Machine Learning for Detection and Risk Assessment of Lifting Action

Brennan Thomas, Ming-Lun Lu, Rashmi Jha, and Joseph Bertrand

***Abstract*—Repetitive occupational lifting has been shown to create an increased risk for incidence of back pain. Ergonomic workstations that promote proper lifting technique can reduce risk, but it is difficult to assess the workstations without constant risk monitoring. Machine learning systems using inertial measurement unit (IMU) data have been successful in various human activity recognition (HAR) applications, but limited work has been done regarding tasks for which it is difficult to collect significant amounts of data, such as manual lifting tasks. In this article, we discuss why traditional methods of data expansion may fail to improve performance on IMU data, and we present a machine learning system capable of detecting lifting action for assessing the risk for back pain using a relatively small amount of data. The proposed models outperform baseline HAR models and function on raw time-series data with minimal preprocessing for efficient real-time application.**

***Index Terms*—Industrial/organizational/workplace safety, injury/illness prevention, machine learning, risk assessment, signal detection, wearable systems.**

I. Introduction

BACK pain is a leading cause of occupation-induced disability and results in large amounts of lost productivity annually [1]. Most occupational back pain is caused by roles requiring repetitive tasks such as heavy lifting of objects. There are many ergonomic guidelines for lifting with the goal of reducing the risk of back pain, but it is difficult for workers to consistently follow these guidelines in all situations. Many may be unable to lift in the required manner due to poor workstation designs or physical limitations [2]. Monitoring of lifting risk can ensure proper technique as well as determine if workers are consistently required to lift in an unsafe manner as part of their daily tasks.

Common ergonomic risk assessment methods involve a safety professional observing a worker's lifting postures for numerous manual materials handling tasks during the entire work shift. This type of assessment requires intensive labor and cannot provide sufficient information for an unbiased risk assessment. An automated system that can both detect lifting action and classify its risk would scale up to more workers and provide real-time feedback to both reduce the risk of health complications for the worker and provide long-term safety data for supervisors.

Both lift detection and lift classification are human activity recognition (HAR) problems, which have been extensively studied in the machine learning field under several different sensor modalities [3]. While HAR systems have been successful for video-based activity recognition [4], automatic lifting risk assessment would require permanent installation of cameras in every possible lift location. This is prohibitively expensive, raises concerns about worker privacy, and is impractical in temporary work sites such as construction zones. In contrast, wearable inertial measurement unit (IMU) sensors can detect motion regardless of position and have fewer privacy concerns compared to video surveillance. IMU sensors have already seen widespread use in technology such as smart watches and therefore provide a tremendous advantage for HAR tasks.

A significant challenge in designing a HAR system for lift risk assessment is the amount of available data. Many HAR systems are designed for common, repetitive action such as walking, running, standing, and sitting, of which there is an abundance of publicly available data [5], [6], [7]. Lifting, in contrast, is a very specific action that does not have a publicly available large dataset, as collecting data for lifting action requires a more specific setup and time-consuming controlled trials for proper labeling. As a result, the dataset used for this study has significantly less training data available for a HAR system built to assess lift risk compared to larger, more general datasets [8]. While there are some strategies for reducing the negative effects of a limited dataset, such as transfer learning and data augmentation, they have unique challenges when used for IMU data compared to the more common image domain, as discussed in Section VI.

Thirty-three studies using machine learning algorithms for musculo-skeletal disorder (MSD) risk assessments were identified in a recent review by Chan et al. [9], however, none of the identified studies used neural network techniques to predict the physical risk factors for MSDs, in particular the hand location of the load in relation to the body and the ground. These two factors are two important variables used in the revised NIOSH lifting equation (RNLE) and the American Conference of Governmental Industrial Hygienists (ACGIH) Threshold Limit

Manuscript received 3 October 2021; revised 22 January 2022 and 24 June 2022; accepted 25 September 2022. Date of publication 20 October 2022; date of current version 15 November 2022. This work was supported in part by the National Institute for Occupational Safety and Health (NIOSH) and National Science Foundation (NSF) under Grant ECCS 1926465 and in part by REU supplement for NSF under Grant SHF 1718428. This article was recommended by Associate Editor M. L. Bolton. *(Corresponding author: Brennan Thomas.)*

Brennan Thomas, Rashmi Jha, and Joseph Bertrand are with the Department of Electrical Engineering and Computer Science, University of Cincinnati, Cincinnati, OH 45219 USA (e-mail: thoma2bm@mail.uc.edu; jhari@ucmail.uc.edu; bertrajh@mail.uc.edu).

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

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2022.3212666.

Digital Object Identifier 10.1109/THMS.2022.3212666

Values (TLV) for Lifting [10], [11]. The minor differences in the distance between the load and the body to classify the risk of a lifting task may result in large differences in the risk of back pain. The similarity between two different risk classes of lifts means that classifying them is particularly difficult. Other HAR systems that focus on classifying actions such as walking, running, and sitting work with inherently more separable data, and show degraded performance on activities that are more similar, such as walking upstairs versus walking downstairs [7]. Therefore, a risk assessment system must be able to learn minute differences between similar activities from a substantially small dataset.

We present a machine learning framework for detecting lifting action and classifying the risk level of the lift. The proposed models perform well despite limitations of available data, and could be extended to other physical actions that need a more in-depth analysis in the HAR domain.

II. Related Work

There has been significant research into machine learning methods for HAR on common daily activities. Chen and Xue [6] collected tri-axial accelerometer data of common activities such as walking, running, jumping, falling, etc. and attempted to classify them using a convolutional neural network (CNN). The kernel of the CNN was sized to span two out of three of the axes, to capture axial relationships of the data.

Weiss et al. [5] introduced Actitracker, a smartphone-based activity monitoring system that provided activity reports to the user with nearly 90% accuracy based on user reports. Actitracker employed a universal model that became more personalized to the user after an initial training phase.

