Individual Assignment

Eren Chaglar (S245735)

June 26, 2025

Task 1

Introduction

I choose paper X as "Machine Learning With Neuroimaging: Evaluating Its Applications in Psychiatry", and as paper Y "Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images"

Introduction to Papers

Paper X presents a practical, tutorial style guide to evaluating machine learning applications in psychiatry. It first describes best practice steps for developing and validating supervised models that rely on MRI based brain connectivity data to make individual level predictions that could aid in diagnosing psychiatric disorders. It then turns from predictive performance to scientific insight, explaining how to assess and interpret the neural mechanism features uncovered by these models. Together, these two sections provide a coherent framework for judging both the clinical utility and the biological interpretability of machine learning approaches in mental health research.

Paper Y is a study that shows a deep neural network trained on face features from ca. 35 thousand dating site photos classifies sexual orientation far better than people can. Researchers also generalize the classifier to Facebook photos, to highlight serious privacy risks, because the model can reveal an intimate personal trait without consent.

Evaluation Analysis

Why the Evaluation in Paper X Is Appropriate

Their training pipeline is solid. Data are split into three clean sets: one for fitting, one held out just for tuning, and a third, completely separate group of people for the final check. The model never sees its own test data, so any accuracy bump should carry over to new scanners and new participants. They also rethink the labels. It is easy to predict today's diagnosis from the scan, but more useful to predict age gaps, later symptoms, or treatment response. To do that they scan people before the outcome appears and train the model on these future targets.

Performance is probed from every angle. They report several metrics, wrap each in bootstrap confidence limits, and run label shuffle tests to show the numbers sit well above chance. Head motion, scanner brand, and age are matched or removed, so the model is not chasing noise. Finally, they retest on the outside group, map which brain links drive the decisions, and publish all code and settings plus a head-to-head comparison with clinicians. This mix of clean data

splits, rich labels, tough stats, confound control, and open reporting makes the evaluation fit for real clinical use.

Why the Evaluation in Paper Y Is Inadequate

The test data are not diverse. It uses only have users that have chosen their location as US and also it is from one dating site. The extra test set has only gay Facebook users, so we cannot tell how the model behaves with other races or cultures. The labels come from how people describe themselves on their profiles. Some straight users could actually be gay or bi, yet the study treats every label as one hundred percent correct.

Many extra factors are left unchecked. Dating photos include style cues like grooming, lighting, and phone filters that link to orientation. The paper hides the background, but it does not match haircut, facial hair, or pose. They score the model almost only with AUC from twenty user splits. There are no confidence intervals. They do talk about ethics and privacy, but they run no formal tests. We still do not know how easy it is to fool the system or how much private information it could leak.

Recommendations for Paper Y

The recommendations I would make here would be to collect broader data, like adding other races, other countries, and people. Include straight user from Facebook too. Then report the scores for each group on its own. Do a follow up survey that looks how people really identify, instead of assuming that they are homosexual. Pairing gay and straight faces that share the same beard status, glasses, and image quality and head pose, to show the model still works after you level these factors. They could also have added confidence intervals, to give out some kind of statistical test rather then only relying on AUC.

Task 2

Data Overview

In this project I will be working with "HR_dataset". Data has been collected by an experiment with total of 26 participants over three cohorts(D1_1, D1_2, D1_3). Because of the how experiment is designed, same persons heart rate measured in 4 different rounds and also 3 different phases in every round. So for each person we have total of 12 measurements for each person. I only have access to following cohorts D1_1 and D1_2, that means total of 14 participants and 168 data points. The raw data has total of 12 features. I will implementing a logistic regression and a random forest to classify different levels of frustration using 8 of the features in the dataset. The other 3 features are excluded because they are anonymised ID, or a time-stamp. Two of the four will be used for the cross validation scheme.

Frustration is a number between 0-10, therefore I made three labels, 0-2 will be "Low" level of frustration, 3-5 will be "Medium" level of frustration and 6-10 will be "High" level of frustration, so it is easier to interpret where we are on the scale. Resulting label distribution is 62.5% "Low", 32.5% "Medium" and 6% "High".

Some of the features are numerical, and some other are categorical. The numerical data has been standardized, and categorical data has been encoded with one-out-of-K encoded.

Cross-Validation Design

Because the dataset contains repeated measures and two distinct cohorts, a **nested cross-validation** scheme is used:

- **Outer loop**: Leave-One-Cohort-Out (2 folds) – train on D1_1, test on D1_2 and vice-versa.
- **Inner loop**: GroupKFold ($k = 3$), grouping by `Individual`, for hyper-parameter tuning via `GridSearchCV`.
- **Metric**: Macro-averaged F1 (equal class weight) for both tuning and reporting.
- **Statistical comparison**: Each outer fold yields paired Macro-F1 scores. Since normality is doubtful with $n = 2$, the **Wilcoxon signed-rank test** is applied; it assesses the median of the pairwise differences without distributional assumptions.

