Activity Classification using Myo Gesture Control Armband data through Machine Learning

Kuntal Kumar Pal[1], Pratyay Banerjee[2], Sandipan Choudhuri[3], Shailaja Sampat[4]
{kkpal[1], pbanerj6[2], schoud13[3], ssampa17[4]}@asu.edu
Arizona State University
Tempe, AZ, USA

ABSTRACT

Activity classification is a task where we need to identify a sequence of gestures for a period of time. It is a challenging task without visual cues and only based on hand movements. There are several applications of activity classification without visual cues in science and technology, and in this paper we propose a solution based on EMG and IMU features from Myo Gesture Control Armband. We try to capture the temporal features of different hand gestures in multiple ways and apply machine learning and new deep learning techniques. Our approach is very promising and we are able to distinguish Eating activity from other activities with 94.76% accuracy.

Author Keywords

Eating Activity Recognition; Activity Classification; Myo Gesture Control Armband; Electromyography (EMG); Inertial Measurement Unit (IMU); Statistical Machine Learning; Deep Learning;

INTRODUCTION

Gesture recognition have been a difficult task for automated systems in the past. But with the developments of gadgets to measure different features needed for the task, have paved the way for successful gesture recognition. electromyography(EMG) helps to monitor the health of muscles and nerve cells controlling them. Inertial measurement units(IMU) records data for bodily functions. Gadgets like myo wrist bands measures exactly same kinds of data from which various bodily gestures can be recognized with proper machine learning techniques.

However any kind of gesture recognition from various parts of our body can be a tricky task and it needs proper devices. So we focus on a smaller subset of the problem with hand gesture recognition task. The Myo gesture control armband can collect all the necessary kinds of data needed for the purpose.

Again, there can be multiple kinds of hand gestures which can be seen while performing different kinds of activities. Some of the activities like eating and brushing can have similar kinds of muscle tension(EMG) data but may have different orientations which is recorded by IMU data. So it is essential to use both kinds of data to recognize an activity.

We worked on a particular subset of all such activities, to simplify the close similarities between many similar activities. Our work uses both the electromyogram and inertial measurement units to classify eating and non-eating gestures as each of the gestures are very much different from the other.

RELATED WORKS

There have been a number of works in hand gesture recognition using only EMG data[12]. Some researchers used artificial neural network[3] on EMG data to recognize hand motion while others have used unsupervised learning[5] for the same purpose. Apart from EMG data, some have used simple LSTM on IMU sensor data to recognize human activity [2]. Some other works like gesture recognition across all devices[11] along with eating and drinking gestures by body worn devices[1] and food intake gesture monitoring system based-on depth sensor [4] are also noteworthy. Apart from eating gestures, some other works like applying gesture of hand exercises for Post-Stroke Rehabilitation using myo band [14], recognizing grasped object via forearm electromyograpy data [8] and body activity tracking using wearable accelerometer[7] are also related to our work.

SYSTEM OVERVIEW

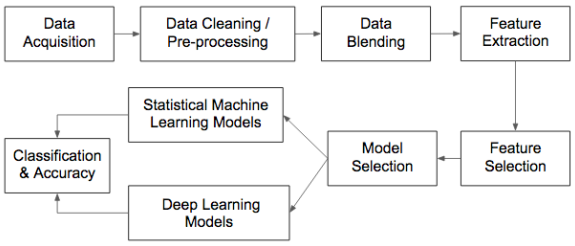

Figure 1. Proposed computing system for classification of eating and non-eating activities

The following sections cover each component of our overall system.

DATA ACQUISITION

First step towards performing activity classification is the data collection as machine learning models relies on significant data to derive inference from data patterns. Each member in our team wore the Myo Gesture Control Armband (referred to as myo device for the convenience in all subsequent sections) at the muscular area under elbow for two days. The arm band is paired with an android device using Bluetooth which records armband data every 5 minutes into files. Myo device collects data about human gestures through two embedded devices- EMG and IMU. EMG records the electrical activity produced by skeletal muscles roughly every 2 ms which is 1x8 sized integer tensor. Whereas, IMU measures body's specific force, angular rates using a combination of accelerometers

and gyroscopes as a 1x10 sized real tensor every 15-25ms. Data for both devices are logged with respective unix epoch timestamp format.

