Categorizing Data Acquisition System for Head Response (DASHR) Head Impact Data with Machine Learning

GWDD Candidate: Sarah Glomski, *Supervisor:* Jason F. Luck

Injury Biomechanics Laboratory, Department of Biomedical Engineering,
Duke University, Durham, N.C.

Introduction

Head Injury

Fig 1. DASHR on a dummy head.

- Football athletes are disproportionately affected by head injuries, making up 12.8% of all sports-related head trauma and 19.4% of all sports-related concussions[1].
- Wearable sensors, such as the Data Acquisition System for Head Response (DASHR), can be used to characterize head impact exposure (HIE) and injury by recording the head's linear acceleration and rotational velocity during activity.
- The high prevalence of false positives in recorded head impact data means that data processing is critical to drawing accurate conclusions about athletic exposure and injury risk.[2]

Classifiers

- Some researchers use basic thresholding algorithms to eliminate false positive impacts. 10 Gs is a common linear acceleration threshold (LAT) for wearable sensors.[2]
- 1-D Convolutional Neural Networks (CNNs) are a common machine learning (ML) model used to classify temporal activity data, such as valid and invalid head impacts. One such model had been developed previously for this task.
- Gradient-weighted Class Activation Mapping (Grad-CAM) can be used to identify which features of the data the model has deemed to be relevant during the decision-making process.

Data Collection

Simulated Data

- The Data Acquisition System for Head Response (DASHR) was used to record linear acceleration and rotational velocity data for 4 activity classes:
  1. Valid impact (post-mortem human surrogate drop tests)
  2. High-g non-impact (flicking, re-seating the DASHR)
  3. Running/walking
  4. Standing still

Composition: 70% | 18% | 8% | 7%

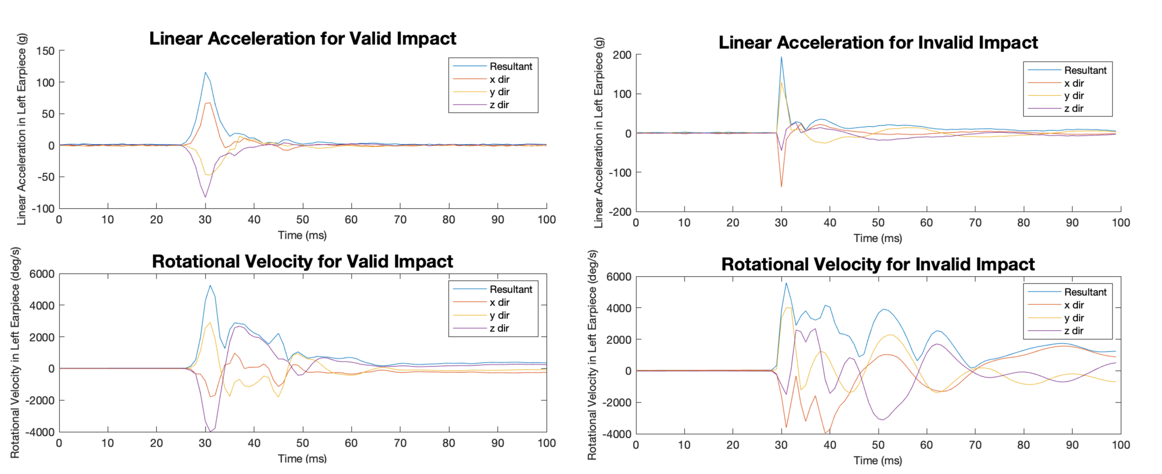

Fig 2. Linear acceleration and rotational velocity over time for a valid impact (frontal drop test from 50 cm) (left) and high-g non-impact (flicking DASHR) (right).

Field Data

- The DASHR was used to record kinematic activity data for high school football players during practices.
- Player activities were tracked and categorized into the same 4 classes. These data serve as verified ground truths.

Composition: 0.3% | 70% | 28.7%

Classification Methods

Binary Classification Task

2 classes: Valid impact | High-g non-impact

Thresholding Algorithms

- 5 algorithms were developed—each with a different threshold(s) that was optimized with respect to accuracy.
  i. Lower bound linear acceleration threshold (LB LAT)
  ii. Upper bound linear acceleration threshold (UB LAT)
  iii. Lower bound duration threshold (LB DT)
  iv. Combined LB LAT and LB DT
  v. Combined UB LAT and LB DT
- The algorithms were tested on 1) simulated data and 2) field data to compare the optimized threshold values.

Machine Learning

- A 1D-CNN (referred to as the "ML model") was trained on simulated data, as had been previously achieved[2].
- The ML model was tested on 1) simulated data and 2) field data to determine if the dataset could be reliably augmented.

