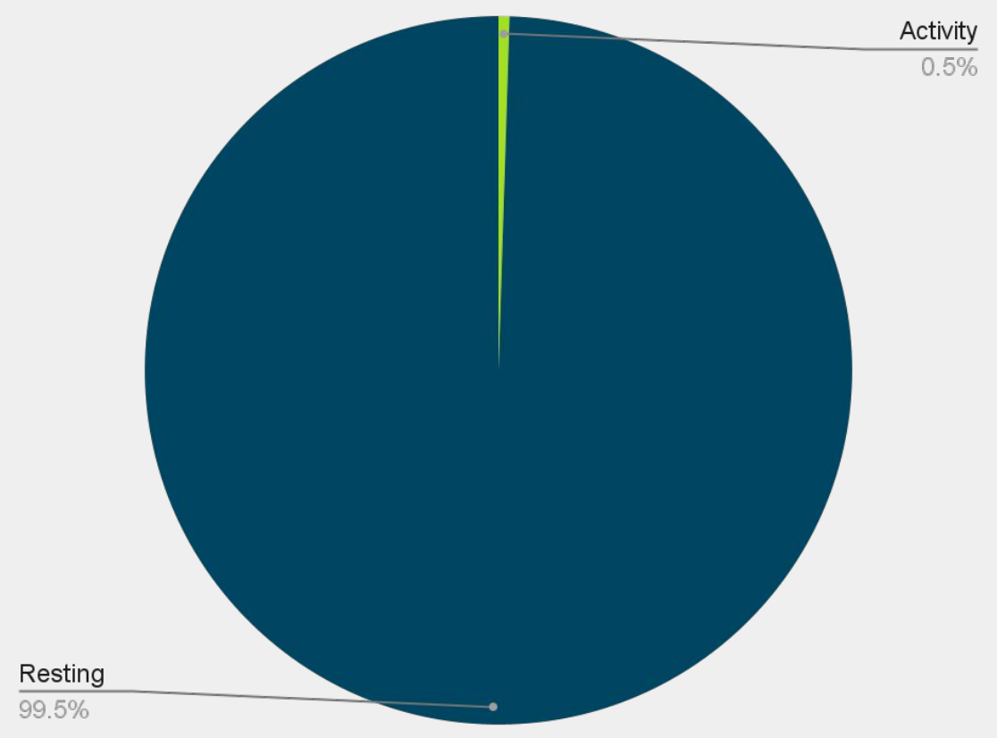
“It looks like you’re working out.”: Using Machine Learning Classification Techniques to Predict Physical Stress

