Analyzing Feature Importance for ML Models Predicting Aggression Among Inpatient Youths With Autism

Briggs Twitchell

August 2024

1 Abstract

Understanding how a machine learning (ML) model functions is a critical step in assuring it behaves as expected. Models deployed in a real-world clinical setting may encounter data that differs (sometimes significantly) from that with which they were developed. This can reduce their performance and diminish their overall utility to healthcare providers, especially when the features from which they obtain most of their predictive power are affected by domain change. In this work, we analyze the importance of features for a select set of ML models, developed by Imbiriba et al. [6],predicting the onset of aggression for inpatient youths with autism. To measure feature importance, we measured the changes in performance when providing the model different feature combinations and also examined the model coefficients and SHAP values. Our findings suggest 1) that the measurement encoding the time since the onset of the previous aggressive episode inform the models only whether aggression has or has not yet occurred within an observation session and 2) that the models obtain the majority of their predictive power from this information.

2 Background

The *JAMA Network Open* article "Wearable Biosensing to Predict Imminent Aggressive Behavior in Psychiatric Inpatient Youths With Autism" [6] assesses the predictive power of machine learning models used to predict the onset of aggression among 86 psychiatric inpatients diagnosed with Autism Spectrum Disorder.

The inpatients wore commercially available biosensors on their wrists, capturing time-series measurements for cardiovascular activity, electrodermal activity, and motion. Aggressive episodes were recorded and independently validated by two onsite research staff, and a total of 429 observational sessions generated 497 hours of data. The researchers processed the time-series data in windows ranging from 1 to 3 minutes, which formed the datasets passed into three different machine learning models: logistic regression, support vector machine, and a neural network.

2.1 Performance

The performance of the models varied depending on their specific configuration, such as the length of the past time window used to predict the onset of aggression. The authors found logistic regression was the best performing model, predicting aggressive behavior 3 minutes prior to onset with a mean Area Under the Receiver Operating Characteristics (AUROC) curve of 0.80. They concluded that changes in peripheral physiology may be useful for predicting aggression of inpatient youths with ASD before the aggression occurs.

The figures below display the ROC and precision-recall curves reconstructed for the two models that are the target of this analysis: the logistic regression (LR) and support vector machine (SVM) population models with session splits on the augmented feature vector (onset) using 3 minutes of prior data to predict aggression onset 3 minutes into the future. See eTable 1 in the supplemental content of Imbirba et al. for further details on this and other model configurations.

The AUROC for the LR model reported Imbiriba et al., 0.80, differed slightly from our findings when re-running the model, which yielded an AUROC 0.76. We deemed this difference sufficiently small for

