Mobility Scooter Riding Behavior Stability Analysis Based on Multimodal Contrastive Learning

Justin Chung
Computer Science Department
Cal Poly Pomona
Pomona, CA, USA
justinchung2@cpp.edu

Chenrui Zhang
Computer Science Department
Cal Poly Pomona
Pomona, CA, USA
chenruizhang@cpp.edu

Tingting Chen
Computer Science Department
Cal Poly Pomona
Pomona, CA, USA
tingtingchen@cpp.edu

2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10825478

***Abstract*—Mobility scooters are popular among people with limited mobility, providing convenient transportation in their communities and thus enhancing their life quality. However, due to various reasons, a high number of mobility scooter accidents have been reported. In this paper, we tackle its safety issues by performing riding behavior stability analysis using both video and motion sensor data. We design a transformer-based cross-modal encoder to generate the embeddings representing the riding stability and develop a binary classifier based on the embeddings to classify riding behaviors as stable or unstable. The advantage of this work is that the multimodal contrastive learning approach enables the neural network to understand the correlations across two modalities for the same events, and to distinguish the unstable riding behaviors from stable, so that in-context stability analysis becomes possible. We have conducted extensive experiments based on real-world mobility scooter riding data labeled by medical practitioners. The results have shown a high level of classification accuracy of our system, across different settings, in ablation study as well as comparative study.**

***Index Terms*—Multimodal learning; Contrastive learning; Mobility Scooter Safety; Vision and Motion Sensor Data Fusion**

I. Introduction

Mobility scooters are a popular assistive tool that provides convenient transportation for people with limited mobility [17]. However, due to various reasons such as scooter design limitations, environmental and road conditions, there are safety hazards (e.g., tipping over, collisions with pedestrians) associated with their use [4], and a high number of mobility scooter accidents have been reported [2], [6].

One factor in mobility scooter safety issues is related to the riders. Mobility scooter users usually suffer from different medical conditions such as stroke, neuropathy, and brain injury that may lead to upper and/or lower extremity impairments and affect their ability to safely ride a mobility scooter. For example, a user with neuropathy may have slower response speed due to weak muscles and pain, causing accidents especially in crowded or dynamic environments. Hence it is critical to have effective safety assessments of the user's riding behavior, especially from the medical practitioner's perspective. As many mobility scooter users' symptoms are progressive, it is desirable to provide such assessments in a timely manner so that unsafe riding can be intervened and accidents can be prevented. Therefore in this work we consider to perform safety assessments by analyzing riders' upperbody movements while riding the mobility scooter.

To analyze rider's body motions, unlike existing study in Kinesiology using multiple wearable inertial sensors [19], we mainly rely on cameras (e.g., those on smart phones), trying to achieve a higher level of usability. The camera mounted on the scooter handle facing the rider allows us to collect real-time riding behavior data and perform deep learning based analysis. In our prior work [11], an LSTM based autoencoder was built to learn the embedding representation for stable riding behaviors of upperbody in the video frames labelled by kinesiologists. In tests, the autoencoder can effectively distinguish video frames with stable riding behaviors from unstable riding based on the loss values.

Although using video data has been shown effective in assessing the stability of mobility scooter riding, it has limitations in reflecting the riders' true ability. For example, a significant posture sway of the rider will make the kinesiologists to label the video frames as unstable, indicating the observed unstable riding behavior. However, it does not distinguish whether the posture sway is caused by a speed bump on the road, or because of the rider's mild cognitive impairment, which should be categorized differently in riding safety assessment.

To solve this problem, in this paper, in addition to video data, we bring in the gyroscope and accelerometer sensor data of the mobility scooter, to provide the context information of the upperbody movements, so that stability analysis can be more accurate. The challenge is to find a way to generate a representation of riding stability in different contexts while using multimodal data. The representation in the latent space should not only reflect the correlations across the two modalities (such as body shakes when riding on uneven road surface), but also discriminate between stable riding and unstable riding.

To tackle this challenge, we apply multimodal contrastive learning in this work. In particular, we train a transformer-based encoder using the sequential data from the two modalities (i.e., videos and motion sensors). The cross-modal training is enabled by a custom contrastive loss function, where positive pairs are those that appear in the same time window, and negative pairs as those with opposite labels. In this way, the encoder can generate embedding vectors that are effective

representations of in-context riding behavior stability.