Analysis

- Receiver operating characteristic (ROC) curves were generated for each classifier.

Table 1. Binary classification task outcomes, where TP=true positive, FP=false positive, TN=true negative, FN=false negative.

		Predicted	
		Valid impact	High-g non-impact
Actual	Valid impact	TP	FN
	High-g non-impact	FP	TN

Classification Metrics:

$$FPR = \frac{FP}{FP+TN}$$

$$TPR = \frac{TP}{TP+FN}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Advanced Classification Task

4 classes: Valid impact | High-g non-impact | Running/walking | Standing still

Machine Learning

- Grad-CAM was used to create heatmaps of relevant features during the following experiments:
  1) The length of the input data was varied from 100 ms to 2100 ms to determine the effects of overfitting.
  2) The kernel size was varied from 5 to 11 to determine the effects of model scope.
  3) The number of feature maps was varied from 128 to 256 to determine the effects of model complexity.
- Grad-CAM revealed that the ML model sometimes overfit to features that were specific to the simulated dataset.

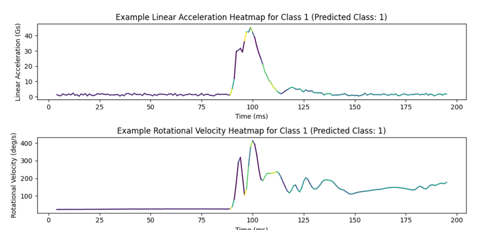

Fig 3. Correct heatmap for simulated valid impact.

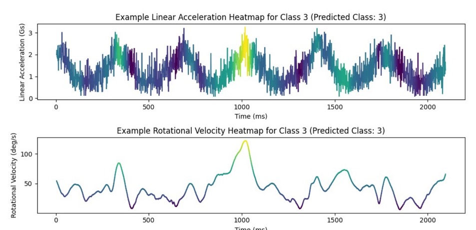

Fig 4. Correct heatmap for simulated walking.

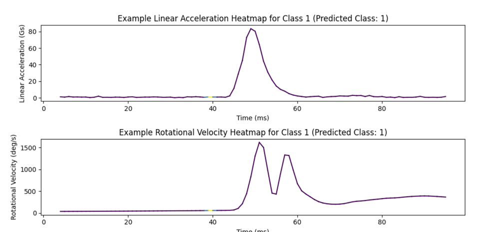

Fig 5. Incorrect heatmap for simulated valid impact.

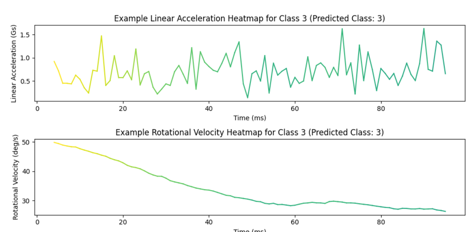

Fig 6. Incorrect heatmap for simulated walking.

- A multi-headed model (k=5 and k=11) was developed based on insights from Grad-CAM.

Results

Binary Classification Performance

- Accuracy and ROC curves were plotted for each of the 6 classifiers, as shown in Figures 7 and 8 below.

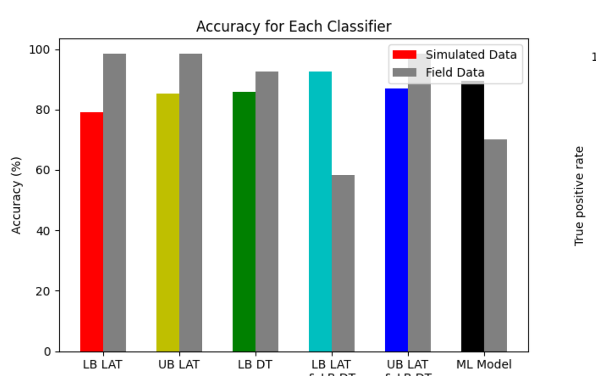

Fig 7. Accuracy for each classifier for simulated and field data.

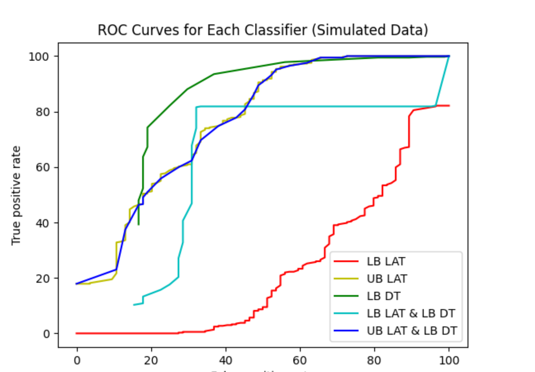

Fig 8. ROC curve for each classifier for simulated data.

