning–based models
- Another major challenge involved training stability, particularly when using embedding-based loss functions. Losses such as supervised contrastive and triplet loss were sensitive to batch composition and sampling strategy. In some cases, improper batch structure led to unstable convergence or slow training, requiring careful tuning and validation.
- GPU-related debugging issues were also encountered during implementation. Errors related to label formatting and tensor indexing caused runtime failures during training. These issues were resolved by validating the data pipeline on CPU and ensuring consistency between model outputs and loss function requirements.
- Additionally, the limited size of the IMU dataset posed a challenge for model generalization. With a relatively small number of samples, the risk of overfitting was high, necessitating the use of early stopping and consistent hyperparameter settings across experiments.

10) Conclusion:

This internship project explored **IMU-based gesture recognition** using a deep learning framework, with a specific focus on representation learning through embedding-based loss functions. A fixed convolutional neural network architecture was used throughout the study to ensure that the effect of different loss functions could be analyzed in a controlled and fair manner.

Four experimental cases were evaluated by varying only the loss functions while keeping the model architecture, preprocessing steps, and training strategy unchanged. Among the evaluated approaches, **supervised contrastive loss** achieved the best classification performance and produced well-structured embedding spaces, as observed through class-wise color plot visualizations. **Triplet loss** and **margin-based loss** also improved embedding structure but showed greater sensitivity to training conditions and sampling strategies.

The results demonstrate that **loss function design plays a significant role in shaping learned representations for IMU time-series data**. While embedding-based loss functions improve feature clustering and interpretability, overall classification performance remains constrained by the noisy and user-dependent nature of IMU signals. This highlights the importance of focusing not only on model architecture but also on optimization objectives when designing deep learning–based IMU gesture recognition systems.

Overall, the project provides practical insights into **representation learning for motion sensor time-series data** and emphasizes the value of embedding analysis in understanding and improving deep learning model behavior.