The contributions of this paper include the following:

- We develop a multimodal mobility scooter riding stability analysis system based on video and motion sensor data, which has high usability with the sensors easily mountable on scooters.
- Our system effectively aligns and fuses the data from the two modalities after the riders' upperbody keypoints coordinates are extracted from the video frames by applying the human pose estimation model Yolov7-pose.
- We design a transformer-based cross-modal encoder that captures the correlation between the data in two modalities and discriminates the stable and unstable riding in the latent space, by using a custom loss function to enable contrastive learning.
- Our extensive experiments based on real-world mobility scooter riding data collected from 8 patients have verified that our multimodal contrastive learning model has achieved a higher level of classification accuracy than the transformer-based single modality models and other models with alternative architectures.

The remaining of this paper is organized as follows. After we describe in Section II the related works in multimodal contrastive learning approaches and multimodal learning used in healthcare, we present the details of our multimodal method to perform stability analysis for mobility scooter users' riding behaviors in Section III. Section IV covers the extensive experiments on our multimodal system to evaluate its accuracy in stability behavior classification using real-world data, comparing our model with single-modality models and state-of-the-art alternatives. We finally conclude this paper in Section V.

II. Related Work

A. Multimodal Contrastive Learning

Multimodal contrastive learning is an advanced machine learning approach that understands the associations among multimodality data and builds representations by maximizing the agreement among positive pairs and pushing negative samples apart in the latent space [15], [28]. The positive pairs consists of data from different modalities but satisfy certain common criteria (such as an image and its corresponding text caption). Negative pairs represent different concepts or are randomly generated. Usually the multimodal contrastive learning is applied on video, audio, and text data for applications such as multimodal generation and multimodal retrieval. Our multimodal contrastive learning model is applied to motion sensor and video data in order to build a more accurate mobility scooter riding behavior classifier by leveraging more representative embeddings based on multimodal data.

Popular loss functions used in contrastive learning include Noise Contrastive Estimation (NCE) [10] and its variants like InfoNCE [26], which aim to maximize the similarity within the positive pair, and minimize it for negative pairs. Triplet loss function is also widely applied in contrastive learning [25], where samples are groups into triplets with one anchor sample, one positive sample (in the same category as the anchor) and one negative sample (in different category than the anchor). Triplet loss encourages that dissimilar pairs be distant from any similar pairs by at least a certain margin value. In this paper, we define positive pairs across modalities that appear in the same time window, and negative pairs as those with opposite labels. Our loss function is hence defined based on both the cross-modality correlation and the class discriminator.

B. Multimodal Learning for Healthcare Data

Multimodal learning in healthcare is an emerging field that leverages data of various modalities, such as images, clinical notes, Electronic Health Records (EHR) and genomic data, in the process of diagnosis, treatment, and medical research [3], [12], [22]. For example, in [16], a multimodal transformer to fuse clinical notes and structured EHR data is proposed for better prediction of in-hospital mortality. Another multimodal large language model for radiograph representation learning is proposed in [14], to learn broad medical knowledge (e.g., image understanding, text semantics, and clinical phenotypes) from unlabelled data. These works frequently use convolutional neural networks (CNNs) [13], recurrent neural networks (RNNs) [21], and transformers [27] to process data from different modalities, which have been shown to be effective. The existing works of multimodal learning in healthcare are mostly language models especially large language models. Our work does not involve language models, but instead takes video and motion sensor data as input to improve the decision-making in kinematic analysis.

III. Riding Behavior Stability Analysis based on Multimodal Data

In this section, we first introduce the overall pipeline of the mobility scooter stability analysis system. Then we describe the data preprocesssing and alignment process. After the detailed presentation of our multimodal encoder training and its loss function in Section III-C, we list the complete information of model configuration.

A. System Framework

The system is designed to make binary classification decisions about mobility scooter riding stability, which classifies scooter riding behavior as being stable or unstable from two measures: rider state captured in videos and scooter state reflected by motion sensor data. Within our model, we represent rider state as 18 pose features derived using YOLOv7 [24] with the front-camera view of the rider, and scooter state from 6 accelerometer and gyroscope values. The proposed model combines these modalities to effectively measure riding stability. The overview of the system components is shown in Figure 1.

