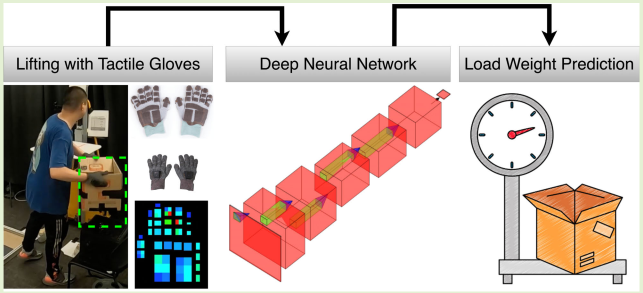
Tactile Gloves Predict Load Weight During Lifting With Deep Neural Networks

Guoyang Zhou, Ming-Lun Lu, and Denny Yu, *Member, IEEE*

***Abstract*—Overexertion in lifting tasks is one of the leading causes of occupational injuries. The load weight is the key information required to evaluate the risk of a lifting task. However, weight varies across different objects and is unknown in many circumstances. Existing methods of estimating the load weight without manual weighing focused on analyzing body kinematics or muscle activations, which either utilize indirect indicators or require intrusive sensors. This study proposed using tactile gloves as a new modality to predict the load weight. Hand pressure data measured by tactile gloves during each lift were formulated as a 2-D matrix containing spatial and temporal information. Different types of deep neural networks were adopted, and a ResNet 18 regression model achieved the best performance. Specifically, it achieved a predicted *R*-squared of 0.821 and a mean absolute error of 1.579 kg. In addition, to understand the model's decision-making logic and the hand force exertion pattern during lifting, the Shapley additive explanations (SHAPs) technique was utilized to determine the importance of each sensor at each frame. The results demonstrated that the right hand was more important than the left hand for the model to predict the load weight. Additionally, fingers were more important than palms, and the middle phase of a lifting task was more important than its beginning and ending phases. Overall, this study demonstrated the feasibility of using tactile gloves to predict the load weight and provided new scientific insights on hand force exertion patterns during lifting.**



***Index Terms*—Load weight prediction, neural networks, tactile gloves.**

I. Introduction

LIFTING tasks are one of the major causes of work-related musculoskeletal disorders (WMSDs) [1]. Load weight, which is the weight of the object to be lifted, is a major risk factor of lifting tasks; all existing lifting risk assessment tools or methods take the load weight as a critical input (see [2], [3], [4]). However, the weight varies across different objects, and it can be unknown and hard to measure in unordered and fast-paced working environments, such as unloading packages from trucks.

Existing methods of predicting the load weight primarily focus on utilizing indirect cues such as body kinematics, postures, and muscle activations (see [5], [6], [7]). In addition, these methods require instrumentations [e.g., inertia measurement units (IMUs), surface electromyography (sEMG)] that are not scalable for application because they are time-consuming to set up and intrusive to workers.

Motivated by the identified research gaps and needs, this study proposed utilizing tactile gloves as a new modality to predict the load weight. Tactile gloves are embedded with pressure sensors capable of measuring pressure at multiple hand regions. They have been used in many different applications, such as hand tool design and rehabilitation [8], [9], [10], [11]. To the best of our knowledge, previous work using tactile gloves focused on one-handed grasp prediction of lightweight objects. This work focuses on workplace applications that require lifting heavy loads in dynamic conditions.

The application of the tactile glove in predicting lifting weights is challenging because lifting requires varying gripping efforts and different contact areas in response to the required force distributions during the course of a lifting task. In addition, the gripping force can potentially be affected by other factors, such as inertia factors and muscular strength.

Manuscript received 16 April 2023; revised 13 June 2023; accepted 16 June 2023. Date of publication 30 June 2023; date of current version 15 August 2023. This work was supported by the National Institute for Occupational Safety and Health and through the Pilot Research Project Training Program of the University of Cincinnati Education and Research Center under Grant T42OH008432. The associate editor coordinating the review of this article and approving it for publication was Dr. Olga Fink. *(Corresponding author: Denny Yu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Purdue University Institutional Review Board.

Guoyang Zhou and Denny Yu are with the School of Industrial Engineering, Purdue University, West Lafayette, IN 47906 USA (e-mail: zhou595@purdue.edu; dennyyu@purdue.edu).

Ming-Lun Lu is with the National Institute of Occupational Safety and Health, Cincinnati, OH 45226 USA (e-mail: uzl5@cdc.gov).

Digital Object Identifier 10.1109/JSEN.2023.3289670

An ergonomics study has demonstrated the complexity of gripping force in lifting tasks [12] and showed that more than 40% of the variance of the gripping force measured by tactile gloves during lifting could not be explained by the variance of the load weight through simple linear regression models.

In this study, two modeling strategies were proposed to predict the load weight using hand pressure. First, motivated by the physical and linear correlation between pressure and force, two lasso regression (LR) models were trained; one was built upon the average pressure of each sensor, and the other was built upon the peak. The second strategy was based on deep neural networks. Specifically, the hand pressure data recorded for each lifting task were formulated as a single-channel and 2-D matrix. Each data point in the matrix represented a sensor's reading at a certain frame; we named this matrix as spatial–temporal pressure matrix (STP matrix). Classic image feature extractors were transferred to our prediction task because of their strong capabilities in learning representations from 2-D data. Specifically, we modified the Visual Geometry Group (VGG) [13], ResNet [14], and vision transformer (ViT) [15] models and utilized them for extracting representative features from the STP matrices. The best-performing model is a ResNet 18 model, which achieved a predicted R-squared of 0.821 and a mean absolute error of 1.579 kg in predicting the weight of loads.

In addition, this study also investigated the importance of each hand region and each frame through explainable AI (XAI) techniques. Understanding the force distribution across the left and right hands is essential for utilizing biomechanical models to analyze two-handed lifting tasks; most practitioners using biomechanical models simply assume that the load weight is equally distributed across the left and right hands because of the lack of related knowledge. Moreover, the dense sensor layout used by tactile gloves can increase the design challenge and production cost. The chance of sensor malfunction can also be boosted because heavy and repetitive lifting tasks can cause damage to gloves. Understanding the importance of each hand region for predicting the load weight can be beneficial for optimizing the hardware design, reducing the cost, and improving the device's durability and practicability. This investigation was performed on the ResNet 18 model. Specifically, we utilized the Shapley additive explanations (SHAPs) technique [16] on this model and investigated the SHAP value of each data point inside the STP matrix. The SHAP technique is a well-validated XAI technique that allows researchers and developers to leverage deep-learning models with interpretability [17], [18].

In summary, the major contributions of this study were listed as follows.

1) This study proposed utilizing tactile gloves as a new modality to predict the load weight with machine learning techniques.
2) This study conducted a lifting and computational experiment with a custom data preparation and augmentation pipeline and deep-learning models; these experiments demonstrated the feasibility of utilizing tactile gloves to predict lifting loads.
3) This study investigated the decision-making logic of our best prediction model, which is helpful for developers and researchers to understand the hand force exertion pattern and optimize the design of tactile gloves for ergonomics risk assessments.

II. Related Work

A. Lifting Risk Assessment

The revised NIOSH lifting equation (RNLE) is among the most widely used ergonomic assessment tools for manual lifting [19]. The RNLE assesses the risk of low back injury by calculating a lifting index (LI), which is the ratio of the load weight to the recommended weight limit (RWL) for a particular lifting task [2]. The RWL is calculated by a load constant (23 kg) and six task variables including the horizontal distance between the hands and the body, vertical distance between the center of the hands and the ground, the lift asymmetry angle, the displacement of the load during the lift, hand coupling, and the frequency of performing the lifting task [2]. The LI has been proven to be significantly correlated with low back disorders in the field of ergonomics and incorporated into the International Organization for Standardization (ISO) ergonomic standard 11228-1 [20], [21], [22].

The variables for calculating the RWL are manually measured by practitioners in the workplace. Specifically, the six RNLE task variables can be determined by visual observations in the field or by reviewing video recordings of the task. Recent studies have focused on developing methods for automatically measuring some RNLE task variables without manual measurements. These methods consisted of wearable IMU sensors [23], [24], [25], computer vision techniques, or marker-based motion capture systems to track body postural information for estimating the task variables [26], [27], [28]. However, the load weight remains a key challenge in automating the data collection process using the RNLE. Measuring the weight of many objects for a work shift is time-consuming and intrusive to the worker's job.

