Variational Latent Clustering with Sample-Weighted Classification for Cross-User EMG Gesture Recognition

Daehee Koo (20211011)
CSE, UNIST
Ulsan, South Korea
goodday@unist.ac.kr

Abstract

In electromyography (EMG)-based hand gesture recognition, inter-user variability poses a major challenge, often leading to poor generalization across subjects. This paper presents a novel framework that enhances cross-user robustness by combining feature compression, clustering, and weighted classification. Time-domain and frequency-domain features are first extracted from raw EMG signals and compressed into a latent space via a variational autoencoder (VAE). In the learned latent space, data points are clustered to capture common gesture patterns across users. Each sample is assigned a weight inversely proportional to its distance from the corresponding cluster centroid, which emphasizes prototypical instances and down-weights outliers.

A Random Forest classifier is then trained using these weights, effectively focusing the model on representative examples. Experiments on a public cross-user EMG gesture dataset (NinaPro) demonstrate that the proposed method significantly improves recognition performance. In cross-user evaluation, The proposed approach achieved a performance improvement of 5.07% in macro-averaged precision(37.29% versus 32.22%) and 1.07% in macro-averaged recall(32.37% versus 31.30%). These results indicate that leveraging latent clustering and sample weighting can enhance generalization across users in EMG gesture recognition.

1 INTRODUCTION

Gesture recognition plays a vital role in the development of natural and seamless human-computer interaction, particularly in applications involving wearable devices, assistive systems, and virtual environments. Among various sensing modalities, surface electromyography (sEMG) is widely used due to its ability to capture neuromuscular activity directly at the source of movement. Electromyography (EMG) is a technique for recording the electrical activity generated by skeletal muscles during contraction. And Surface electromyography (sEMG) is a non-invasive technique that measures the electrical activity of muscles using electrodes placed on the skin surface.

This physiological basis enables real-time and low-latency gesture interpretation, offering advantages over vision-based methods which are susceptible to lighting and occlusion. Furthermore, sEMG sensors can be embedded in compact wearable systems, making them highly suitable for practical use in prosthetics and ubiquitous computing. As such, EMG-based gesture recognition has become a promising approach for intuitive control interfaces.

EMG-based gesture recognition systems typically achieve very high accuracy when trained and tested on the *same* user under controlled conditions. However, this within-user performance sharply contrasts with *cross-user* results: when a model trained on one individual is applied to another, accuracy often drops dramatically. In one early study, the authors observed an average drop of approximately 57% in classification accuracy on unseen users [7]. More recent evaluations confirm this performance gap: for instance, LSTM-based cross-subject models achieved only around 93% accuracy on new users compared to about 98% for within-subject testing [2]. This substantial performance degradation clearly illustrates the generalization challenge in myoelectric gesture recognition.

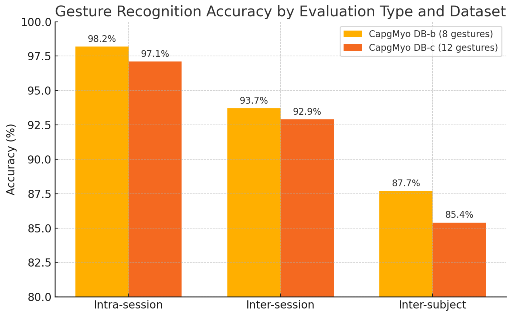

Figure 1: Comparing gesture recognition accuracy under three evaluation settings using CapgMyo DB-b and DB-c. [3]

Several factors contribute to the poor cross-user generalization of EMG classifiers:

- **Physiological and anatomical variability between users.** Different individuals have different muscle properties and limb anatomies. EMG signals are known to vary with fiber type, blood flow, skin impedance, and other physiological factors [2].
- **Variability in sensor placement and environmental conditions.** Small changes in electrode location, orientation, contact pressure, or limb posture can alter the recorded EMG waveform. In practice, it is difficult to place electrodes in exactly the same positions across users or even across sessions for the same user. Studies show that classification accuracy degrades when the arm or sensor position shifts [4].