DATA PREPROCESSING

There are several possible reasons which results in malformed data collection such as loosened arm band, data logging discrepancies, bluetooth connection errors etc. Therefore, it is necessary to clean up such erroneous data samples in order to avoid adversarial impact of such data while learning process of a model. We incorporated two major steps here- removal of datapoints which contains missing (NaN) values and processing of malformed/incomplete data recordings. After this step, we have prepared the data in a structured format which will be processed through the proposed computing pipeline to train the machine learning models for activity classification which we are taking into consideration.

DATA BLENDING

Both the IMU and EMG data have been used for the purpose of eating activity recognition. Although same numbers of files are generated of each type with a particular time epoch, the data logged individually are of varying time-stamp. Since we plan on working with individual time-stamps logged in those files rather than the file level, we used an aggregation interval of one second time-frame for our work. This helped us to blend both IMU and EMG data together in a particular time-frame. We aggregated all the data across all 18 sensor readings using simple mean. A detailed **Feature Analysis** is present in the Appendix Section of the paper.

FEATURE GENERATION

For the statistical models, two types of features are used. First, the raw data generated by the sensors that is eight features from EMG and ten features from IMU data are used after preprocessing and blending. Second, we generated features by concatenating raw sensor data for a continuous time-frame of 5s, 15s, 60s each. This helps us to keep the temporal information of the sensor data.

FEATURE SELECTION

For the statistical machine learning algorithms we attempted two feature selection approaches.

All raw features

In this approach we kept all the features we generated.

L1 Distance Based Feature selection

We used model based feature selection techniques. The model we chose is linear Support vector classifier with L1 distance.

Tree-based Feature selection

In this model based feature selection techniques we used Extra tree classifier with 50 estimators as parameters.

DATASETS

We created two types of datasets for each of three time-intervals of 5s, 15s and 60s on which ran our statistical models. First a balanced dataset is created by taking as many non-eating gesture samples as there are eating gesture samples for the raw sensor data. There were 3181 eating gesture samples in our dataset. So the balanced dataset is of size 6362. An unbalanced dataset is also prepared with size 77670 with extra non-eating gesture samples which were not present in the balanced dataset.

MODELS

Statistical Machine Learning Models

Among the statistical machine learning models we used the following basic models with the various dataset prepared.

- **Support Vector Machines (SVM)**
  We used SVM with linear, polynomial and radial-basis function kernels for our experiments.
- **Random Forest**
  This is an ensemble model with decision tree classifier as base weak classifier.
- **XgBoost**
  Extreme Gradient boosting algorithm as defined in [6]. We use gbdt as weak classifiers.
- **LightGBM**
  Another Gradient boosting algorithm as defined in [10]. We use the default gbdt as weak classifiers.
- **Logistic Regression**
  We performed grid search to obtain optimal parameters, used them as argument to Logistic Regression function.
- **Gaussian Process Classifier (GPC)**
  GPC is a classifier based on Laplace approximation to get estimated value of non-Gaussian posterior by a Gaussian.

Deep Learning Models:

Since an event, like eating food, that occur in certain times of the day can depend on both *short* and *long-term* physical habits (for example, exercising might trigger the craving for food sooner than the regular time of meal), it will be wise to consider the temporal dependency aspect of the data as well. In other words, constructing the dataset in a time-distributed fashion will prove beneficial for capturing these temporal correlations. Prior research on time-series data [9],[13] suggests that along with identifying temporal correlations between different time-steps, *capturing links between pairs of time series (spatial)* is also essential to draw inference from a multi-variate time-series data. Therefore, in this section, we formulate methodologies that take both temporal and spatio-temporal aspects into consideration.

- **LSTM**
  Long Short-Term Memory (LSTM) shares resemblance with a recurrent neural network, (RNN) architecture utilized in deep-learning. Unlike standard fully-connected feed-forward neural networks, LSTM has feedback connections, using which it can not only process single data points, but also a sequences of data instances, such as audio signals, power-grid PMU data, etc. An LSTM unit comprises of a cell, an input gate, a forget gate and an output gate. The cell memorizes values over arbitrary time intervals, while the three gates controls the information flow in

and out of the cell. Since there are possibilities of lags of unknown duration between significant events in time series, LSTM networks are well-suited for such tasks. Relative insensitivity to gap length is an advantage of LSTMs over Recurrent Neural Networks, Hidden Markov models and other sequence learning methods in many tasks.