B. Load Weight Prediction

A few studies have been focused on classifying the lifted objects' weight using features extracted from body kinematics and physiological signals. Body kinematics features (e.g., the moving velocity and moving jerk) have been extracted by either computer vision techniques [6], [29] or wearable motion sensors [5] to classify the weights or the resulting risks into two or three different levels. In general, the performance of these kinematics-based models depended on their mission (e.g., binary classification or multi-class classification), and their accuracy ranged from 89% to 47%. Specifically, the kinematic-based models can perform well when used to classify weights that are significantly different (e.g., light (26.7–53.4 N) versus heavy (106.8–133.4 N), as introduced in [29]). However, they were not developed to provide continuous or precise weight estimations. In addition to body kinematics, sEMG signals, which indicate muscle activations, have also been used to predict the load weight. Specifically, a study utilized the EMG signals from the lower back to classify three different weights (0-, 10-, and 24-lbs) with an accuracy of around 80% [30]. Similarly, another study utilized the EMG signals from various upper body regions to predict five weights ranging from 5 to 25 kg using a nonlinear parallel cascade algorithm [31]. The leave-one-subject-out testing results of this study demonstrated that the algorithm's performance could be affected by the lifting methods (i.e., squat or stoop); specifically, the mean absolute error of the algorithm on squat and stoop lifts was ±1 and ±2.3 kg, respectively.

The existing techniques for predicting the load weight still have two major limitations. First, most physiological and motion sensors require well-trained practitioners to operate. Attaching these sensors to human bodies is time-consuming for practitioners and intrusive to workers. Second, most of these techniques focused on utilizing physiological signals or body kinematics (e.g., muscle activation and joints' moving velocity) that indirectly correlated with the load weight. These indirect indicators can potentially be affected by factors irrelevant to the load weight, such as the lifting methods. Therefore, there is a need to find a new solution for predicting the load weight; this solution should be relatively non-intrusive and based on cues or indicators strongly correlated with the load weight.

C. Tactile Gloves

State-of-art wireless tactile gloves can be a practical tool because workers commonly wear gloves. Tactile gloves have been successfully used in different application areas, such as hand tool design and rehabilitation [2], [3], [4]. A recent study has demonstrated that gripping force measured by tactile gloves is a significant indicator of the load weight [12]. Specifically, the regression analyses in this study illustrated that the load weight could explain up to 72% of the variance of the gripping force when including random effects. When random effects were not included, only 58% of the variance could be explained by the load weight. Moreover, this study illustrated that task conditions, which include the lifting height and handle type, could impact the measured gripping force, but these task conditions could only explain approximately 1% of the variance.

Similarly, another study presented a prototype of a low-cost tactile glove and used it to recognize objects statically held by hands and predict these objects' weight [32]. The objects being handled in this study were lightweight (≤700 g), and their convolution neural network (CNN)-based weight prediction model demonstrated that the prediction error increased as the object's weight increased. Specifically, when the object's weight was between 151 and 700 g, the mean absolute error reached 110.97 g.

Although these studies have demonstrated the potential of tactile gloves, the current modeling methods for tactile gloves still have two major limitations that prevent them from being used at the workplace for lifting risk assessments. First, with simple linear regression models, the load weight factor could only explain 58% of the variance of the gripping force. Second, only one prediction model was developed for using hand pressure to estimate the object's weight [32], and it was only used for lightweight objects. In addition, this prediction model was developed on static handling tasks (e.g., holding a calculator) and did not consider the information in the temporal domain. However, lifting is a dynamic process, and the gripping effort required to complete a lifting task can be affected by many factors [12]. Therefore, the feasibility of utilizing tactile gloves for predicting the load weight in dynamic lifting tasks remains a scientific gap.

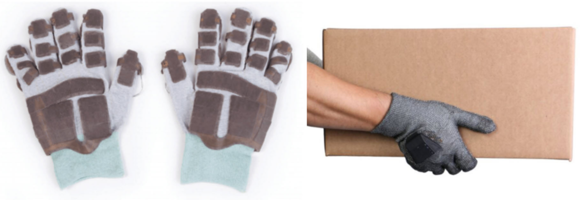


Fig. 1. Tactile gloves used in this study (permission to use photographs from the Pressure Profile System Company).

III. Human Subjects Experiment

This experiment was approved by the Institutional Review Board (IRB) of Purdue University.

A. Device

The tactile gloves utilized in this study were developed by the PPS, Inc. (Hawthorne, CA). Each glove had 65 embedded pressure sensors for measuring the pressure exerted on the palmar side of the hand Fig. 1. Each pressure sensor's full-scale range and minimum sensitivity was 80 psi (55 N/cm^2) and 0.04 N, respectively. The tactile gloves were designed to scan at a rate of 25–40 Hz and were connected to a computer via Bluetooth.

B. Participants

The conducted lifting experiment recruited 30 participants who wore tactile gloves and performed lifting tasks with different task conditions. The participants were recruited from a college student population; 60% of the participants ($n = 18$) were male, and the rest ($n = 12$) were female. In addition, 80% of the participants ($n = 24$) were right-handed. In total, 3147 lifts were recorded for analysis, and each lift's start and end frames were manually labeled and included.

C. Tasks

During the experiment, the participants were required to repetitively move a box between a fixed-height chair, a height-adjustable platform, and the floor using both hands. Each lifting task can be described by four conditions: height level of the platform, required body rotation degree, lifting direction, and weight level. These task conditions were systematically adjusted during the experiment to increase the variability and generalizability of the dataset.

Specifically, there were three height levels: high (1.1 m), middle (0.9 m), and low (0.7 m); three body rotation degrees: −90°, 0°, and 90°; two lifting directions: up and down; three load weight levels: light (1.1–4.5 kg), medium (7.9–12.5 kg),

TABLE I
EXACT WEIGHTS ASSIGNED IN THE EXPERIMENT

Weight Level	Exact Weights (unit: kg)
Light	1.1, 2.3, 3.4, 4.5
Medium	7.9, 9.1, 10.2, 11.3, 12.5
Heavy	13.6, 14.7, 15.9, 17.0

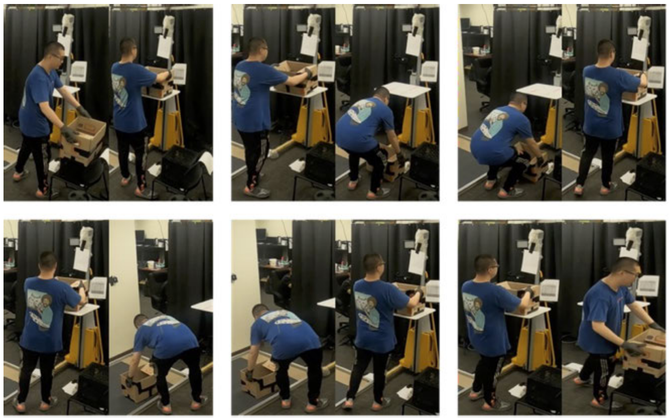


Fig. 2. This figure demonstrates six lifting tasks with three rotation degrees and two directions; the platform's height level and the box's weight were constant in these tasks.

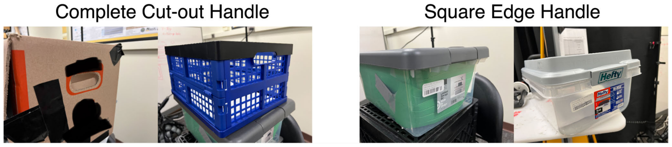


Fig. 3. Boxes assigned in the experiment.

and heavy (13.6–17.0 kg). The weight levels depended on the LIs of the RNLE. Specifically, weights leading to LIs less than 1 were categorized as the light level, and weights leading to LIs between 1 and 2 were categorized as the medium level. Finally, weights leading to LIs larger than 2 were categorized as the heavy level. An exact weight would be randomly selected from each weight level for each task during the experiment, and there were 13 different weights being assigned in total (Table I).

