CS 5805 Fall 2025 Final Project Report

Lingyu Li
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061-0002
Lingyu99@vt.edu

Abstract

Work-related musculoskeletal disorders are closely tied to manual material handling, yet hand loads are rarely monitored continuously in real work environments. In this project, I explored whether forearm surface EMG, collected via a wireless armband, can be used to estimate hand load during lifting–lowering tasks using random forest models. Participants performed varied lifting tasks that differed in object type, weight, and lifting height. EMG signals were preprocessed, segmented into short windows and converted into time- and frequency-domain feature vectors, which were then used to train both regression and classification models. I evaluated performance using held-out trials so that entire lifting sequences, rather than individual windows, were unseen during testing. While the models captured some relationship between EMG patterns and hand load on the training data, their performance on held-out trials was modest and substantially lower than what has been reported in previous studies, which used a larger data set with more controlled lifting motions. These findings suggest that, with the current small and unconstrained dataset, forearm EMG alone is not sufficient for accurate hand-load estimation, but the proposed pipeline provides a reusable baseline for future work with larger datasets and additional sensors.

1. Introduction

(10 points) What did you try to do? What problem did you try to solve? Articulate your objectives using absolutely no jargon.

Work-related musculoskeletal disorders (WMSDs) are conditions gradually developed in the workplace that affect the supporting structures of the human body. In the occupations involved manual material handling, back WMSD is one of the most prevalent WMSD types, resulting in heavy economics burden to society from medical expenses and loss of production [2]. Repetitive lifting is one of the major risk factors to back WMSD, and existing evidence has shown a strong association between high physical demands and back WMSD [3, 7, 8]. In order to mitigate this risk factor, one important step is to identify the lifting tasks that involve high back WMSD risk factor. Compared to manual risk factor assessment (observation and evaluation by humans), automated risk factor evaluation is more efficient and cost effective.

2. Background

(5 points) How is it done today, and what are the limits of current practice?

Recent work has increasingly used wearable biomechanical measurements to monitor posture and external loading in manual material handling tasks. For example, inertial measurement units (IMUs) have been used to detect lifting phases and estimate workstation height, handling mode, and relative hand load from continuous kinematics, enabling automated risk assessment for low-back disorders [5, 6]. In addition, several studies have explored surface EMG as a complementary or alternative modality, using armband EMG to detect lifting-lowering activities and classify hand loads, as well as to compare different EMG-based classification algorithms across diverse lifting trials [9, 10]. EMG-driven models have also been used to estimate dynamic hand contact forces during overhead work with and without arm-support exoskeletons, demonstrating that EMG-based random forest regression can provide reasonably accurate force estimates across task conditions and interventions [1].

These studies report reasonably good accuracy for hand-load estimation, but many of the tasks were still relatively controlled. For example, Taori and Lim's experiments [9, 10] involved lifting objects between two nearby shelves on the left and right sides of the subject, with some trials constraining foot position. Their 2025 results also showed that when foot position was not controlled, hand-load estimation became less accurate. In contrast, the dataset used in this current project was collected during lifting tasks with unconstrained lifting style; due to the distance between origin and destination, walking could occur during the lifts. As a result, these tasks are closer to real-world work situa-

tions than the controlled tasks in prior studies. Developing models that perform well under these more complex conditions is important for achieving more accurate ergonomic risk assessment in practice.

3. Motivation

(5 points) Who cares? If you are successful, what difference will it make?

There are a number of ergonomics risk assessment tools available. For example, TuMeke uses an AI-based computer-vision system to estimate risk from workers' postures. However, postural information alone provides a limited view of ergonomic exposure. Incorporating additional sensor modalities, such as EMG or IMU sensors, would provide richer input to risk assessment models and may improve prediction accuracy. In this project, the EMG sensors are integrated into a wireless armband that can be easily fitted to the forearm. If these sensors can be made more low-profile, workers could wear them during normal tasks, enabling continuous monitoring of ergonomic risk in real work environments. If models similar to what is proposed here can be developed to accurately estimate hand load from EMG signals, AI-based ergonomics tools could provide more precise and individualized risk assessments.

4. Approach

(10 points) What did you do exactly? How did you solve the problem? Why did you think it would be successful? Is anything new in your approach?