- **Attention-based Convolutional LSTM (Conv-LSTM)** ConvLSTM [15] is a deep learning architecture used when an application needs to capture spatio-temporal information in a time-series data, such as video sequence. However, the performance of Conv-LSTM is observed to deteriorate as the sequence length grows. To tackle this performance bottleneck, [16] developed an attention mechanism based Conv-LSTM which can adaptively select relevant hidden states (feature maps) across different time steps. Motivated by their work, we formulated a Conv-LSTM architecture that is augmented with attention layers. These layers assign greater priorities to relevant features discriminant for the classification problem.

  We formulated the task of prediction as: every label Y_s (eating/ non eating) of a specific profile is predicted using a series of X vectors $< X_{s-w_{max}}, X_{s-w_{max}-1}, .., X_s >$ with w_{max} being the largest time-frame of monitoring. The dataset, thus constructed, can be easily modified using feature repetitions for each volunteer to meet this purpose. Specifically, to capture the spatial information from n time-series, we construct feature-affinity matrices by utilizing pairwise inner-product of two time series within a variable time-frame w. To elaborate further, for two time-series $x_i^w = (x_i^{t-w}, x_i^{t-w-1}, , x_i^t)$ and $x_j^w = (x_j^{t-w}, x_j^{t-w-1}, , x_j^t)$ in a multivariate time series segment X_w $(i, j \in \{1, ..., b \times \frac{24}{p} \times d\})$, their correlation $m_{ij}^t \in M^t$ is calculated as:

$$m_{ij}^t = \frac{\sum_{\delta=0}^{w} x_i^{t-\delta} x_j^{t-\delta}}{w} \quad (1)$$

  Transforming features representations in this technique not only aids in Convolutional-Neural-Network processing, but also ensures that small-scale noise perturbations, that are not representative of human actions, exhibit lesser impact on the prediction mechanism. Additionally, framing the input features in this pattern bolsters capturing the compound effects of spatial, as well as temporal correlations. In order to characterize multi-scale feature information, we work with multiple value of w ($w = 2, 4, 6$ and consequently, $w_{max} = 6$), thereby producing three feature-affinity matrices for each instance of a prediction label.

- **Attention-based Prediction Model:**

  Given 10+8 IMU and EMG features at each time instance, the inputs to our model are $18 \times 18 \times 4$ affinity matrices. On the constructed matrices, a convolutional encoder is employed to encode the spatial patterns. An encoder architecture generally comprises of a series of convolutional and max pooling layers, with increasing number of filters at each convolution operation. Scaled Exponential Linear Unit (SELU) is used as the activation function in each convolutional block. Filtering temporal information from the spatial encodings of the convolutional layers is achieved via Convolutional-LSTMs. This is followed by an adaptive choosing of relevant hidden steps h, of the LSTMs that aid to form a refined output of informative feature maps. This is done by adding a temporal attention mechanism. We follow a variant of Zhang et. al.'s work [16] where the group-level context vector (the last hidden state) is used in a softmax function to compute the influence weights α of previous hidden states. Using these weights, aggregation of the informative feature maps are performed to generate refined output of features. The output feature map of every attention layer is concatenated with the feature maps of the previous layer. Finally, final concatenated feature map is passed to subsequent convolutional and fully-connected layers and a soft-max is performed over the outputs of the last layer. We use categorical cross-entropy loss to train our model over the supervised dataset and perform end-to-end learning. The efficacy of prediction can be measured using *percentage of accuracy*. A schematic diagram of the model is given in figure 4.

EXPERIMENTS

Statistical Machine Learning-Based Approaches

For the statistical machine learning algorithms we used accuracy, precision, recall and F1-measure as evaluation metric. We also represented the results of the binary classifier by its confusion matrix. We trained the classifiers on balanced data and tested on both balanced and unbalanced data. The intention that the model should learn to classify between both eating and non-eating data samples better with the balanced set and also learn to generalize on the bigger unbalanced data.