Moreover, each lifting task was completed twice by the participants. In total, each participant completed 108 lifting tasks during the experiment (3 height levels × 3 body rotation degrees × 2 lifting directions × 3 weight levels × 2 repetitions) (Fig. 2); the participants would complete these tasks in a controlled order, and details of the experimental procedure were presented in [12].

In addition, four different types of boxes were assigned during the experiment (Fig. 3). These boxes had two types of commonly observed handles: the square edge handle and the complete cut-out handle. The research team randomly selected a box for each participant. 60% of the participants ($n = 18$) completed their tasks with the square edge handle, while 40% ($n = 12$) completed their tasks with the complete cut-out handle.

In total, 3260 lifting tasks performed by 30 participants were recorded by the end of the experiment.

IV. Computational Experiment

A. Data Preparation and Augmentation

The dataset contains the start and end frames of each lift, which were manually labeled; the pressure recording of each lift was well-trimmed based on these time labels. There are two major challenges in utilizing the well-trimmed recording of each lift for modeling. First, the length of the well-trimmed recordings varied between 39 and 125 frames (71.7% of them were between 60 and 100 frames), and the tactile gloves used in this study are embedded with 130 sensors. The resulting data became high-dimensional but inconsistent in terms of size, which is challenging for spatial–temporal modeling. In addition, from a practicable standpoint, locating each lift's start and end frames requires action detection algorithms, which always have prediction errors. The predicted start and end frames of lifting tasks will not be the same as the time labels provided by humans. Therefore, the practicability of a prediction model can potentially be reduced if it is trained on well-trimmed recordings.

To solve these challenges and develop prediction models that can tolerate different start and end frames, we formulated and augmented the data by using a sliding window with a fixed size and stride. The raw dataset contained untrimmed recordings, including various activities not associated with lifting. These non-lifting activities included hand pressure data of the participant momentarily preparing for or leaving the task, such as touching the handle while not exerting force to lift it or pulling the hands off the handles. The portion of recording included in the sliding window but not related to lifting (i.e., the proportion of the sliding window not being covered in Fig. 4) was handled by the model as noises to increase the model's generalizability.

The sliding window techniques have been proven to be an efficient technique for augmenting different types of time-series sensor data for modeling (e.g., audio and electroencephalogram (EEG) signals [33], [34]). Specifically, the sliding window in this study had a length of 90 frames, which is the average length of the well-trimmed recordings; the stride of the sliding window was 3. The sliding window generated multiple proposals from each lift; each proposal was a 90 × 130 matrix, where the first dimension indicated the temporal domain (90 frames) and the second dimension indicated the spatial domain (130 sensors). Those proposals whose coverage rate was at least 70% (1) were selected as valid spatial–temporal pressure (STP) matrices for model training, validating, and testing

$$\text{Coverage Rate} = \frac{\text{Number of frames covered}}{\text{Duration of the original lift}}. \quad (1)$$

A graphic illustration of how the sliding window generated a proposal is presented in Fig. 4. The proposal extracted by the sliding window in this figure was a valid STP matrix because it covered most of a trimmed recording (i.e., coverage rate $\geq$ 70%). In addition, since the stride of the sliding window is equal to 3, the following proposal will start at frame 83 and end at frame 173, which will also be selected because of a sufficient coverage rate. With this sampling technique, the

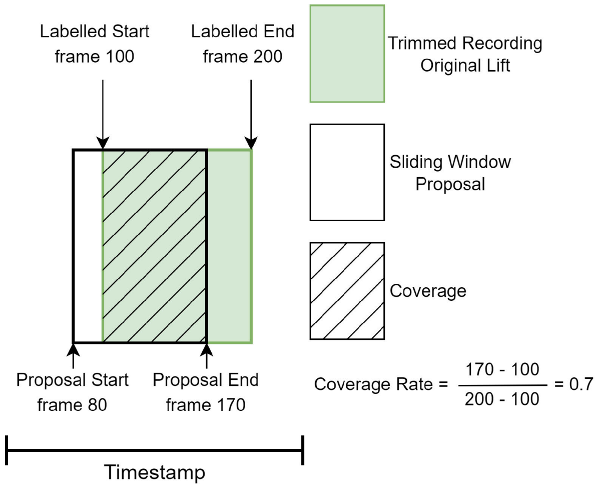


Fig. 4. Graphic illustration of the sampling technique.

dataset was significantly expanded. In total, 35 517 valid STP matrices were extracted for the augmented dataset.

A mathematical illustration of the STP matrix was presented in (2). Specifically, the $P_{y,x}$ in the STP matrix represents the pressure reading of sensor x at frame y. In this study, the x ranged from 0 to 129 because there were 130 sensors, and the y ranged from 0 to 89 because of the sliding window

$$\text{STP Matrix} = \begin{pmatrix} P_{0,0} & P_{0,1} & \cdots & P_{0,x} \\ P_{1,0} & P_{1,1} & \cdots & P_{1,x} \\ \vdots & \vdots & \ddots & \vdots \\ P_{y,0} & P_{y,1} & \cdots & P_{y,x} \end{pmatrix}. \quad (2)$$

The proposed method of preparing and augmenting the collected data had two major advantages. First, the size was unified, and the data were formulated as a 2-D matrix, which is suitable for powerful 2-D representation learning techniques, such as CNN and ViT. More importantly, most 2-D representation learning techniques were designed for learning features in the spatial domain; with the utilized data preparation and augmentation method, 2-D representation learning techniques can now learn representations not only in the spatial but also in the temporal domain from the hand pressure data. Second, models trained on the augmented data could potentially be more generalizable than models trained on the well-trimmed data because they could tolerate irrelevant information induced by varying start and end frames.

B. LR Models

One of the LR models was built upon the average pressure of each sensor, and the other one was built upon the peak pressure of each sensor. The LR model was selected because the L1 regularization technique can perform the variable selection. Both LR models utilized the coordinate descent approach [35] to fit the coefficients, and different alpha values were tested. The objective function of the LR is presented in (3). In this equation, x represents the input features (i.e., the calculated average or peak pressure); y represents the ground-truth load weight; w represents the coefficients; α defines the penalty term, which denotes the amount of shrinkage or constraint that will be implemented

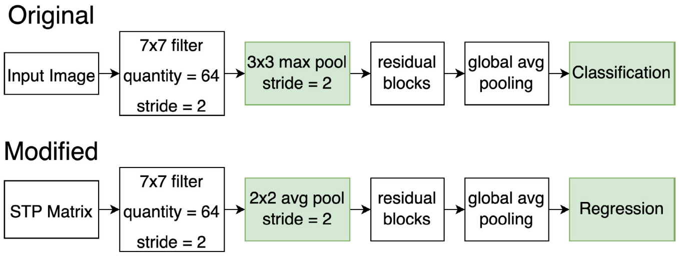


Fig. 5. Original and modified ResNet models. The green blocks highlight the modifications made to adopt the ResNet models for the load weight prediction task.

$$\min_{w} \frac{1}{2n}||xw-y||_2^2 + \alpha||w||_1. \quad (3)$$

C. Convolutional Neural Networks

The VGG and ResNet techniques were transferred to our prediction task [13], [14]. For the VGG technique, we adopted the VGG 11, VGG 13, VGG 16, and VGG 19 models. Four modifications were made to the original VGG models. First, the original VGG models were composed of convolutional, pooling, and fully connected layers. Our study replaced the fully connected layers with the global average pooling layer. Compared to the fully connected layers, the global average pooling layer is more native to the convolution structure and is less prone to overfitting because of fewer parameters to train [36]. Second, we replaced the max pooling layers with the average pooling layers. The key rationale of our decision was that subtle features should be learned and retained because every data point in the STP matrix was a reading of a sensor that could be in contact with the load. Third, the batch normalization layer was added to the VGG models to increase the training speed and reduce overfitting [37]. To avoid overfitting, we also tried regularizing the VGG models with the dropout technique during the fine-tuning process; a dropout layer (0.3) was added after each batch normalization layer. Our preliminary analyses demonstrated that adding the dropout layers did not significantly impact the performance of the VGG models in this study; in fact, it slightly degraded the performance of some VGG models. Therefore, we decided not to include the VGG models with dropout layers in this study. Finally, the original VGG models were developed for image classification tasks. To turn them into regression models, the final layer was modified to have a single neuron activated by the linear function. Details of our modified VGG models are presented in Table II, and details of the original VGG models can be found in [13].

