Cook Engineering Design Center

Final Report

Submitted in partial fulfillment of the requirements for

ENGS 90: Engineering Design Methodology and Project Initiation

Eye Tracking for Seizure Detection

3/7/2022

Sponsored by

Dr. Erik Kobylarz

Project Team #592

Brian Provost & David Vonderheide

Faculty Adviser

Professor Kofi Odame

EXECUTIVE SUMMARY

Epilepsy is a neurological condition characterized by unpredictable, recurring seizures which affects over 65 million people worldwide[1,2]. Epilepsy treatment plans depend on seizure frequency, duration, and type; however, no reliable methods exist to monitor seizure activity outside of the hospital and self-reported logs are notoriously unreliable[4]. The golden standard for seizure monitoring is combined video and electroencephalography (vEEG)[5] in the Epilepsy Monitoring Unit (EMU) – an expensive and disruptive process. Our sponsors have developed a multi-modal seizure detection system using a suite of non-cerebral sensors for eventual use in non-clinical settings[5]. Preliminary results show that while the system detects seizures with a high level of sensitivity and specificity, there remains variability between patients. Sensor suite expansion and personalization may optimize results. Currently, the system does not include sensors to monitor head and eye movement, prominent indicators of certain seizure types[6,7,8].

Research has shown the efficacy of using commercial eye-tracking glasses to detect seizures[9], though these devices have high-power consumption and limit movement, which may be incompatible with extended, non-clinical use. Additionally, the high precision afforded by state-of-the-art technology may be unnecessary to detect the eye behavior of interest.

Our sponsors would like to develop novel, low-power hardware to collect eye and head tracking data for eventual incorporation into their sensor suite, should the data prove useful. Our objective is to use a state-of-the-art eye tracker to investigate the separability of typical EMU behaviors and simulated seizure behavior with a low-resolution eye-tracking grid and head-tracking data, provide benchmark tests, and inform specifications for future hardware development.

We developed and ran an experiment to collect eye and head motion data during typical EMU behaviors (technology use, eating, conversation) and simulated seizure behavior with an off-the-shelf commercial eye tracker (Pupil Core) and Inertial Measurement Unit (IMU). Data processing using linear discriminant analysis (LDA) showed high separability between seizure and non-seizure data with features derived from head accelerometry and low-resolution eye position classification (area under the receiver operating characteristic curve was .98 and F1 score was .86 for eye position mapped to a 3x3 grid). The experimental protocol and data processing scripts we have provided allow for easy replication of our methodology to evaluate future hardware against these benchmark results. Additionally, adding artificial noise to Pupil Core data revealed that tracking the pupil within 2.75mm was allowable before separability scores significantly decreased. These results are promising for the use of eye and head motion data to identify certain seizure presentations, without requiring the high precision and power consumption of current technology.

Successful incorporation of eye and head tracking data may improve our sponsor's seizure detection system. The ability to detect seizures outside of the clinic could improve epilepsy treatment efficacy, as well as save patients tens of thousands of dollars in healthcare spending, and up to a week of clinical monitoring[14

TABLE OF CONTENTS

ABSTRACT

Treatment plans for epilepsy depend on seizure frequency, duration, and type. Seizure monitoring in non-clinical settings relies on unreliable self-reports[4]. Our sponsors have developed a suite of non-cerebral sensors which detects seizures with high specificity, though personalization of the suite may optimize results[5]. Eye and head motion, not currently measured in the sensor suite, are indicative of certain seizure types[6,7,8], though state-of-the-art eye tracking devices have high power consumption and may not be suitable for extended monitoring. To inform future development of a low-power prototype for potential addition to the sensor suite, we gathered data from 8 subjects using a commercial eye-tracker (Pupil Core) and Inertial Measurement Unit (IMU), for non-seizure behaviors (technology use, eating, and conversation) and simulated seizure activity. Data processing using linear discriminant analysis (LDA) showed high separability between seizure and non-seizure data with features derived from head accelerometry and low-resolution eye position classification (area under the receiver operating characteristic curve was .98 and F1 score was .86 for eye position mapped to a 3x3 grid). The experimental protocol and data processing scripts we have provided allow for easy replication of our methodology to evaluate future hardware against these benchmark results. Additionally, adding artificial noise to Pupil Core data revealed that tracking the pupil within 2.75 mm was tolerable before separability scores significantly decreased. These results are promising for the use of eye and head motion data to identify certain seizure presentations, without requiring the high precision of current technology.

1. PROJECT DEFINITION and VALUE

1.1 - Problem Statement and Significance

Epilepsy is a broad term used to describe a range of neurological disorders characterized by recurrent, unpredictable seizures as the core symptom[2], which affects over 65 million people worldwide[1]. Epilepsy disorders are highly dynamic and variable, with symptom profiles (frequency, duration, and presentation of seizures) varying from person to person. Diagnosis is obtained through assessment of the patient's symptom profile and medical history, in addition to any supplementary tests that may inform cause (e.g. fMRI, etc)[1]. Thus, accurately monitoring seizure activity is crucial for efficiently arriving at a diagnosis, and equally important in the assessment of treatment plans once medication is prescribed. In the early stages of diagnosis and treatment, doctors rely on patients to self-monitor seizures. Though self-reported seizure logs are notoriously unreliable[3,4], this blunt measure is often sufficient to arrive at an effective treatment plan. However, approximately 1 in 3 patients with epilepsy experience refractory, or treatment-resistant, epilepsy[10]. If seizures remain unmanageable after a year, more extensive evaluations are recommended. This often entails around the clock monitoring of the patient's seizure activity for several days, using combined video and electroencephalography (vEEG) in an Epilepsy Monitoring Unit (EMU)[5], the current gold standard for seizure monitoring. Yet, there are several drawbacks to this method: (1) It is taxing to the patient (invasive, obstructive, exhausting), (2) inaccessible (expensive, requires specialized staff, and only available at approximately 170 locations nationally[10]), and (3) provides a limited view of the patient's natural seizure activity. In fact, it is sometimes necessary to induce seizures during EMU stays by limiting antiepileptic drugs and exposing the patient to seizure-provoking conditions such as sleep deprivation and hyperventilation[6]. Evidently, **doctors need accurate ways to detect seizures in non-clinical settings to improve patient care**.

1.2 - Value Proposition

There are currently no reliable ways to monitor seizures outside of an EMU. Our sponsors have assembled a multi-modal suite of non-cerebral sensors which collects electrocardiography (ECG), electromyography (EMG), electrodermal activity (EDA), accelerometry, and audio signals which shows a high level of sensitivity and specificity in detecting seizure data (.96 average area under the receiver operating characteristic curve (AUROC) for all patients studied[5]). However, variability in results between patients with different seizure types indicates sensor personalization could optimize results while reducing the number of total sensors. Eye and head motion during seizures can be indicative of seizure type[6,7,8] and witness accounts of eye motion are used to inform diagnosis[6]. Neither of these behaviors are captured in our sponsor's system. Commercial eye-trackers have been shown to detect seizures via eye-deviation behavior[9], though high power consumption as a result of high precision may be incompatible with extended non-clinical use. Using Pupil Labs' Pupil Core, an off-the-shelf eye-tracker, we plan to inform future low-power prototype design by investigating low-resolution eye classification grids and providing benchmark tests. This prototype could eventually be incorporated into a system that provides medical staff with accurate day-time seizure activity, which would impact three main aspects of epilepsy care: (1) accessibility, (2) patient comfort, and (3) care efficiency through a reduction in the number of EMU stays and in-person specialist consultations. This is elaborated upon in Section 6.