Ignatov [7] combined hand-crafted statistical features such as mean and variance with features extracted via a shallow CNN as input to a fully connected layer for classification. They evaluated the model on two different datasets and performed a cross-dataset evaluation by training on one dataset and evaluating the model performance on the other. Inclusion of the statistical features resulted in a significant boost in performance.

Anguita et al. [8] introduced a dataset of daily living activities collected via smartphone and evaluated the performance of a support vector machine (SVM) used to classify the activities. The SVM achieved excellent performance on most classes but showed relatively worse performance on the sitting class, often misclassifying it as standing.

Ordoñez and Roggen [12] utilized stacked 1-D convolutional and long short-term memory (LSTM) layers to perform automatic feature extraction and captured temporal relationships between these features, achieving state-of-the-art scores on the OPPORTUNITY [13] and Skoda [14] datasets.

Hammerla et al. [15] compared several different model architectures across three different datasets, including standard deep neural networks, CNNs, and recurrent neural networks (RNNs) including LSTM, achieving best results utilizing a bi-directional LSTM network.

Ravì et al. [16] introduced a low-power model for real-time classification on wearable devices. Features were extracted from the spectral domain to reduce variation due to sensor placements or rotation.

Bhattacharya and Lane [17] implemented a Restricted Boltzmann Machines on a smartwatch, showing competitive performance on a range of classification tasks while utilizing an acceptable level of resource consumption for the smartwatch platform.

Guan and Ploetz [18] utilized an ensemble of LSTM models and achieved performance improvements over several recent methods across multiple datasets, showing that ensemble methods can improve HAR capabilities in many situations.

Finally, Snyder et al. [19] performed risk analysis of the NIOSH lifting dataset by reformatting the time-series IMU data as a 2-D image that is filtered and then used as input to a CNN utilizing average pooling rather than max pooling. They achieved 90.6% accuracy in classifying the risk level of a lift. The current work expands on this research by introducing a method for risk analysis requiring less preprocessing and including the detection step of the task.

The majority of past works were evaluated with common publicly available datasets, such as OPPORTUNITY and Skoda. While this allows the performance of different models to be directly compared, it represents a hole in current HAR research, where models are developed for common household tasks and not much else. There is very little research regarding utilizing machine learning for specialized tasks.

This work addresses the gap in current research by focusing on detecting a specialized task and separating instances within that task.

III. NIOSH Lifting Dataset

The IMU dataset used in this study was collected by NIOSH [20]. It consists of data collected from six motion sensors (Kinetic Inc.) placed in specific locations on subjects' bodies while each subject performed lifting actions at various levels of risk for back pain. Six sensors were used to cover a wide range of placements and help determine, which sensors would be most useful in settings with fewer sensors.

Each lift was performed in one of the 12 zones defined by the ACGIH TLV for Lifting [11]. This TLV defines limits for movement in both safe and unsafe lifting, and can be mapped to zones in various levels of low, medium, or high risk. These zones, shown in Fig. 1, are defined relative to the lifter's body in the sagittal plane. Subjects lifted a 36 cm × 12 cm wire grid with handles to simulate a crate. To prevent injury, the weight was kept small at 0.45 kg. After lifting, subjects turned 180°, walked a short distance and placed the wire grid in a designated area. Subject demographics are shown in Table I.

Data were collected from ten subjects fitted with six IMU sensors on the dominant thigh, dominant upper arm, waist, upper back, and each wrist, as shown in Fig. 2. Each IMU sensor sampled tri-axial accelerometer and gyroscope data at a rate of 25 Hz. Sensors were synchronized and calibrated prior to each trial. Each subject performed six trials in each zone, for a total of 720 lifting trials. However, one of the subjects had incorrectly aligned wrist sensors during half of the trials, and

TABLE I
PARTICIPANT DEMOGRAPHICS

Gender	Count	Age	Height (cm)	Weight (kg)	BMI
Male	5	55 ± 1.87	176.28 ± 1.39	101.8 ± 14.65	32.77 ± 4.81
Female	5	48 ± 13.55	163.58 ± 4.61	69.66 ± 7.39	26.06 ± 2.95
Total	10	52 ± 9.83	169.93 ± 7.42	85.72 ± 20.15	29.42 ± 5.16

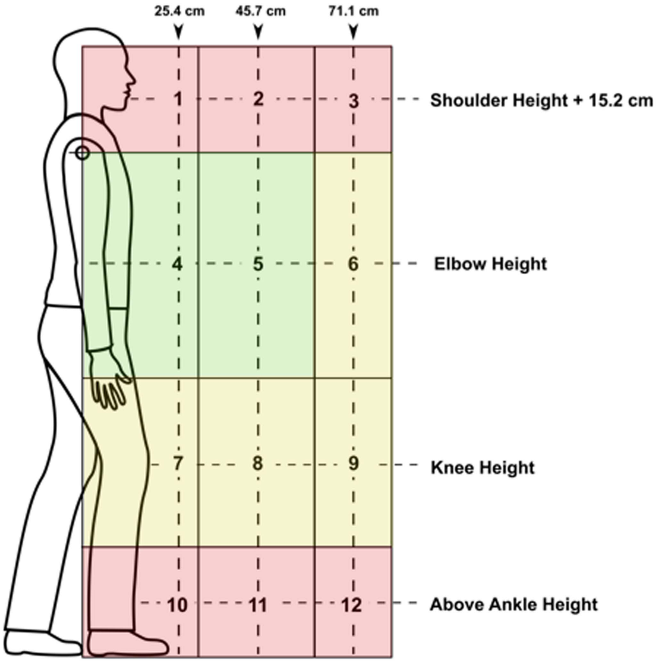

Fig. 1. Lifting zones defined by ACGIH Lifting TLV. Green zones (4-5) are considered Low Risk, yellow zones (6-9) are Medium Risk, and red zones (1-3, 10-12) are High Risk. Intersections of dashed lines indicate where in the zone the object is lifted from. (Source: NIOSH).

was removed from the dataset, resulting in 684 usable trials. Due to the distribution of risk assignments to each zone, there were twice as many medium-risk lifts as low-risk, and three times as many high-risk. Each subject (with the exception of one due to the misaligned sensors) also performed 11 trials involving activities such as jumping, walking, standing, and sitting, for an additional 99 nonlifting trials.