For the ResNet techniques, we adopted the ResNet 18, ResNet 34, and ResNet 50 models. Only two modifications were made to the original ResNet models as shown in Fig. 5.

Specifically, we first replaced the top 3 × 3 max pooling layer with a 2 × 2 average pooling layer to encourage the models to learn and retain subtle features. In addition, we replaced the original output layer with a single-neuron layer activated by the linear function to turn the ResNet models

TABLE II
CONFIGURATION OF THE VGG MODELS

VGG 11	VGG 13	VGG 16	VGG 19
Input: 90x130 STP matrix			
Conv3 - 64	Conv3 - 64 Conv3 - 64	Conv3 - 64 Conv3 - 64	Conv3 - 64 Conv3 - 64
average pooling (size = 2), batch normalization			
Conv3 - 128	Conv3 - 128 Conv3 - 128	Conv3 - 128 Conv3 - 128	Conv3 - 128 Conv3 - 128
average pooling (size = 2), batch normalization			
Conv3 - 256 Conv3 - 256	Conv3 - 256 Conv3 - 256	Conv3 - 256 Conv3 - 256 Conv3 - 256	Conv3 - 256 Conv3 - 256 Conv3 - 256 Conv3 - 256
average pooling (size = 2), batch normalization			
Conv3 - 512 Conv3 - 512	Conv3 - 512 Conv3 - 512	Conv3 - 512 Conv3 - 512 Conv3 - 512	Conv3 - 512 Conv3 - 512 Conv3 - 512 Conv3 - 512
average pooling (size = 2), batch normalization			
Conv3 - 512 Conv3 - 512	Conv3 - 512 Conv3 - 512	Conv3 - 512 Conv3 - 512 Conv3 - 512	Conv3 - 512 Conv3 - 512 Conv3 - 512 Conv3 - 512
average pooling (size = 2), batch normalization			
global average pooling			
one neuron, linear activation			

* conv3 represents a convolution filter with a size of (3,3). The number after the conv3 represents the number of filters.

into regression models. Details of the residual blocks can be found in the original paper of ResNet [14].

D. Vision Transformers

Motivated by the recent successes of utilizing the ViT for image classification tasks, this study explored the potential of utilizing ViT for analyzing the STP matrix. In this study, we adopted the ViT architecture introduced in [15], which was the first ViT architecture. All ViT models in this study were trained without pretrained parameters, and we optimized the hyperparameters using the grid search method. Specifically, we prepared a few sets of hyperparameters containing different combinations of hyperparameters. We trained the models on each set and chose to report the one that achieved the best performance.

The first part of the ViT was a patch encoder. Specifically, the patch encoder separated each 90 × 130 STP matrix into 117 patches; each patch had a size of 10 × 10 (Fig. 6). Then, each patch was flattened and linearly projected into a low-dimensional vector with a size of 32. The positional embeddings were added to the projected vectors to retain positional information.

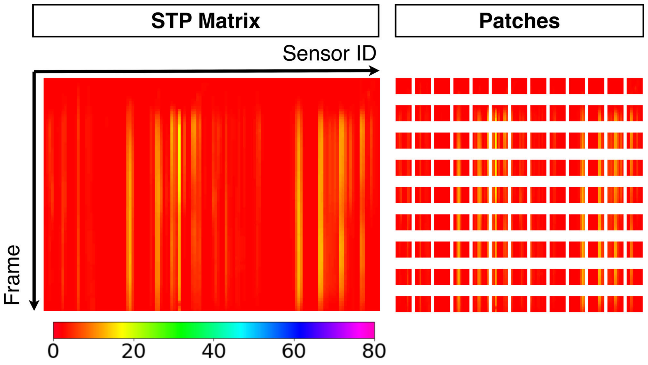


Fig. 6. Figure presents a visualization of an STP matrix (left) and the separated patches of the STP matrix (right). The horizontal axis represents the sensor id, and the vertical axis represents the frame number. The pressure values ranging from 0 to 80 psi were visualized by colors.

TABLE III
CONFIGURATION OF THE TRANSFORMER ENCODERS

Model	Layer	Attention Head
ViT–Light 1	8	8
ViT–Light 2	8	12
ViT–Base 1	12	8
ViT–Base 2	12	12

After the positional embeddings were added, the projected vectors were then sent to the transformer encoders. The transformer encoders contained multiple layers of multiheaded self-attention and multilayer perceptron (MLP) blocks. Different numbers of transformer layers and attention heads were tested (Table III). The normalization layer was added before each block, and the residual connection was added after each block. Each MLP block contained two dense layers activated by the GELU function. The first and second dense layers in the MLP block had 64 and 32 neurons, respectively; a dropout layer (0.2) was added after each dense layer to reduce overfitting.

The output of the transformer encoders was first normalized and then flattened. A dropout layer (0.2) was added after the flatten layer to reduce overfitting. Three fully connected layers were added at the end. The first two layers both had 4096 neurons activated by the ReLu function and regularized by L2. A dropout layer (0.5) was added after each fully connected layer. The final layer was a single-neuron output layer activated by the linear function.

E. Training

The VGG, ResNet, ViT models were trained without pretrained parameters. The Adam optimizer [38] was chosen to train each with $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate = 0.001, batch size = 256, and weight decay = 0; we considered the setups in [15] and used the grid search method to finalize these hyperparameters for the Adam optimizer. In addition, the models were compiled to minimize the mean-squared error.

The models were originally set to be trained for 450 epochs. The early stopping technique was applied to stop the training process on time to avoid overfitting. Specifically, the training process would end early if the validation loss was not further reduced within 30 epochs.

F. Validation and Test

All models were tested using the leave-one-subject-out method. Specifically, each participant's data would be left out as testing data in turn for testing the model trained on other participants' data. Utilizing the leave-one-subject-out method can examine a mode's robustness when facing a completely unseen person.

For the deep-learning-based models, 33% of the training data would be selected as validation data for determining the best model checkpoint and stopping the training process early. For the LR models, a fivefold cross-validation was performed on the training data for hyperparameter tuning. Specifically, the fivefold cross-validation was only for searching the optimal alpha value between 0.1 and 3.

G. Metrics

The overall performance of each model was evaluated using the mean absolute value (MAE) (4) and predicted R-squared value. In this equation, y represents the ground truth weight, $\hat{y}$ represents the predicted weight, and N represents the number of samples. We further evaluated the MAE of each model under different weight levels (Table I). The MAE was chosen because it was found to be a natural measure of average error and is relatively unambiguous [39]. The predicted R-squared was chosen to represent the proportion of the variance in the load weight, which was explained by our prediction models

$$\mathrm{MAE}(y, \hat{y}) = \frac{\sum_{i=0}^{N-1} |y_i - \hat{y}_i|}{N}. \quad (4)$$

V. Feature Importance Analyses

The SHAP technique calculated the SHAP value of each feature. The SHAP value was a concept in game theory used to determine each player's contribution in a coalition or a cooperative game [40]. In a machine learning context, a cooperative game can be considered as a machine learning model which is making a prediction. The players can be considered the features used by the machine learning model to make the prediction. The SHAP value of a feature indicates how much the feature increases or decreases the model's predicted value from the baseline value. In this study, the features were the 11 700 (90 × 130) data points inside the STP matrix, and the sum of their SHAP values represents the difference between the baseline load (i.e., the expected model out on selected background samples) and the model's predicted load

$$\text{Prediction} = \text{Baseline} + \sum_{i=1}^{90\times 130} \text{SHAP}_i. \quad (5)$$

The Deep SHAP explainer, an enhanced version of the DeepLift algorithm [41], was picked to interpret the ResNet 19 model. Specifically, the explainer was first fit with the trained model with 100 background samples randomly drawn from the testing data. Then, the explainer would take each sample from the testing dataset as input in turn and estimate the SHAP value of each data point in the sample. The output of the explainer would also be a 90 × 130 matrix (6), where $S_{y,x}$ represents the SHAP value of sensor x at frame y

$$\text{SHAP Matrix} = \begin{pmatrix} |S_{0,0}| & |S_{0,1}| & \cdots & |S_{0,x}| \\ |S_{1,0}| & |S_{1,1}| & \cdots & |S_{1,x}| \\ \vdots & \vdots & \ddots & \vdots \\ |S_{y,0}| & |S_{y,1}| & \cdots & |S_{y,x}| \end{pmatrix}. \quad (6)$$