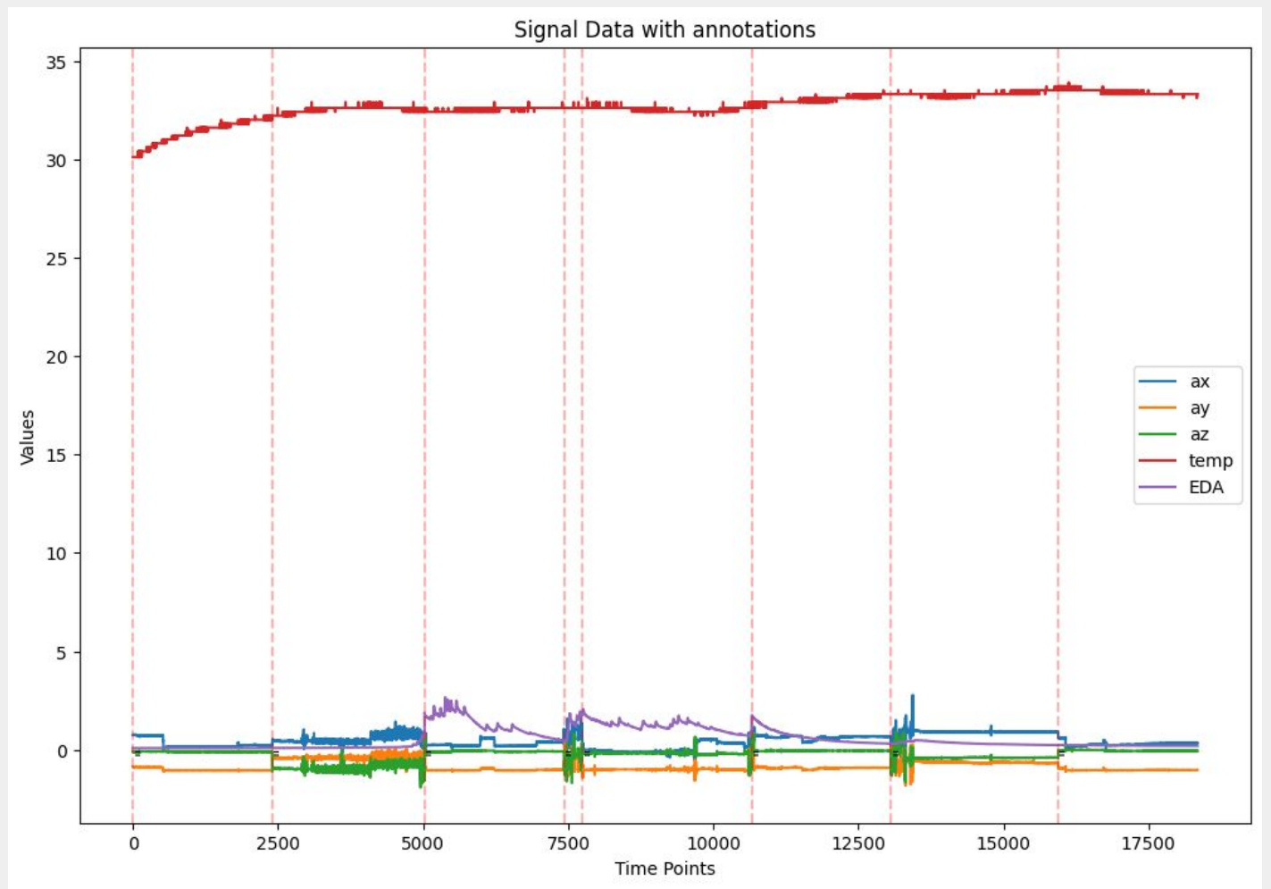
Eva Schwartz
DATS 4001: Data Science Capstone,
George Washington University

Dataset and Methods



Dataset courtesy of PhysioNet, collected using the Empatica 4 fitness wearable, with extreme class imbalance for active state v. resting state as indicated by pie chart to the left

Initial variables included heart rate, electrodermal activity (EDA), body temperature, respiratory rate, and z, y, and x-axis acceleration