Balanced Dataset

Model	Raw	5s	15s	60s
SVM	87.17	83.24	85.18	85.28
Random Forest	84.23	83.97	85.18	88.48
Logistic Regression	68.52	67.89	67.57	68.52
Gaussian Process	87.48	79.78	79.61	85.61
XGBoost	89.31	92.46	93.19	**94.76**
LightGBM	89.47	93.03	93.92	93.92

Unalanced Dataset

Model	Raw	5s	15s	60s
SVM	85.39	88.83	79.02	77.27
Random Forest	81.13	87.58	87.46	88.72
Logistic Regression	65.25	66.49	67.11	68.67
Gaussian Process	85.21	77.61	76.98	86.50
XGBoost	89.95	93.51	93.17	**95.35**
LightGBM	89.95	93.83	93.68	93.88

Table 1. The evaluation scores of various statistical machine learning models for both balanced and unbalanced test data (All values are in %)

The results of our experiments with statistical models can be seen from the table 1 It can be seen that *Support Vector*, *Gaussian Process* and *Random Forest* classifiers works well(**~85%**)

across all the four types of data that we created for both balanced and unbalanced set. The decent performance can be attributed to the discriminatory features which we have plotted in Appendix. On the other hand the logistic regression is performing poorly with on ~65% on average over all the data. However we have achieved much better accuracy(**~95%**) with *boosting* since it uses separate weak classifiers fitted across various features internally. *XGBoost* outperforms all the statistical models across both the balanced and unbalanced data having temporal relations for 60s. This shows that the eating actions are well identified over a longer time interval. The confusion of the best statistical model have been provided on both the balanced(Figure 2) and unbalanced(Figure 3) dataset.

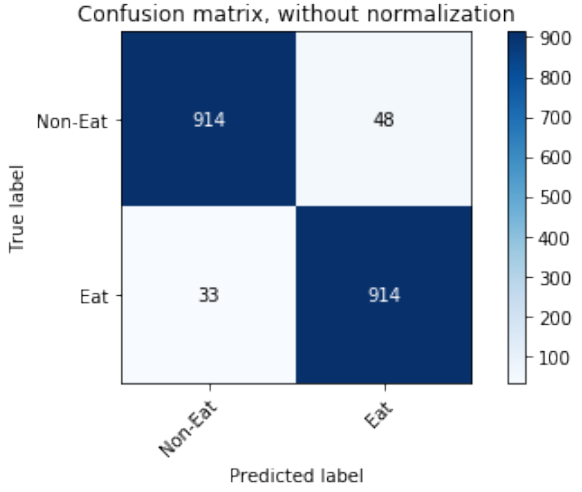

Figure 2. XGBoost (best performing statistical machine learning method) confusion Matrix on Test data from Raw balanced dataset

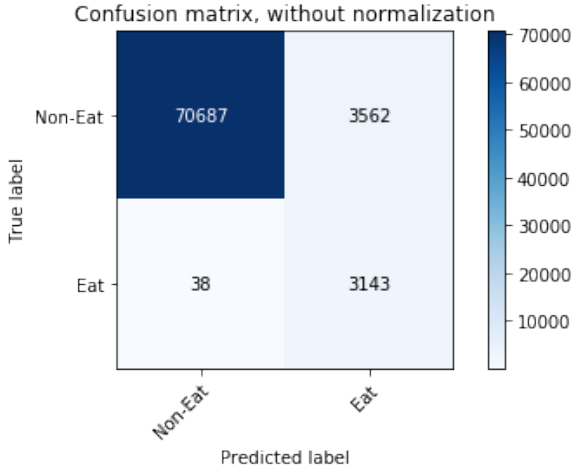

Figure 3. XGBoost (best performing statistical machine learning method) confusion Matrix on Test data from Raw unbalanced dataset

Deep-Learning-Based Approaches

For the deep-learning-based approaches we reported accuracy, precision, recall and F1-measure as metric. We also plotted the confusion matrices of some experiments with the binary classifiers. As deep learning algorithms require large amount of data to train, We trained and tested the classifiers on both balanced and unbalanced data. Similar to the motivation highlighted when experimenting with statistical-machine learning approaches, we went further to investigate whether the models learn to classify between both eating and non-eating data samples better with the balanced set and whether they can generalize the same on the bigger unbalanced data.