The system consists of three parts: preprocessing and alignment, the cross-modal encoder and the classifier. The preprocessing and alignment module extracts 9 keypoints' 2D coordinates of riders' upperbody in the video frames, and

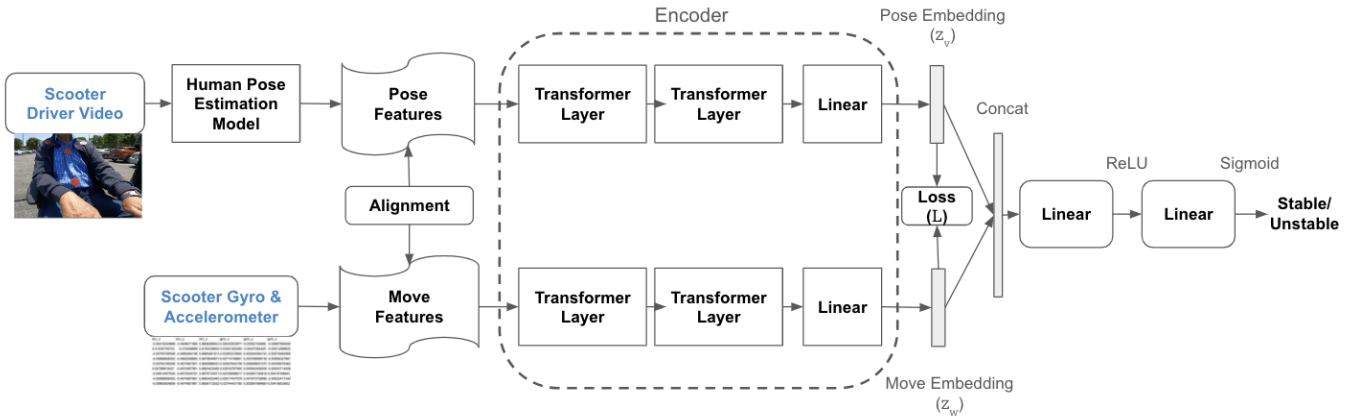

Fig. 1. Overview of Multimodal Mobility Scooter Riding Stability Analysis System

align them with the motion sensor readings by down-sampling the video data to match the sampling rate of motion sensor readings. The pose and motion data matched in sequence length and aligned in time is pushed to the model which is comprised of two central parts: an encoder and classifier.

The encoder serves to represent each modality in the same dimension with an embedding. Our hypothesis is differing modalities of the same label have similar representations, so by representing the modalities in the same space it can form a correlation. We promote this behavior within our encoder training detailed in Section III-C. The classifier then coalesces modality embeddings to output a binary classification of stable or unstable.

B. Data Preprocessing and Alignment

Data pre-processing and alignment are important before the multimodal data can be fed to the neural network architecture. As a first step, the video frames are passed to YOLOv7 for human pose estimation. The pose estimations are 9 keypoints' 2-d coordinates on the rider's upperbody, i.e., neck, left/right shoulders, left/right elbows, left/right wrists and left/right hips within each video frame. This leads to pose features to total in 18 values with a sampling rate of 30 frames per second. Conversely, accelerometer and gyroscope data total in 6 values but sampled at about 1.66 readings per second. In order to align modalities, we first align their initial values to the same time stamp and down-sample pose features by only taking every 18th value. Note that 18 is the rounded ratio of pose to motion sampling rate. This effectively deflates pose samples by a factor of 18 to have modalities be represented by two 2d arrays of matching lengths and index. These 2d arrays are then sequenced to be fed into the model.

C. Multimodal Encoder Training

We utilize two separate training steps for the classifier and encoder. Before the classifier training, the encoder is first trained with its own specialized loss function heavily inspired by the loss implementation in [8]; the intuition is that the loaded encoder will better represent input data by maximizing the correlation of cross-modality embeddings of the same time step and minimizing inter-modality embeddings correlation of differing labels.

The encoder is trained in batches with data being balanced by having an equal number of stable and unstable sequences. Within each batch, samples are passed through the model and outputs are separated into positive and negative pairs for loss calculation. Positive pairs are the combination of every sample who differ in modalities and match in time step. Negative pairs are the combination of every unstable-labeled sample with a stable-labeled sample who match in modality.