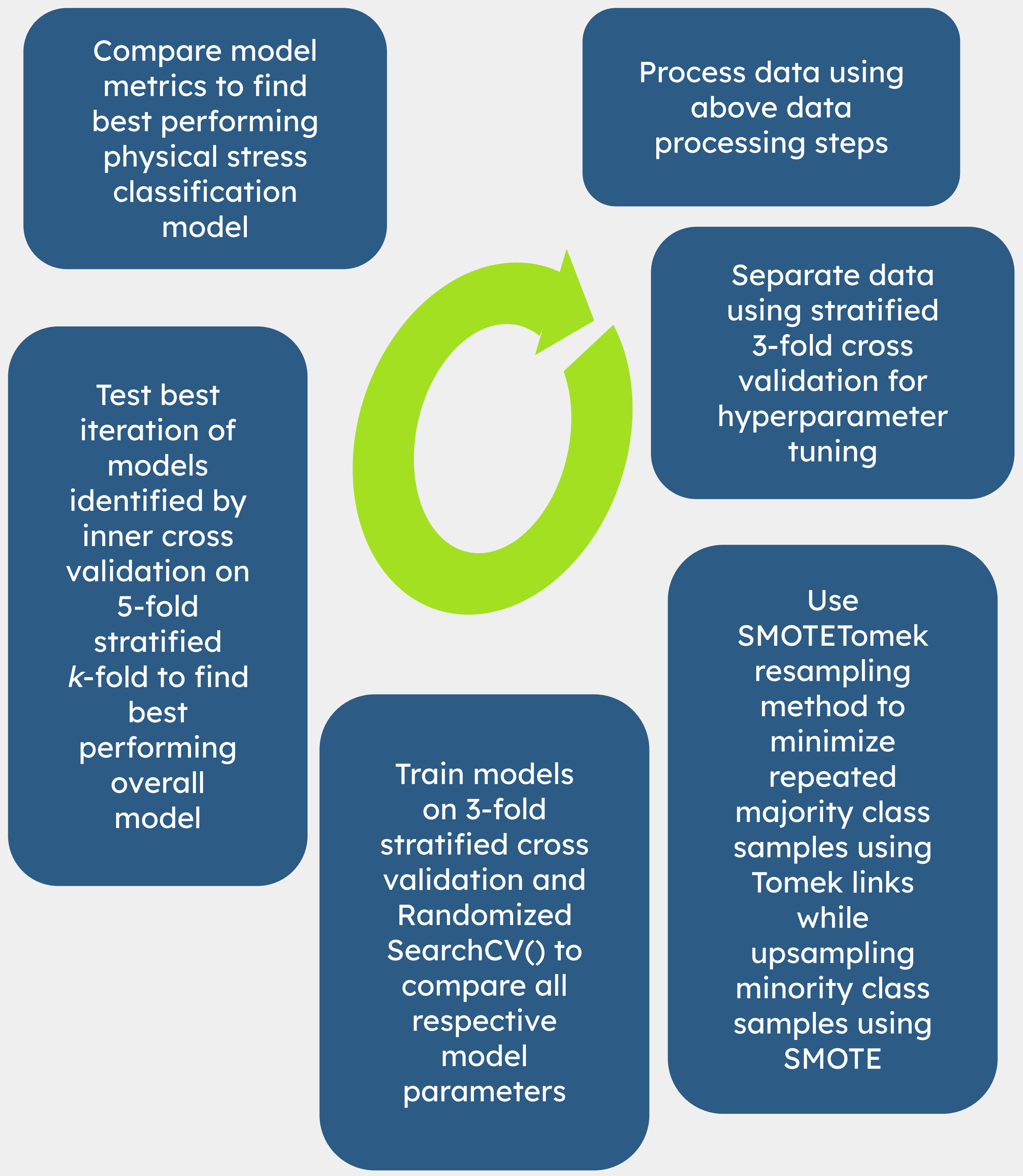


Data processing steps:

1. Raw signal processing, with raw signal example to the right
2. Merging of dataframes
3. Time synchronization of dataframes:
    a. Some data was collected in 1-second intervals and some in ⅛ - second intervals therefore all data was synced to ⅛ second intervals to avoid data loss
4. Feature extraction:
    a. Features extracted include temperature/EDA ratio, acceleration magnitude with the formula for this feature to the bottom right, and activity adjusted heart rate

$$\|\vec{a}\| = \sqrt{a_x^2 + a_y^2 + a_z^2}$$

Implementation



Process data using above data processing steps

Separate data using stratified 3-fold cross validation for hyperparameter tuning

Use SMOTETomek resampling method to minimize repeated majority class samples using Tomek links while upsampling minority class samples using SMOTE

Train models on 3-fold stratified cross validation and RandomizedSearchCV() to compare all respective model parameters