Models & Hyper parameters

Logistic regression has been chosen as the simple, interpretable baseline and RF is chosen to capture non linear interactions. For the baseline model we only tune the regularization strength λ over the values 0.1, 1, and 10. A higher λ means the model will fit to data more closely. For the random forest to keep the search minimal I varied the maximum depth of tree, minimum number of samples per leaf and the number of features tested at each split.

Both grids are explored inside the inner 3 fold GroupKFold. The combination that yields the highest macro F1 on the inner validation folds is refitted on the full outer training set before final testing.

Results

Figure 1 summarises the macro F1 obtained on two leave one cohort out folds. Logistic regression achieves mean of 0.446 ± 0.06, which is slightly higher than random forest. A paired Wilcoxon test is applied to the two outer fold scores which returns the $p = 0.750$ which mean the null hypothesis cannot be rejected. So there is no significant difference between models.

Model	**Macro-F1 (mean ± std)**
Logistic Regression	0.446 ± 0.058
Random Forest	0.414 ± 0.100
Wilcoxon signed–rank (LR <RF): $W = 2.000$, $p = 0.750$	

Table 1: Cross-cohort LOCO F1 scores and Wilcoxon comparison.

Both models classify the "Low" frustration state reliably 1, but they both struggle with "Medium". Over 30% of "Medium" instances are mislabeled as "Low". Performance on "High" is volatile probably because of only having 6% of the samples.

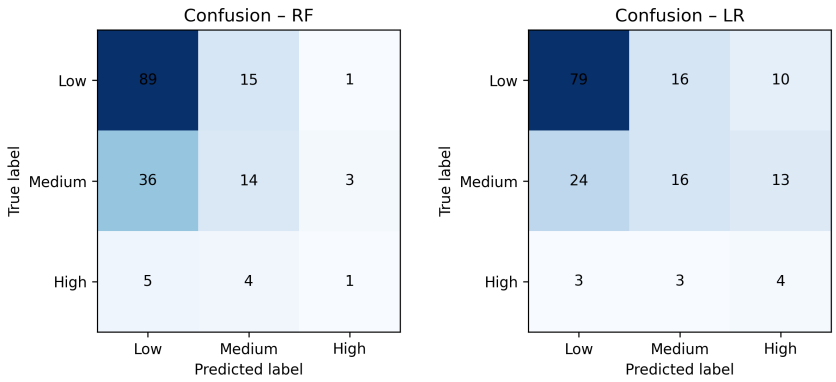

Figure 1: Side-by-side comparison of the two models' confusion matrices

The random forest feature importance plot shows which features model actually relies on. We can see that forest lean mainly on HR_std (heart rate variability) and HR_Max. It was expected as the change in heart rate can show change in frustration, and maximum heart rate can also indicate that someone is frustrated. The fun part is to see that model relies on "phase2", which could be indicating that most of the people are frustrated in phase 2.

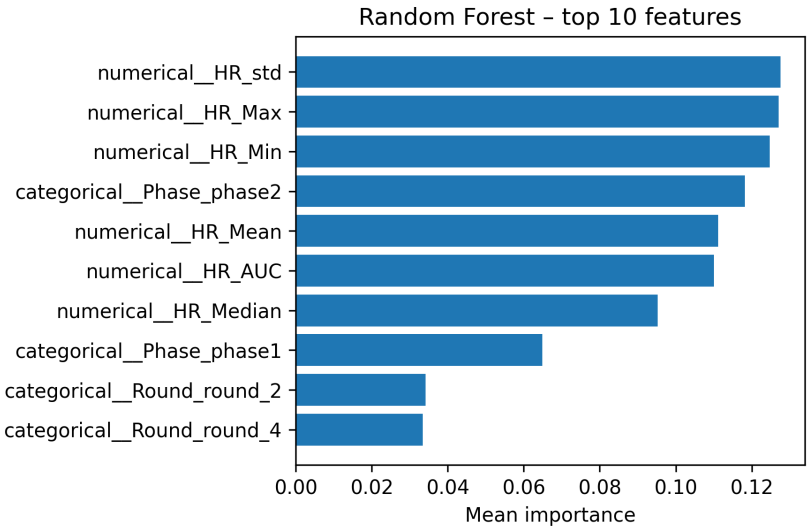

Figure 2: Enter Caption

Conclusion

In this study I implemented two ML models to predict self reported frustration from heart rate data collected in small, repeated measures experiment. Using a nested Leave-one-cohort-out design ensured that neither round nor person specific signals leaked into the test set.

Logistic regression baseline performed on par with and in fact a little bit better than a random forest classifier. A paired Wilcoxon test confirmed that the difference between the two models are not statistically significant.