White Paper

Preliminary Research to Determine Whether Data Generated by DRUID Can Distinguish Between Alcohol and Cannabis Impairment

Titi Ala'ilima, Impairment Science, Inc.
Max Daniller-Varghese, PhD, Impairment Science, Inc.

Summary
This report presents evidence that the DRUID app's cognitive and psychomotor testing platform can go beyond quantifying a person's level of impairment and distinguish between alcohol-related impairment and cannabis-related impairment.

Background DRUID, developed by Impairment Science, Inc. ("ISI"), has demonstrated effectiveness for measuring impairment, whatever its cause, by evaluating user performance on a series of short psychomotor and cognition tasks. Historically, the app has assessed the severity of cognitive and psychomotor impairment without identifying its cause. However, scientific literature and years of internal testing by ISI have suggested that different chemical substances may affect balance, reaction time, decision-making, time estimation, and motor control in distinct ways. Accordingly, we set out to determine whether DRUID can reliably detect these differences between alcohol- and cannabis-induced impairment.

Method Building on a small pilot study conducted in February 2025, we followed up in August 2025 with a more rigorous study with a greater number of participants, more DRUID tests per testing session, and more data per test.

Nineteen participants completed closely supervised alcohol and cannabis consumption sessions, producing 568 DRUID tests suitable for advanced machine-learning analysis.

We applied the data modeling method, Random Forest Classifier, across 1,890 feature combinations, with each model executed 1,000 times to establish its median performance.

Results We found that hundreds of the tested models achieved strong precision, recall, and overall classification accuracy. Specifically, the F1 scores (a machine-learning evaluation metric that measures accuracy) exceeded 0.85 for both substances in 315 unique models, a result that is both highly significant and clinically meaningful.

Conclusion These results demonstrate that alcohol- and cannabis-induced impairment leave reliably distinct data signatures. Importantly, the study provides strong evidence that with further development, DRUID can move beyond measuring general impairment alone to identify its underlying cause.

The study's success sets the stage for pursuing several paths forward. Researchers can now examine whether certain user groups are easier or harder to classify, investigate which specific test features are most informative, and explore how this capability could be integrated into DRUID's standard testing procedure.

Introduction

DRUID has successfully demonstrated its capacity to assess the degree to which test subjects, commercial users, and athletes are impaired due to any of several causes (alcohol, cannabis, concussion, fatigue, etc.). To date, however, the app has not yet been designed to identify the source of a user's impairment. The app calculates an impairment score and simultaneously records thousands of component measurements. ISI has had a long-standing interest in determining whether this rich store of data might make it possible to conclusively identify the cause of a user's impairment.

Model Performance

To analyze the data from February and August 2025 studies, we used the machine learning concepts of *Precision* and *Recall*.

Precision, also known as *Positive Predictive Value*, is the fraction of correctly identified cases among all cases that were flagged as "positive," including those misidentified as positive ("false positives"). In other words, this metric answers the question, "Of all the people the model flagged as positive for alcohol [or cannabis], what proportion were actually positive?" Fewer false positive cases mean greater precision.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall, also known as *Sensitivity*, is the fraction of actual positive cases that were correctly flagged as "positive." This metric answers the question, "Out of all the people who are actually positive for cannabis [or alcohol], what proportion were correctly classified as positive?" Fewer "false negative" cases mean greater sensitivity.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Note that a model with high precision (i.e., fewer false positives) may miss some true positives (low recall), while a model with a high recall (i.e., fewer false negatives) may produce more false positives (low precision).

A blended metric between these two values is known as the "F1" score, which is calculated as the harmonic mean of Precision and Recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The closer F1 is to 1.0, the better the model is at discerning between the different classes (in this instance, alcohol and cannabis) and the more skillful it is at identifying a given class without

compromising its ability to identify the other.

August 2025 Study

For this study, we modified the research design of our pilot February 2025 study by increasing the number of participants and tightening up the eligibility criteria. In addition, we utilized the 3-minute version of the app, DRUID*Benchmark*, instead of the 1-minute version, DRUID*Rapid*, which was used in the February study. The longer version of DRUID incorporates a fourth subtest not included in the 1-minute version and increases the amount of test data that can be used to build and tune the machine learning model.