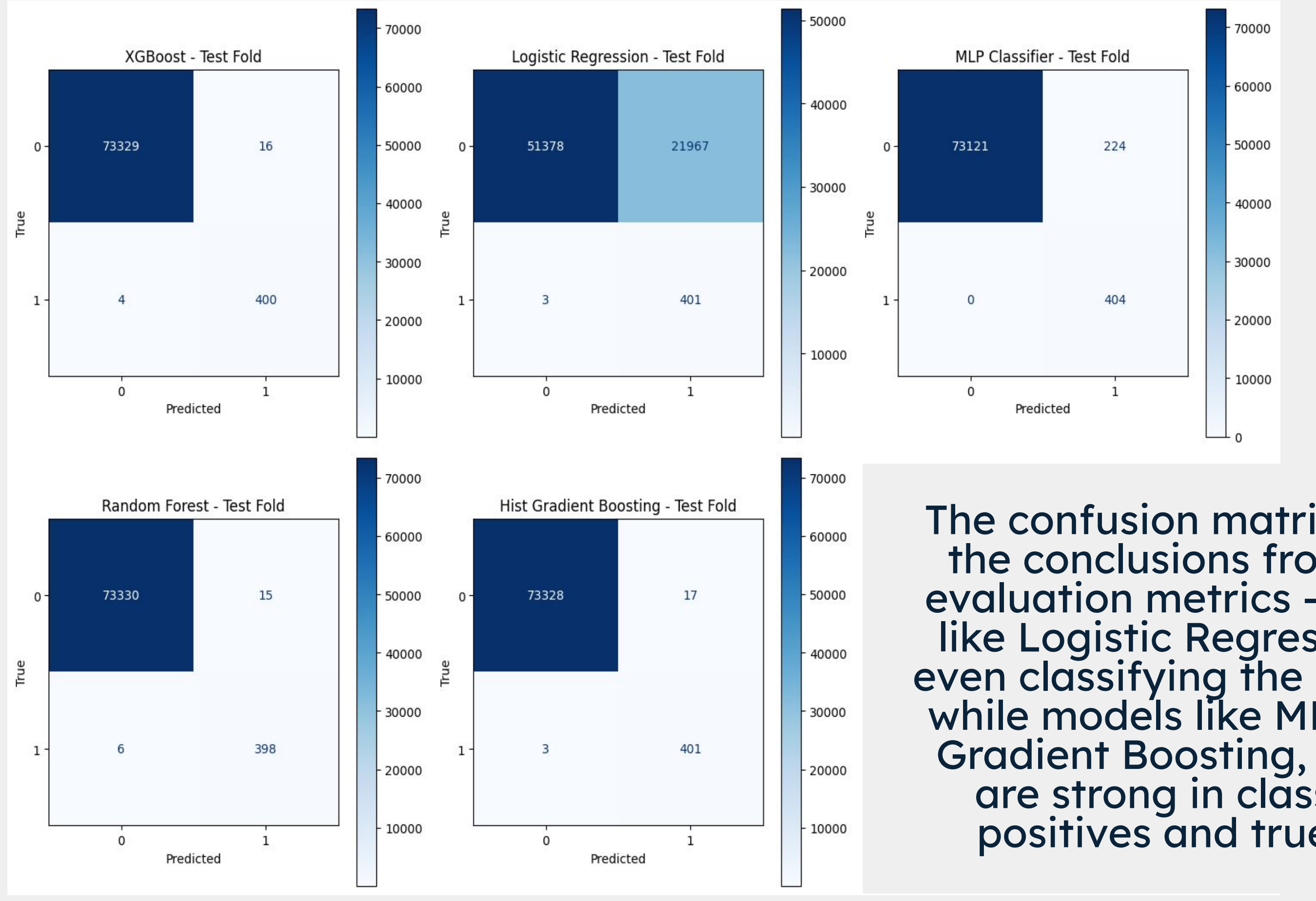
Test best iteration of models identified by inner cross validation on 5-fold stratified *k*-fold to find best performing overall model

Compare model metrics to find best performing physical stress classification model