2. RESEARCH GOALS, SPECIFICATIONS, and DELIVERABLES

2.1 - Research Goals

The goal of our project is to inform future low-power hardware development using a commercial eye-tracking device (Pupil Core) as "ground-truth" hardware – in other words, this device will remove the imprecision and uncertainty in eye data that would be present with a novel prototype . See Appendix A.1 for Pupil Core hardware specifications. This goal can be split into two sub-goals:

(1) *Proof of concept* that non-seizure and seizure data are separable using features extracted from a low-resolution eye-tracking grid. In discussion with our sponsors and considering the biological behavior they hope to capture (eye and head deviation in the “fencing posture”), we planned to classify eye position to 3x3 grid across the field of vision as a starting point. Features will be extracted from the low-resolution space and characterized based on their overall importance to data separation. Experimental protocol and supporting hardware and software will allow for easy replication of our methodology and results with novel hardware, for comparison against the Pupil Core device.

(2) *Inform hardware specifications.* More technically, we want to inform where trade-offs can be made to reduce power consumption This includes characterizing the effect the resolution of the eye-tracking grid has on data separability (starting with 3x3 and exploring more possibilities), and characterizing the effect the of increasing error in pupil position by artificially creating noise in Pupil Core data.

2.2 - Specifications

To determine whether pursuing eye and head tracking is useful for seizure detection, we first constructed an experiment involving 6-10 subjects, lasting 45-60 minutes per subject to gather data from non-seizure and simulated seizure activity (Section 3.3). These specifications were set to maximize total data collected within our given timeframe. To ensure that the experiment can be recreated during future prototype development, we developed written protocols and checklists (Appendix B), such that non-involved proctors can learn to run the experiment in under 1 hour.

To quantify results of our experiment, we used an F1-macro and area under the receiver operating characteristic (AUROC) score. AUROC evaluates the maximum separability of the given data and F1 evaluates the predictive quality of the Linear Discriminant Analysis (LDA) (Section 4.4). We chose thresholds for each score (AUROC > .95, F1 > .85) to determine if a given LDA separated data to an acceptable degree, each threshold considered a promising result for its respective metric[5,20].

Additionally, we plan to investigate the importance of each extracted feature to overall data separation with LDA. We also plan to investigate the separability of data using only head or eye data, to get a sense of the overall importance of each.

Proof of Concept	
Component	**Specification**
Experiment	6-10 subjects
	45 - 60 minutes per subject
	Non-involved proctors can be taught to run the experiment in <1 hour
Separability	F1 > .85 AUROC > .95[5]
Features	Rank by importance
	Eye only vs Head only

Table 1: Proof of Concept Specifications

To inform future hardware design, we have identified specific grid resolutions to investigate as well as a method to identify maximum pupil position error rate.

Informing Hardware	
Hardware Spec.	**Method**
Classifier Size and Shape (mxn)	1x3, 3x1, 2x2, 3x3
	NxN for N>3
Position Error	Add artificial noise to collected Pupil Core data and characterize effect on LDA score

Table 2: Informing Hardware Specifications

2.3 - Deliverables

Three deliverables have been outlined to meet the above goals and specifications.

1. Experimental protocol, code documentation, hardware, and Python script to calibrate and collect data from the hardware
2. Well-documented, hardware-agnostic software pipeline for data processing
3. Written analysis of data processing, benchmark tests and hardware recommendations to cover all specified areas of interested

All software is located at https://github.com/MatterBaby/ENGS90_Software. The top-level README for this repository is located in Appendix C. The Experimental protocol and checklist is provided in Appendix B, and our written analysis is provided in Section 5 (Results) and Section 7 (Discussion) of this report. The written analysis is contained within this report, Section 3 and 4 (Methodology), Section 5 (Results), and Section 7 (Discussion).

3. METHODOLOGY: *Data Collection*

3.1 - Hardware

Eye and head tracking data were collected by the commercial eye-tracking device, Pupil Core, and Arduino compatible Inertial Measurement Unit (IMU) which contains an accelerometer and gyroscope. See exact hardware specifications in Appendix A.

3.2 - Calibration and Classification

The Pupil Core device outputs normalized x and y coordinate values (x position of the pupil in pixels divided by image width, y position of pupil in pixels divided by image height), along with a confidence value between 0 and 1. With a 3x3 grid across the subject's field of view as our initial eye-tracking resolution, we needed a method to map this raw output to one of the nine possible cells.

Given that the Pupil Core output is the relative pupil position within the camera's frame, the data is quite sensitive to the way the device fits to a subject, as well as variations in a subject's eye shape and size. Because of this, it was necessary to calibrate the classification system to each subject individually.

To calibrate our classifier, a monitor was placed in front of the subject such that it took up as much of their field of vision as possible while still ensuring the Pupil Core camera could detect the subject's pupil even with the subject looking to the edges of the monitor. The subject was then directed to look at a series of dots, each of which correspond to the center of a cell in a low-resolution grid, and a centroid point was generated for each grid cell by averaging the samples. Eye position was then simply classified by calculating the closest centroid.

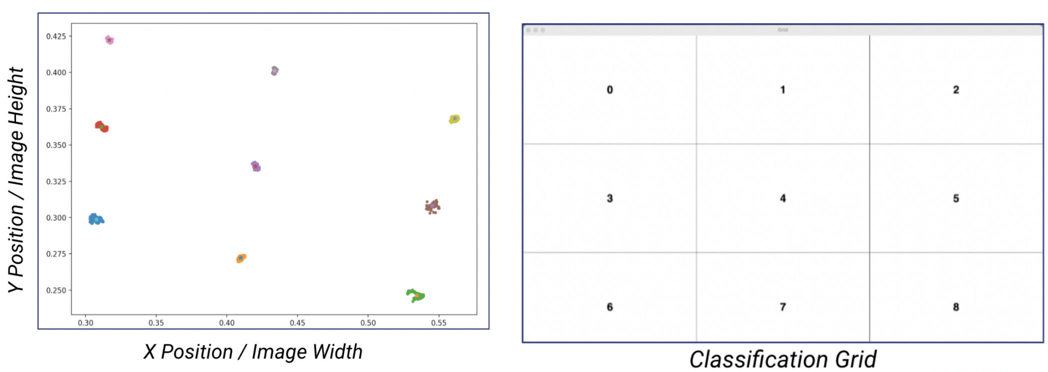


Figure 1: Raw Pupil Core output corresponding to cells on a 3x3 grid

To initially evaluate this methodology for calibration and classification, a subject followed a dot tracing a horizontal and vertical line through the central row and column respectively, colored by classification value.

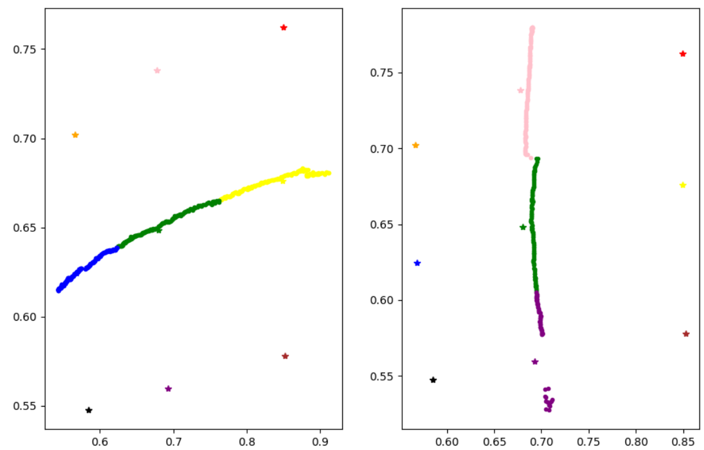


Figure 2: Results of horizontal and vertical line trace

3.3 - Experiment

A total of eight subjects participated in the experiment, for an average total time of 48 minutes. Each subject first calibrated three sets of centroids – 3x3 (mounted), 3x3 (unmounted), 2x2 (mounted). The centroids were plotted and recalibrated on an as-needed basis (ie. if errant eye samples interfere with the integrity of the calibration). In the data collection phase, subjects performed three typical behaviors seen in the EMU for ten minutes each – using technology such as a phone or computer, eating, and holding a conversation. These behaviors were recommended by our sponsor, Dr. Kobylarz, a neurologist at Dartmouth-Hitchcock Medical Center who has extensive experience monitoring patients in the EMU. Finally, the subjects were instructed to hold the "fencing posture" [7,8], head turned to one direction with eyes fixed the same direction, for one minute on each side. In reality, this position may also present with arm extension to the same side as eye and head fixation, though this aspect of the behavior was not recreated for the purposes of our experiment.