I propose a pipeline that estimates hand load based on forearm muscle activities (Figure 1). This pipeline takes normalized EMG data (in percentage of maximum muscle contraction) recorded by an EMG armband as input and estimate the hand load based on EMG signals of the forearm muscles. The EMG signals were notch filtered with a built-in tool in the armband software, and I applied a zero-lag high-pass fourth-order Butterworth filter with a cutoff frequency of 5 Hz to remove low frequency noise, followed by a zero-lag low-pass fourth-order Butterworth filter with a cutoff frequency of 10 Hz to extract the envelope of the signal. Subsequently, the EMG signals were segmented into short windows (100ms with 50% overlap) and converted into a feature vector per window. I derived features from the EMG signals in time and frequency domains [11, 4]. In the time domain, I computed mean absolute value, variance, mean absolute deviation, and waveform length for each window and channel to capture the overall amplitude and complexity of muscle activation. In the frequency domain, I extracted total power and the first, second, and third spectral moments. In total, eight features were extracted for each of the eight EMG channels. These feature vectors then served as inputs to machine learning models that either performed regression (predicting continuous box weight) or classification (predicting discrete load levels).

Participants performed several lifting-related activities, including lifting weights from a shelf to a cart, pushing and pulling the cart, and lifting weights from the cart back to the shelf. The start and end points of each activity type were labeled manually using an RGB video that was synchronized with the EMG recordings. For this project, I only analyzed the time intervals when the subject was actively lifting the object (from lift-off until drop-off). The lifting tasks included three object types (horizontal box, vertical box, sandbag) and two shelf heights (floor height and waist height). I included two modeling approaches: training a single model on all object types together, and training separate models for each object type. This idea was inspired by Lim (2024). That study first categorized data by height and handling mode and then trained separate models for each condition, and the hierarchical model outperformed the baseline model with no categorization. In my case, I used forearm EMG data, and different object types are likely to produce different grip strategies and muscle activation patterns, which may affect model performance. For the regression formulation, each EMG window was labeled with the corresponding numeric box weight (in kg), whereas for the classification formulation I treated hand load as a categorical label, using either five classes (2.3, 4.5, 6.8, 9.1, and 11.3 kg) or a reduced three-class subset (2.3, 6.8, and 11.3 kg) to examine whether separating light, medium, and heavy loads improved discriminability.

I initialized my random forest model with hyperparameters similar to those used by Behjati Ashtiani et al.[1], their work systematically tested different combinations and reported settings that performed well. I then adjusted these hyperparameters to better fit my data. Taori and Lim evaluated several feature configurations: time-domain only, frequency-domain only, and combined time- and frequency-domain features. They found that using both domains together yielded the best performance. Following their findings, I also included both time and frequency domain features in my models.

In summary, my modeling approach was informed by these prior studies: I adopted configurations that had previously performed well and used them as baselines for comparison in my own project. In practice, I implemented both a random forest regressor (for continuous load estimation) and a random forest classifier (for discrete load prediction) using scikit-learn, initializing key hyperparameters such as n_estimators, max_depth, min_samples_split, and min_samples_leaf from the literature and then exploring a small set of alternatives to balance underfitting and overfitting.

I used 80% of the data as a training set and 20% as a testing set. To reduce information leakage from temporally