TABLE IV
Results of the LR Models

Models	Predicted R^2	MAE (Overall)	MAE (Light)	MAE (Medium)	MAE (Heavy)
LR (Average)	0.594	2.734	2.599	2.277	3.411
LR (Peak)	0.538	2.938	2.681	2.659	3.537

Using the SHAP values, three steps were conducted to summarize the importance of different sensors and frames. First, each SHAP value had a direction (positive or negative); however, since we focused primarily on the overall importance, we utilized the absolute value. Second, to determine the importance of each sensor, we summed up the absolute SHAP values of each sensor across frames, which was the sum of each column in (6). Similarly, to determine the importance of each frame, we summed up the absolute SHAP values at each frame across sensors, which was the sum of each row in (6). As a result of the previous calculations, for each sample, each sensor and each frame had a cumulative SHAP value, respectively. Finally, we determined each sensor's and each frame's mean cumulative SHAP value by taking the average across samples.

To further demonstrate the importance of each sensor, we followed the procedure in [12] to calculate the gripping force of each hand at each frame. Then, we determined the average and peak gripping force of each hand across frames and visualized them through box plots in Appendix (Figs. 9 and 10); these visualizations served as supplementary materials for researchers to further understand the hand force exertion pattern during lifting.

VI. Results

A. Lasso Regression Models

The results of the LR models are presented in Table IV. Specifically, the LR model built upon the average pressure achieved a predicted R^2 of 0.594 and an overall MAE of 2.734 kg. On the other hand, the LR model built upon the peak pressure achieved a predicted R^2 of 0.538 and an overall MAE of 2.938 kg. The average pressure model outperformed the peak pressure model. In addition, the medium weight lifts had the lowest MAE, and both models performed the best with the medium weight lifts in terms of MAE.

B. Convolutional Neural Networks

The results of the VGG models are presented in Table V. Specifically, the VGG 11 model, which is the shallowest model, achieved the best performance; specifically, the VGG 11 model achieved a predicted R^2 of 0.789 and an MAE of 1.737 kg. On the other hand, the VGG 19 model, which is the deepest model, achieved the worst performance. All VGG models performed the best under the lightweight level and performed the worst in the heavy weight lifts.

The results of the ResNet models are presented in Table VI. The ResNet 18 model, which is the shallowest model, achieved the best performance among the ResNet models (Table VI); specifically, it achieved a predicted R^2 of 0.821 and an MAE of 1.579 kg. On the other hand, the ResNet 34 model

TABLE V
RESULTS OF THE VGG MODELS

Models	Predicted R^2	MAE (Overall)	MAE (Light)	MAE (Medium)	MAE (Heavy)
VGG 11	0.789	1.737	1.289	1.808	2.136
VGG 13	0.780	1.773	1.354	1.887	2.091
VGG 16	0.783	1.750	1.324	1.715	2.249
VGG 19	0.777	1.802	1.384	1.756	2.303

TABLE VI
RESULTS OF THE RESNET MODELS

Models	Predicted R^2	MAE (Overall)	MAE (Light)	MAE (Medium)	MAE (Heavy)
ResNet 18	0.821	1.579	1.189	1.503	2.086
ResNet 34	0.799	1.643	1.068	1.666	2.235
ResNet 50	0.804	1.574	0.998	1.609	2.151

TABLE VII
RESULTS OF THE VIT MODELS

Models	Predicted R^2	MAE (Overall)	MAE (Light)	MAE (Medium)	MAE (Heavy)
ViT-Light 1	0.751	1.829	1.427	1.897	2.181
ViT-Light 2	0.728	1.889	1.441	2.011	2.227
ViT-Base 1	0.736	1.865	1.364	1.972	2.278
ViT-Base 2	0.714	1.928	1.487	1.970	2.352

achieved the worst performance. Similar to the previous deep-learning models, all ResNet models performed the best in the lightweight lifts and performed the worst in the heavy weight lifts.

C. Vision Transformers

The ViT—Light one model, which had eight transformer layers and eight attention heads, achieved the best performance among all ViT models (Table VII); specifically, it achieved a predicted R2 of 0.751 and an MAE of 1.829 kg. On the other hand, the ViT—Base two model, which had 12 transformer layers and 12 attention heads, achieved the worst performance. All ViT models performed the best in the lightweight lifts and performed the worst in the heavy weight lifts.

D. Feature Importance

Fig. 7 presents a visualization of the mean cumulative SHAP value of each sensor through color coding. Three trends can be observed from this figure. First, the right hand was more important to the model than the left hand. On average, the mean cumulative SHAP values of the right hand were 160.69% higher than that of the left hand.

Second, fingers were slightly more important than palms. The most important sensor on the right hand was on the middle finger, and the most important sensor on the left hand was on the index finger. Specifically, the mean cumulative SHAP values of the fingers were 7.12% higher than that of the palms. Finally, the left palm and both thumbs were weak contributors to the model's predictions.

Fig. 8 presents a visualization of the mean cumulative SHAP value of each frame through a scatter plot. The original data were processed by the moving average algorithm so that a clear trend could be observed. The model clearly relied more on the data in the middle phase than the beginning and ending phases. Specifically, the first and last few frames were the least important moments for the model. The data became increasingly important to the model as time went on. The data fall between frames 35 and 50 had cumulative SHAP values significantly higher than others, indicating that the model relied heavily on the data fall in this time interval to make predictions. After frame 50, the data became less and less important to the model.

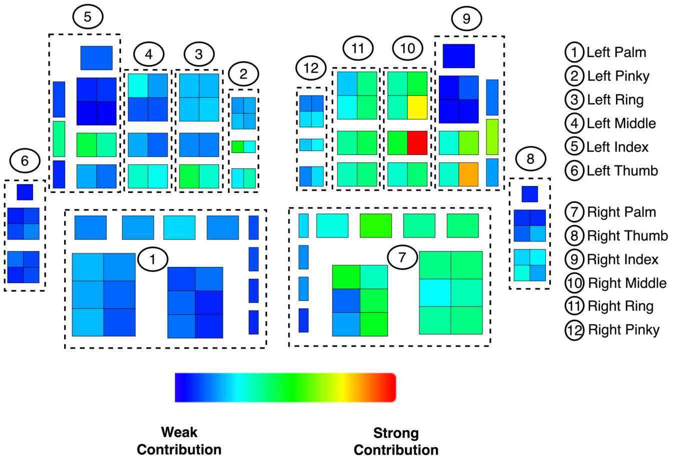


Fig. 7. Visualization of the importance of each sensor. The BGR color map was chosen; the smallest mean cumulative SHAP value was coded as blue, and the largest mean cumulative SHAP value was coded as red.

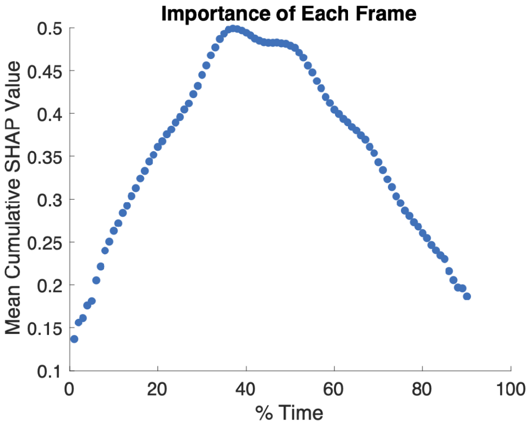


Fig. 8. Visualization of the importance of each frame. The original data were processed by the moving average algorithm and the peak value was found at frame 37.