The layers used in building attention-based Conv-LSTM model are given sequentially as follows: 64 3×3 Conv-LSTMS with stride 1×1, max pooling with stride 2×2, 128 3×3 Conv-LSTMS with stride 1×1, max pooling with stride 2×2, 256 3×3 Conv-LSTMS with stride 1×1, attention layer, 5 1-dimensional convolutions of 512 3×3 filters, interleaved with max-pooling layers with filter size 2×2. This is followed by a fully-connected layer. The last layer produces a softmax output. *Categorical cross-entropy* is used as the loss function. Each model is trained for 80 epochs using a train-test split of 30%. The validation size was set to 20% of the training set.

From the experiments, it was noticed that DL models faced difficulties to generalize from scarce data. The training accuracy varied within range 96 to 99.8% for all the models, while the testing accuracies dropped considerably. Furthermore, the precision recall and F1-score were reasonably low for both the models when trained and tested on unbalanced data. One reason for this is the models learn to predict the majority class (true negative, in this case) frequently, thereby maintaining the accuracy, but affecting the remaining metrics (refer tables 2 and figures 5,6. In order to mitigate this issue, we trained our models on balanced dataset, but to no avail- the prediction accuracies over the test set were badly affected (ranging between 65-70%) even when the training accuracies were within 97-99.8%. To reduce overfitting, we introduced regularizers to the CNN and LSTM models, but it reduced the accuracies further. A major problem that was highlighted during the experiments were that the volume of data played a major role in data overfitting. Therefore, in order to improve the generalization power of the proposed models, a sophisticated dataset with large number of data instances are required.

Balanced Dataset

DL Model	Config	Accur	Preci	Recall	F1
LSTM	T=6, N=20	89%	.87	.90	.89
ConvLSTM	T=4	80%	.79	.81	.80

Unbalanced Dataset

DL Model	Config	Accur	Preci	Recall	F1
LSTM	T=6, N=20	97%	.78	.49	.60
ConvLSTM	T=4	96%	.55	.58	.57

Table 2. The evaluation scores of LSTM and Conv-LSTM model for both balanced and unbalanced test data for different values of time frames (T) and sizes of the cell (N) state.

Figure 4. Proposed Convolutional LSTM Based deep learning model architecture

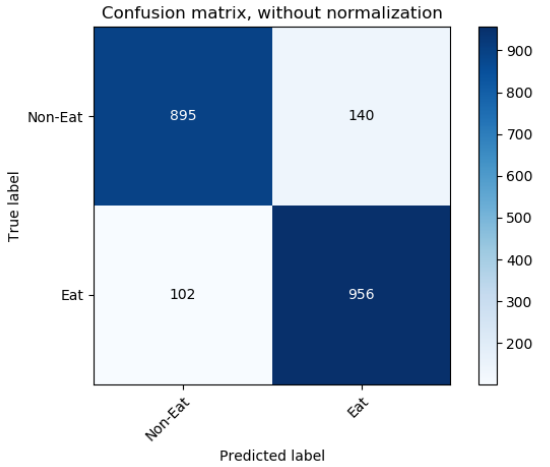

Figure 5. LSTM (best performing deep learning method) confusion Matrix on Test data from balanced dataset (T=6 and N=20).

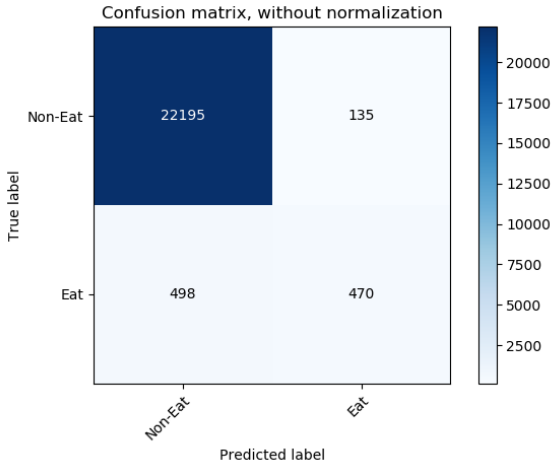

Figure 6. LSTM (best performing deep learning method) confusion Matrix on Test data from unbalanced dataset (T=6 and N=20)

CHALLENGES & LIMITATIONS

There were many challenges in correctly recognizing the eating gestures.