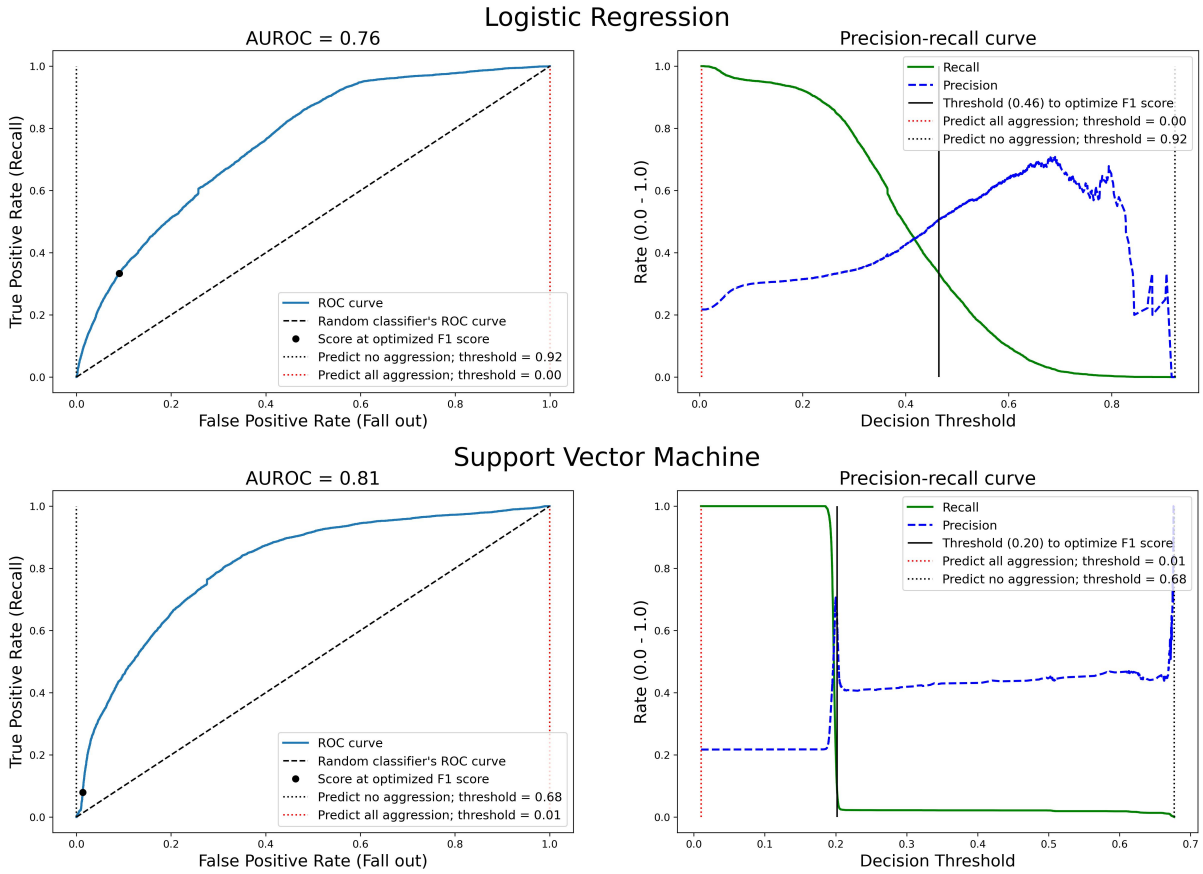

Figure 1: Performance results for the LR (top) and SVM (bottom) showing the ROC and Precision-Recall curves on the left and right, respectively.

the post-hoc analysis of the LR model to be representative. No difference was noted for the SVM model (AUROC 0.81)[1].

2.2 Description of Dataset

The inpatient physiological measurements were captured via a wearable biosensor that tracked blood volume pulse (BVP), electrodermal activity (EDA), and three-axis acceleration (ACCx, ACCy, ACCz) within a session. A 15-second sliding window calculated the following summary statistics for each measurement: first, last, maximum, minimum, mean, and median value; the number of unique values; and sum, standard deviation, and variance of values falling in a window. The researchers also added a binary observation of aggression flag (AOF), the time since previous aggression onset (TPA), as well as a calculation of the standard deviation for each summary statistic within the 3-minute series of adjacent 15-second windows. The Augmented Feature Vector (AFV) used by the LR and SVM models includes the summary statistics and standard deviations of all physiological measures, TPA, and AOF, yielding 676 values in total for each 3-minute observation time block. For the configuration predicting aggression 3 minutes into the future, there were ultimately 101,929 instances of 3 minute time blocks, of which 80,496 (80%) was used to train the models and 21,433 (20%) was used to test the models.

In this analysis, we named features according to their measurement, summary statistic, and time window. For example, the mean BVP over the first 15-second window in the 3-minute time block is named `BVP_mean_t-180s`, while the mean BVP for the last 15-second window is named `BVP_mean_t-15s`. The standard deviation of summary statistics were named according to their measurement, summary statistic, and the total time block for each observation — for example, the standard deviation of the median values for z-axis acceleration with the 3-minute time block is `ACC_Zmedian_180s_window_std_dev`.

[1]Unexpectedly, we noted a significant decrease in performance, particularly for the SVM, when training the model without calling the Scikit-learn-Intelex patch function [3], suggesting that the current version (2024.0.0) of the Scikit-learn-Intelex library affects not only training time but also (in certain instances) model parameters.

We refer to values that are part of the feature vector as features, whereas we refer to the original data from which the features originated as measurements.