IV. Lift Detection

Lift detection shares considerations and difficulties with anomaly detection, a field in which machine learning has seen extensive use [21]. In both situations, the detection system must detect an event that is considered exceptional, and therefore occurs much less often in training data than nonexceptional events. This can often lead to high rates of false positives, as the system must be sensitive enough to detect these events, but may be too sensitive and detect an event where there is none. In lift detection specifically, the system is trying to detect a specific event, rather than any anomalous event. Therefore, traditional unsupervised methods used in anomaly detection that detect any deviation from the baseline, such as autoencoders, would detect actions that are not included in the training data but are also not lifts, and are not useful for lift detection.

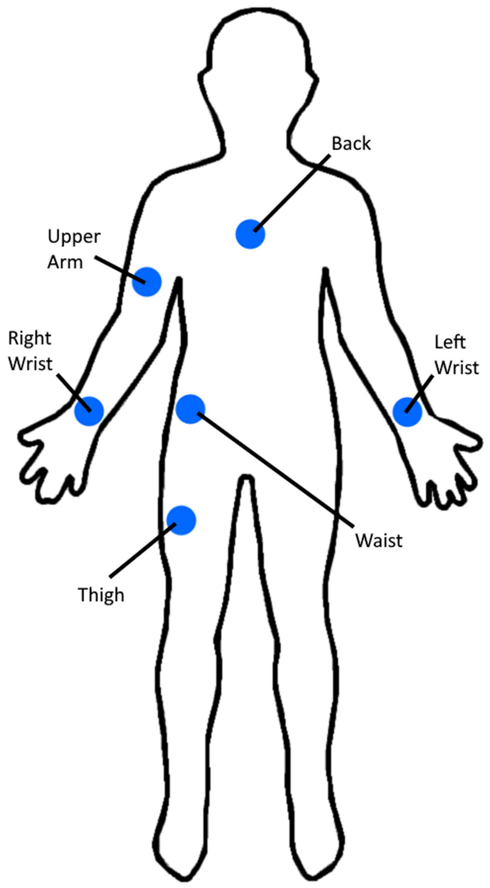

Fig. 2. Placements of six IMU sensors on subjects during data collection.

The proposed lift detection system utilizes a sliding window to segment lifting trials into discrete samples. The label of a segment is determined in a probabilistic manner based on the proximity of the last time step in the segment to the start or end of the lift. For example, a segment that ends exactly at the start of the lift is given a hard label of 100% lift start. A segment that ends some number of frames before the lift is assigned a fuzzy label giving probabilities for both lift start and neutral. A visual depiction of these labels for an entire trial is shown in Fig. 3. Labeling based on the last frame in the segment allows the model to operate in an "on-line" fashion, where a prediction for a specific moment in time uses only information that is already available. For instance, the model can make a prediction at time t using time segment $[t-24, t]$, and then continue on to predict the next time step $t + 1$ using $[t-23, t + 1]$. In this way, the system can perform lift detection in real-time, with delay on the order of a few milliseconds depending on hardware. Raw sensor data with no filtering were used as input for the model. Preliminary tests showed that the models performed better with raw data, likely

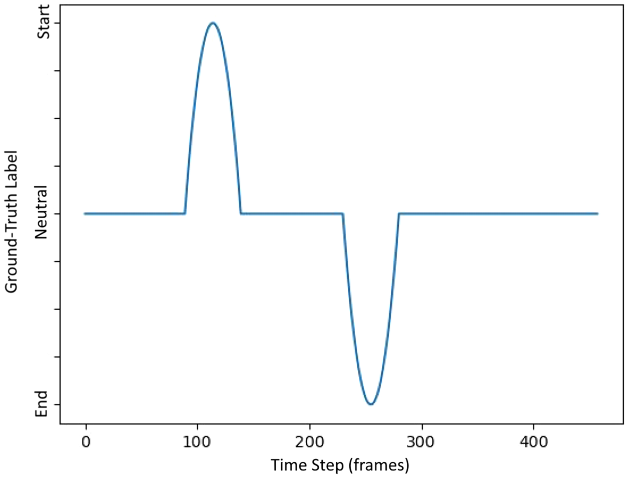

Fig. 3. Ground-truth labels (blue) generated by the sliding window segmentation method.

TABLE II
DETECTION MODEL ARCHITECTURE

Layer	Shape	Activation
Input	25 x 36	N/A
Conv1D	21 x 64	ReLU
MaxPool1D	10 x 64	None
Conv1D	6 x 128	ReLU
MaxPool1D	3 x 128	None
LSTM	3 x 128	tanh
LSTM	128	tanh
Output	3	softmax

because filtering and normalizing the data reduced relationships of magnitude between sensor axes. Additionally, the reduced need for preprocessing increased the response time of the model.

The lift detection model consisted of an LSTM layer followed by a fully connected softmax output layer that classified the sample as a lift start, lift end, or neither. The model architecture is shown in Table II. The model was implemented in Python using Tensorflow 2.0 with the Keras application programming interface (API). Training was performed using a categorical cross-entropy loss function, with Adam [22] as the gradient descent algorithm. To ensure the model was fully trained, an early stopping module that monitored validation loss was used. Training was therefore halted if there was no improvement in validation loss after 50 epochs. After training, the weights corresponding to the lowest validation loss from the last 50 epochs were kept. This prevented overfitting while ensuring that the model was trained completely. To reduce variance caused by the relatively small dataset, a leave-one-subject-out (LOSO) cross-validation (CV) was considered for evaluation of the model [23]. In LOSO CV, a separate model is trained for each subject, holding that subject out for validation and using the remaining subjects for training. The results of all models' performances are then averaged for the final evaluation, simulating the model's performance on an unseen subject. However, an examination of the video data of certain trials showed that there is often an offset between the actual lift sequence in the IMU data and the location the start of the lift is marked at, and this offset is consistent within a single subject. Since there is a systemic difference between the labeling patterns of different subjects, LOSO CV could not be used for evaluation. Therefore, a standard K-Fold CV was employed instead, though LOSO CV should be preferred in cases where there is no systemic labeling difference between subjects to reduce subject bias. Additionally, to avoid poisoning of the model with samples labeled incorrectly, samples with lift start labels several seconds away from the observed lift were removed from the training data for this model.