This cross-user generalization problem has motivated new approaches. Campbell et al. noted that "although confounding factors like electrode shift make EMG pattern recognition unstable, some common information exists between users" [2]. Zheng et al. likewise emphasized that without adaptation, models cannot generalize due to individual variability [7]. In summary, robust multi-user EMG gesture recognition critically depends on identifying or learning features that capture the underlying muscle activation invariantly, despite the anatomical, attachment, and temporal differences between users. This necessity motivates the present investigation of methods (e.g., dimensionality reduction and clustering-based weighting) aimed at extracting such user-invariant representations.

This paper proposes a new framework for extracting user-invariant representations from EMG signals by combining variational autoencoder (VAE)-based dimensionality reduction with a clustering-guided weighting mechanism. The VAE is employed to retain the most salient features of the EMG data while reducing dimensionality, and clustering is used to assign weights to each data point inversely proportional to its distance from the cluster centroid. This reduces the influence of outlier or atypical samples during model training. Experimental results on a common EMG dataset demonstrate that the proposed method outperforms existing approaches in terms of generalization and recognition performance.

2 RELATED WORK

Zheng et al. [8] address the user-independence problem with an adaptive learning method. They extract muscle-synergy features from a source user to form an initial training set, and use an adaptive k-nearest neighbors classifier to label new user samples. Qualified new samples (filtered by a risk evaluator) are then incrementally added to the training set to update the KNN model [8]. On Ninapro benchmark datasets, this method achieved roughly 68–83% recognition accuracy across gesture classes [8], comparable to user-specific models. However, adaptation is gradual: accuracy improves only slowly as more data are collected [8]. Moreover, the method depends critically on the risk-evaluator threshold to select new samples and does not include any additional dimensionality-reduction step beyond the muscle-synergy features.

Zhang et al. [6] take a different approach by fusing multimodal signals. They combine surface EMG with hand acceleration data using a dual-stream convolutional neural network equipped with a spatial-attention module (SA-CSAM) to integrate features from both modalities. This spatial-attention CNN yields higher gesture classification accuracy on their dataset compared to single-modality baselines [6]. However, the model is relatively complex and computationally expensive, and the experiments were limited to single-subject data, so cross-user generalizability was not demonstrated. The approach also relies on obtaining high-quality signals from both sensors, which may be difficult to guarantee in practice.

In summary, existing methods either perform adaptive updating [8] or fuse additional sensor data [6], but they do not explicitly learn a user-invariant representation with principled dimensionality reduction and robust outlier handling. Unlike Zheng et al. [8] or Zhang et al. [6], our approach explicitly enforces a compact, user-invariant latent embedding via a variational autoencoder (VAE) and incorporates a clustering-based weighting scheme to mitigate outlier influence. This combination of representation learning, dimensionality reduction, and systematic outlier management is designed to improve cross-user generalization beyond prior methods.

3 PROBLEM STATEMENT

Electromyography (EMG)-based gesture recognition has shown high accuracy in within-user settings, but suffers a significant drop in performance when models are applied to unseen users. This cross-user degradation arises from inter-subject variability in muscle anatomy, skin impedance, and electrode placement, which causes the same gesture to produce different EMG patterns across individuals. While various adaptation and domain-transfer methods have been proposed, they often require additional calibration data or impose high computational costs. Furthermore, most existing models operate on high-dimensional features without explicitly reducing noise or emphasizing prototypical gesture patterns.

To address this gap, this paper propose a novel framework that combines variational autoencoder (VAE)-based dimensionality reduction with clustering-guided sample weighting. This method compresses EMG features into a low-dimensional latent space using a VAE, then assigns weights to each training instance based on its distance to gesture-specific cluster centroids. This paper hypothesize that this structure-aware, relevance-weighted approach improves the generalization ability of classifiers across users.

In summary, this study aims to experimentally verify the impact of VAE-based dimensionality reduction and clustering-based sample weighting on generalized gesture recognition.

4 ALGORITHM

4.1 EMG Feature Extraction

I segment the raw multichannel EMG data into fixed-length windows and compute standard time- and frequency-domain features for each window and channel. Specifically, I extract the following:

- **Time-domain features:** Mean absolute value (MAV), root mean square (RMS), variance (VAR), zero-crossing count (ZC), slope sign changes (SSC), waveform length (WL), and Willison amplitude (WAMP), as well as skewness and kurtosis.
- **Frequency-domain features:** Mean frequency (MNF), median frequency (MDF), total power M_0, and spectral moments M_2 and M_4.