- People tend to do some unavoidable tasks during eating which obfuscate the original unique gestures of eating.
- The data we collected have 50% of eating gestures of left-handed candidates. So the patterns of data we collected for the left-handed candidates vary with those of the right-handed candidates.
- Also there were challenges like bluetooth connections getting disconnected and data collected needed proper cleansing which led to elimination of few data. This in turn lead to the loss of temporal coherence of data for which performance of our convolutional LSTM model got hampered.
- We collected data for two days which involved only five-six eating activities per person. With such a limited amount of data is very hard to train the deep neural models like traditional LSTMs and convolutional LSTMs.

COMPLETION OF TASKS

Sr.	Task	Assignee	%
1	Data Collection	All	5%
2	Data Preprocessing	All	10%
3	Correlation Analysis	All	10%
5-9	Feature Engineering	All	15%
10-14	SML Models Implementation - Model Selection - Model Preparation - Hyperparameter Tuning - Accuracy/Error Analysis	All	30%
15-19	Deep Models Implementation - Model Selection - Model Preparation - Hyperparameter Tuning - Accuracy/Error Analysis	All	25%
20	Final code & Report	All	5%
Member Contribution: Kuntal (25%), Pratyay (25%), Sandipan(25%), Shailaja (25%)			

Table 3. Task completion

****HEREBY WE AGREE THAT ALL TEAM MEMBERS HAVE EQUAL CONTRIBUTION FOR THIS PROJECT**

CONCLUSION AND FUTURE WORKS

The eating gesture recognitions get extremely hard if data quality and quantity requirements are not met. This is because

the need of precise distinction and relation between the data. So in our future work we intend to gather more data to balance eating and non-eating time-frames which we hope would improve our classification tasks. We would also like to use other preprocessing techniques to extract extra features which is essential for statistical models. We wish to precisely annotate the samples in time boundaries, that is the time-frames where the eating starts and where it ends which would improve the precision and recalls of our models.

ACKNOWLEDGEMENT

We would like to thank Dr. Ayan Banerjee and TA Junghyo Lee for their support, guidance and feedback for this project.

REFERENCES

1. Oliver Amft, Holger Junker, and Gerhard Troster. 2005. Detection of eating and drinking arm gestures using inertial body-worn sensors. In *Ninth IEEE International Symposium on Wearable Computers (ISWC'05)*. IEEE, 160–163.
2. Sara Ashry, Reda Elbasiony, and Walid Gomaa. An lstm-based descriptor for human activities recognition using imu sensors. In *Proceedings of the 15th International Conference on Informatics in Control, Automation and Robotics, ICINCO*, Vol. 1. 494–501.
3. Habiba Aziz, Mohsin Khan, Sajid Hussain, and Farah Haroon. Hand Motion Recognition from EMG using Artificial Neural Network. (????).
4. Muhammad Fuad bin Kassim11 and Mohd Norzali Haji Mohd. 2019. Food Intake Gesture Monitoring System Based-On Depth Sensor. *Bulletin of Electrical Engineering and Informatics* 8, 1 (2019).
5. Sushmalekha Shankar Birur. 2018. *Hand gesture recognition using unsupervised learning*. Ph.D. Dissertation. San Francisco State University.
6. Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 785–794. DOI: http://dx.doi.org/10.1145/2939672.2939785
7. Liang Dong, Jiankang Wu, and Xiang Chen. 2007. A body activity tracking system using wearable accelerometers. In *2007 IEEE international conference on multimedia and expo*. IEEE, 1011–1014.
8. Junjun Fan, Xiangmin Fan, Feng Tian, Yang Li, Zitao Liu, Wei Sun, and Hongan Wang. 2018. What is That in Your Hand?: Recognizing Grasped Objects via Forearm Electromyography Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 161.
9. David Hallac, Sagar Vare, Stephen Boyd, and Jure Leskovec. 2017. Toeplitz inverse covariance-based clustering of multivariate time series data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 215–223.
10. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*. 3146–3154.
11. Bryce Kellogg, Vamsi Talla, and Shyamnath Gollakota. 2014. Bringing gesture recognition to all devices. In *11th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 14)*. 303–316.
12. Ali Moin, Andy Zhou, Simone Benatti, Abbas Rahimi, George Alexandrov, Alisha Menon, Senam Tamakloe, Jonathan Ting, Natasha Yamamoto, Yasser Khan, and others. 2019. Adaptive EMG-based hand gesture recognition using hyperdimensional computing. *arXiv preprint arXiv:1901.00234* (2019).
13. Dongjin Song, Ning Xia, Wei Cheng, Haifeng Chen, and Dacheng Tao. 2018. Deep r-th root of rank supervised joint binary embedding for multivariate time series retrieval. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2229–2238.
14. Muhammad Syahrizal Widodo, Moh Zikky, and Artiarini Kusuma Nurindiyani. 2018. Guide Gesture Application of Hand Exercises for Post-Stroke Rehabilitation Using Myo Armband. In *2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)*. IEEE, 120–124.
15. SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*. 802–810.
16. Chuxu Zhang, Dongjin Song, Yuncong Chen, Xinyang Feng, Cristian Lumezanu, Wei Cheng, Jingchao Ni, Bo Zong, Haifeng Chen, and Nitesh V Chawla. 2018. A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data. *arXiv preprint arXiv:1811.08055* (2018).