Figure 3: Fencing Posture [7]

A Python script (*remote_run.py*) controlled data collection through a Graphical User Interface. Using keystrokes, the investigator can begin or end data recording ('r') which is then saved out to a .csv file in a designated data directory, cancel a current recording ('c') such that no data is written out, or quit the experiment entirely ('q'). While recording, the data was sampled at an average rate of 26.1 Hz, as measured by Python's built-in timer function. Each sample includes the normalized pupil position generated by the Pupil Core (x and y), the confidence value from the Pupil Core, and the accelerometer and gyroscope readings from the x,y and z-axis. Sample eye-tracking output from the experiment is shown in Figure 4 (below).

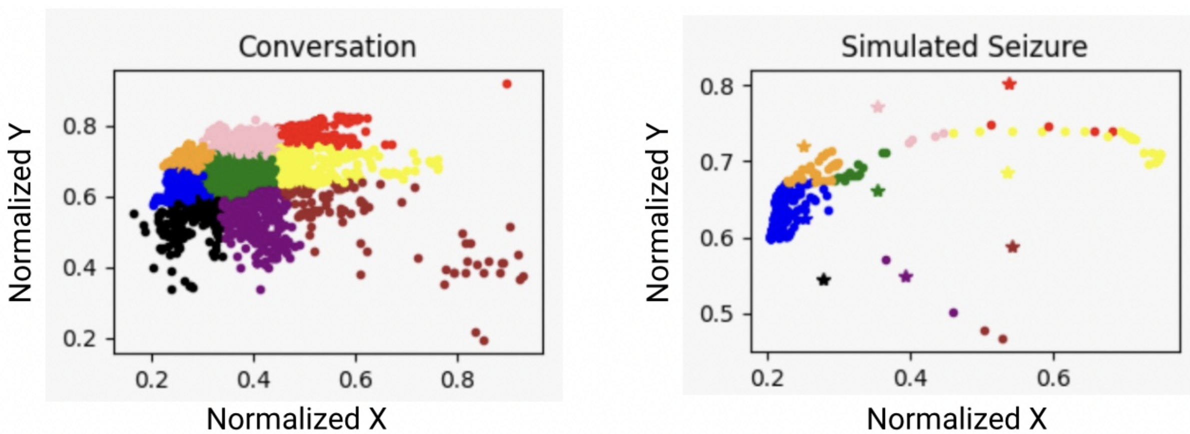


Figure 4: Sample eye-tracking data from the experiment, colored by classification value

See Appendix B for full experimental protocols as well as a checklist to assist non-knowledgeable proctors in setting up and administering the test. We tested this protocol by having a subject and proctor, unfamiliar with the project, attempt to conduct the experiment from beginning to end. This process took just under an hour. Their feedback was used to modify and improve the protocol.

3.4 - Validation

Given the nature of the grid classification system, accuracy of any single sample's classified value will be affected by how close the point is to the edge of the grid. To quantitatively judge the accuracy of the classifier, each subject underwent a validation procedure following 3x3 mounted calibration. During validation, the subject follows a moving dot on-screen. The known grid number of the moving dot is compared with the grid number output by the classifier. An accuracy score is then calculated from the comparison. Across the eight subjects, there was an average 92% classifier accuracy score. Assuming the Pupil Core is ground truth, this score can be used as a benchmark to compare against a future hardware's classifier accuracy.

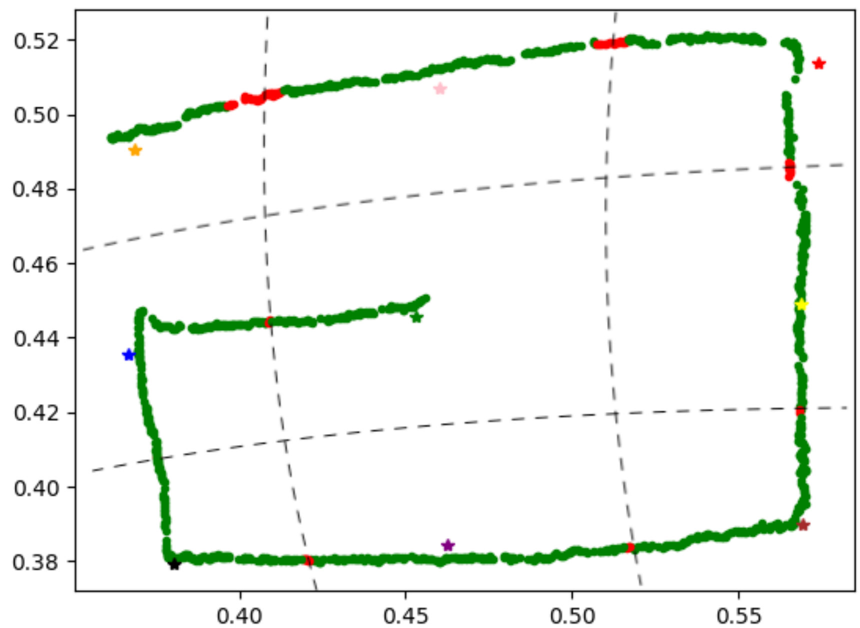


Figure 5: Example of validation output. Green is correctly classified samples. The incorrectly classified samples are mainly at the borders between grids, and may be due to the subject not following the center of the moving dot perfectly, or the pupil being obscured by the eyelid and/or eyelashes.

4. METHODOLOGY: *Data Processing*

4.1 - Hardware Agnostic Pipeline

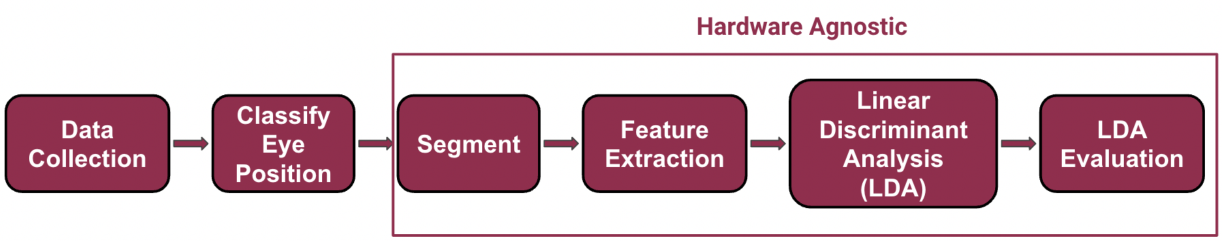


Figure 6: Overview of Data Processing Pipeline

Once data is collected (*remote_run.py*) and classified with a given centroid file (*process_raw.py*), it is saved to a hardware agnostic .csv file of the form: "Eye position classification, Acceleration_X, Acceleration_Y, Acceleration_Z, Gyroscope_X, Gyroscope_Y, Gyroscope_Z", for each row. A .csv of this format, generated by any piece of hardware, can enter the hardware-agnostic portion of pipeline (*analysis.py*) which includes segmenting into epochs (4.2), feature extraction (4.4), and LDA (4.5). Section 3.2 includes hardware-specific data processing undertaken for this project, but the methodology described may provide insight for future hardware development should similar issues arise.