The loss function utilizes these positive and negative pairs correspondingly in the calculation of a correlation and discriminator factor.

The correlation factor ($\mathcal{L}_C^t$) maximizes the similarity of different modal embeddings of the same time step. For a singular value x^t, it is defined as:

$$\mathcal{L}_C^t = \sum_{w \neq v} e^{\left(\frac{1 - S_{v,w}^{t,t}}{\tau}\right)}$$

$$\mathcal{S}_{v,w}^{t,t} = \frac{z_v^t \cdot z_w^t}{\|z_v^t\| \cdot \|z_w^t\|}$$

where τ is an adjustable hyper-parameter (also known as temperature parameter [8]) and $\mathcal{S}_{v,w}^{t,t}$ is the normalized dot-product of embedding representations for modality v (pose) and w (movement/motion).

The discriminator factor ($\mathcal{L}_D^v$) maximizes the difference of opposite labeled embeddings of the same modality. With stable sample x^t and unstable sample $x^{t'}$, the discriminator is defined as:

$$\mathcal{L}_D^v = \frac{1}{T} \sum_{t,t'}^{T} e^{\left(\frac{S_{v,v}^{t,t'}}{\tau}\right)}$$

$$\mathcal{S}_{v,v}^{t,t'} = \frac{z_v^t \cdot z_v^{t'}}{\|z_v^t\| \cdot \|z_v^{t'}\|}$$

where T is the batch size and $\mathcal{S}_{v,v}^{t,t'}$ is the normalized dot-product of embedding representations for samples of the same modality but differing labels.

The loss functions is the sum of the cross-modality correlation (positive pairs) and discriminator (negative pairs) with an added hyper-parameter scalar λ to adjust the weight of the discriminator term:

$$\mathcal{L} = \sum_{t} \mathcal{L}_C^t + \lambda \sum_{v \in V} \mathcal{L}_D^v$$

D. Model Configuration

The cross-modal encoder consists of two consecutive transformer layers for each modality. After preprocessing and alignment, inputs are first re-arranged and squeezed to one-dimension then passed to their own corresponding encoder layers; these transformer layers implemented with PyTorch's nn.TransformerEncoder followed by a linear layer map input features to a specified embedding dimension. The number of heads for each encoder is 8. The two modalities' embeddings are coalesced in a final linear layer returning an concatenation of both modalities. The encoder is trained with the method as detailed in Section III-C.

Then the generated embedding is passed to the classifier layers which comprise of two linear layers with the final layer mapping to a singular value. A sigmoid is used as the activation function to create a binary classification with unstable being in the range of [0, 0.5) and stable being between [0.5, 1]. For classifier training, the loss function is binary-cross entropy loss (BCE) on the sigmoid output, with stochastic gradient descent (SGD) as the optimizer.

IV. Evaluation

Our multimodal riding behavior stability analysis system is developed in Python, with PyTorch and torchvision [1] as the core machine learning framework. Pandas, Numpy and OpenCV [18] libraries are used for data preprocessing and analysis. We perform experiments using real-world mobility scooter data to evaluate both the accuracy and computational efficiency of the models. The training and testing of the models are carried out on the Delta system [9], which is equipped with NVIDIA A100 GPUs, each having 40 GB of HBM2 memory.

A. Dataset

We collect mobility scooter riding behavior data from 8 patients at Casa Colina Hospital and Centers for Healthcare [5] and Center of Achievement at California State University Northridge [7]. The patients have upper extremity challenges caused by different medical conditions, e.g., stroke, neuropathy, brain injury, and arthritis. These patients are instructed to complete a set of riding tasks on a Drive Medical Phoenix LT 4 Wheel Mobility Scooter. As illustrated in Figure 2, there is a IMX219 120° HD camera mounted on the riding handle facing the rider's upperbody. A raspberry Pi 4B connected with a sense HAT (including Gyroscope and Accelerometer) [23] is placed on the foot-mat of the mobility scooter. The experiment dataset consists of 7891.7 seconds of video, with 236,751 frames of patients' upper body movements. Video frames are labeled into two classes (i.e., stable and unstable) by kinesiologists using our web-based labeling tool [20]. In addition to the video data, the dataset includes motion data captured. The motion data consists of 24,677 readings from accelerometers and gyroscopes, capturing three-axis acceleration and angular velocity over time. It is important to note, the distribution of the labels, including the pre-process deflation step, is a 85% to 15% split of stable and unstable labels respectively; there is a total of 2308 unstable and 13631 stables samples which means with balancing there is a total of 4616 samples. Our data collection and experiment procedures have been approved by Cal Poly Pomona's Institutional Review Board (CPP-IRB 22-88).