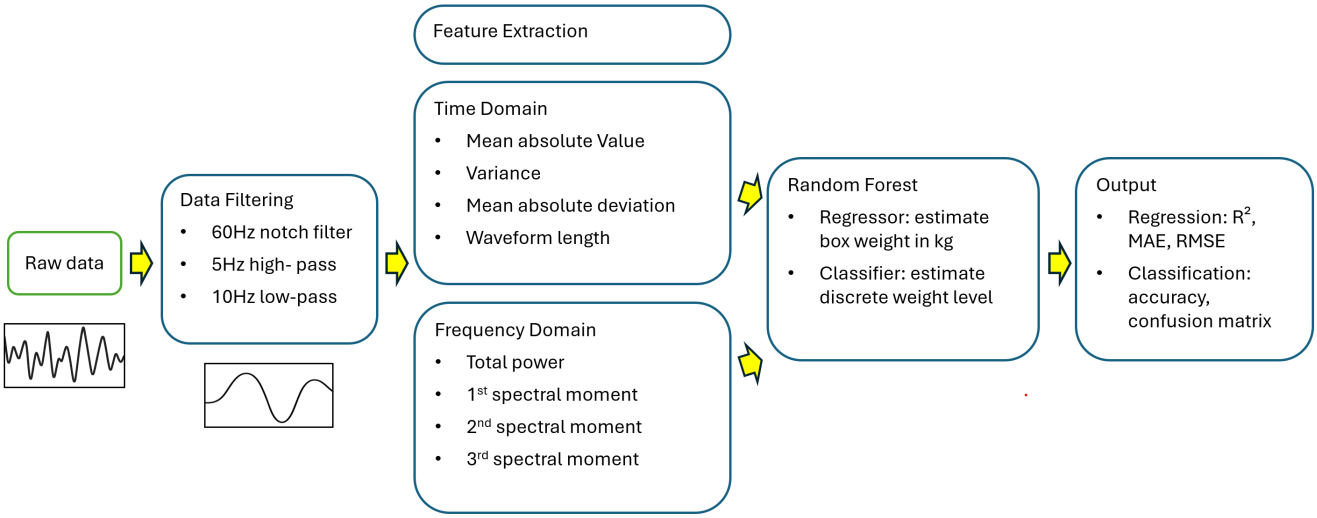

Figure 1. Overview of the EMG-based hand-load estimation pipeline.

adjacent windows, I performed this split at the trial level by grouping on the "Timeline" identifier, so that all windows from a given trial were assigned entirely to either the training or the test set. I did not include a separate validation set because the overall dataset is small, and further splitting would worsen the data sparsity problem. In random forests, a validation set is primarily useful for hyperparameter tuning rather than for training itself, so instead I started from hyperparameters reported in previous work, adjusted them based on training performance, and then evaluated the final model on the 20% held-out test set. For regression experiments, I summarized performance on this held-out set using R^2, mean absolute error (MAE), and root mean squared error (RMSE), whereas for classification experiments I used overall test accuracy.

(5 points) What problems did you anticipate? What problems did you encounter? Did the very first thing you tried work?

I anticipated that model performance might be low, because the dataset is sparse, and the EMG data might be noisy. When I implemented the model, I did see the issues I expected. At the beginning of the project, I planned to use a raw dataset previously collected in our lab that includes EMG, IMU, depth camera-based motion, and RGB video from 40 subjects. A key step for using this dataset was manually labeling the task intervals based on the RGB video. Since the other data were synchronized with the video, this approach would give accurate time labels. However, each recording was long, so the labeling process was very time-consuming. In addition, labeling errors occurred in some files, and we had to restart the labeling process from the beginning. By late November, data for only two subjects were fully labeled and ready to use, which made the data sparsity problem especially severe for this project.

The low performance of my model could also be explained by noisy EMG signals and limitations in the features I extracted. EMG is sensitive to electrode placement, skin impedance, movement artifacts, and changes in grip strategy. During dynamic motion such as lifting, these factors introduce noise into the signal, so the same hand load can produce quite different activation patterns across trials and participants. In addition, I used a relatively simple set of window-based time and frequency domain features, which may not fully capture the temporal structure or subtle spatial patterns across channels that relate to hand load. With more subjects, cleaner signals, and richer feature representations (or sequence models such as CNNs on raw or minimally processed EMG), the model performance would likely improve.

The data pre-processing and feature extraction procedures were performed smoothly. However, I encountered an issue when splitting the data into training and testing sets. I initially trained a classification model with five object weights, but the random grouped splitting did not preserve all five weights in both sets; in some cases, certain weights appeared only in the training data or only in the test data. This made the performance metrics hard to interpret and, in extreme cases, led to degenerate results (for example, the model was never trained on a given weight but was evaluated on it). To address this, I changed the splitting procedure so that, for each box weight, entire trials (timelines) were split into training and test subsets separately and then combined. This ensured that every weight level was represented in both training and test data while still preventing data leakage across windows from the same trial.

5. Experiment and results

(10 points) How did you measure success? What experiments were used? What were the results, quantitative or qualitative, or both? Did you succeed? Did you fail? Why?