VII. DISCUSSION

The load weight is a major risk factor significantly affecting workers' occupational health and safety across various industries. However, estimating the load weight is a challenging task for safety practitioners. This study utilized tactile gloves as a new modality to estimate the load weight. Furthermore, we proposed a custom data preparation and augmentation technique to formulate the pressure data and transferred image embedding techniques to learn representations from the pressure data in both temporal and spatial domains. Compared to existing modalities used to predict the weight load (i.e., EMG sensors and wearable motion sensors), gloves are pervasive and non-intrusive—commonly used by workers across many industries. In addition, by definition, pressure strongly correlates with force. Our experiments demonstrated that the proposed method could efficiently estimate the load weight. In SectionsVII-A–VII-E, we will discuss the modeling results, limitations, and the potential future works of this study.

A. Modeling Results—Lasso Regression Models

In a gripping force study [12], the authors used two simple linear regression models to investigate how much variance in the load weight could be explained by the average and peak gripping force, respectively. The gripping force was defined as the sum of each pressure sensor's exerted force, calculated as the product of pressure and sensor area. The gripping force was calculated at each frame first, and the peak and average gripping force were then found across all frames. Their results demonstrated that the average peak gripping force could explain 58.1% ($R^2 = 0.581$) and 57.4% ($R^2 = 0.574$) of the variance in the load weight, respectively.

Unlike the gripping force study accumulating sensors' readings at each frame, our study extracted features (i.e., peak and average) from each sensor separately. The R^2 of our LR model built upon the average pressure of each sensor was slightly higher than the R^2 of the average gripping force-based model (0.594 versus 0.581); however, the R^2 of our LR model built upon the peak pressure of each sensor was significantly lower than the R^2 of the peak gripping force-based model (0.538 versus 0.574). The degraded performance of our LR model could imply the importance of spatial information at each frame because it was not incorporated into our model as the peak gripping force-based model did; the peak reading of different sensors could be found at different frames.

Overall, the predictions of the LR models were not as accurate as the predictions of the CNN or ViT models. Nevertheless, these LR models only require a small amount of computational power to infer; hence, they are potentially suitable for embedding or edge computing (e.g., smart gloves embedded with a microcomputer), especially of our feature importance findings are used to optimize sensor designs.

B. Modeling Results—CNN and ViT

The CNN models were the most effective models in our study. By comparing the performance of the VGG and ResNet models individually, we observed that a deeper structure could not improve the performance of either the VGG or the ResNet models. Specifically, the VGG 11 model achieved a better performance than the other VGG models. Similarly, the ResNet 18 model was better than the other ResNet models. However, by comparing the performance of the VGG and the ResNet models, we observed that the residual connection technique had a positive impact because the ResNet models performed better than the VGG models. Our comparisons and observations on these models did not completely agree with the convention that a deeper network could lead to a better performance, which had been demonstrated in image classification tasks. The scarcity of our data could be a potential cause for this discrepancy because it could induce an overfitting issue for large models. Specifically, compared to RGB images used to train the original VGG and ResNet models (224 × 224 × 3), the STP matrix is relatively small (90 × 130). In addition, even with our proposed data augmentation technique, only 35 517 samples were available in the dataset, which is relatively tiny compared to most image datasets.

Similarly, the lightweight ViT models outperformed the other ViT models in our experiment. In addition, the CNN models clearly outperformed the ViT models in our study. Compared to the CNN models, the ViT models require more data to attain better performance and are less efficient on small datasets because of the lack of inherent inductive biases [42]. However, numerous researchers have been improving the ViT models with different techniques (see [43], [44]). Therefore, future studies can be focused on exploring how these improved ViT models can be modified or transferred to our prediction task.

C. Feature Importance

By using the SHAP technique, this study discovered that the ResNet 18 model relied more on the right hand than the left hand to make load weight predictions. This result could be explained by the participants' handedness and grip strength. Specifically, studies have demonstrated that the grip strength of the right hand was 12.72% larger than that of the left hand for right-handed people [45]. In contrast, the grip strength tended to be similar between the left and the right hand for left-handed people. Specifically, a study in [46] conducted an experiment that demonstrated that 10.93% of the right-handed subjects had their nondominant hand stronger than their dominant hand. In comparison, 33.33% of the left-handed subjects had their nondominant hand stronger than their dominant hand. Based on these grip strength studies, there are indications that right-hand grip strength is higher than left-hand grip strength regardless of handedness. This could be one potential explanation for why our model relied more on the right hand, as more weight could be absorbed or counterbalanced by the stronger side. Further studies should be conducted to investigate the impact of handedness during lifting using tactile gloves.

Moreover, the model relied slightly more on the fingers than the palms to predict load weight. The handles' shape could increase the fingers' importance because the fingers could primarily wrap around the complete cut-out and square edge handles during lifting. In addition, we visualized and compared the SHAP values of different hand regions across handles, illustrating the difference in hand pressure distribution caused by different handles (see Appendix). These findings are inconclusive for comprehending the hand pressure pattern during lifting since they only indicated each hand region's importance to the model. In addition, they may not be generalizable to all lifting tasks because two types of handles were considered. Further studies and experiments should be conducted to analyze lifting tasks involving other types of handles.

Finally, this study found that the pressure data lying in the middle phase of lifting tasks are important for the model to predict the load weight. There are two potential causes of this trend. First, the sensors' readings may be relatively steady in the middle phase, while the beginning and ending phases involve rapid changes. The relatively steady middle phase may be more helpful for predicting load weight than the force patterns in other phases of the lifting tasks. Second, the middle phase of the recordings was scanned more often than the other phases, leading to a relatively imbalanced input distribution. Specifically, the middle phase of the recordings was included in most of the samples used to develop the models, while the starting and ending phases were only included in a relatively small proportion of the samples. For instance, for a lift that lasted for 100 frames, most samples generated by the sliding window would include the 50th frame, while only a few would include the first frame. Models trained on such imbalanced input distribution could exhibit bias toward the pressure readings in the middle phase [47].

In summary, the feature importance analyses increased the transparency of the ResNet 18 model by revealing some rationales for the model's decision-making process. These analyses also provided new scientific insights regarding the hand force exertion pattern during lifting.

D. Practical Applications

For our target application for assessing lifting injury risks, the proposed load weight estimation method can directly provide load weight estimations for practitioners to input for the NIOSH lifting equation. However, lifting at the workplace can be highly varied, and several task conditions may affect the accuracy of our pipeline. These factors include lifts with boxes without handles, load-carrying tasks, one-handed lifts, and lifting at paces shorter than 39 or more than 125 frames. These factors may affect our observed accuracy since the proposed model has not seen the hand pressure pattern in such tasks. The model will have to be trained for these additional factors; however, the lifts studied in this work are representative of common lifts observed in the workplace. To apply our model, workers will need to wear the glove system, and safety practitioners will need to apply the pipeline and model to monitor load weights over time.

In addition, many weights, a person can lift, were not included in the training dataset. Although not in our train set, the LR model presented in the results can potentially infer random weights unseen by our model, especially if the random weight is within the range of weights in our train dataset, which is a key strength of our presented LR model. For our deep-learning-based regression models (i.e., the CNN and ViT models), the accuracy will likely be furthered lower than the accuracy presented in our results because the model was optimized for the weights in our dataset, which means the model could be overfit to the existing weights. Further experiments with additional weights are needed to demonstrate or investigate the models' performances on those unseen weights.

E. Limitation and Future Work

Based on the presented results and previous discussions, five major limitations were found, and the corresponding future solutions were proposed.

First, the conducted experiment and the resulting dataset needed to be more comprehensive to develop generalizable load weight prediction models because the task conditions vary at the workplace. For example, lifting tasks involving objects without handles or single-handed lifting tasks were missing in the experiment and dataset, while they can be commonly observed at the workplace. In addition, the current models did not consider lifting shaped objects and lifting with physical contact between the lifted objects and workers. Future studies can focus on collecting data covering a wide range of lifting tasks and conditions for improving generalizability.

Second, the sliding window utilized in this study may not be suitable for lifting tasks with extremely short or long recordings (e.g., shorter than 39 or longer than 125 frames). For short lifts, too many unrelated pressure readings can be included and potentially affect the model's prediction performance; for long lifts, the information extracted by the sliding window can potentially be insufficient for the model to make reliable predictions. Future studies can explore different techniques for handling these corner cases; some potential methods include utilizing random downsampling and interpolation techniques.