Fig. 2. Mobility Scooter with Camera and Motion Sensors Setup

B. Experiment Setup

Using this dataset, we conduct a set of experiments to evaluate the system's performance in behavior classification accuracy, including overall performance evaluation varying different parameters, ablation study to investigate the benefit of the first training step for the encoder, and a comparative experiment with the models using an alternative backbone and with single modality inputs only. We use an 85/15 train-test split for all experiments. Each model is trained with 10 epochs, 0.0001 learning rate and 50 batch size. In the loss function, $\tau = 5$, and $\lambda = 2$. The reported results are averaged over 50 independent runs.

To measure performance, we use the metrics of average classification accuracy (the proportion of correctly classified instances in the test dataset) when threshold is set to 0.5. We also calculate the true positive rate (TPR) and false positive rate (FPR) for each threshold and plot ROC curves varying the thresholds. To quantify the model's overall classification ability for each run, we compute the area under the curve (AUC), which condenses the information from the ROC curve into a single number, with 0.5 indicating random guessing and 1.0 meaning perfect classification. To reflect the model's discriminative ability over multiple experiments we also calculate the mean AUC value over 50 runs.

C. Overall Performance

We evaluate the overall accuracy performance of our multimodal stability analysis model by investigating the impact of two parameters in the system, i.e., the embedding dimension and the sequence length. The embedding dimension is the length of embedding vector for each modality as the output of the encoder. The sequence length is the number of consecutive readings defined as one sample to feed to the transformer layer, as the unit size to explore the temporal representation of the upperbody movements. With a sampling rate of $2/sec$, a sequence length of 4 means the data collected from about 2 seconds will form a segment as one sample.

TABLE I
AVERAGE CLASSIFICATION ACCURACY AND ROCAUC VALUES WHEN VARYING PARAMETERS

Metrics	**Accuracy (thresh=0.5)**			**ROCAUC**		
	Embed. Dim.			Embed. Dim.		
Seq. Len.	8	16	32	8	16	32
4	0.928	0.931	**0.934**	0.965	0.968	**0.971**
6	0.916	0.917	0.915	0.951	0.954	0.950
8	0.904	0.904	0.909	0.942	0.945	0.955
10	0.890	0.895	0.901	0.932	0.942	0.947

We present the result of average accuracy and ROCAUC of our multimodal system in Table I, varying the sequence length from 4 to 10, and the embedding dimensions from 8 to 32. The sequence length was chosen based on the length of a single frame being about 0.5 seconds, we believed an unstable sequence length would at maximum be limited to a 5 second window: 10 frame long. Table I indicates our multimodal system achieved a high accuracy level overall. The classification accuracy and ROCAUC values are consistently higher for shorter sequence lengths, with the best performance observed when the sequence length is set to 4. This trend is thought to be attributed to the higher concentration of unstable samples in shorter sequences. As the sequence length increases, the proportion of unstable part decreases. We believe the dilution leads to reduced performance in accuracy. In terms of embedding dimensions, the results show a slight but consistent improvement as the embedding dimension increases from 8 to 32. It suggests that higher-dimensional embeddings provide a more detailed representation of the input features, yielding enhanced model performance. However, the improvements from increasing the embedding dimensions appear to be incremental, with diminishing returns beyond a certain point, as seen in the marginal differences between dimensions 16 and 32.

The rest of the test results reported in this paper are collected with embedding dimension 32 and sequence length 4.

D. Ablation Study

To further study the alternative model structure and the benefits of the encoder going through the first training step for both video and motion modalities, we conduct the ablation experiments with different ways to integrate the encoder in classifier training. The option without loading the encoder means the trained encoder is not used in building the classifier. Instead, random weights and biases are initialized at the beginning of the classier training process. Another alternative option is to freeze the trained encoder when training the classifier. Table II shows the accuracy results for these options together with those for our original system (loading the trained encoder and not freezing it while training the classifier).