4.2 - Segmenting

The collected data was first segmented into 5 second "epochs" with 50% overlap (each epoch overlapping the previous epoch by 2.5 seconds), the methodology used by our sponsors for examining other non-cerebral sensors in seizure detection[5]. As the Python script sampled at a rate of ~26Hz, 130 samples were included in each epoch with a 65 sample overlap between epochs.

4.3 - Handling Sparse Data

As we began to collect data, we noticed that many samples from the Pupil Core were returned as "nan." While we expected some samples to fail due to extreme eye position or eyelid blocking the camera from detecting the pupil, our data was sparse enough so that some epochs contained very minimal or no eye data altogether.

Overall, 37.8% of samples contained no eye data information. To investigate the effect of these samples on overall classifier performance, we created a tolerable failed sample threshold for each epoch, such that the epoch was not included in our analysis if the percentage of failed samples within the epoch exceeded the threshold. We swept this parameter from 1% to 100%, evaluating the classifier at each threshold (see Section 4.4).

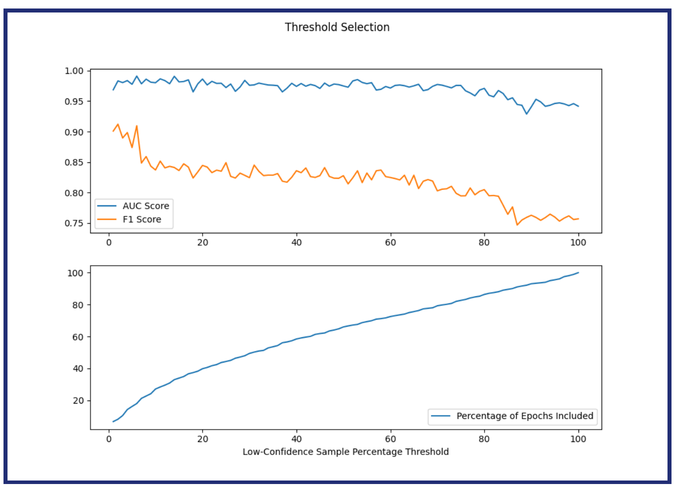


Figure 7: Effect of failed sample threshold on classifier performance

These results showed that increasing the failed sample threshold generally decreased the classifier performance, as there was less information to use to classify samples. While failed samples may contain information about an extreme pupil position, epochs with too many failed samples were too sparse for accurate classification.

Upon closer examination of the collected data, two distinct cases were noticed: failed samples interspersed within dense successful samples and long stretches of failed samples. Techniques to handle such cases were researched.

0.715437998989738	0.7694042690988004	1.0
0.7198695134517874	0.7697885496885695	0.9855856597423553
0.7270250507608202	0.7780162247753403	0.9776228070259094
0.7323823061154834	0.7829779077241246	0.9587594568729401
0.7320353127395732	0.7832660308391037	0.9661943316459656
0.7302023233163455	0.784177681674129	1.0
0.7264364832294125	0.7871625556948384	1.0
0.7242585906515095	0.7866840877815369	0.912664145231247
0.7256359976281099	0.7846414171060546	0.988778293132782
nan	nan	nan
nan	nan	nan
0.42467542916428147	0.7527855468448205	0.9673289656639099
0.4174063067017716	0.7465303964546284	1.0
0.4306038273285025	0.7435335707489761	1.0
0.444825846207219	0.7403747864367963	1.0
0.4556484817082509	0.7380562656871017	1.0

nan	nan	nan
nan	nan	nan
nan	nan	nan
0.659691529497033	0.5095616606498665	0.7070621848106384
nan	nan	nan
nan	nan	nan
nan	nan	nan
nan	nan	nan
nan	nan	nan
nan	nan	nan
nan	nan	nan
nan	nan	nan
nan	nan	nan
nan	nan	nan
nan	nan	nan
nan	nan	nan

Figure 8: Different cases of failed sample stretches. Columns are normalized x, normalized y, confidence

A simple data imputation algorithm algorithm was implemented using Pandas Python library (Appendix F) – fill k-consecutive failed samples with values from the nearest successful sample. This algorithm was run for k=1,5,10,25,100,Max (full results in Appendix D.1). Figure 9 (below) shows the results of k=5 against k=Max.

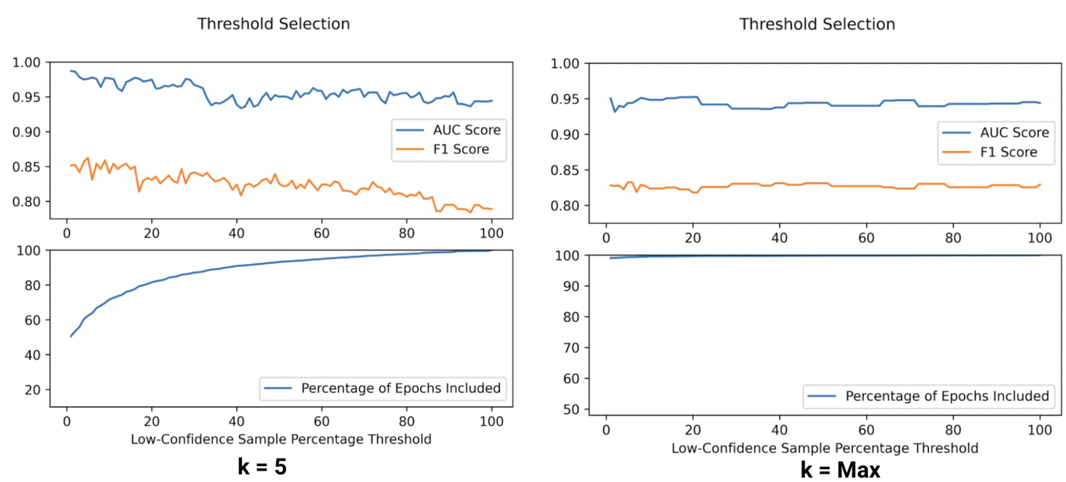


Figure 9: k-Fill Algorithm for k=5 and k=Max

The results of data imputation showed that filling a limited amount of failed samples resulted in the optimal results, such that small gaps in eye information were filled and used for analysis but long stretches of failed samples were discarded. 5-Fill left 11.8% of total samples failed, while Max-Fill left .2% of samples failed. 5-Fill with a failed sample threshold of 5% was chosen as a baseline for the remaining data analysis given the increased classifier scores while still retaining a high percentage of total epochs (62.4% of total epochs remained).

4.4 - Feature Extraction

The following set of features were extracted from each epoch:

Sensor	Feature
Pupil Core (Eye tracking)	Average Eye Classification
	Number of Visited Cells
	Number of Cell Changes
	Number of Failed Samples
	% Samples in the Top/Bottom Row of Grid
	% Samples in the Left/Right Column of Grid
Accelerometer (Head tracking)	Average magnitude
	Standard deviation of magnitude
	Maximum magnitude
	Minimum magnitude
	Range of magnitude
Gyroscope (Head tracking)	Average magnitude
	Standard deviation of magnitude
	Maximum magnitude
	Minimum magnitude
	Range of magnitude

Table 3: Extracted features

We found no literature regarding feature extraction for low-resolution eye-tracking and had to invent our own features. Average eye classification was aimed to capture relative eye position in the epoch, number of visited cells was to capture range of eye movement, number of cell changes was to capture quickly moving eye behavior, failed samples captured eyes in extreme positions, beyond the view of Pupil Core. The % Samples in the Top/Bottom Row and % Samples in the Left/Right Column features were devised to capture bilateral eye deviation behavior. Head tracking features were a subset of the accelerometer features used by our sponsors[5].