3 Goals For This Analysis

The primary goal for this analysis are to 1) understand which features in the dataset contain predictive power and 2) to examine how the selected (already trained) models make their binary predictions for aggression onset. That is, we seek to identify which input features contribute most to the models' predictions. This, in-turn, has the potential to inform researchers about methods to improve the models' performance or additional data or feature engineering techniques to render the models more robust. It may also provide an explanation to caretakers regarding why the model(s) do or do not predict an oncoming aggressive episode and even help inform them about policies or interventions to better manage or prevent aggression.

We chose to focus our analysis primarily on these specific LR and SVM models because they were identified by the authors as among the best performing and most relevant models. Additionally, limiting the analysis to models with a common set of features controls for multiple confounding factors, enabling more useful comparison among them.

Importantly, this analysis examines only a subset of the models. Though we did perform a test to examine if our findings applied to another model configuration, they do not necessarily extrapolate to every model detailed by Imbiriba et al., particularly the neural network models. Moreover, this analysis does not seek to identify underlying *causes* of aggression for inpatients but rather understand how certain existing models make their predictions.

4 Results

To determine feature importance, we employed two post-hoc methods, summed model coefficient values and SHAP values. Additionally, we measured the change in model performance when retraining the models on different subsets of the AFV. Subsequent sections detail these methods and their related challenges.

Our analysis identifies TPA as the most important measurement, as it consistently contributed the most predictive information across each of the three methods employed. ACC also displayed importance for the LR model, but this result was not consistent across each method. AOF displayed importance, but, as explained in Section 5, this measurement encodes the same information as TPA. We conclude that TPA reduces to the same information provided by AOF and that together they direct the model as follows: *aggression not yet observed within an observation session guides the models toward a future prediction of non-aggression, and, once aggression has been observed in an observation session, the models are more likely to predict future aggression.*

4.1 Global Perspective

Figures 2 and 3 depict a global quantification of feature importance. The absolute value of coefficients summed by measurement group and the top mean absolute SHAP values for both the LR and SVM models identify TPA as contributing significantly to model predictions. For the LR model, ACC had the largest value for the coefficients summed by measurement group, but ACC did not have significant mean absolute SHAP values. TPA also made up 50% and 30% of the 10 largest absolute coefficients for the LR and SVM models, respectively.

We noted that TPA has the largest absolute value of coefficients summed by measurement group and largest SHAP values for the same LR and SVM models using just 1 minute of past data to predict 1 minute into the future, suggesting that the importance of this measurement may hold for the other model configurations.

Section 5 details the TPA and AOF measurements, demonstrating that they encode the same information. Though Figures 2 and 3 show that the LR model treats the two measurements consistently (i.e. a low TPA and true AOF value correspond with a prediction of aggression, and the inverse correspond with a non-aggression prediction), the SVM treats them inconsistently (i.e. a low TPA corresponds with an aggression prediction, but for certain AOF features a true value corresponds with a *non*-aggression