These features capture the amplitude distribution, waveform complexity, and spectral characteristics of the EMG signal, and are commonly used in EMG pattern recognition.

4.2 Variational Autoencoder for Dimensionality Reduction

I train a variational autoencoder (VAE) on the extracted features to obtain a low-dimensional, structured latent representation. The encoder network maps an input feature vector x to a latent Gaussian distribution $q_\phi(z|x)$ with mean and variance parameters, while the decoder reconstructs the input from a latent sample z. In my implementation, the encoder has two hidden layers (e.g. 128 and 64 units) that produce μ and $\log \sigma^2$ for a latent dimension d_z, and the decoder is a symmetric multilayer perceptron. The VAE is trained by minimizing the evidence lower bound (ELBO), which combines a reconstruction loss with a KL-divergence regularization. Concretely, I minimize

$$\mathcal{L}_{\text{VAE}}(x) = \mathbb{E}_{q_\phi(z|x)}\left[\|x - \hat{x}\|^2\right] + D_{KL}\big(q_\phi(z|x) \parallel p(z)\big),$$

where $\hat{x}$ is the decoder output and $p(z) = \mathcal{N}(0, I)$ is the prior. This loss encourages the model to accurately reconstruct input features while imposing a smooth Gaussian structure on the latent space.

4.3 Latent-Space Clustering

After training, I encode all samples into their latent means z_i. For each gesture class, I then perform several clustering method on the latent vectors belonging to that class. The objective of this study is to investigate the impact of weight assignment through clustering. Therefore, various clustering methods are applied to verify whether this approach is a generally applicable idea. This step partitions each class's latent vectors into K clusters, identifying prototypical centroids for each gesture. Then compute the Euclidean distance d_i from each latent vector to the centroid of its assigned cluster.

After clustering, I assess whether the data points are evenly distributed across the clusters. If not, I sort the clusters in descending order based on the number of data points they contain, retain only those clusters whose cumulative sum accounts for 80% of the total data, and exclude data points belonging to the remaining clusters from the training model. This has a positive impact on the generalized gesture recognition model by excluding data that are considered outliers.

4.4 Sample Weighting by Proximity

Using the gesture-wise cluster assignments, I compute a weight for each sample based on its normalized distance to the cluster centroid. For each sample i, let d_i be its Euclidean distance to the assigned cluster centroid, and let $\bar{d}$ be the maximum observed distance in its class. I define the normalized distance as $\tilde{d_i} = d_i / \bar{d}$, and assign the weight as:

$$w_i = \frac{1}{1 + \exp\left(20 \cdot (\tilde{d_i} - 1.0)\right)}.$$

This sigmoid-based weighting function sharply decreases the weight for samples whose normalized distance exceeds 1.0, effectively downweighting outliers while preserving the influence of representative data near the centroid. Unlike simple inverse-distance schemes, this formulation introduces a soft boundary that suppresses the impact of outliers in a continuous and bounded manner. The constant in this weighting function was empirically determined to yield the most effective performance in practice. It has the added advantage of numerical stability and consistent interpretability in weighting gesture prototypes for classification training, as similarly advocated in soft-margin latent space regularization schemes.

4.5 Classification with Weighted Random Forest

Finally, I train a Random Forest classifier using the latent vectors z_i as input features and the gesture labels as targets, incorporating the sample weights w_i during training. In gesture recognition, real-time processing speed is also a factor that must be considered, so the number of decision trees was designed to be a relatively small number, 100. The Random Forest are trained with weight so that samples with larger w_i (closer to centroids) have greater influence on the model. This ensures that the classifier is primarily guided by the prototypical latent examples, potentially improving robustness to within-class variability.