Extracting these features turns each epoch into a point in 16-dimensional feature space. A total of n epochs can be stacked into an n by 16 matrix for separation with LDA.

4.5 - Classification and Evaluation

Following the methodology of previous work on this project[5,11], LDA was implemented on the feature matrix to determine separability of seizure epochs from non-seizure epochs. LDA identifies a linear combination of features to project *n* dimensional points onto a one-dimensional axis of maximal separation, such that a threshold on the axis exists to best divide one class from another[21]. LDA was implemented using the Python scikit-learn library (Appendix F).

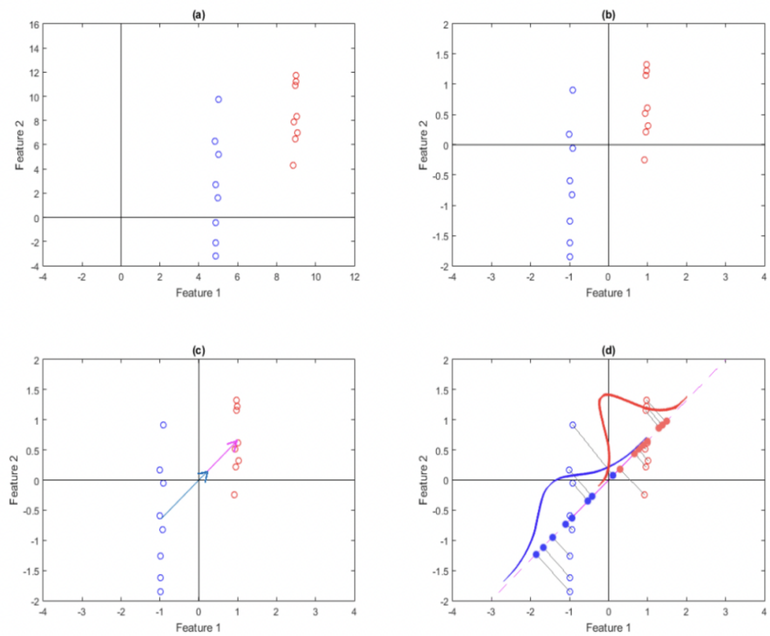


Figure 10 - LDA in 2-Dimensional Space[11]

Previous work scored LDA with area under the receiver operating characteristic curve (AUROC) [5,11] to quantify maximal separability of two classes, where the ROC curve plots 1 - specificity (FP / FP + TN) on the x axis and sensitivity (TP / TP + FN).

To assess the predictive nature of the LDA (how well a calculated axis of maximal separation classifies new data), we also used an F1-macro score to evaluate each LDA. An F1 score is the harmonic average of precision (TP/TP+FP) and sensitivity (TP/TP+FN), with an F1-macro score averaged across classes. See Appendix D.2 for more supplementary figures of these scoring metrics. To calculate the F1 score, the data was partitioned into a 70/30 training/testing split, where 70% of the data was used in LDA and the resulting axis of maximal separation was used to predict classification values for the 30% testing data and generate an F1-macro score. F1-macro scores were averaged across 5-Fold Cross Validation (the training/testing split was performed randomly five times).

5. RESULTS

5.1 - Baseline Classification Evaluation

Using 5-Fill data imputation, LDA was performed for each subject individually and for all subjects together, separating seizure epochs from technology, eating, and conversation epochs and seizure epochs from all non-seizure epochs.

Subject #	*Technology*	*Eating*	*Conversation*	**Non-Seizure**
1	1.0/1.0	1.0/.98	1.0/.98	1.0/.96
2	1.0/.97	1.0/.98	1.0/.94	.99/.94
3	1.0/1.0	1.0/.98	.99/.89	1.0/.90
4	1.0/1.0	--	1.0/1.0	1.0/1.0
5	--	--	--	--
6	1.0/.97	1.0/.98	1.0/.93	.96/.80
7	1.0/1.0	.98/1.0	.98/.84	1.0/.78
8	.98/.78	1.0/.94	.92/.82	.96/.77
All	.99/.93	.99/.93	.95/.85	.98/.860

Table 4: LDA Scores for All Behaviors for Individual Subjects and Total (AUROC/F1)

Subject 5 had no seizure epochs and Subject 4 had no eating epochs after processing, so scores could not be generated.

The LDA for non-seizure behavior versus seizure behavior for all subjects exceeds our feasibility threshold specifications (Section 2.2), showing high separability (AUROC of .98) and predictability (F1 of .86). The average AUROC and F1 scores for seizure vs non-seizure LDA for all patients was .99 and .88 respectively.

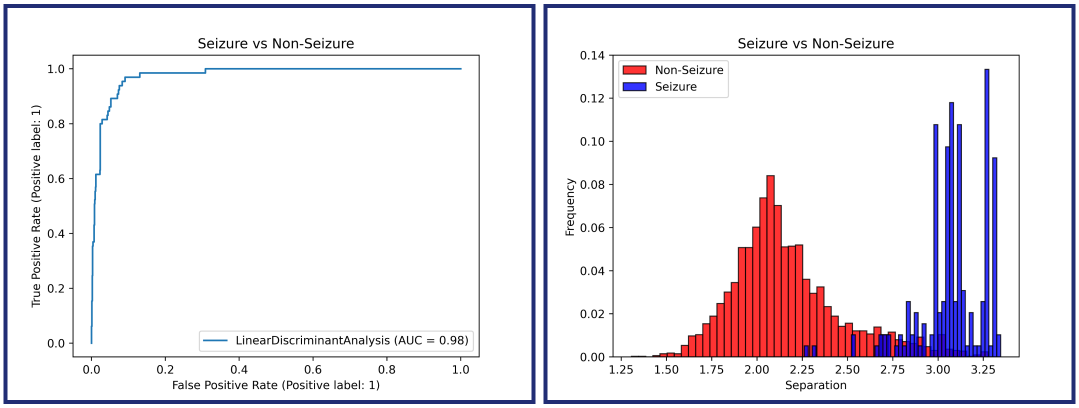


Figure 11: ROC Curve and LDA Visualization for Seizure vs Non-Seizure Behavior for all Subjects

While not all subjects exceed the set F1 threshold, they all exceed the AUROC threshold, showing high separability. See Appendix E for supplementary result figures, including drop-one-subject analysis and analysis without data imputation.

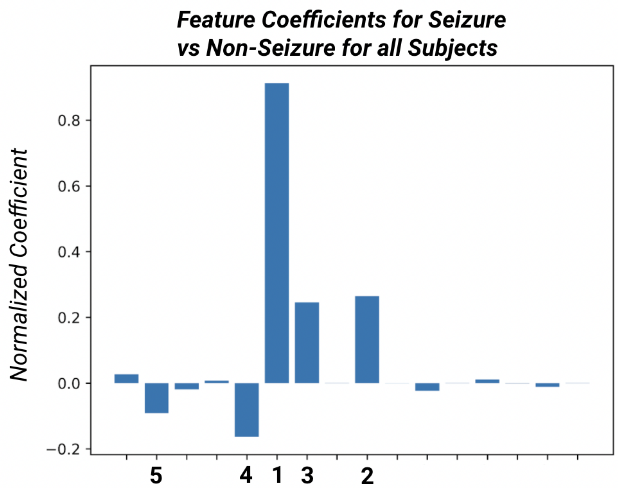
5.2 - Feature Analysis

LDA generates a weight vector *w*, containing coefficients corresponding to each feature for use in mapping each epoch point to the axis of maximal separation. Normalizing *w* to unit vector *u* allows us to rank each feature by separation importance. The top five features for seizure vs non-seizure LDA for subjects individually and all subjects together are collated in Table 5, with blue cells indicating eye tracking features and red cells indicating head tracking features.