Results

Model	Accuracy (mean ± std)	Recall (mean ± std)	Precision (mean ± std)	ROC AUC (mean ± std)	PR AUC (mean ± std)	Specificity (mean ± std)	NPV (mean ± std)
XGBoost	0.9934 ± 0.0018	0.9871 ± 0.0036	0.9537 ± 0.0071	1.0000 ± 0.0000	0.9970 ± 0.0013	0.9997 ± 0.0000	0.9999 ± 0.0000
Logistic Regression	0.8476 ± 0.0010	0.9852 ± 0.0056	0.0184 ± 0.0004	0.9196 ± 0.0030	0.0319 ± 0.0004	0.7100 ± 0.0072	0.9999 ± 0.0000
MLP Classifier	0.9981 ± 0.0005	0.9995 ± 0.0010	0.6297 ± 0.0209	0.9996 ± 0.0001	0.9301 ± 0.0250	0.9967 ± 0.0003	1.0000 ± 0.0000
Random Forest	0.9937 ± 0.0026	0.9876 ± 0.0052	0.9551 ± 0.0050	1.0000 ± 0.0000	0.9949 ± 0.0012	0.9997 ± 0.0000	0.9999 ± 0.0000
Hist Gradient Boosting	0.9951 ± 0.0028	0.9906 ± 0.0055	0.9414 ± 0.0104	1.0000 ± 0.0001	0.9971 ± 0.0011	0.9997 ± 0.0001	0.9999 ± 0.0000

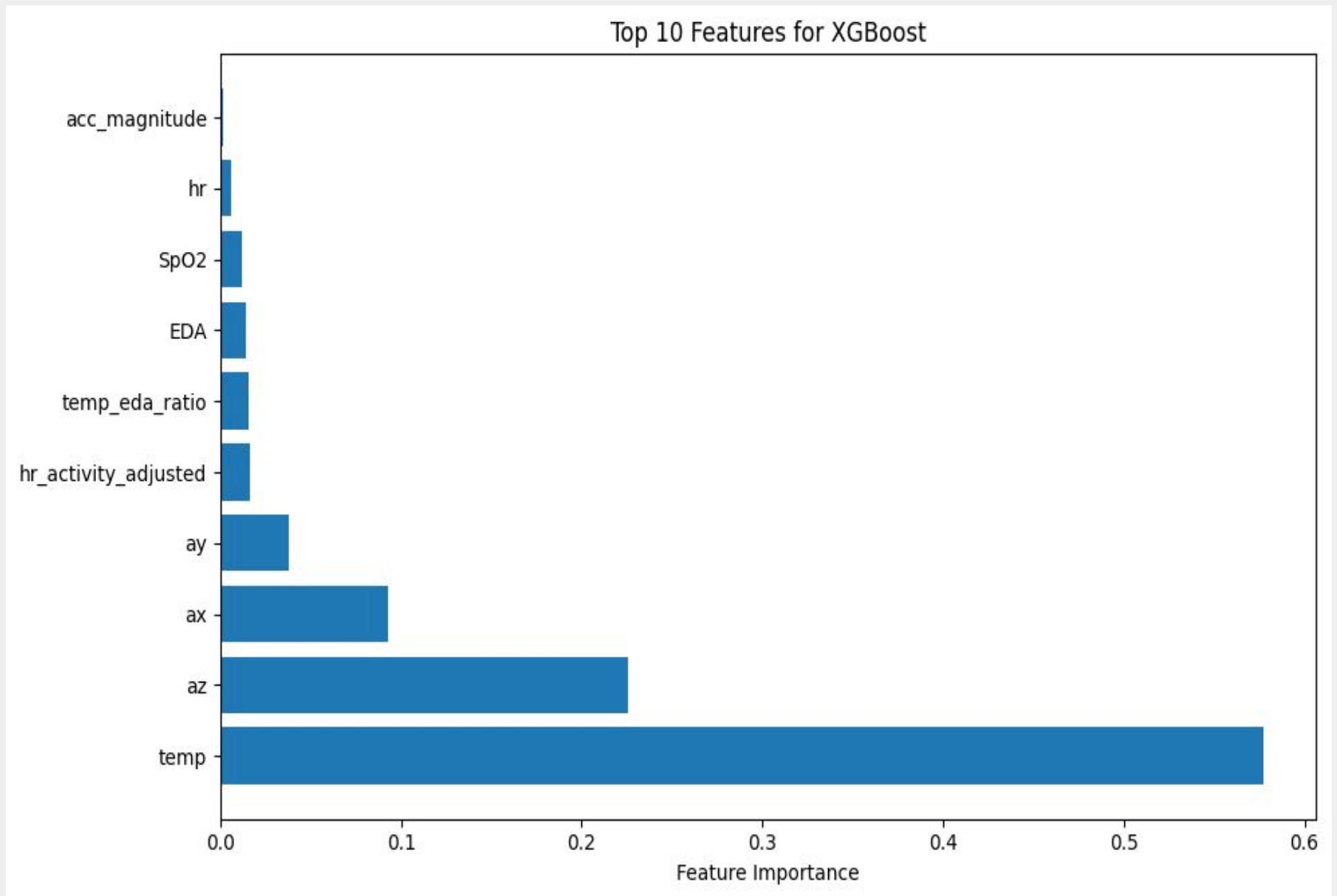
Random Forest and XGBoost both had strong specificity and negative predictive value metrics.
The Multi-Layer Perceptron model had the highest negative predictive value, though the difference between models in this metric is very small.
The weakness of Logistic Regression when comparing metrics is fairly clear, though this is to be expected given the simplicity of this model.