APPENDIX

FEATURE SET VISUALIZATION

EMG Feature Set Visualization

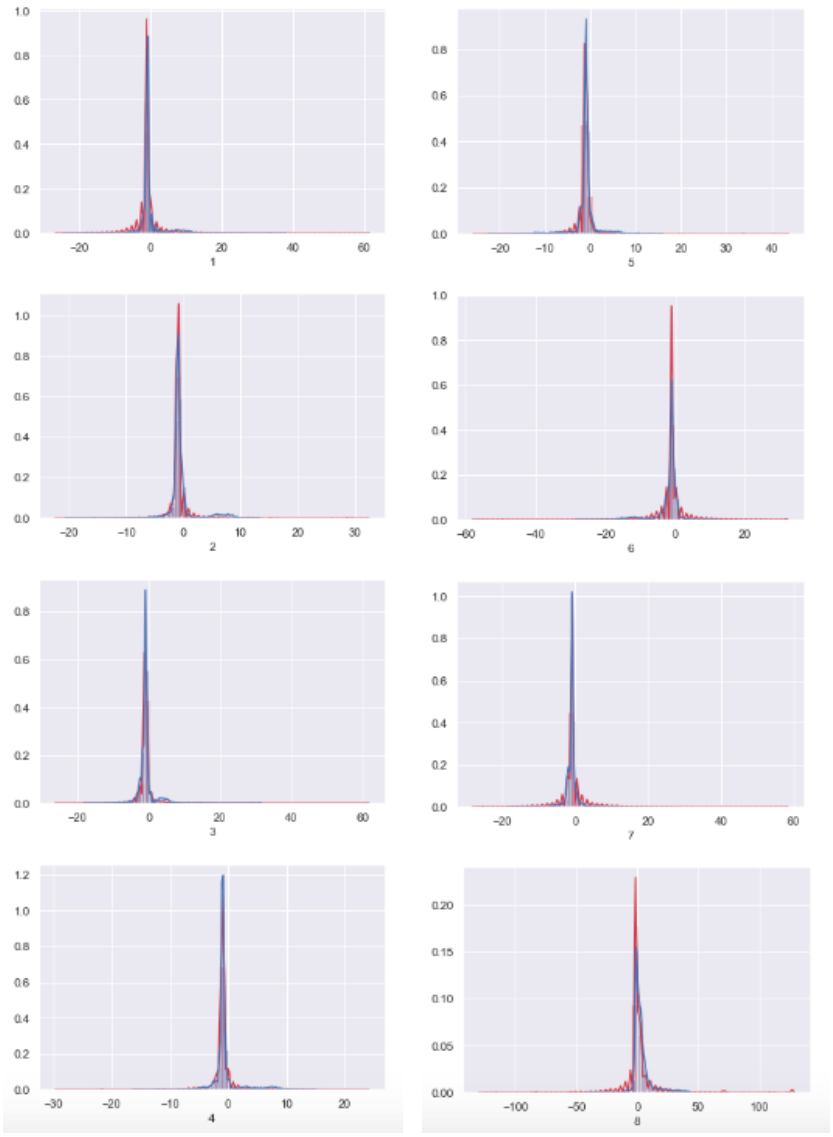

Figure 7. Visualization of 8 EMG Features: red values denote non-eating activity, whereas blue denote eating activity.