	Sub. 1	*Sub. 2*	*Sub. 3*	*Sub. 4*	*Sub. 5*	*Sub. 6*	*Sub. 7*	*Sub. 8*	**All**
1	%L/R Column	Avg Accel.	%L/R Column	Avg Accel.	--	Std Accel.	Avg Accel.	Std Accel.	**%L/R Column**
2	Avg Accel.	%L/R Column	Avg Accel.	Std Accel.	--	Avg Accel.	%L/R Column	Avg Accel.	**Std Accel.**
3	%T/B Row	Std Accel.	Std Accel.	%L/R Column	--	%L/R Column	%T/B Row	Max Accel.	**Avg Accel.**
4	#Visited	%T/B Row	%T/B Row	%T/B Row	--	%T/B Row	Std Accel.	%T/B Row	**%T/B Row**
5	Std Accel.	Max Accel.	Min Accel.	Min Accel.	--	Max Accel.	#Visited	%L/R Column	**#Visited**

Table 5: Top Five Features for Seizure vs Non-Seizure Behavior

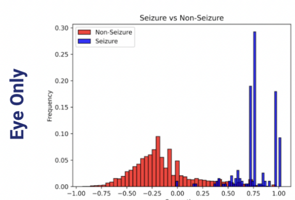
Of the 16 total features extracted, 7 unique features comprised the entirety of the top five features for individual subjects and all subjects together. Each top five feature list contains four of the same features, %Left/Right Column, %Top/Bottom Row, Standard Deviation of Accelerometer Magnitude, and Average Accelerometer Magnitude, though in different orders.



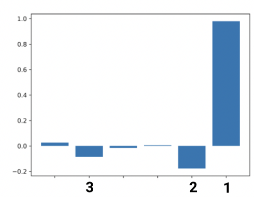
	All
1	%L/R Column
2	Std Accel.
3	Avg Accel.
4	%T/B Row
5	#Visited

Figure 12: Normalized Feature Coefficients for All Subjects

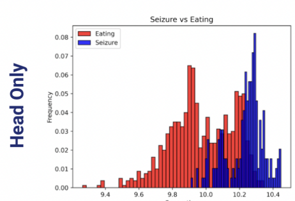
The normalized feature coefficients for all subjects were visualized with a bar chart to gain a sense of scale, given that the top five features are not linearly spaced in terms of importance.



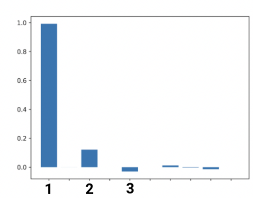
AUC	.97
F1	.858



	Feature
1	%L/R Column
2	%T/B Row
3	# Visited



AUC	.83
F1	.49



	Feature
1	Avg Accel.
2	Std Accel.
3	Max Accel.

Figure 13: Seizure vs Non-Seizure LDA for All Subjects with Only Eye or Head Features

LDA was also performed with only eye or head features for seizure vs non-seizure for all subjects. LDA with only eye features performed similarly to LDA for all features, while only head features performed significantly worse. This indicates that head tracking features may not be useful for detecting this presentation of seizure activity, or that we have not extracted the proper features.

5.3 - Grid Resolution and Shape

Given the success of the 3x3 grid in meeting the separation and prediction requirements laid out in our Section 2.1, smaller classification grids were explored: 2x2, 3x1, and 1x3. 2x2 centroids were calibrated during our experiment and the corresponding .csv file was used to classify eye positions. For 3x1 grid, the center row of centroids from the calibrated 3x3 grid were used for classification, and for the 1x3 grid, the center column of centroids were used. Eye tracking features were altered slightly for these grids, as the percentage of samples in the top/bottom row or left/right column don't translate neatly. For the 2x2 grid, the percentage of samples in each individual grid were extracted, the 3x1 grid used percent of samples classified to the left and right sides, and the 1x3 grid used percent of samples classified to the top and bottom.

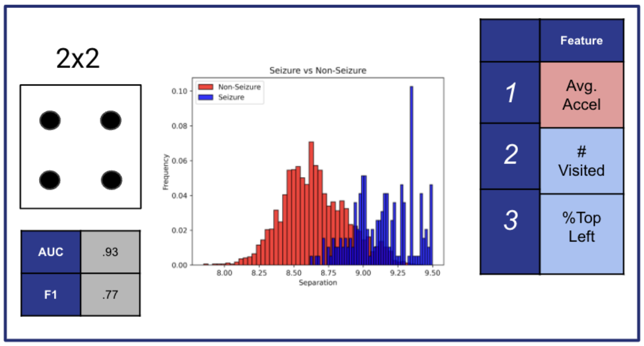


Figure 14: LDA Results for 2x2 Grid

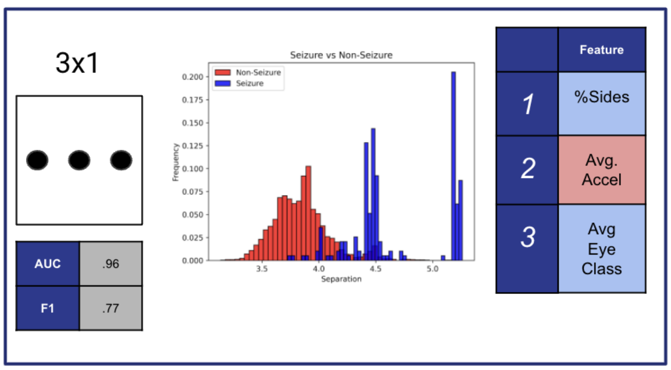


Figure 15: LDA Results for 3x1 Grid

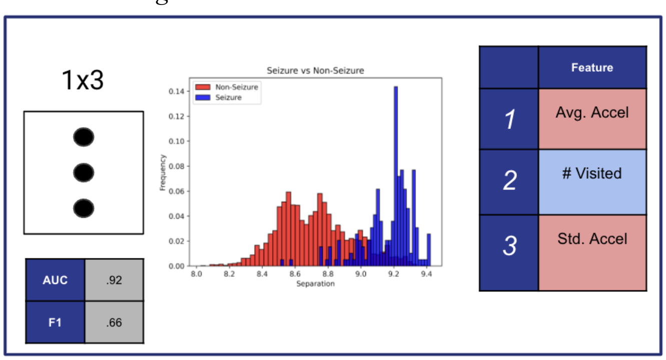


Figure 16: LDA Results for 1x3 Grid

None of these grids met both the F1 and AUROC score specifications. The 3x1 grid met the AUROC specification, intuitively performing the best out of all the options given that it used the closest approximation to the % Left/Right Column feature with % Sides.

Given that current low-power eye tracking research prototypes have been shown to approximate eye position to within one millimeter[12,13], it is likely that a grid with a higher resolution than 3x3 is achievable. To investigate the effect a higher resolution grid may have on LDA scores, an algorithm was developed to place NxN evenly spaced centroids using the calibrated 3x3 grid as a starting place. Figure 17 (below) shows the calibrated 3x3 centroids (green stars) and extrapolated NxN centroids (purple stars).