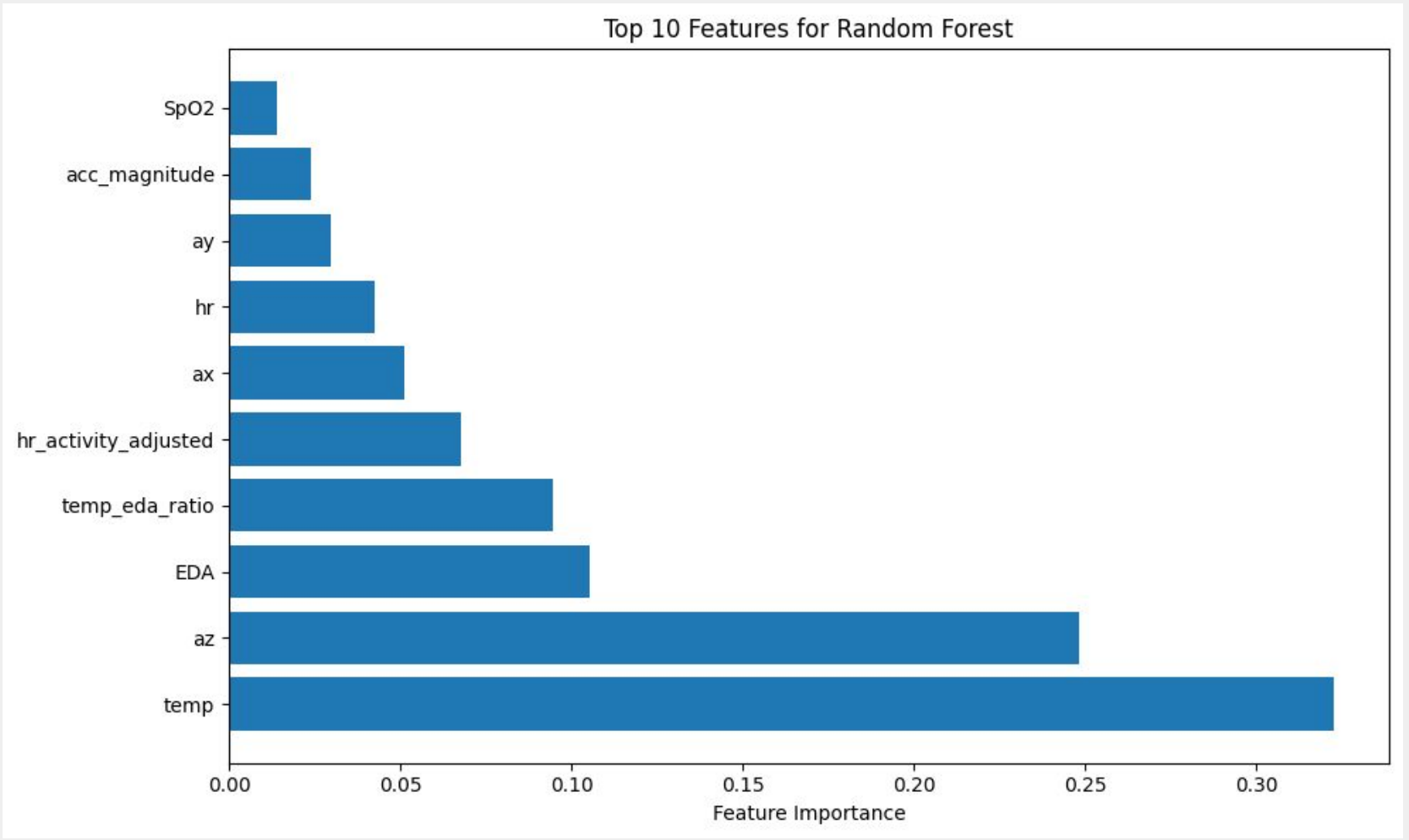


Confusion matrices of the best performing models for each respective model can be seen to the left, with darker blue shades indicating higher class counts

The confusion matrices align with the conclusions from the model evaluation metrics - basic models like Logistic Regression suffer in even classifying the majority class, while models like MLP, Histogram Gradient Boosting, and XGBoost are strong in classifying true positives and true negatives