Finally, we also created a new version of DRUID*Benchmark* – DRUID*ai* - that increased the data-sampling frequency for certain parts of the test by a factor of ten. Whereas the "heartbeat" of the commercial versions of DRUID is ten measurements per second, DRUID*ai* takes measurements at 100 Hz, that is, 100 per second, without changing the app's fundamental algorithm.

After securing IRB approval, we recruited 19 participants to use alcohol on one day and cannabis on another, all under close supervision. They consumed alcohol to attain a breath alcohol concentration (BrAC) of ≥ 0.10 mg/dL, and on a separate occasion inhaled cannabis to the point of self-reported intoxication. We accumulated a total of 568 tests across all study participants. While this amount of data is not in the realm of "big data," we were firmly in the realm of "medium data," sufficient for meaningful machine learning.

Results

After much feature engineering – the process of creating, transforming, selecting, and representing input variables (features) – and after extracting from the data meaningful relationships that appeared to be correlated, we subjected the data to a large-scale ensemble of models using the most likely informative combinations of features.

Each unique combination (1,890 models in the most recent run) received 1,000 unique executions. The median results from each model's ensemble of unique executions are reported here. Sometimes a model does exceptionally well or poorly due to the initial random allocation of the feature tree, which may not be truly reflective of the model's skill at discerning between the two classes at hand, and therefore reporting median results for the executions is optimal.

Of the many models performed with different feature sets, there are several that perform extremely well. The distribution of median model performances is reported below in the histograms. While there are models with feature sets whose median performance is not performant, there are several hundred with high Precision and Recall for both alcohol and cannabis.

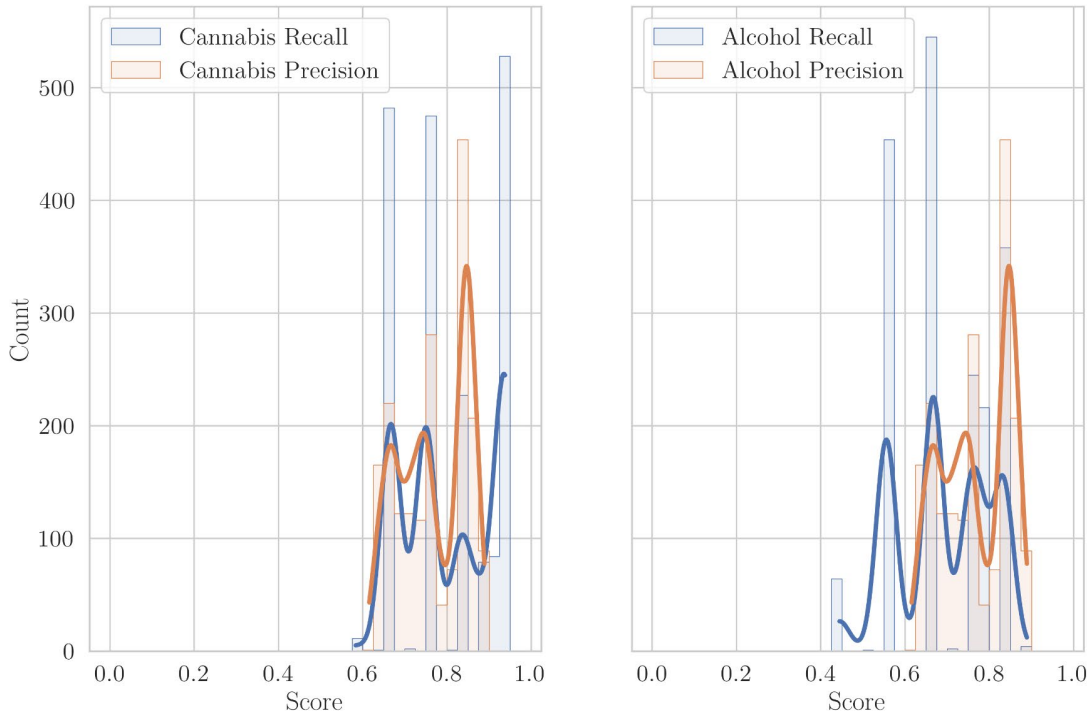
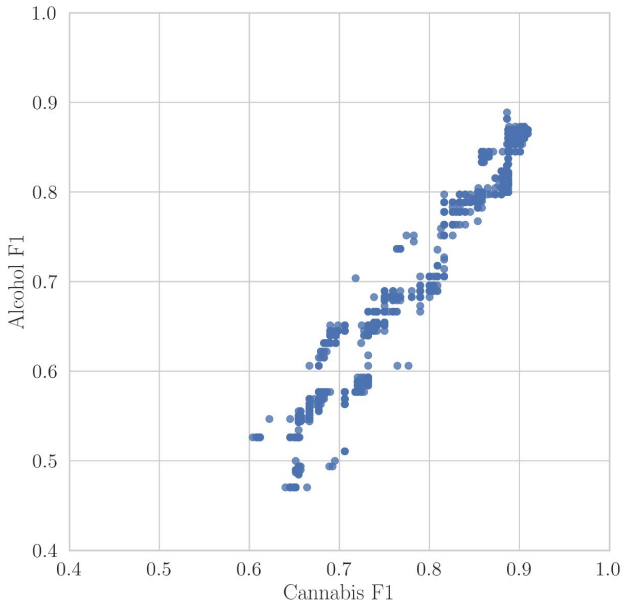