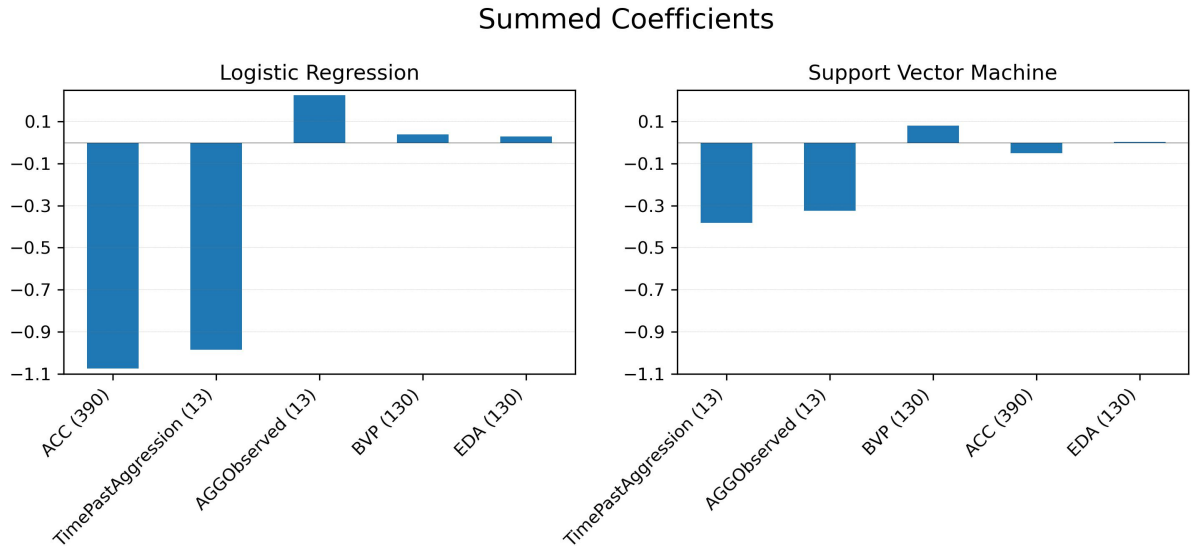

Figure 2: The model coefficients summed by measurement group and ordered by largest absolute value. Left: ACC and TPA stand out as the largest for the LR model. Right: TPA and AOF stand out as the largest for the SVM model. The value next to each label indicates the number of features summed for each measurement.

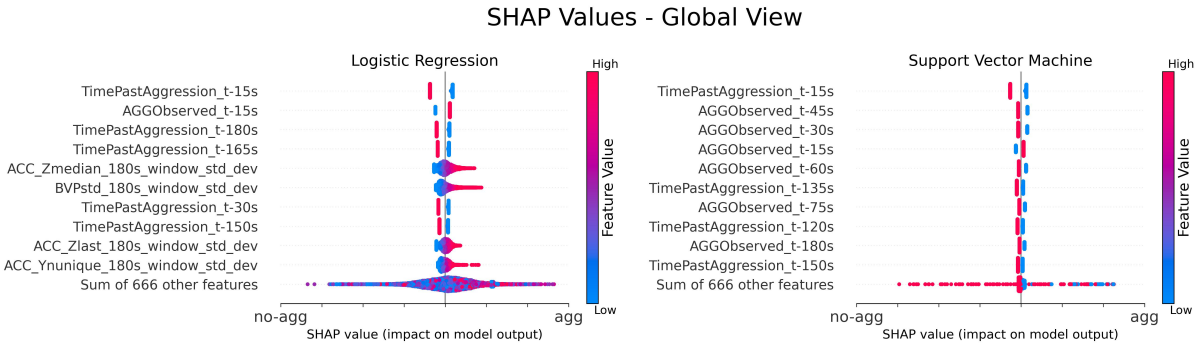

Figure 3: A beeswarm plot of SHAP values, ordered by mean absolute value of the SHAP values. Left: TPA makes up 5 of the 10 largest mean absolute SHAP values. Right: the 10 largest mean absolute SHAP values consists of TPA or AOF features. Note that for both models, a low TPA pushes the models toward a prediction of aggression.

prediction). This finding may reflect how the models respond differently to multi-collinearity[2], and in Section 5 we discuss our methods to control for this.

4.2 Local Perspective

Figure 4 displays an instance of a single 3-minute observation block that resulted in a true, non-aggression prediction by the LR model. This plot shows that TPA at the *T - 15-sec.* window has the largest absolute SHAP value. Also note that the features that are part of the TPA measurement make up 5 of the 10 largest absolute SHAP values. This pattern was largely consistent across all model prediction outcomes (i.e. true-negative, false-negative, true-positive, and false-positive), though it was less pronounced for the true-positives instances of the LR model, which displayed relatively higher SHAP values among the physiological variables, particularly ACC and standard deviation of physiological summary statistics.

[2]Figure 4 and other force plots we examined showed greater unidirectionality among the TPA and AOF SHAP values for the LR model. Conversely, for the SVM model we noted multiple instances (included in appendices) in which the `TimePastAggression_t-15s` alone displayed the greatest feature importance with the remaining TPA and AOF SHAP features effectively cancelled out their prediction contributions. Figure 5 demonstrates that the SVM model using the most recent 15-second window alone outperformed those using the AFV. Moreover, `TimePastAggression_t-15s` had the single largest coefficient and SHAP value for *all* models in which it was included as an input feature.