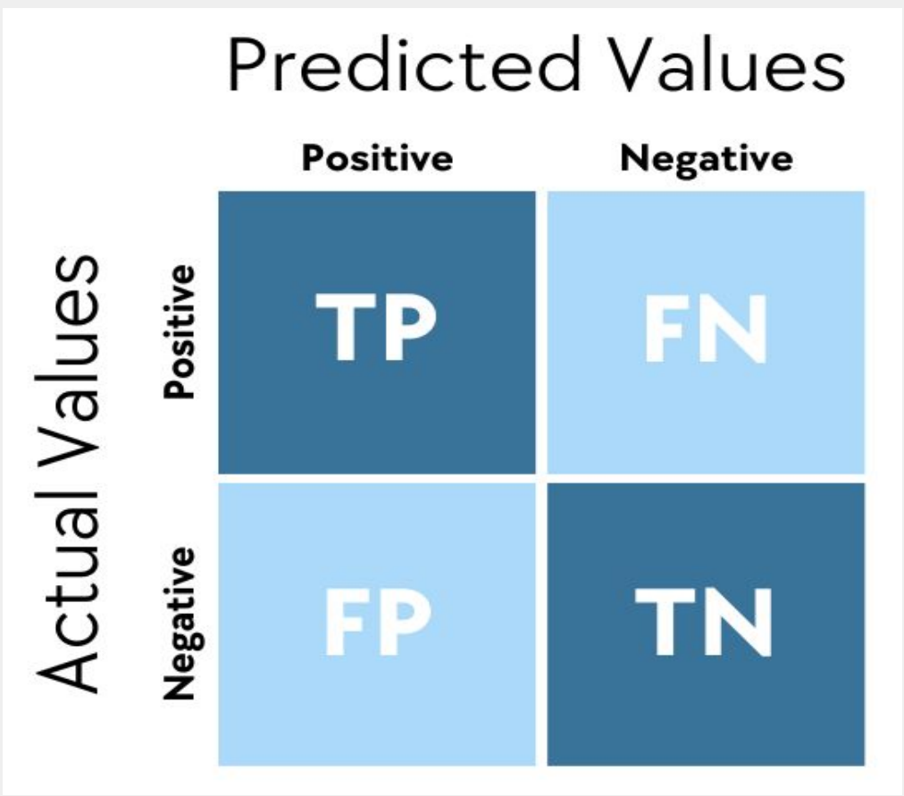
Below is a feature importance plot for the best performing XGBoost model. Body temperature indicates very strong importance in predicting physical stress, followed by z-axis acceleration, measuring forward and backward movement.





Above is a feature importance plot for the best performing Random Forest model. Like XGBoost, body temperature and z-axis movement are both very important in predicting physical stress, followed by electrodermal activity.