V. LIFT CLASSIFICATION

The proposed lift classification system used an entire trial sequence as input. The impetus for this was twofold; first, using the entire sequence improved the translational invariance of the model. Since the lifting action could occur anywhere within the sequence of data, the model had to learn to classify the lift regardless of its location. Second, human movement is affected by the intent of the actor, and therefore there may be trainable data in the motion both before and after the lift [24]. Using the entire sequence ensures that this information is captured and used. Since the length of the lifting trials varied, each trial was zero-padded to the length of the longest trial, resulting in samples with a length of 750 frames and 36 channels.

The architecture of the model was based on DeepConvLSTM, an LSTM-based activity recognition model consisting of convolutional layers acting as feature extractors followed by LSTM layers modeling temporal relationships between features. This combination allows the model to accept IMU acceleration and gyroscopic angular velocity data directly, thus, avoiding the need for deriving additional features from the data. This simplicity allows the model to be easily applied to real time systems [12].

The model architecture was modified from the standard DeepConvLSTM by replacing the convolutional layers with residual blocks. Residual blocks are formed of convolutional layers with the addition of skip connections, which prevent vanishing or exploding gradients and allow for easier training of deeper networks [25]. In this model, residual blocks with preactivation are used [26], as they have been shown to improve regularization and ease optimization compared to traditional residual blocks. A total of six residual blocks and two LSTM layers with 128 units each form the base of the model, with an additional fully connected feedforward layer leading to a fully connected softmax output with four neurons representing the probability of the sample as an instance of one of the three risk levels or nonlifting. The activation function used for LSTM layers was the hyperbolic tangent function (tanh), while the function for all other layers was the rectified linear unit (ReLU). Table III shows a description of the full model architecture. Training of the model was performed in the same way as the detection model, but LOSO CV could be employed for evaluation since each subject performed the same number of trials in each ACGIH TLV risk zone, so there was no difference between subjects' labeling.

VI. DATASET CONCERNS

A. Transfer Learning

Transfer learning is the process of applying information gained in one domain to a problem in a different, but similar, domain. In a machine learning context, this can be achieved by

TABLE III
CLASSIFICATION MODEL ARCHITECTURE

Layer	Shape	Activation
Input	750 x 36	N/A
Residual Block	750 x 128	ReLU
Residual Block	750 x 128	ReLU
Residual Block	750 x 128	ReLU
Residual Block	750 x 128	ReLU
Residual Block	750 x 128	ReLU
Residual Block	750 x 128	ReLU
LSTM	750 x 128	tanh
LSTM	750 x 128	tanh
Flatten	96000	N/A
Dense w/ BatchNorm	512	ReLU
Output	4	softmax

training a model on one dataset (the source), and then fine-tuning the model on another dataset (the target) [27]. This can be used to improve the performance of a machine learning model on a small dataset, by training the model on a large, more general dataset, and fine-tuning on the smaller, more specific dataset. Transfer learning is commonly used in the image domain to train a model to recognize generic concepts in images before fine-tuning on specific subjects. This can significantly improve performance of models working with a limited amount of data. However, the IMU domain poses some roadblocks to successful transfer learning.

Most image recognition systems work in the same input space, i.e., RGB pixels arranged in a 2-D format. While inputs may be different sizes, images can generally be resized to fit a new input size while retaining its original subject label. In contrast, IMU datasets have significant variances in sensor placements, data ranges, and frequency. This means that one IMU dataset may look significantly different compared to another one, and any knowledge gained on the source dataset cannot be applied to the target. Since it is unlikely to find a larger, more general dataset that contains the exact same sensors placements and frequencies, the benefits of transfer learning do not apply.

To assess the applicability of transfer learning to lift risk assessment, the classification model was first trained on an IMU dataset collected by Kasebzadeh et al. [28], containing several hours of activity including walking, running, and standing still. After successfully training, the model's hidden layers were frozen to preserve the learned information, and the output layer was replaced with a 4-class output for the NIOSH dataset. The model was then fine-tuned on the NIOSH dataset. The intent was for the model to learn general IMU information from the larger dataset, and then apply this learned knowledge to the NIOSH problem.

B. Data Augmentation

Another strategy for improving model performance on small datasets is to use data augmentation. Data augmentation is a process through which training samples are transformed to simulate real-world variation in samples, resulting in a larger dataset with more variation for the model to train on [29]. This often results in better performance as the model can generalize more to samples that may not have been reflected in the original training dataset. Augmentation is often applied in the image domain in the form of rotations and scaling of images, and often significantly improves recognition of a larger variety of images. While this works well in the image domain, it poses some challenges in domains that cannot be visually inspected, such as IMU. It is difficult to determine which transformations result in a sample that is still within the original training domain, and how much of this transformation can be applied before the original label of the sample is no longer preserved. One method of determining these thresholds is label-preserving (LP) augmentation, in which validation data is augmented and the models performance on the validation data is compared to a baseline model to ensure the samples retain their labels [30]. For the NIOSH dataset, LP augmentation was performed using the augmentation strategies defined in [31] to determine which augmentations and how much of them could be applied to the training data without changing the label. The tested augmentations are as follows.

1) Jitter: Adding random noise to each point of data.
2) Magnitude Warping: Scaling the magnitude of each point by a parameter defined by a continuous curve.
3) Time Warping: Scaling the time dimension of the sample by a parameter defined by a continuous curve.
4) Permutation: Splitting the sample into a number of sections and reordering these sections.
5) Scaling: Scaling the magnitude of the sample by a constant factor.