Figure 1: Cannabis and Alcohol Precision and Recall for all 1,890 models. Each model was executed 1,000 times to create the reported median value.

Additionally, the model performance (F1 score) for alcohol was not produced at the expense of performance for cannabis, and vice versa (as shown below in Figure 2). The linear relationship between the two model performances indicates good model calibration.



.

Figure 2: Cannabis and Alcohol F1 comparison.

The data show that there is a large ensemble of machine learning models that performed extremely well, with over 315 models that have an F1 from 0.85 - 0.90+ for both cannabis and alcohol, a metric that indicates good-to-excellent model performance and clinical significance.

Future Work

The overall strength of the models for our test data demonstrates the value of continuing this line of research and development. Next steps include the following:

Cohort analysis: While overall metrics are reported for each class, further investigation is warranted of particular users or cohorts of users who are identified unusually well or poorly. This analysis could yield more sophisticated models and incorporate demographics as a feature of testers.

Examining success and failure modes: Still needed across our models is a detailed feature importance analysis. Certain models may achieve good performance with a range of different features, and understanding why this is the case will lead to more robust models in the future.

Implementation: AI analyses could be incorporated in a future version of DRUID*Benchmark*, allowing it not only to measure cognitive and motor function but potentially distinguish between various causes of detected impairment.

Technical Appendix

Analysis The backbone of the data analysis is a modeling method called the Random Forest Classifier (RFC), a machine learning algorithm that combines multiple models to produce a single, stronger predictive model that is commonly used for predictive tasks such as classification and regression.

The RFC builds decision trees, simple models that are independent of one another, that split data into branches based on feature thresholds that are predictive of the classes of interest. Each tree is trained on a "bootstrap" sample of the data which can be reused by other trees, and at each "split" the algorithm considers a random subset of features, a process called "bagging."

By executing many iterations of decision trees with a unique ensemble of features and then averaging their predictions, the model pipeline reduces variance and avoids overfitting for a given set of features, while still capturing complex interactions in the data. This method is useful because it is robust, stable, and generally requires less tuning (adjusting a model's settings to improve its predictive accuracy). It handles nonlinear relationships, mixed data types, and noisy features effectively. The algorithm is also computationally efficient, especially on medium-sized datasets, and can be parallelized easily because each tree is fit independently.

For datasets derived from our studies, the RFC offers a good analytical balance. It can model complex patterns and provides helpful interpretability tools such as feature importance rankings that allow us to understand which variables most strongly influence model predictions.

Pragmatically, the RFC often performs well as a reliable baseline or first-choice model in many applied machine learning tasks. Its mix of statistical stability, computational efficiency, and practical interpretability makes it particularly well-suited for problems where data are structured, moderately sized, and the goal is to capture meaningful patterns without requiring highly specialized models with complicated architectures. Further exploration of the data, and any additional data, will use these results as a benchmark.