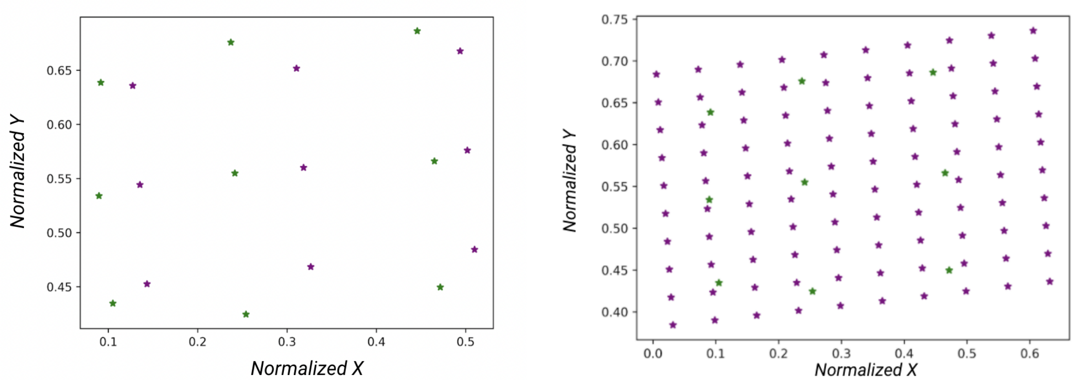


Figure 17: Extrapolated 3x3 Centroids (left), 10x10 Centroids (Right)

Centroids were extrapolated for a grid of resolution 2x2 through 15x15. F1 and AUROC scores were generated for LDA using eye data classified to the new centroid grid with and without supporting head features.

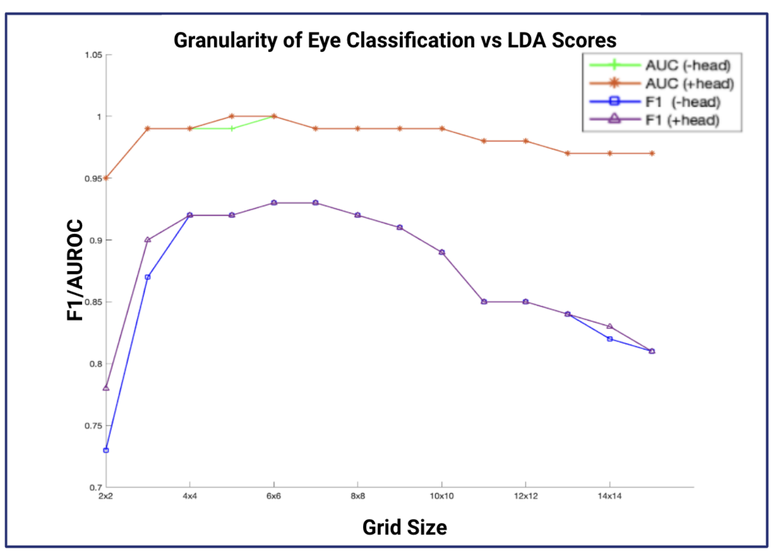


Figure 18: Grid Resolution Effect on F1 and AUROC Score for LDA with and without Head Features

The results of this analysis show that increasing eye classification resolution can increase the separability and predictive quality of the LDA, to a point. After maximal scores with a 6x6 grid (F1 .93, AUROC 1.0), scores begin to gradually taper as resolution is increased. From the feature analysis (Section 5.2), one of the top eye features was the percentage of samples in the left/right column of the grid. These results indicate that there is an ideal perimeter size with which to capture eye deviation behavior.

5.4 - Noise

While testing subjects in our experiment, we noticed that occasionally the pupil would be obscured by eyelashes or the eyelid. In this state, the Pupil Core hardware was not able to accurately track the eye. Figures 19 and 20 show the difference in tracking between a subject with an unobscured pupil and a subject that is squinting. The associated validation scores are paired with each eye image. Additionally, as the subject looks further down, more of the pupil is obscured by the eyelid, which is why the bottom half of the validation for the squinting case is noticeably worse.

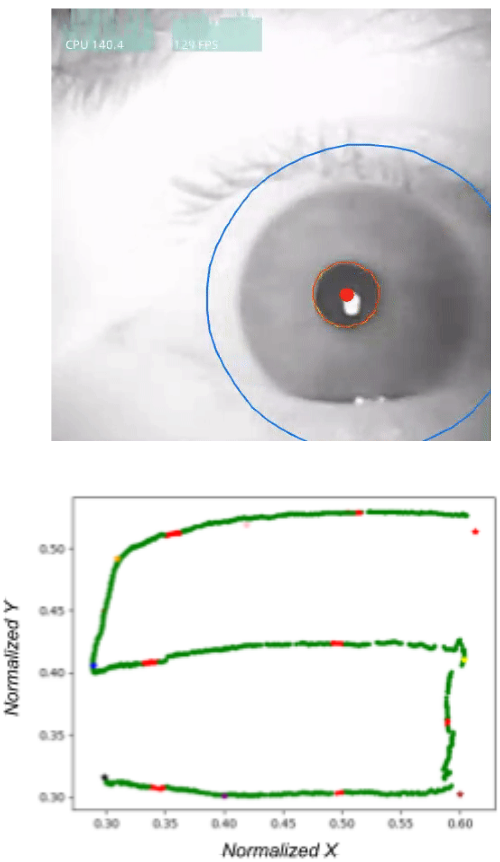


Figure 19: Ideal pupil tracking is 94% accurate

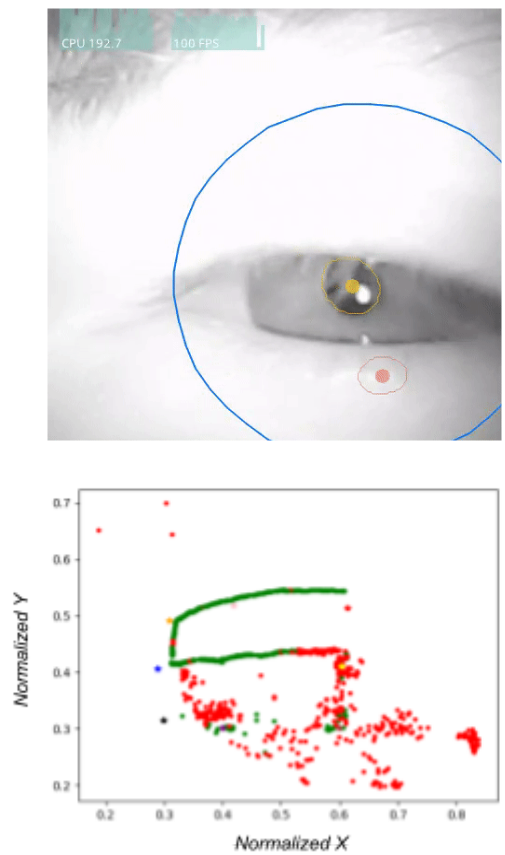


Figure 20: Obscured pupil tracking is 62% accurate

We replicated this noise artificially. By adding varying levels of noise and seeing how F1 and AUROC scores respond, we could find the required level of accuracy for future hardware. This method assumes that the Pupil Core is ground truth, assuming that the pupil is not overly obscured. Evidence from the unobscured pupil validation score (Figure 19) and the line trace tests (Figure 2) give us reason to make this assumption.