The parameters learned by LP augmentation determine the magnitude at which these augmentations should be applied, except in the case of permutation, where the parameter specifies the number of segments that should be permuted.

These parameters were used to apply transformations to the training data. The transformed data and the original data were used as the training set for model learning, and only one transformation was applied at a time to assess the efficacy of each transformation. For example, the training data were augmented using the "jitter" transformation, and the model was trained using this data and evaluated. This process was repeated for each transformation. Since the format and distribution of data differs between the detection and classification models, the LP augmentation process was performed independently for each, resulting in two sets of optimal parameters. The transformations for the detection data were applied after the segmentation preprocessing.

VII. RESULTS AND DISCUSSION

A. Lift Detection

Fig. 4 shows a sample output from the detection model. When successful, the model is able to detect both the start and end of the lift within a small margin of error. Decreasing the distance from the time stamp for a positive labeling of a given sample reduces the error in seconds, by making the labeling and thus the predictions more precise, but also reduces the average recall of the model, making it more likely to miss a lift instance. Including the samples with significant labeling offset makes this drop in recall significantly more impactful. Since missing multiple high-risk lifts over a long period of time reduces the capabilities of

TABLE IV
LIFT DETECTION RESULTS

Metric	10 Frames	20 Frames	30 Frames	40 Frames	50 Frames
F1 score	0.468 ± 0.226	0.967 ± 0.020	0.971 ± 0.013	0.981 ± 0.016	0.980 ± 0.009
Start Time Error (s)	0.266 ± 0.667	0.551 ± 0.719	0.604 ± 1.035	0.691 ± 0.805	0.832 ± 0.889
End Time Error (s)	0.326 ± 1.309	0.700 ± 0.945	0.552 ± 0.949	0.700 ± 0.945	0.919 ± 1.034

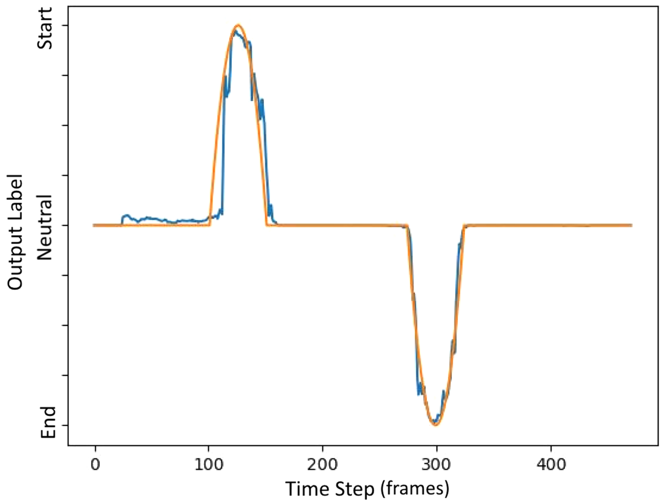

Fig. 4. Ground-truth labels (orange) generated by the sliding window and predictions (blue) from the detection model.

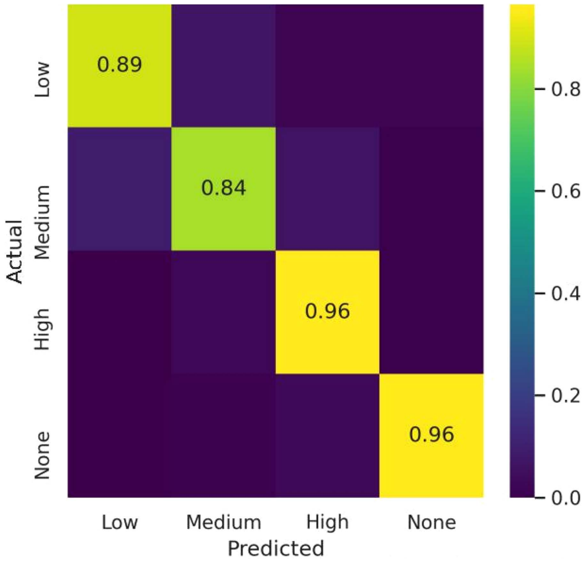

Fig. 5. *Heatmap confusion matrix of the testing results by the proposed model.* Numbers on the diagonal denote classification recall for that class. Rows are normalized so that every class has the same color scale.

the model to accurately assess risk, using a wider labeling range is likely the better option. Table IV shows the results achieved in both cases. The F1 score was calculated considering only lifts detected within 2 s of the labeled start as true positive samples.

B. *Lift Classification*

Fig. 5 shows a heatmap comparing the risk level predicted by the classification model to the ground truth risk. Table V compares the performance of the proposed model to other baseline models [6], [7], [12], as well as a standard linear kernel SVM. The F1 score is used as a metric for model performance on the individual classes, and a balanced accuracy score [32] is used to account for imbalance in classes in the overall assessment of the models. The proposed model is able to clearly differentiate between nonlift and lift action (helping to mitigate false positives from the lift detection model) and shows a 22 point improvement over the next highest baseline for overall balanced accuracy.

The model shows a bias toward the more prevalent high-risk lifts, a known issue with imbalanced datasets [33]. Despite the bias, the model still detects low- and medium-risk lifts at an acceptable rate. Additionally, this bias was considered acceptable because incorrectly classifying a high-risk lift as low-risk has a greater cost than the alternative due to risk of injury. Therefore, increasing recall on high-risk samples should be a priority, even at the cost of precision with regard to false positives.

An ideal lift assessment system is minimally invasive and does not impede daily tasks, so it is important to determine, which sensors are the most important to assess lift risk, and which can be safely removed. Saliency mapping [34] was used to determine which sensors contribute the most to the decision of the model over all classes. The saliency map for high-risk lifts (see Fig. 6) shows that the the upper arm, back, and right wrist sensors are contributing the most to the final decision of the model. This suggests that the inertial measurements provided by those sensors are the most important features for the model decision, so these sensors should be prioritized in settings where fewer sensors are available. Saliency maps for low- and medium-risk lifts showed similar results.