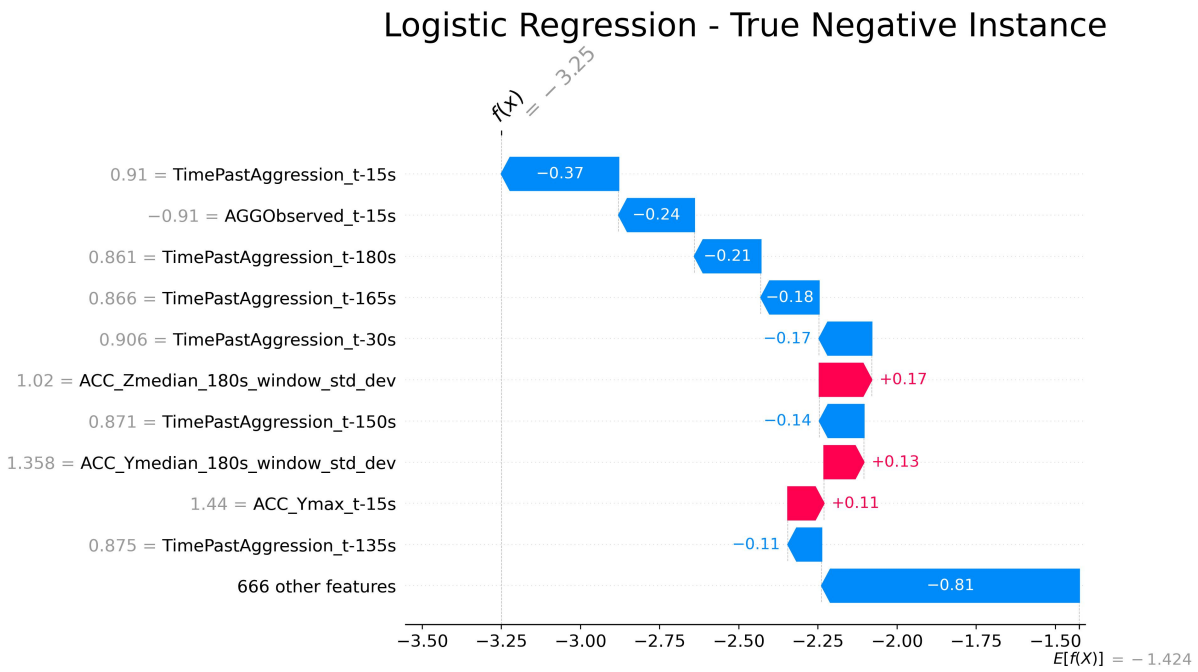

Figure 4: The top 10 largest SHAP values for an instance of a 3-minute observation block that resulted in a true, non-aggression prediction by the LR model. Note that variables from the TPA measurement make up 5 of the 10 largest absolute SHAP values.

The SHAP plots demonstrate that high TPA values tend to push the models toward a prediction of non-aggression, and low TPA values tend to guide the models toward an aggression prediction. This indicates that once aggression has been observed *within* a clinical observation session, the likelihood that the model will predict aggression in the future increases.

4.3 Model Performance According to Input Features

For the LR and SVM models, we apply a decision threshold that maximizes the weighted F1 score[3]. Among the feature combinations listed in Figure 5, we note that the best performing models use the TPA and AOF measurements alone, excluding all of the physiological data. Adding the physiological measurements to the feature space only adds noise, on average decreasing their weighted F1 scores and AUROCs by 0.13 and 0.07, respectively (3 minutes of prior data). A baseline dummy model that randomly predicts aggression in line with the distribution of aggression in the train data (20.58%) yielded a macro and weighted F1 score of 0.49 and (due to the imbalances in classes) 0.65, respectively.