Third, the results demonstrated that the tactile gloves were not suitable for analyzing heavy tasks (i.e., load weight > 13.6 kg), potentially because of the dead spaces between sensors [12]; however, heavy tasks can bring more risks to workers than tasks involving medium or light weights. There is a need to explore a solution for increasing the model's performance for heavy lifting tasks. A potential method is to develop a specialized regression model for analyzing heavy lifts only. Specifically, the specialized regression model should be trained only on heavy lifts.

Fourth, this study only transferred and implemented a handful of classic CNN and ViT architectures, while there are many other architectures of CNN and ViT [42], [48]. Future studies can be focused on transferring different architectures of CNN and ViT or developing custom architectures to further improve the prediction performance.

Finally, in addition to the load weight, there are other factors that can also affect the lifting risks, such as body kinematics and posture [49], [50]. This study only focused on predicting the load weight while did not account for other risk factors. For practitioners to conduct comprehensive risk assessments, the tactile glove should be integrated with other modalities capable of measuring other risk factors in future studies. For example, body kinematics and posture can be measured by computer vision techniques and wearable motion sensors [6], [25]; thus, integrating cameras or motion sensors with tactile gloves can potentially help practitioners conduct comprehensive lifting risk assessments.

VIII. Conclusion

This study proposed utilizing tactile gloves as a new modality for predicting the load weight, one of the most important risk factors of lifting tasks. To validate the feasibility, a human subjects experiment was first conducted to collect hand pressure data of participants performing various lifting tasks. Then, a computational experiment was performed. A custom data preparation and augmentation technique was proposed to formulate the hand pressure data into 2-D STP matrices. Convolutional neural networks and ViTs were modified and

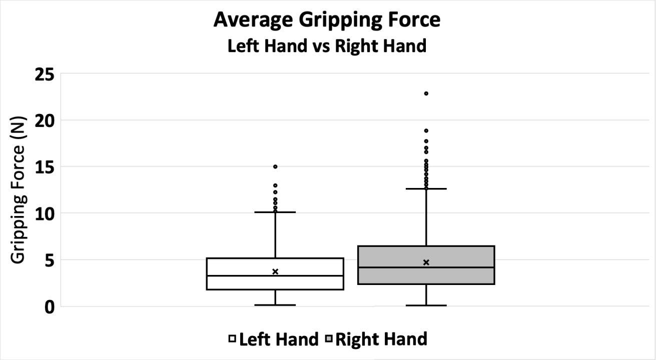


Fig. 9. Comparison of the average gripping force between the left and right hand.

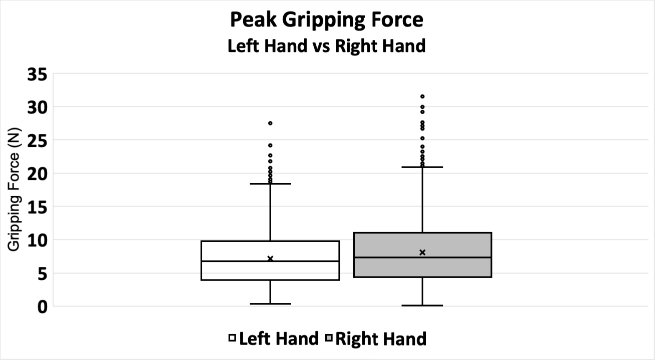


Fig. 10. Comparison of the peak gripping force between the left and right hand.

transferred to learn representations from the STP matrices for predicting the load weight. The ResNet 18 model was the best-performing model, and it achieved a predicted *R*-squared of 0.821 and a mean absolute error of 1.579 kg.

In addition, the SHAPs technique was used to investigate the decision-making logic of the ResNet 18 model. Specifically, the importance of each sensor and the importance of each frame were determined. From the spatial perspective, the data from the right hand were more important than the data from the left hand, and the data from the fingers were more important than the data from the palms. From the temporal perspective, the data lying in the middle phase of lifting tasks were more important than the data lying in the beginning and ending phases.

In summary, this study demonstrated the efficiency of utilizing tactile gloves with the ResNet 18 model to predict the load weight, and the SHAP technique increased the transparency of the model and enhanced our understanding of hand force exertion during lifting tasks.

Appendix

A. Comparison of Gripping Force Between Hands

We visualized and compared the average and peak gripping force across the left and right hands in Figs. 9 and 10. These two figures illustrated that the right hand exerted more force than the left hand on average. In general, this trend aligned with the decision-making logic, where the SHAP values of the right hand were larger than that of the left hand.

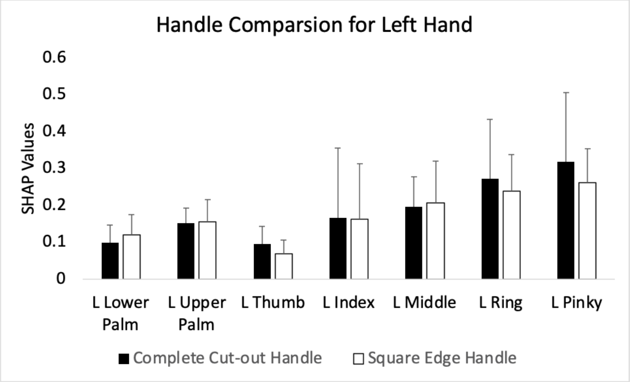


Fig. 11. Comparison of the SHAP values between handles for the left hand.

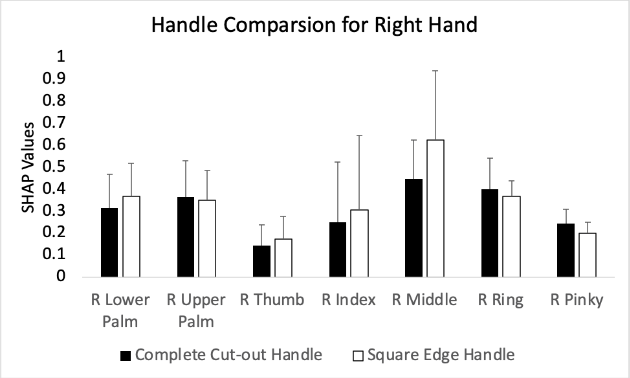


Fig. 12. Comparison of the SHAP values between handles for the right hand.

B. Comparison of SHAP Values Between Handles

We visualized and compared the SHAP values of different hand regions between handle types in Figs. 11 and 12. Average SHAP values for both handles were observed to be similar regardless of handle type. However, the right middle fingers appeared more important for the square edge handle than the complete cut-out handle.

Acknowledgment

Findings and conclusions in this report are those of the authors and do not necessarily represent the official positions of the National Institute for Occupational Safety and Health (NIOSH) and Centers for Disease Control and Prevention (CDC). Mention of any company or product does not constitute endorsement by NIOSH, CDC.

References

[1] V. Anderson et al., "Musculoskeletal disorders and workplace factors: A critical review of epidemiologic evidence for work-related musculoskeletal disorders of the neck, upper extremity, and low back," Nat. Inst. Occupat. Saf. Health, Cincinnati, OH, USA, Tech. Rep. DHHS (NIOSH) Publication 97B141, 1997.

[2] T. R. Waters and A. Garg, "Applications manual for the revised NIOSH lifting equation," DHHS, Washington, DC, USA, Tech. Rep., 94-110, 1994.

[3] J. R. Potvin, V. M. Ciriello, S. H. Snook, W. S. Maynard, and G. E. Brogmus, "The liberty mutual manual materials handling (LM-MMH) equations," *Ergonomics*, vol. 64, no. 8, pp. 955–970, Aug. 2021.

[4] S. Gallagher, R. F. Sesek, M. C. Schall, and R. Huangfu, "Development and validation of an easy-to-use risk assessment tool for cumulative low back loading: The lifting fatigue failure tool (LiFFT)," *Appl. Ergonom.*, vol. 63, pp. 142–150, Sep. 2017.

[5] S. D. Hlucny and D. Novak, "Characterizing human box-lifting behavior using wearable inertial motion sensors," *Sensors*, vol. 20, no. 8, p. 2323, Apr. 2020.

[6] G. Zhou, V. Aggarwal, M. Yin, and D. Yu, "A computer vision approach for estimating lifting load contributors to injury risk," *IEEE Trans. Hum.-Mach. Syst.*, vol. 52, no. 2, pp. 207–219, Apr. 2022.