C. *Transfer Learning*

The classification model was able to successfully train on the source dataset, achieving an accuracy of 96%. However, the model suffered from severely reduced performance compared to the baseline when fine-tuned on the target dataset, as shown in Fig. 7. This failure to learn is likely due to the differences between the source and target datasets. Despite best efforts to match the format of the two (similar sensors in similar locations), any deviation between them reduces the ability of the model to apply its learned knowledge to the new domain. This suggests that transfer learning is not an applicable solution to the small dataset problem for IMU data. It is unlikely to find a large enough dataset that uses the exact same sensor modalities for the model to gain knowledge from.

D. *Data Augmentation*

The results of the LP augmentation for time warping can be seen in Fig. 8. The formulation of LP augmentation suggests choosing parameters prior to a sharp drop in accuracy on the

TABLE V
CLASSIFICATION RESULTS FOR PROPOSED MODEL VERSUS BASELINE

Measure	SVM	Chen et al.	Ignatov et al.	DeepConvLSTM	Proposed
Low F1 Score	0.500	0.503	0.545	0.519	**0.861**
Medium F1 Score	0.599	0.612	0.659	0.683	**0.875**
High F1 Score	0.695	0.818	0.849	0.873	**0.955**
None F1 Score	0.779	0.747	0.912	0.911	**0.950**
Overall Balanced Accuracy	0.555	0.558	0.643	0.660	**0.886**

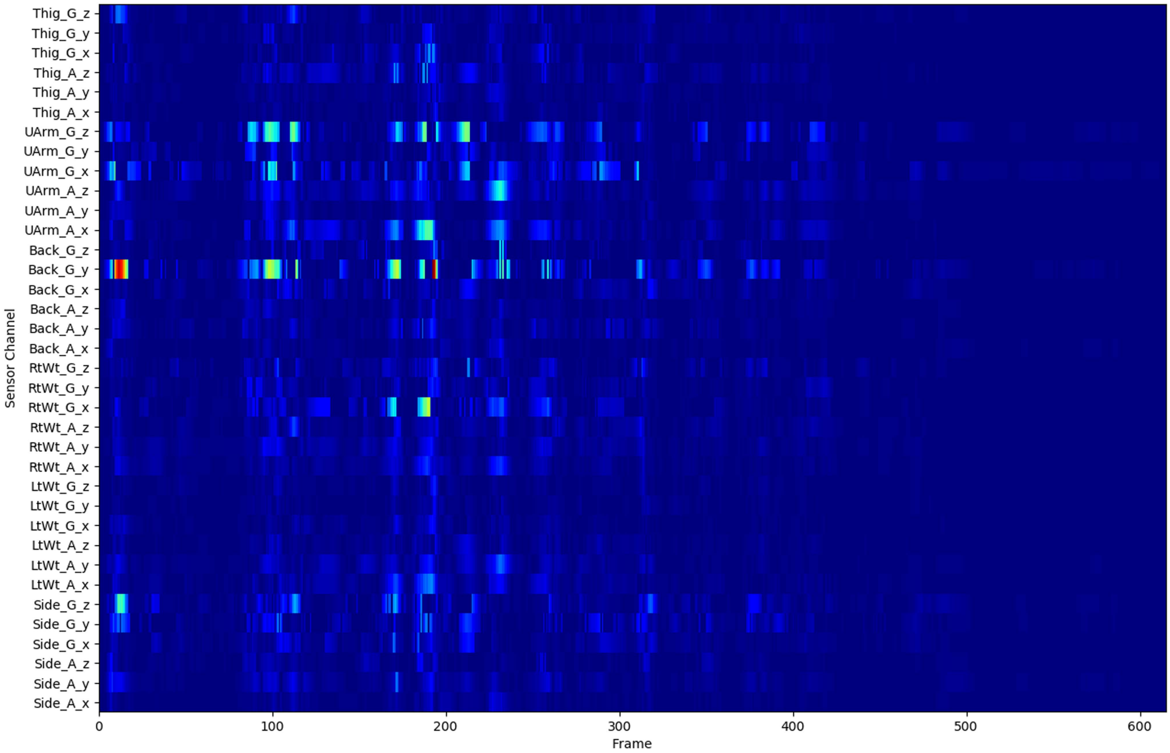

Fig. 6. *Average saliency map generated for the high-risk class.* Warmer colors correspond to more effect on the output classification for the given sensor at that frame. Sensor channels are named based on the sensor and axis, with A and G representing accelerometer and gyroscope, respectively, and *x*, *y*, and *z* representing the axis.

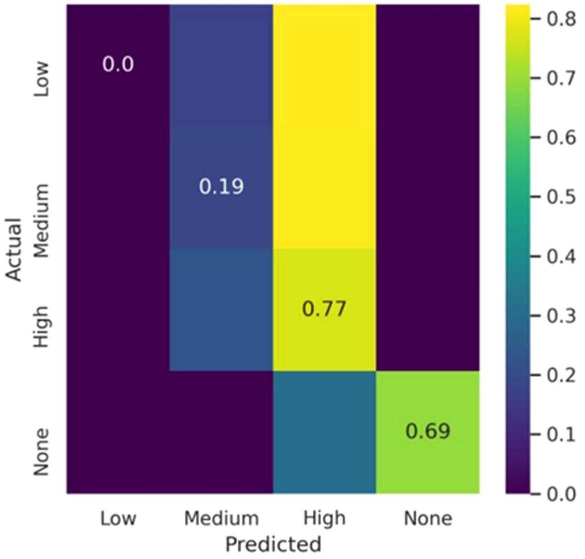

Fig. 7. Results from the model transfer learned on a larger IMU dataset.