We also created a modified dummy model, referred to as the Single Condition Dummy Model (SCD). If none of the TPA features in the 3-minute time block indicate that aggression began to or already occurred within the inpatient's observation session, the SCD predicts non-aggression. Otherwise, it randomly predicts aggression according to the distribution of aggression in the train data, as the regular Dummy model does. The SCD achieved a weighted F1 score of 0.72, outperforming the LR and SVM models that exclude the TPA and AOF variables. Reducing the time block of prior data and the future prediction time from 3 minutes to 1 minute increased the SCD's weighted F1 score to 0.77.

We also note that the ACC variables alone perform slightly better than the AFV and are comparable to the SCD.

[3]This decision threshold would likely differ to some degree from the one that is clinically or operationally desirable, as it assumes equal consideration of precision and recall. Though, unlike AUROC, the F1 score is sensitive to prior class imbalances [4, 8], Fawcett notes that AUROC can give false-positives too little importance, rendering precision based methods such as F1 score preferrable. Regardless, the F1 score offers a means to compare models to a baseline Dummy classifier, so we use it in combination with the AUROC.

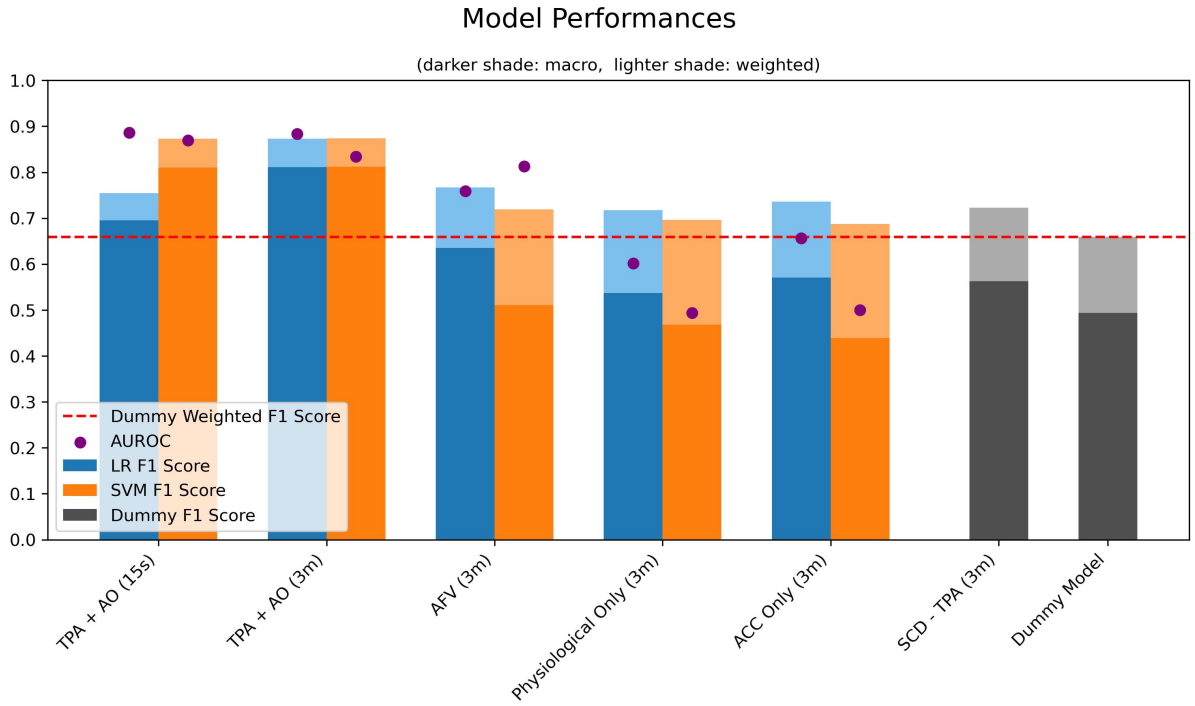

Figure 5: A comparison of macro and weighted F1 scores and AUROCs for different input feature vectors. The 3m models represent those using 3 minutes of prior data to predict aggression onset 3 minutes into the future. The 15s model (far left) uses 15s of prior data to predict aggression onset 3 minutes into the future.