- The combined LB LAT & LB DT had the highest accuracy at 92.5%. This outperformed the ML model, which had an accuracy of 89.5%.
- The LB DT had the highest area under the curve (AUC) with 76.4%.
- The two datasets also had drastically different compositions, which led to skewing of the accuracy data.
- The uneven dataset composition led to skewed accuracy for both datasets, but especially for the field data.

Grad-CAM Insights

- Grad-CAM could be used to diagnose both specific classification decisions and general trends in the dataset.

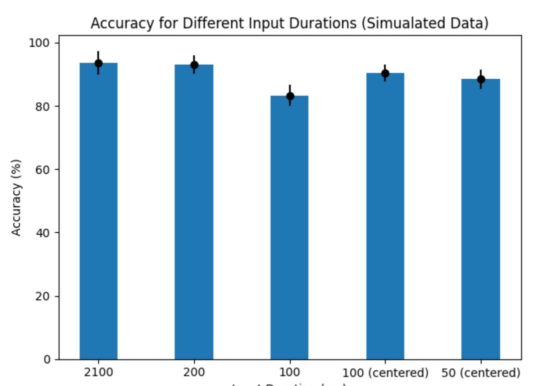

Fig 9. Accuracy for different input durations for simulated data.

Table 2. Mean TPR for different kernel sizes and number of feature maps (input length=100 centered, n=10).

		Number of feature maps			
		128		256	
Kernel size	5	Class 1	Class 2	Class 1	Class 2
		73.3%	98.4%	78.0%	98.1%
		Class 3	Class 4	Class 3	Class 4
		66.2%	25.3%	81.8%	25.0%
	11	Class 1	Class 2	Class 1	Class 2
		63.9%	95.8%	75.0%	99.3%
		Class 3	Class 4	Class 3	Class 4
		55.7%	44.6%	54.9%	38.2%

- Reducing the input length from 2100 ms to 100 ms reduced this effect, as shown in Figure 9, but led to a TPR of 25.3% for running/walking, as shown in Table 2.
- Increasing the kernel size from 5 to 11 helped increase the model's TPR for running/walking to 44.6%.
- Changing the number of feature maps did not significantly affect model accuracy or TPR.
- The multiheaded model did not perform significantly differently than the original ML model.

Discussion and Future Work

Discussion

- The performance of the ML model in comparison to simpler methods indicates that a more complex model may be needed.
- Implementing out-of-set classification should be the next step towards increasing model accuracy.
- Currently, these classifiers could help to reduce the inflation of HIE and support improved injury risk development.
- However, with a maximum accuracy of 92.5%, the false positive rate (FPR) is still higher than ideal. The lower the FPR, the most accurately injury risk curves can be created to quantify the HIE in high school football.
- Applying a combined LB LAT and LB DT thresholder to wearable sensing devices in the field would be a simple solution to lowering the number of false positive events.

Limitations

- Both training sets were unbalanced and had drastically different compositions, which skewed the accuracy of the classifiers and subsequent comparisons between them.

Future Work

- Improvements to the model should be made cautiously, with the risk of overfitting increasing as the model accuracy approaches 100%.
- The addition of out-of-set classification would allow researchers to discard outliers, creating a pipeline for more reliable data.
- Implementing multi-class classification may improve the results for the advanced classification task.
- Grad-CAM should be used to further improve the transparency of the model's confidence in decision making, and researchers could use video footage to reexamine the classifications made with low confidence.
- Addition of more data—both to the existing classes and new classes—for increased model robustness and advanced activity classification. This includes low-g behavioral activities associated with head impact.
- Other moderately complex classifiers (ie. Support Vector Machine) and machine learning models (ie. General Adversarial Network, or GAN, Long Short-Term Memory model, or LSTM) should be explored in parallel.

Acknowledgements & References

We acknowledge Duke Bass Connections - Brain & Society, Duke Institute for Brain Sciences (DIBS), Duke Pratt School of Engineering, and the Department of Biomedical Engineering.

References

1. Gaw, C. E. et al., (2016). Emergency department visits for head injury in the United States. BMC Emergency Medicine, 16(1).
2. Wu, L. C. et al., (2021). Head impact sensor triggering bias introduced by linear acceleration thresholding. Annals of Biomedical Engineering, 49(12), 3189–3199.
3. Liu, P., (2021). Graduation With Departmental Distinction Thesis.