[7] M. H. Jali, T. A. Izzuddin, Z. H. Bohari, H. I. Jaafar, and M. N. M. Nasir, "Pattern recognition of EMG signal during load lifting using artificial neural network (ANN)," in *Proc. IEEE Int. Conf. Control Syst., Comput. Eng. (ICCSCE)*, Nov. 2015, pp. 172–177.

[8] M. Caeiro-Rodríguez, I. Otero-González, F. A. Mikic-Fonte, and M. Llamas-Nistal, "A systematic review of commercial smart gloves: Current status and applications," *Sensors*, vol. 21, no. 8, p. 2667, Apr. 2021.

[9] Q. Ye, M. Seyedi, Z. Cai, and D. T. H. Lai, "Force-sensing glove system for measurement of hand forces during motorbike riding," *Int. J. Distrib. Sensor Netw.*, vol. 11, no. 11, Nov. 2015, Art. no. 545643.

[10] Y.-K. Kong and B. D. Lowe, "Optimal cylindrical handle diameter for grip force tasks," *Int. J. Ind. Ergonom.*, vol. 35, no. 6, pp. 495–507, Jun. 2005.

[11] M.-L. Lu, T. James, B. Lowe, M. Barrero, and Y.-K. Kong, "An investigation of hand forces and postures for using selected mechanical pipettes," *Int. J. Ind. Ergonom.*, vol. 38, no. 1, pp. 18–29, Jan. 2008.

[12] G. Zhou, M.-L. Lu, and D. Yu, "Investigating gripping force during lifting tasks using a pressure sensing glove system," *Appl. Ergonom.*, vol. 107, Feb. 2023, Art. no. 103917.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[15] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent.*, May 2021, pp. 1–12.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 4768–4777.

[17] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[18] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: A comprehensive review," *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 3503–3568, Jun. 2022.

[19] B. D. Lowe, P. G. Dempsey, and E. M. Jones, "Ergonomics assessment methods used by ergonomics professionals," *Appl. Ergonom.*, vol. 81, Nov. 2019, Art. no. 102882.

[20] M.-L. Lu, T. R. Waters, E. Krieg, and D. Werren, "Efficacy of the revised NIOSH lifting equation to predict risk of low-back pain associated with manual lifting: A one-year prospective study," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 56, no. 1, pp. 73–85, Feb. 2014.

[21] M.-L. Lu, V. Putz-Anderson, A. Garg, and K. G. Davis, "Evaluation of the impact of the revised national institute for occupational safety and health lifting equation," *Hum. Factors, J. Human Factors Ergonom. Soc.*, vol. 58, no. 5, pp. 667–682, Aug. 2016.

[22] R. R. Fox, M.-L. Lu, E. Occhipinti, and M. Jaeger, "Understanding outcome metrics of the revised NIOSH lifting equation," *Appl. Ergonom.*, vol. 81, Nov. 2019, Art. no. 102897.

[23] M. S. Barim, M.-L. Lu, S. Feng, G. Hughes, M. Hayden, and D. Werren, "Accuracy of an algorithm using motion data of five wearable IMU sensors for estimating lifting duration and lifting risk factors," *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 63, no. 1, pp. 1105–1111, 2019.

[24] M.-L. Lu, M. S. Barim, S. Feng, G. Hughes, M. Hayden, and D. Werren, "Development of a wearable IMU system for automatically assessing lifting risk factors," in *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Posture, Motion and Health*, V. G. Duffy, Ed. Cham, Switzerland: Springer, 2020, pp. 194–213.

[25] L. Donisi, G. Cesarelli, A. Coccia, M. Panigazzi, E. M. Capodaglio, and G. D'Addio, "Work-related risk assessment according to the revised NIOSH lifting equation: A preliminary study using a wearable inertial sensor and machine learning," *Sensors*, vol. 21, no. 8, p. 2593, Apr. 2021.

[26] A. Patrizi, E. Pennestrì, and P. P. Valentini, "Comparison between low-cost marker-less and high-end marker-based motion capture systems for the computer-aided assessment of working ergonomics," *Ergonomics*, vol. 59, no. 1, pp. 155–162, Jan. 2016.

[27] J. T. Spector, M. Lieblich, S. Bao, K. McQuade, and M. Hughes, "Automation of workplace lifting hazard assessment for musculoskeletal injury prevention," *Ann. Occupational Environ. Med.*, vol. 26, no. 1, p. 15, Dec. 2014.

[28] X. Wang, Y. H. Hu, M.-L. Lu, and R. G. Radwin, "The accuracy of a 2D video-based lifting monitor," *Ergonomics*, vol. 62, no. 8, pp. 1043–1054, Aug. 2019.

[29] Y. Li, R. L. Greene, F. Mu, Y. H. Hu, and R. G. Radwin, "Towards video-based automatic lifting load prediction," *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 64, no. 1, pp. 962–963, Dec. 2020.

[30] D. Totah, L. Ojeda, D. D. Johnson, D. Gates, E. M. Provost, and K. Barton, "Low-back electromyography (EMG) data-driven load classification for dynamic lifting tasks," *PLoS ONE*, vol. 13, no. 2, Feb. 2018, Art. no. e0192938.

[31] S. Chan, "The use of EMG for load prediction during manual lifting," Ph.D. dissertation, Dept. Elect. Comput. Eng., Queen's Univ., Kingston, ON, Canada, 2007.

[32] S. Sundaram, P. Kellnhofer, Y. Li, J.-Y. Zhu, A. Torralba, and W. Matusik, "Learning the signatures of the human grasp using a scalable tactile glove," *Nature*, vol. 569, no. 7758, pp. 698–702, May 2019.

[33] S. Garg, R. K. Patro, S. Behera, N. P. Tigga, and R. Pandey, "An overlapping sliding window and combined features based emotion recognition system for EEG signals," *Appl. Comput. Informat.*, vol. 2021, pp. 1–12, Aug. 2021.

[34] M. A. F. da Silva, R. L. De Carvalho, and T. D. S. Almeida, "Evaluation of a sliding window mechanism as DataAugmentation over emotion detection on speech," *Academic J. Comput., Eng. Appl. Math.*, vol. 2, no. 1, pp. 11–18, Apr. 2021.

[35] S. J. Wright, "Coordinate descent algorithms," *Math. Program.*, vol. 151, no. 1, pp. 3–34, Jun. 2015, doi: 10.1007/s10107-015-0892-3.

[36] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.

[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[39] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, pp. 79–82, 2005.

[40] S. Hart, "Shapley value," in *Game Theory*, J. Eatwell, M. Milgate, and P. Newman, Eds. London, U.K.: Palgrave, 1989, pp. 210–216.

[41] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3145–3153.

[42] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[43] S. Lee, S. Lee, and B. Cheol Song, "Improving vision transformers to learn small-size dataset from scratch," *IEEE Access*, vol. 10, pp. 123212–123224, 2022.

[44] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[45] P. Petersen, M. Petrick, H. Connor, and D. Conklin, "Grip strength and hand dominance: Challenging the 10% rule," *Amer. J. Occupational Therapy*, vol. 43, no. 7, pp. 444–447, Jul. 1989.

[46] N. Incel, E. Ceceli, P. Durukan, H. Erdem, and Z. Yorgancioglu, "Grip strength: Effect of hand dominance," *Singap. Med. J.*, vol. 43, pp. 234–237, Jan. 2002.

[47] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, p. 27, Mar. 2019, doi: 10.1186/s40537-019-0192-5.

[48] L. Alzubaidi et al., "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[49] K. O. Greenland, A. S. Merryweather, and D. S. Bloswick, "The effect of lifting speed on cumulative and peak biomechanical loading for symmetric lifting tasks," *Saf. Health at Work*, vol. 4, no. 2, pp. 105–110, Jun. 2013.

[50] M. F. Antwi-Afari, H. Li, D. J. Edwards, E. A. Pärn, J. Seo, and A. Y. L. Wong, "Biomechanical analysis of risk factors for work-related musculoskeletal disorders during repetitive lifting task in construction workers," *Autom. Construct.*, vol. 83, pp. 41–47, Nov. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926580517303898