TABLE VI
CLASSIFICATION LP AUGMENTATION PARAMETERS

Augmentation	Parameter Value	Balanced Accuracy
None	–	0.886
Jitter	0.50	0.907
Magnitude Warp	0.35	0.920
Time Warp	0.83	0.849
Permutation	1	–
Scaling	0.3	0.928

augmented validation data. If there is no sharp drop, then a threshold is chosen where the accuracy drops to 10% of the original baseline accuracy. Since there is no sudden drop in validation accuracy for time warping, the parameter for time warping is set where the accuracy drops to 83.7%, a 10% drop from the baseline 93%. This corresponds to a time warping parameter of about 0.1. The LP derived parameters for each augmentation on the classification dataset are listed in Table VI. The table also shows the balanced accuracy achieved by applying the augmentations to the training data at those parameter values, doubling the size

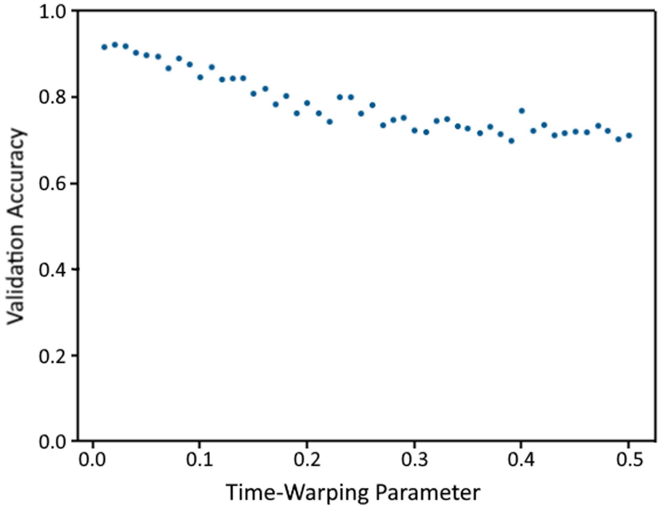

Fig. 8. Results from LP augmentation on time warping. Validation accuracy gradually decreases as the augmenting transformation gets stronger.

TABLE VII
Detection LP Augmentation Parameters

Augmentation	Parameter Value	F-Score	Start Error (s)
None	–	0.980	0.832
Jitter	0.80	0.975	0.899
Magnitude Warp	0.30	0.973	0.849
Time Warp	0.80	0.970	0.918
Permutation	4	0.975	0.847
Scaling	0.30	0.975	0.874

of the available data. Permutation is not applied, as a parameter of 1 corresponds to 1 segment, a permutation of which is the sample itself. All augmentations aside from time warping result in an increase in model performance, up to 4 points in the case of scaling.

Data augmentation had remarkably less of an effect on the detection model's performance. While the LP parameters showed the same general trend during the LP process of slowly dropping model performance as they increased, applying the transformations at that level did not have any impact on the model's performance. This is possibly because the detection dataset consists of many small samples due to the segmentation process, rather than a few large samples. Since there is a large number of samples in this dataset, the primary benefit of data augmentation (increasing the dataset size) does not have much of an impact. The results of the LP augmentation process for the detection model are shown in Table VII.

VIII. Conclusion

The proposed solution is able to both detect and classify lifting action with a high degree of accuracy, providing real-time feedback about the injury risk of a particular lift. This performance is achieved despite the size limitations of the dataset. Data augmentation and transfer learning are considered for additional enhancements; label-preserving augmentation allows for increased performance despite common augmentation difficulties, but transfer learning is not particularly appropriate for the IMU domain and does not provide meaningful gains. While the proposed solution uses all six sensors, future work may focus on using fewer sensors for increased comfort of the user. This solution serves to assist workers, safety professionals and supervisors in automatic lifting risk monitoring in the workplace.

The proposed models may be able to be improved by utilizing body kinematic data. Such data were used to estimate biomechanical risk factors for lifting in two previous studies [35], [36]. The Varrecchia et al. study classified the lifting index of the RNLE using kinematic features collected by a reflective marker-based motion capture system, while the Dopnisi et al. study predicted one lifting index level using one IMU device on the lower back. In the Vareecchia et al. study, the center of mass and motion jerk features were extracted to estimate the LI of three levels using a back-propagation neural network algorithm, while in the Donisi et al. study, random forest, decision tree, and gradient boots were used to estimate a binary lifting risk greater or less than 1.0. Additionally, Donisi et al.'s study suggests the use of deep neural networks as an area for future work. In the current study, the kinematic information came from a greater variation of body posture (12 different lifting zones) combining three horizontal distances between the load and the body and the four vertical distances between the load and the ground. Moreover, the detection of lifting events was identified in this study using a probability model and an LSTM layer for classification, whereas the lift detection was not assessed in the previous studies. Detecting lifting events during normal work duties is the first step for any practical lifting risk assessments. The proposed lifting detection model will contribute to the development of robust lifting detection algorithms for practical applications. The proposed lifting risk model, however, has many limitations with regards to the weight lifted, body asymmetry during lifting and the frequency of lifting. These factors are the variables used by the RNLE for estimating the overall lifting risk, but were not evaluated in the present study. To verify the generalizability of the proposed model, future machine learning research considering these factors is recommended.

Acknowledgment

The authors would like to thank Marie Hayden, Menekse Barim and Dwight Werren for their support in creating the dataset.

Conflict of Interest

Disclaimer: findings and conclusions in this report are those of the authors and do not necessarily represent the official positions of NIOSH, Centers for Disease Control and Prevention (CDC) or the NSF. Mention of any company or product does not constitute endorsement by NIOSH, CDC or NSF.

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*, 1997.

[2] "Work practices guide for manual lifting," Tech. Rep., Mar. 1981. [Online]. Available: https://doi.org/10.26616/nioshpub81122

[3] D. Cook, K. D. Feuz, and N. C. Krishnan, "Transfer learning for activity recognition: A survey," *Knowl. Inf. Syst.*, vol. 36, no. 3, pp. 537–556, Jun. 2013. [Online]. Available: https://doi.org/10.1007/s10115-013-0665-3

[4] I. Rodríguez-Moreno, J. M. Martínez-Otzeta, B. Sierra, I. Rodriguez, and E. Jauregi, "Video activity recognition: State-of-the-art," *Sensors*, vol. 19, no. 14, Jul. 2019, Art. no. 3160. [Online]. Available: https://doi.org/10.3390/s19143160

[5] G. M. Weiss, J. W. Lockhart, T. T. Pulickal, P. T. McHugh, I. H. Ronan, and J. L. Timko, "Actitracker: A smartphone-based activity recognition system for improving health and well-being," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal.*, 2016, pp. 682–688.