5 Methods for Feature Importance Identification

We used additive feature attribution methods to define feature importance. Both the LR and SVM models are linear. The SVM decision function is:

$$f(x) = \mathbf{w} \cdot \mathbf{x} + b, \tag{1}$$

where $\mathbf{w}$ is a vector of coefficients, $\mathbf{x}$ is the input feature vector, and b is a bias term. The LR model applies the sigmoid function to (1), yielding:

$$P(y = 1 \mid x) = \sigma(\mathbf{w} \cdot \mathbf{x} + b). \tag{2}$$

Thus, the coefficient w_i corresponds to the importance assigned to input feature x_i. Since coefficient w_i captures the average effect of feature x_i across the entire train data, it offers a global perspective of feature importance [1, 5]. However, to quantify feature importance at the local (individual observation) level, we used Kernel SHAP [11], which, given an observation and target feature, approximates the difference in a model's prediction with and without that feature, averaged over all possible feature subset combinations of the non-target feature[4]. The Shapley value for a target feature $\phi_i(v)$ is

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} \cdot \left(v(S \cup \{i\}) - v(S)\right), \tag{3}$$

where N is the set of input features, S is a subset of features excluding i, $v(S)$ is the model's prediction provided the subset of features S that exclude feature i, and $v(S \cup \{i\})$ is the model's prediction including feature i. Because calculating Shapley values is NP-hard, Kernel SHAP approximates (3) via a weighted linear regression problem:

$$\hat{\phi} = \arg\min_{\phi} \sum_{S \subseteq N} w(S) \cdot \left(v(S) - \phi_0 - \sum_{i=1}^{n} \phi_i z_i\right)^2, \tag{4}$$

[4]When input features are independent, the SHAP value of the i^{th} feature for the prediction $f(x)$ is $\phi_i = w_i \cdot (x_i - E[x_i])$ [10].

where z_i is a binary indicator for the inclusion of feature i, and $w(S)$ are the Shapley kernel weights defined as:

$$w(S) = \frac{(n-1)}{n|S| \cdot |S| \cdot (n-|S|)}, \tag{5}$$

for which $n = |N|$. Kernel SHAP samples the subsets S, weighting them according to (5).

5.1 The TPA and AOF Measurements

Our examination of the TPA and AOF measurements shows that TPA reduces to AOF, and they both indicate whether or not aggression has yet been observed within the inpatient's observation session. While TPA is intended to measure the duration since the previous observed onset of aggression within a session, as detailed in eFigure1 of the supplemental content of Imbiriba et al. [6], its initial default value (the value assigned to it before aggression has been observed within a session) of 5,000 creates a binary distribution mirroring that of the AOF measurement. Their Pearson's correlation coefficient is approximately -1.0 for all twelve time windows. This implies that *the models' larger weighting of TPA over AOF is arbitrary and that TPA and AOF carry the same predictive power.* We confirmed this by conditioning the SCD on AOF, which yielded a weighted F1 score just 1% less than the SCD using TPA.

These measurements originate from the ground-truth labels, which were independently coded by two onsite research staff and subsequently validated to ensure interrater reliability. Given their demonstrated importance to the SVM and LR models, we must consider the likelihood that these data would be reliably available and accurate, if either model were deployed in a real-world clinical setting.

6 Challenges in Quantifying Feature Importance

When there is multi-collinearity among input features, the predictor model may assign arbitrary weights to some of them[5]. Additionally, Kernel SHAP's method to approximate Shapley values perturbs the input vector, which, without feature independence, can produce improbable or logically impossible feature combinations [12, 13]. Given that the AFV consists of several highly correlated features (for example, `BVP_mean_t-180s` and `BVP_median_t-180s`) this may distort individual coefficients or SHAP values.

To address this issue, we considered features from the aggregate perspective of their originating measurements. Hence, Figure 2 sums coefficients by measurement. We applied a similar process to the SHAP values, evaluating their influence by their standard deviation [7], noting the following results:

	TPA	ACC	BVP	EDA	AOF
LR	1.05	0.41	0.29	0.20	0.11
SVM	0.55	0.08	0.08	0.01	0.54

This summing method can help cancel out collinearity created by the summary statistic calculations, but it does not account for correlations among the original measurements. Thus, examining feature importance via the differences in model performance remains relevant as additional evidence to suggest that TPA and AOF are the models' most important measurement.