I measured success in terms of how well random forest (RF) models could recover hand load from forearm EMG during lifting, using both regression (predicting box weight in kilograms) and classification (predicting discrete weight levels). For regression, I used the coefficient of determination (R^2), mean absolute error (MAE, in kg), and root mean squared error (RMSE, in kg) computed on a held-out test set. For classification, I used overall test accuracy.

For the regression task, I first trained subject-specific RF models that predicted continuous box weight from time and frequency domain EMG features. For participant 3 alone, the model showed moderate fit on the training data but very poor generalization: R^2 around 0.61 with MAE around 1.36 kg and RMSE around 1.8 kg, but test R^2 slightly negative (worse than just predicting the mean weight) and test MAE/RMSE in the 2-3 kg range. For Participant 4, training R^2 was a bit higher (around 0.7) and training errors similar or slightly smaller, but test R^2 became strongly negative and test errors remained large. I also varied the window length and overlap during feature extraction (for example, 100 ms vs 200 ms, 50% vs 100% overlap), but these changes did not substantially alter the test metrics: the models consistently fit the training data reasonably well and then failed to generalize to new trials.

I then combined the data from participants 3 and 4 and trained separate models by load type. Combining data from two participants increases the number of training samples, and separating load types may reduce the influence of object shape on gripping strategy. The resulting performance is summarized in Table 1. However, even with this setup, the models achieved only moderate fit on the training data (Train R^2 between 0.45 and 0.67, Train MAE around 1.3–1.6 kg) and poor generalization on held-out trials (all Test $R^2 < 0$, Test MAE around 2.5–2.7 kg and Test RMSE above 3 kg across load types). These results indicate that the models overfit the training data and fail to generalize to new trials. Pooling participants increased variability in EMG patterns due to between-subject differences and different lifting styles, but did not provide enough additional independent trials to compensate, so the random forest likely learned subject- and object-specific noise rather than a stable relationship between EMG features and hand load. This pattern is especially evident in the sandbag condition, which shows the most negative Test R^2 and highest Test RMSE compared with the horizontal and vertical box models.

Because continuous prediction was performing poorly, I also formulated the problem as classification, where each window was assigned to one of the discrete box weights.

Metric	Horizontal box	Sand bag	Vertical box
Train R^2	0.454	0.670	0.492
Test R^2	-0.257	-0.350	-0.107
Train MAE	1.615	1.265	1.523
Test MAE	2.662	2.703	2.501
Train RMSE	2.123	1.642	2.024
Test RMSE	3.254	3.408	3.070

Table 1. Random forest regression performance on combined participants 3 and 4 EMG features, reported separately for each load type.

Using all five weights (2.3, 4.5, 6.8, 9.1, and 11.3 kg) as classes and both subjects' features together, the RF classifier achieved training accuracies in the 0.3–0.4 range and test accuracies around 0.26–0.30, only slightly above the 0.20 chance level for five balanced classes. I systematically varied the maximum tree depth. With deeper trees (for example, max_depth = 5), training accuracy increased to roughly 0.5 but test accuracy stayed low (around 0.23), indicating clear overfitting. As I reduced the maximum depth down to 3 , training accuracy decreased but test accuracy increased slightly, and the gap between them narrowed, reflecting a move toward a simpler but better-regularized model. I also tried a simplified three-class problem using only the lightest, medium, and heaviest weights (2.3, 6.8, 11.3 kg) , but due to the limited number of trials per load, the gap between training and testing accuracies increased, indicating that the data overfitting issue was more severe. Example confusion matrices (shown in Figure 2 and Figure 3) indicate that misclassifications is widespread and that the classifier has difficulty distinguishing weights reliably across trials. These results contrast with those reported in related work. IMU-based models for lifting detection and load estimation have reported high performance in more controlled contexts; for example, prior work [5, 6] using multiple inertial sensors during repetitive lifting–lowering tasks has achieved very accurate detection of lifting phases and reasonably small errors when estimating workstation height, handling mode, and relative hand load across participants. Other studies [9, 10] using forearm EMG armbands in controlled lifting setups—where objects were lifted between fixed positions and foot placement was constrained—have reported classification accuracies around 80% or higher for distinguishing different load levels or distinguishing lifting from non-lifting activities. Likewise, EMG-driven random forest regression models have been used to estimate hand contact forces during overhead work [1], with more participants and more repetitions per condition, and those models produced reasonably accurate force estimates that remained relatively stable across arm-support conditions and task variations. Compared to these benchmarks, my best EMG-only models un-