[6] Y. Chen and Y. Xue, "A deep learning approach to human activity recognition based on single accelerometer," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2015, pp. 1488–1492.

[7] A. Ignatov, "Real-time human activity recognition from accelerometer data using convolutional neural networks," *Appl. Soft Comput.*, vol. 62, pp. 915–922, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494617305665

[8] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proc. 21th Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn.*, Bruges, Belgium, 2013, pp. 437–442.

[9] V. C. Chan, G. B. Ross, A. L. Clouthier, S. L. Fischer, and R. B. Graham, "The role of machine learning in the primary prevention of work-related musculoskeletal disorders: A scoping review," *Appl. Ergonom.*, vol. 98, Jan. 2022, Art. no. 103574. [Online]. Available: https://doi.org/10.1016/j.apergo.2021.103574

[10] T. R. Waters, V. Putz-Anderson, A. Garg, and L. J. Fine, "Revised NIOSH equation for the design and evaluation of manual lifting tasks," *Ergonomics*, vol. 36, no. 7, pp. 749–776, Jul. 1993. [Online]. Available: https://doi.org/10.1080/00140139308967940

[11] R. Splittstoesser and D. O' Farrell, "Acgih lifting TLV guidance," *Los Alamos Nat. Lab.* [Online]. Available: https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-16-27981

[12] F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, Jan. 2016, Art. no. 115. [Online]. Available: https://doi.org/10.3390/s16010115

[13] "UCI Machine, 'Learning repository: OPPORTUNITY activity recognition data set'," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/opportunityactivityrecognition

[14] P. Zappi, D. Roggen, E. Farella, G. Tröster, and L. Benini, "Network-level power-performance trade-off in wearable activity recognition: A dynamic sensor selection approach," *ACM Trans. Embedded Comput. Syst.*, vol. 11, no. 3, pp. 68:1–68:30, Sep. 2012. [Online]. Available: https://doi.org/10.1145/2345770.2345781

[15] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1533–1540.

[16] D. Ravi, C. Wong, B. Lo, and G.-Z. Yang, "Deep learning for human activity recognition: A resource efficient implementation on low-power devices," in *Proc. IEEE 13th Int. Conf. Wearable Implantable Body Sensor Netw.*, 2016, pp. 71–76.

[17] S. Bhattacharya and N. D. Lane, "From smart to deep: Robust activity recognition on smartwatches using deep learning," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, 2016, pp. 1–6.

[18] Y. Guan and T. Ploetz, "Ensembles of deep LSTM learners for activity recognition using wearables," *Proc. ACM Interactive, Mobile, Wearable and Ubiquitous Technol.*, vol. 1, no. 2, pp. 1–28, Jun. 2017.

[19] K. Snyder et al., "A deep learning approach for lower back-pain risk prediction during manual lifting," *PloS One*, vol. 16, no. 2, Feb. 2021, Art. no. e0247162. [Online]. Available: https://journals.plos.org/plosone/article?id= doi: 10.1371/journal.pone.0247162.

[20] M. 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," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2019, vol. 63, pp. 1105–1111.

[21] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019, doi: 10.48550/ARXIV.1901.03407.

[22] D. P. Kingma and J. Ba, "Adam: A. method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, May 7–9, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[23] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," *GigaScience*, vol. 6, no. 5, Mar. 2017, Art. no. gix019. [Online]. Available: https://doi.org/10.1093/gigascience/gix019

[24] A. Cavallo, A. Koul, C. Ansuini, F. Capozzi, and C. Becchio, "Decoding intentions from movement kinematics," *Sci. Rep.*, vol. 6, Nov. 2016, Art. no. 37036, doi: 10.1038/srep37036.

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

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Comp. Vis.–ECCV*, 2016, pp. 630–645.

[27] L. Torrey and J. Shavlik, "Transfer learning," in *Proc. Handbook Res. Mach. Learn. Appl. Trends, Algorithms, Methods, Techn. Glob.*, 2010, pp. 242–264.

[28] P. Kasebzadeh, G. Hendeby, C. Fritsche, F. Gunnarsson, and F. Gustafsson, "IMU dataset for motion and device mode classification," in *Proc. Int. Conf. Indoor Positioning Indoor Navigation*, 2017, pp. 1–8.

[29] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[30] M. Kim and C. Y. Jeong, "Label-preserving data augmentation for mobile sensor data," *Multidimensional Syst. Signal Process.*, vol. 32, no. 1, pp. 115–129, May 2020. [Online]. Available: https://doi.org/10.1007/s11045-020-00731-2

[31] T. T. Um et al., "Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 216–220. [Online]. Available: http://dx.doi.org/10.1145/3136755.3136817

[32] L. Mosley, "A balanced approach to the multi-class imbalance problem," Iowa State University, Ames, IA, USA, 2013, doi: 10.31274/etd-180810-3375.

[33] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 30, no. 1, pp. 25–36, 2005.

[34] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. 2nd Int. Conf. Learn. Representations*, Banff, AB, Canada, Apr. 14–16, 2014. [Online]. Available: http://arxiv.org/abs/1312.6034*abs/1312.6034*.

[35] T. Varrecchia, C. D. Marchis, F. Draicchio, M. Schmid, S. Conforto, and A. Ranavolo, "Lifting activity assessment using kinematic features and neural networks," *Appl. Sci.*, vol. 10, no. 6, Mar. 2020, Art. no. 1989. [Online]. Available: https://doi.org/10.3390/app10061989

[36] 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, Apr. 2021, Art. no. 2593. [Online]. Available: https://doi.org/10.3390/s21082593