7 Conclusion and Discussion for Further Work

In summary, among the models targeted by this analysis, the evidence presented suggests that these models obtain the majority of their predictive power from TPA and little from the physiological variables.

Our work does not undermine the original hypothesis by Imbiriba et al., which is that changes in peripheral physiology may be useful for predicting aggression among inpatients in the target population. However, we show that the higher performing models trained on the AFV are likely not evidence supporting the usefulness of the wearable biosensor data for predicting future aggression onset. Nevertheless, the paper by Imbiriba et al. is an extension of prior research that established a sound basis for the aforementioned hypothesis. Further development of existing or new models using the gathered data, especially the physiological data alone, may yield even greater support for it.

Though we deem this collection of evidence presented as sufficient for arriving at our conclusion, further interrogation of the input features, using methods designed to confront the challenges presented by multi-collinearity, such as Cohort SHAP [12, 13], could reveal different results. Additionally, a similar

[5] This is because many different response functions may provide the same good fit [9].

analysis of feature importance could be performed on the models that exclude TPA and AOF. This could help identify which physiological features contain the most predictive power.

As mentioned above, this analysis does not seek to identify underlying *causes* of the observed aggression. Further work could explore methods to discover causal relationships from the time-series data [2,14]. An accurate causal model could help reduce the models' sensitivity to domain shift and inform caregivers about possible preventative interventions.

References

[1] Jenny Balfer and Jürgen Bajorath. Visualization and interpretation of support vector machine activity predictions. *Journal of Chemical Information and Modeling*, 55(6):1136–1147, 2015. PMID: 25988274.

[2] M. Castro, P. R. Mendes Júnior, A. Soriano-Vargas, et al. Time series causal relationships discovery through feature importance and ensemble models. *Scientific Reports*, 13:11402, 2023.

[3] Intel Corporation. scikit-learn-intelex: Accelerated machine learning for intel(r) architecture, 2024. Python package.

[4] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ROC Analysis in Pattern Recognition.

[5] UCLA: Statistical Consulting Group. Faq: How do i interpret odds ratios in logistic regression?, 2024.

[6] Tales Imbiriba, Ahmet Demirkaya, Ashutosh Singh, Deniz Erdogmus, and Matthew S. Goodwin. Wearable Biosensing to Predict Imminent Aggressive Behavior in Psychiatric Inpatient Youths With Autism. *JAMA Network Open*, 6(12):e2348898–e2348898, 12 2023.

[7] Harrison Jansma. You are underutilizing shap values: Feature groups and correlations. `https://towardsdatascience.com/you-are-underutilizing-shap-values-feature-groups-and-correlations-8df1b136e2c2`, 2024. Accessed: 2024-08-08.

[8] László A. Jeni, Jeffrey F. Cohn, and Fernando De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 245–251, 2013.

[9] Michael H. Kutner et al. *Applied Linear Statistical Models*. McGraw-Hill/Irwin series Operations and decision sciences. McGraw-Hill/Irwin, New York, NY, 5th edition, 2005. Includes bibliographical references and index.

[10] Scott Lundberg. Sentiment analysis with logistic regression. `https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/linear_models/Sentiment%20Analysis%20with%20Logistic%20Regression.html`, 2018. SHAP latest documentation.

[11] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017.

[12] Masayoshi Mase, Art B. Owen, and Benjamin Seiler. Explaining black box decisions by shapley cohort refinement. *CoRR*, abs/1911.00467, 2019.

[13] Masayoshi Mase, Art B. Owen, and Benjamin B. Seiler. Cohort shapley value for algorithmic fairness. *CoRR*, abs/2105.07168, 2021.

[14] J. Runge, A. Gerhardus, G. Varando, et al. Causal inference for time series. *Nature Reviews Earth & Environment*, 4:487–505, 2023.