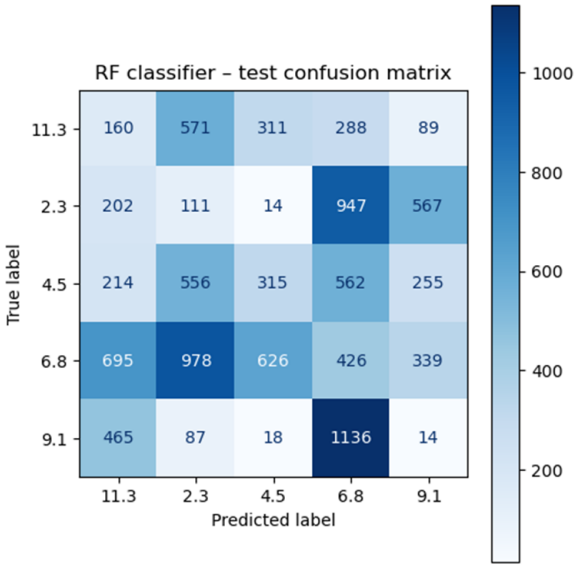

Figure 2. Heatmap of the random forest classifier confusion matrix showing predicted versus true hand-load classes (5 levels).

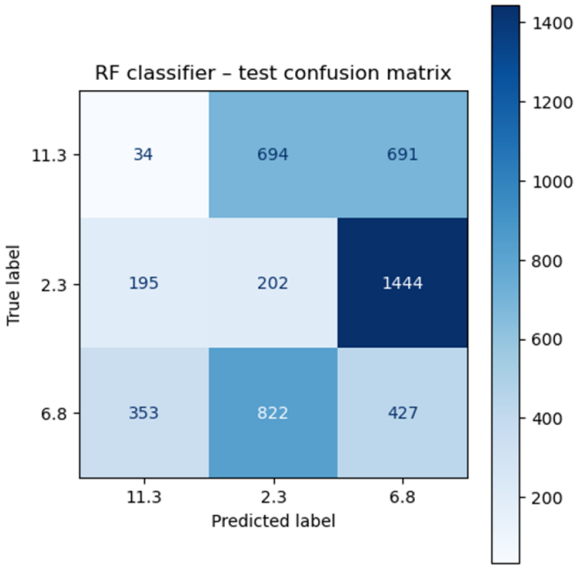

Figure 3. Heatmap of the random forest classifier confusion matrix showing predicted versus true hand-load classes (3 levels).

derperform: the regression models yield negative R^2 on test data, and the classification models achieve only slightly above chance accuracy, even under simplified conditions.

The differences in performance are consistent with differences in the data and task. My dataset consists of only two subjects, and the lifting tasks are relatively unconstrained: participants move boxes between locations with different heights and depths, lifting styles were freely selected by participants, and walking and carrying was involved during the lifting activity. EMG is sensitive to electrode placement, movement artifacts, and small changes in arm and hand posture. Therefore, the same hand load can correspond to quite different EMG patterns across trials and participants, and the dynamic lifting task likely increased noise level and affects of artifact. The feature set I used consists of relatively simple window-based time and frequency domain descriptors averaged over the window, which may not capture more subtle temporal or cross-channel structure. In contrast, prior studies that reported stronger EMG-based performance had more subjects, more repetitions per condition, more controlled movement patterns.

Given these constraints, I would describe the project as only partially successful. On the positive side, I implemented an end-to-end EMG processing and modeling pipeline: filtering and normalization, segmentation, feature extraction in both time and frequency domains, careful grouped train–test splitting by trial, and random forest training for both regression and classification. The models were able to pick up some signal in the training data, suggesting that EMG does contain information about hand load under these tasks. However, on held-out trials, the models generalized poorly, and neither the regression nor classification approaches achieved the level of accuracy that would be required for practical hand-load estimation in real workplaces. The main reasons are severe data sparsity (only two subjects with a limited number of fully labeled trials), noisy and variable EMG signals, unconstrained lifting styles, and relatively simple feature representations.