IMU Feature Set Visualization

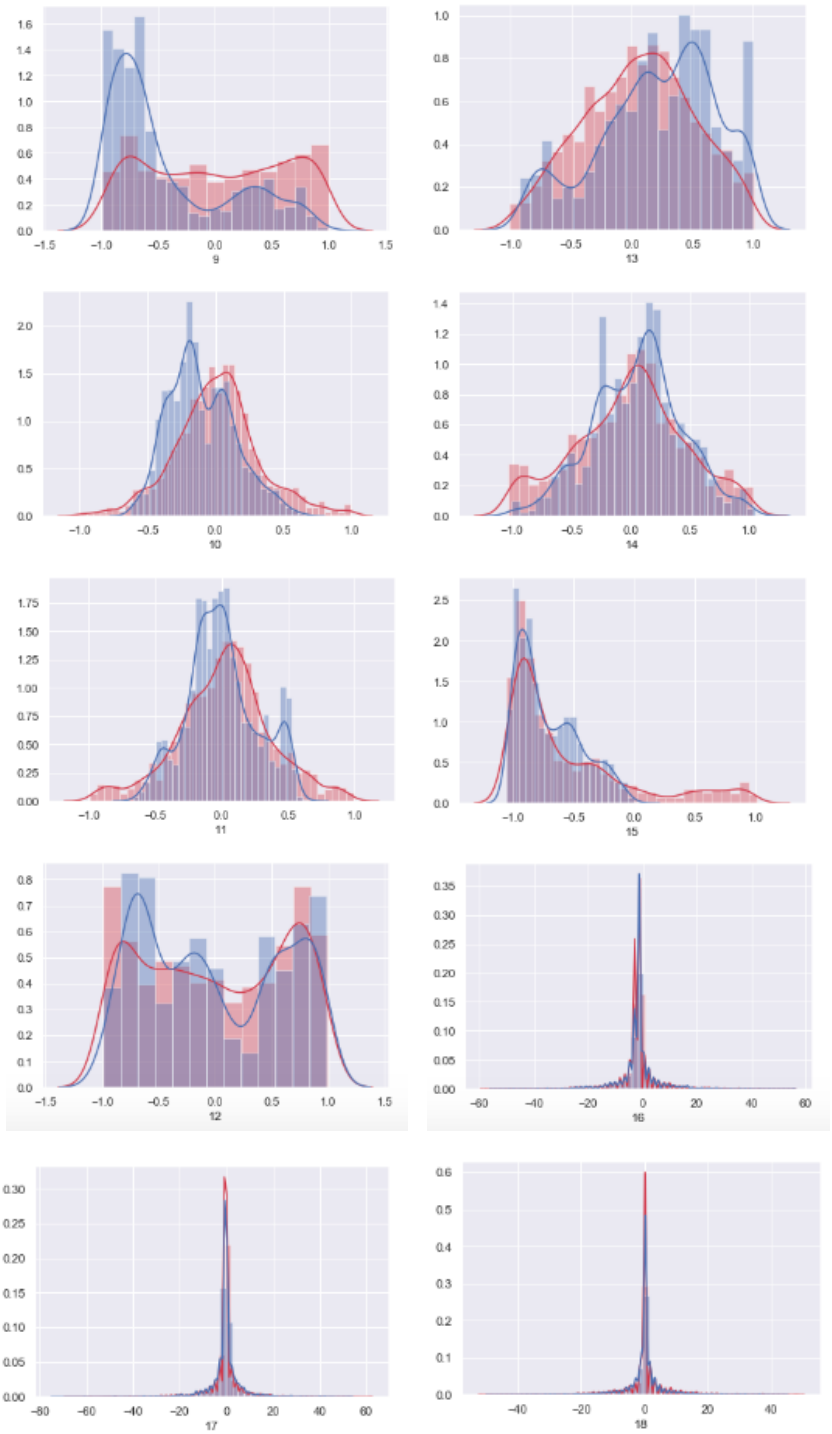

Figure 8. Visualization of 10 IMU Features: red values denote non-eating activity, whereas blue denote eating activity.