TABLE II
ACCURACY AND ROCAUC WITH DIFFERENT ENCODER INTEGRATION OPTIONS IN CLASSIFIER TRAINING

Metrics	**Accuracy (thresh=0.5)**		**ROCAUC**	
Trained	Encoder Option		Encoder Option	
Encoder	freezing	unfreezing	freezing	unfreezing
loading	0.893	0.934	0.946	0.971
unloading	0.884	0.932	0.909	0.968

The results in Table II highlight the observed benefit in loading a pretrained encoder while still updating encoder weights during classifier training on classification accuracy and ROCAUC performance. In general, the loaded pretrained encoder outperforms an encoder with random weights. When the pretrained encoder is loaded and the layers are frozen, both accuracy rate and ROCAUC are worse (0.893 accuracy and 0.946 ROCAUC), compared to the scenario where the encoder is loaded but not frozen (0.934 accuracy and 0.971 ROCAUC). This result suggests that our custom multimodal loss function is critical in improving the model's performance, and continued updates to encoder weights during classifier training allows for further refinement.

E. Comparative Study

To show the strength of our multimodal framework, we conduct two comparative experiments: 1) comparison with single modality models; 2) comparison with multimodal structure using an alternative backbone, i.e., LSTM. We implement two models based on single modality input, i.e., a pose-only model and a motion-only model, both of which share the same structure. Each model consists of a multi-layer transformer-based encoder with the MSE loss function followed by a linear layer with a sigmoid activation to produce the classification output. The main difference between the single modality model and the multimodal counterpart is that it does not have the embedding concatenation and the cross-modal contrastive loss function when training the encoder.

As shown in Figure 3, the multimodal model ROC curve (AUC = 0.954) outperforms both the motion-only (AUC = 0.903) and pose-only ROC curves (AUC = 0.913). These results demonstrate that using both motion and pose data to represent the stability of riders' upperbody movements improves classification performance, as the multimodal model is able to capture more information than using each modality alone.

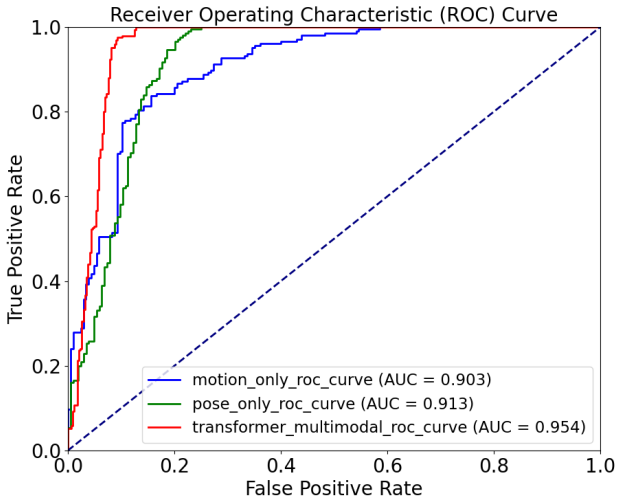

Fig. 3. Comparison with Single Modality Models

We also test the alternative backbone for our multimodal system, by replacing the transformer layers with two consecutive LSTM layers in the encoder and other parts remaining the same.

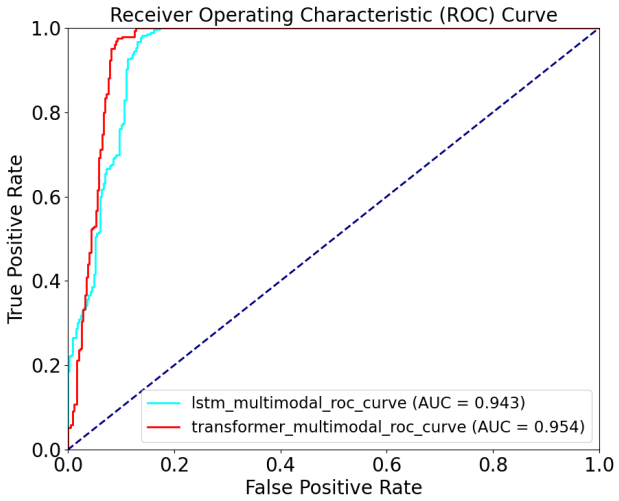

Fig. 4. Comparison with LSTM-based Models

Figure 4 compares the performance of the transformer-based multimodal model with a similar model using an LSTM-based encoder. Despite that both models handle sequential data, the transformer achieves a higher AUC score (0.954) compared to the LSTM-based model (AUC = 0.943). The transformer-based encoder is better at modeling long-range dependencies and interactions between multimodal inputs. While LSTMs are effective at processing sequences, they are less effective in capturing the full complexity of data.