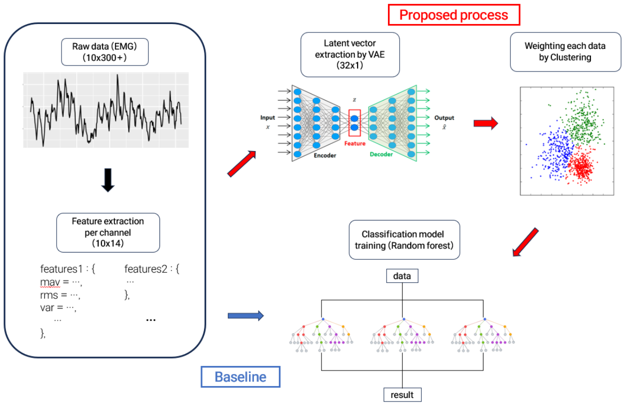

Figure 2: proposed process to training classification model for gesture recognition

4.6 Method Rationale

This pipeline is designed to leverage the VAE's structured latent space and to bias learning towards prototypical data. VAE maps high-dimensional data into a latent space, enabling it to capture complex patterns and correlations more effectively. Furthermore, by regularizing the latent representation through KL divergence, the VAE encourages the model to retain only prototypical features, thereby learning representations that are more robust to noise. [5] reported that VAE is effective for dimensionality reduction of time-series data. Clustering in this space identifies central modes of each class, and inverse-distance weighting then focuses training on these modes. By emphasizing samples near the learned centroids and de-emphasizing outliers, this method aims to reduce noise from atypical data and improve generalization across users.

5 EXPERIMENTS

5.1 Dataset

I used the Ninapro DB1 dataset, which consists of surface EMG (sEMG) recordings from 27 intact subjects performing 12 different gestures (including rest). Each gesture is repeated 10 times, and sEMG signals are recorded from 10 channels at a sampling rate of 2000 Hz. [1] I applied the following preprocessing steps:

- Notch filtering at 50 Hz to remove powerline noise.
- Band-pass filtering from 20–500 Hz using a 4th order Butterworth filter.
- Full-wave rectification.
- Segmentation by gesture label and repetition using MATLAB annotations.

5.2 Experiment Setting

I first extracted time-domain and frequency-domain features from the EMG signals using a sliding window (length: 200 samples, stride: 50). Each windowed segment was processed to compute statistical features such as mean absolute value, RMS, zero-crossing count, and spectral moments.

To evaluate cross-subject generalization, data from subjects 1 through 22 was used for training, and the rest subjects for testing.

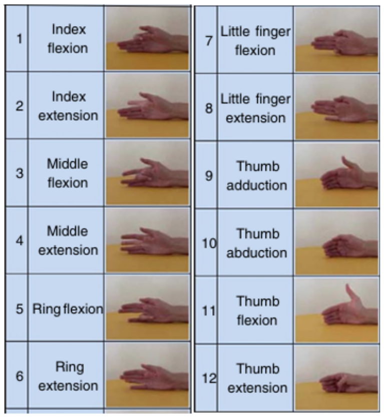

Figure 3: Gestures performed in the ninapro db1

The baseline model is a Random Forest classifier trained on this raw extracted features.

Then I applied a Variational Autoencoder (VAE) to reduce the dimensionality of the features (latent size: 32). The encoder network consists of two fully connected layers with 128 and 64 neurons, respectively, both using ReLU activation, and maps the input to a 32-dimensional latent space through separate mean and log-variance layers. A sampling layer then produces latent vectors via the reparameterization trick. The decoder mirrors the encoder with hidden layers of 64 and 128 neurons (ReLU), and a linear output layer reconstructs the input.

The latent vectors were then clustered using three different algorithms:

- K-means
- Gaussian Mixture Model (GMM)
- HDBSCAN

Since noisy data will be excluded based on the distribution of data after clustering, the number of clusters (in K-Means and GMM) is set to 10, a sufficiently large value to account for potential variability. In KMeans and GMM, clusters containing only a small number of samples were excluded from training based on the distribution of the data. For HDBSCAN, samples that were not assigned to any cluster (i.e., labeled as noise) were excluded from the training process.

Each training sample was assigned a cluster ID and a weight based on its distance to the cluster center, where samples closer to the center received higher weights. These weights were used to train weighted Random Forest classifiers. The random forest classifiers is designed to have 100 decision tree during training.

5.3 Experiment Result

The classification performance on the test subjects is summarized in Table 1.

The observed improvement in precision across all VAE-based methods, particularly with HDBSCAN, suggests that the proposed weighting mechanism effectively suppresses noisy or ambiguous samples during training, leading to more confident and accurate predictions. In contrast, the relatively modest gains in recall indicate that the models still struggle to correctly identify a broader range of true positive instances, possibly due to class imbalance or latent space sparsity. The notably higher precision in the HDBSCAN-based model may stem from its ability to assign more samples to noise, thereby removing uncertain or borderline data from the training set and focusing the classifier on more prototypical examples. While this improves prediction accuracy for included samples, it may limit the model's coverage and contribute to the smaller gains in recall.