Overview of Machine Learning Model Comparison Metrics



		Predicted Values	
		Positive	Negative
Actual Values	Positive	TP	FN
	Negative	FP	TN

Confusion matrices are used in classification tasks to aid in formulation of model evaluation metrics

For this project, the ‘positive’ majority class is the non-active, resting state, and the ‘negative’ minority class is the active, physical stress state

The false negative and false positive values aid in indicating the strength of a model, while also demonstrating where a model may be weaker

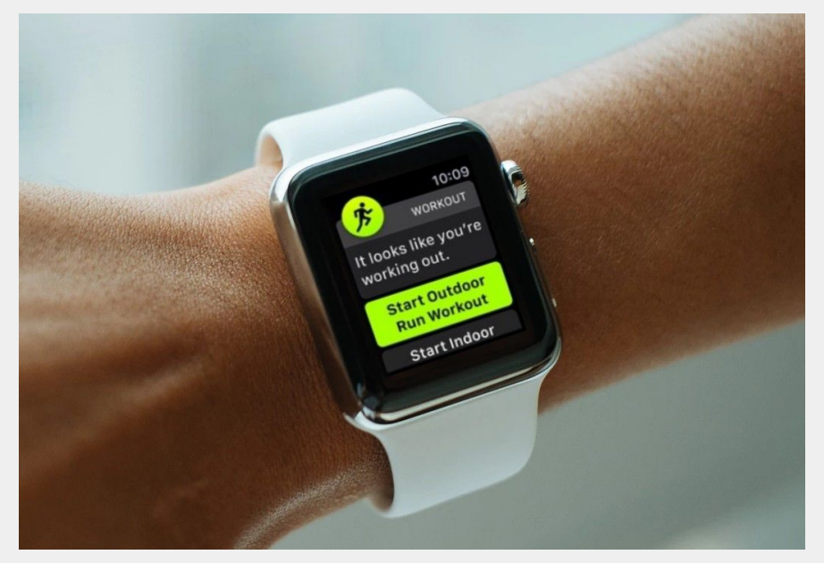
Metrics like **Accuracy, Recall, Precision, AUC ROC,** and **PR AUC** are important for imbalanced datasets, but cater more towards evaluating the strength of models in predicting the majority class

In the case of predicting physical stress, a false negative – predicting no physical stress when physical stress is present – can be more costly than a false positive – predicting physical activity stress when physical stress is not present, indicating the importance of the following metrics

Specificity: TN/(TN + FP) → analyzes the ratio of negative values that are correctly classified as negative, vital metric as this research is focused on predicting is the negative class

Negative Predictive Value (NPV): TN/(TN+FN) → analyzes proportions of true negatives out of all negative predictions, measures the accuracy of negative predictions by models

Limitations and Alternate Applications



- There was a lack of female representation in the dataset, indicating that though these models had strong performance, their performance on a true, diverse dataset may change
- There are inherent difficulties in predicting extremely imbalance datasets → some false positives sneak through even on very well developed models and products
- Multi-layer perceptron is a good model, but it operates as a “black box,” meaning there is not much information on how the hidden layers work

Model comparison research can be used in predicting cyber attacks, especially because cyber attacks also generally suffer from imbalanced data in training

This research can be applied to disease/medical prediction given the health focus of the model development and the focus on optimization of machine learning models for imbalanced datasets

Conclusion

- Multi-layer Perceptron, Random Forest, and XGBoost are all strong models for this task. Implementing lighter, faster models like LightGBM for classifying physical stress could be a powerful next step for this body of research.
- Temperature emerged as a very important factor in predicting physical stress across models therefore correctly calibrating and developing wearables sensors to accurately predict body temperature is vital in deploying this research to fitness wearables.
- An interesting exploration for this research would be to compare the results of these models to those gathered from a female-dominant biosignal dataset. While this would be a fruitful examination, the lack of female inclusion and representation in health datasets could limit this path of research.

Please reference the QR code for complete report, code, presentation, and references:

Contact: Eva Schwartz
eschwartz25@gwu.edu
The George Washington University