V. Conclusion

This paper presents a multimodal system to perform mobility scooter riding behavior stability analysis based on riders' upperbody video and the accelerometer and gyroscope readings from the motion sensors mounted on the bottom of the mobility scooter. The system leverages contrastive learning in training the cross-modal encoder to generate the embedding representing in-context riding stability. A binary classifier is built using the embeddings to produce classification results of being stable or unstable. Experiments on real-world mobility scooter riding data have been conducted to show that our system achieves a high level of classification accuracy, compared with single modality models and LSTM based alternatives.

VI. Acknowledgement

This work is supported partly by grant NSF CNS #2318671. This work used the Delta system at the National Center for Supercomputing Applications through allocation CIS240470 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

We thank our collaborators Dr. Mai Jara, Joshua Rogers and Michihito Ichihara from Department of Kinesiology and Health Promotion at Cal Poly Pomona for their effort and support in mobility scooter riding video data collection and annotation in this project.

References

[1] Ansel et al. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, April 2024.

[2] Emilie S. Bakgaard. Mobility scooter accidents - need for preventative action? *Clinical Medical Reviews and Case Reports*, 4, 02 2017.

[3] Qiong Cai, Hao Wang, Zhenmin Li, and Xiao Liu. A survey on multimodal data-driven smart healthcare systems: approaches and applications. *IEEE Access*, 7:133583–133599, 2019.

[4] Anna Carlsson and Jörgen Lundälv. Acute injuries resulting from accidents involving powered mobility devices (pmds)—development and outcomes of pmd-related accidents in sweden. *Traffic injury prevention*, 20(5):484–491, 2019.

[5] Casa. Casa colina hospital and centers for healthcare. https://www.casacolina.org/.

[6] US Consumer Product Safety Commission et al. National electronic injury surveillance system coding manual. 2019. 2020.

[7] CSUN. Center of achievement, california state university northridge. https://www.csun.edu/center-of-achievement.

[8] Shohreh Deldari, Hao Xue, Aaqib Saeed, Daniel V Smith, and Flora D Salim. Cocoa: Cross modality contrastive learning for sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–28, 2022.

[9] Delta Science Gateway. Delta science gateway documentation. https://gateway.delta.ncsa.illinois.edu/wiki/, 2024.

[10] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[11] Ruoqi Huang, Mai Narasaki-Jara, and Tingting Chen. Deep learning based driving posture stability analysis for people with mobility challenges. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4900–4906. IEEE, 2023.

[12] Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171, 2022.

[13] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[14] Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1):226, 2023.

[15] Yunze Liu, Qingnan Fan, Shanghang Zhang, Hao Dong, Thomas Funkhouser, and Li Yi. Contrastive multimodal fusion with tupleinfonce. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 754–763, 2021.

[16] Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction. In *AMIA Annual Symposium Proceedings*, volume 2022, page 719. American Medical Informatics Association, 2022.

[17] Esther May, Robyne Garrett, and Alison Ballantyne. Being mobile: electric mobility-scooters and their use by older people. *Ageing & Society*, 30(7):1219–1237, 2010.

[18] OpenCV. *The OpenCV Reference Manual*, 2.4.13.7 edition, April 2014.

[19] Pietro Picerno, Andrea Cereatti, and Aurelio Cappozzo. Joint kinematics estimate using wearable inertial and magnetic sensing modules. *Gait & posture*, 28(4):588–595, 2008.

[20] Mobility Scooter Project. Mobility scooter driving video annotation tool. https://mobility-scooter-project.github.io/labeler/, 2024.

[21] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[22] Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

[23] TDK. 6-axis mems motion sensors. https://invensense.tdk.com/products/motion-tracking/6-axis/.

[24] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023.

[25] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

[26] Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. Rethinking infonce: How many negative samples do you need? *arXiv preprint arXiv:2105.13003*, 2021.

[27] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

[28] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004, 2021.