In future work, I would utilize the full set of 40 participants once the fully labeled dataset is available. I expect model performance to improve because the main limitation in the current project is data sparsity. A larger dataset would not only support better training but also make it easier to diagnose where the current model fails and to fine-tune feature choices and hyperparameters. I would also explore a convolutional neural network (CNN)–based model for hand load estimation. Because eight EMG channels were recorded simultaneously on the forearm, the spatial and temporal relationships among these muscles may contain important information that tree-based models cannot fully exploit, whereas CNNs are well suited to learn such multi-channel patterns. However, CNNs are more data-demanding and sensitive to overfitting, so this approach would only be practical once substantially more participants and lifting trials are available (and possibly with additional regularization or data augmentation).

6. Acknowledgment

(5 points) Is your code available? Did you use open-source license to release your code?

My code is available on GitHub with a open-source li-

cense.

(10 points) How do you plan to disseminate your method? Are the findings available via freely accessible project website and/or GitHub?

Materials used in this project is available on GitHub repository. The repository contains code for the full analysis pipeline, including EMG preprocessing, feature extraction, and random forest training and evaluation for both regression and classification. The findings are stated in the previous sections of this report, and this report will be available freely on the GitHub project page.

(10 points) How can others reproduce your results? Are training, validation, and test data freely provided?

This work is reproducible given the dataset, pipeline, and random forest hyperparameters. The raw EMG and other de-identified data are still being processed as part of a larger project, and the full dataset will be released later as part of that work. Due to GitHub's file size limitations, the data set is uploaded in this OneDrive folder.

(5 points) Are model parameters fully reproducible?

Yes. The model parameters are fully reproducible as long as the same data and environment are used. In all random forest models, I fixed the random seed (e.g., *random_state*=42) and used a deterministic grouped 80/20 train–test split. Given the same dataset, splitting procedure, hyperparameters, and library versions (scikit-learn), rerunning the code will produce the same trees, the same predictions, and therefore the same performance metrics.

References

[1] Mohamad Behjati Ashtiani and others. Emg-driven estimates of hand contact forces during overhead work with and without an arm-support exoskeleton using random forest regression. *Journal of Biomechanics*, 2025. Preprint / in press.

[2] Centers for Disease Control, Prevention, et al. Work-related musculoskeletal disorders & ergonomics-workplace health strategies by condition-workplace health promotion, 2020.

[3] WM Keyserling, DS Stetson, BA Silverstein, and ML Brouwer. A checklist for evaluating ergonomic risk factors associated with upper extremity cumulative trauma disorders. *Ergonomics*, 36(7):807–831, 1993.

[4] Elnaz Lashgari and Uri Maoz. Dimensionality reduction for classification of object weight from electromyography. *Plos one*, 16(8):e0255926, 2021.

[5] Sol Lim. Exposures to select risk factors can be estimated from a continuous stream of inertial sensor measurements during a variety of lifting-lowering tasks. *Ergonomics*, 67(11):1596–1611, 2024.

[6] Sol Lim and Clive D'Souza. Gender and parity in statistical prediction of anterior carry hand-loads from inertial sensor data. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 1142–1146. SAGE Publications Sage CA: Los Angeles, CA, 2019.

[7] William S Marras, Steven A Lavender, Sue E Leurgans, Fadi A Fathallah, Sue A Ferguson, W Gary Allread, and Sudhakar L Rajulu. Biomechanical risk factors for occupationally related low back disorders. *Ergonomics*, 38(2):377–410, 1995.

[8] Laura Punnett, Lawrence J Fine, W Monroe Keyserling, Gary D Herrin, and Don B Chaffin. Back disorders and nonneutral trunk postures of automobile assembly workers. *Scandinavian journal of work, environment & health*, pages 337–346, 1991.

[9] Sakshi Taori and Sol Lim. Use of a wearable electromyography armband to detect lift-lower tasks and classify hand loads. *Applied Ergonomics*, 119:104285, 2024.

[10] Sakshi Taori and Sol Lim. Comparing armband emg-based lifting load classification algorithms using various lifting trials. *International Journal of Industrial Ergonomics*, 2025. In press.

[11] Qi Xu, Yazhi Quan, Lei Yang, and Jiping He. An adaptive algorithm for the determination of the onset and offset of muscle contraction by emg signal processing. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 21(1):65–73, 2012.