The experimental data and code can be accessed by following github repository: github.com/DH-Koo/CSE304_20211011

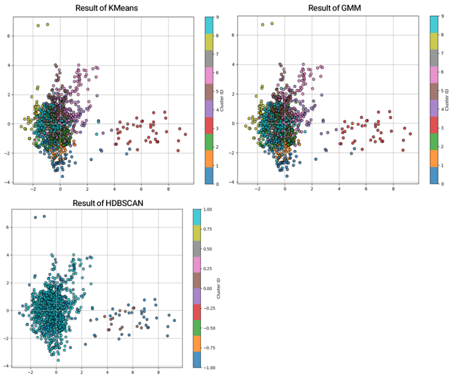

Figure 4: Clustering result of KMeans, GMM, HDBSCAN

Table 1: Classification accuracy and macro F1-score for each method.

Method	**precision(%)**	**recall(%)**
Baseline (RF)	32.22	31.30
VAE + KMeans	35.13	32.37
VAE + GMM	33.75	31.49
VAE + HDBSCAN	37.29	31.94

6 CONCLUSION

In conclusion, this study introduces a sample-weighted classification framework that leverages variational latent clustering to improve cross-user EMG gesture recognition. By emphasizing prototypical samples in the latent space, the proposed method achieved some gains in precision (37.29% VAE+HDBSCAN vs. 32.22% baseline) and recall (32.37% VAE+KMeans vs. 31.30% baseline) on a cross-user evaluation.

Although HDBSCAN yielded the highest precision, its clustering quality was limited, which likely contributed to lower classification performance. This approach is expected to be applicable to existing EMG-based gesture recognition systems. While promising, performance remains affected by gesture variability and was only validated on a single dataset (NinaPro). Future research should explore integrating this method into high-performing models to evaluate its potential for further improvements.

References

[1] Manfredo Atzori, Arjan Gijsberts, Claudio Castellini, Barbara Caputo, Anne-Gabrielle Mittaz Hager, Simone Elsig, Giorgio Giatsidis, Franco Bassetto, and Henning Müller. 2014. Electromyography data for non-invasive naturally-controlled robotic hand prostheses. *Scientific data* 1, 1 (2014), 1–13.

[2] Evan Campbell, Angkoon Phinyomark, and Erik Scheme. 2021. Deep cross-user models reduce the training burden in myoelectric control. *Frontiers in Neuroscience* 15 (2021), 657958.

[3] Md Rabiul Islam, Daniel Massicotte, Philippe Massicotte, and Wei-Ping Zhu. 2024. Surface EMG-based inter-session/inter-subject gesture recognition by leveraging lightweight all-ConvNet and transfer learning. *IEEE Transactions on Instrumentation and Measurement* (2024).

[4] Lihong Jin, Jehan Yang, Douglas J Weber, and Zackory Erickson. [n. d.]. Ref-EMGBench: Benchmarking Reference Normalization for Electromyography Data. ([n. d.]).

[5] Min Hyuk Lim, Young Min Cho, and Sungwan Kim. 2022. Multi-task disentangled autoencoder for time-series data in glucose dynamics. *IEEE Journal of Biomedical and Health Informatics* 26, 9 (2022), 4702–4713.

[6] Shenke Zhang, Wenjie Chen, Xiantao Sun, and Cheng Zhang. 2024. Improving gesture recognition accuracy with multimodal signals based on fusion of surface EMG and acceleration signals. In *Proceedings of the 2024 5th International Symposium on Artificial Intelligence for Medicine Science*. 563–568.

[7] Nan Zheng, Yurong Li, Wenxuan Zhang, and Min Du. 2022. User-independent emg gesture recognition method based on adaptive learning. *Frontiers in Neuroscience* 16 (2022), 847180.

[8] Nan Zheng, Yurong Li, Wenxuan Zhang, and Min Du. 2022. User-independent emg gesture recognition method based on adaptive learning. *Frontiers in Neuroscience* 16 (2022), 847180.