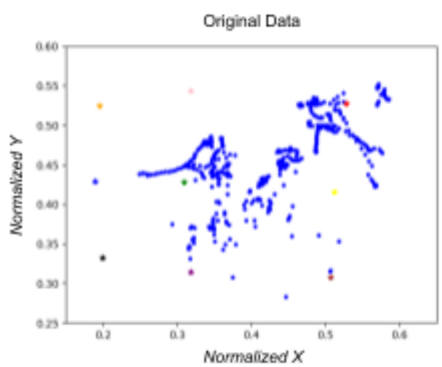


Figure 21: Sample of original data (blue)

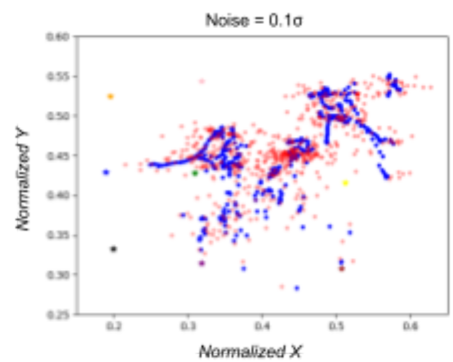


Figure 22: Original data with noisy data (red) overlaid

Noise was added by first finding the X and Y standard deviations of the original data. These standard deviations, multiplied by a noise scalar that we controlled, were used to create two different standard normal distributions. Values were randomly pulled from each distribution to get the X and Y displacements for each individual point. The reason two different distributions for X and Y were created is because some data was more spread out horizontally rather than vertically, so it was necessary for the noise to reflect that as well.

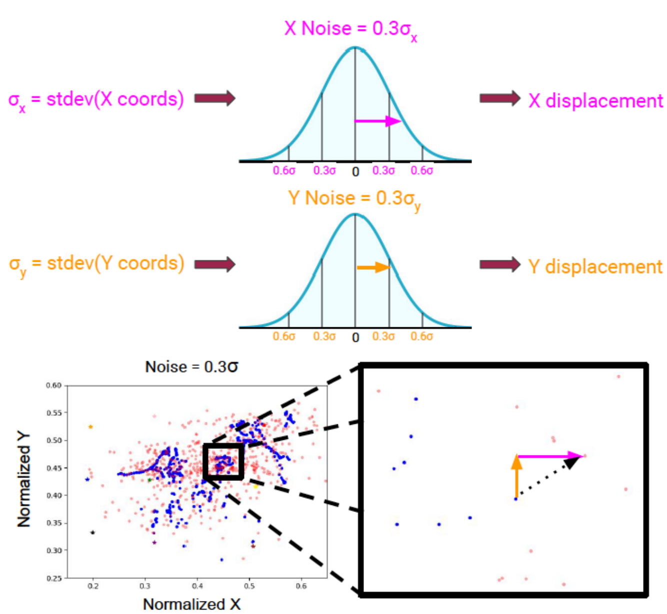


Figure 23: Noise addition technique

5.5 - Future Hardware Accuracy

We ran an LDA on each subject's data with increasing noise levels (see Appendix G), and on all of the subjects' data combined. Overall, we noticed that as noise increased, F1 and AUROC scores decreased. Percent accuracy was determined by comparing the grid position for each sample for the original data with the grid position for each sample for the noisy data. As noise increased, the percent accuracy leveled out at around 50% accurate.

Figure 24 shows the F1 and AUROC scores for all subjects' data as noise increases. Using this information, we recommend that future hardware be able to track the center of the pupil within 2.75 mm. This corresponds to one standard deviation of noise, after converting from normalized units into millimeters. The reason we chose this was because at one standard deviation of noise, we recorded the largest F1 and AUROC scores (0.925 and 0.993 respectively) before we noticed a decline.

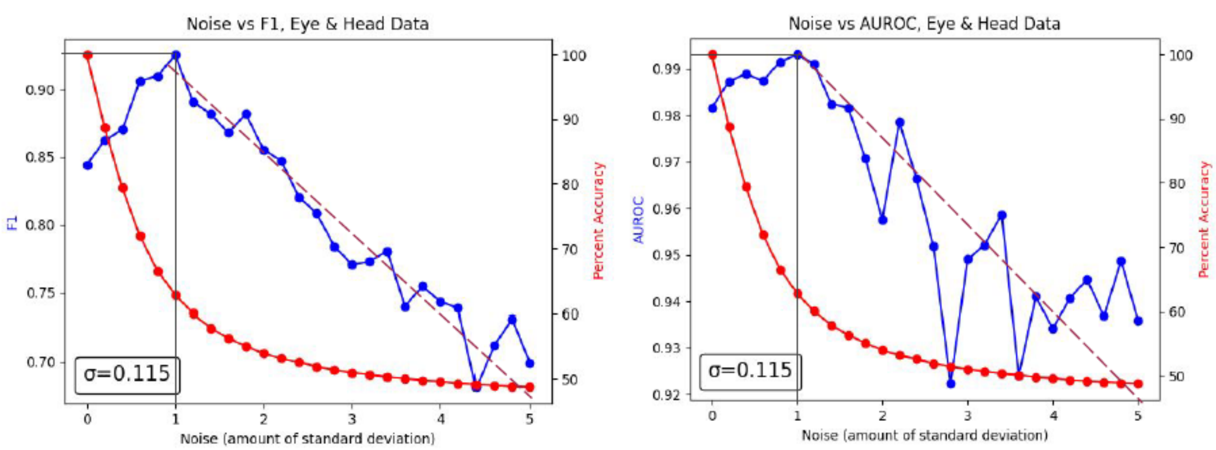


Figure 24: F1 & AUROC vs noise for all subjects' data.The linear portion is defined by the purple dotted line. One standard deviation of noise is identified in each plot.

While tracking the pupil within 2.75 mm is recommended, the maximum allowed threshold is 5.5 mm. This corresponds to 2 standard deviations of noise and is the point at which the F1 and AUROC scores are just above specification values (AUROC > .95, F1 > .85).

Linear Portion	
F1	AUROC
-0.055/σ	-0.014/σ
-0.020/mm	-0.005/mm

Table 6: Slopes of F1 and AUROC vs noise, both in per standard deviation (normalized coordinates) and millimeter error. These slopes correspond to the purple dotted line in Figure 24.

5.6 - Feature Importance vs Noise

In order to better understand what features future hardware might find most useful, the coefficients of eye and head tracking features were recorded as noise increased. While most feature coefficients stayed relatively constant, we noticed that the percentage of samples in the left and right columns and top and bottom rows inverted roughly. We believe this is due to the fact that, while seizure data is mostly concentrated in the central row, the non-seizure data is more spread out among the grids. For the same level of added noise, non-seizure data is being pushed more into the top and bottom rows, which is why we believe the LDA relies more on the top and bottom row feature as noise increases.

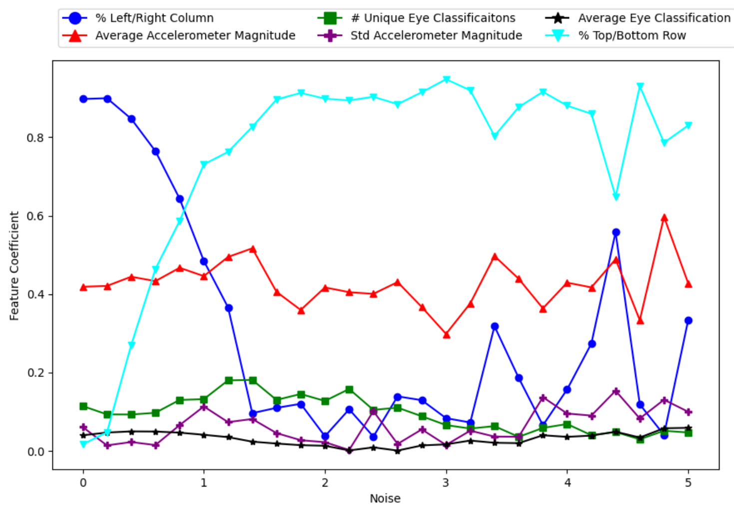


Figure 25: Feature coefficients vs noise

5.7 - Unmounted Calibration

Given the careful precautions needed when performing studies with epileptic patients, our sponsor wanted to avoid future hardware requiring full head stabilization during the calibration process. During our experiment, subjects calibrated a second set of 3x3 centroids without their head fixed to the mount. These centroids were used to classify eye position and LDA was performed on this data.

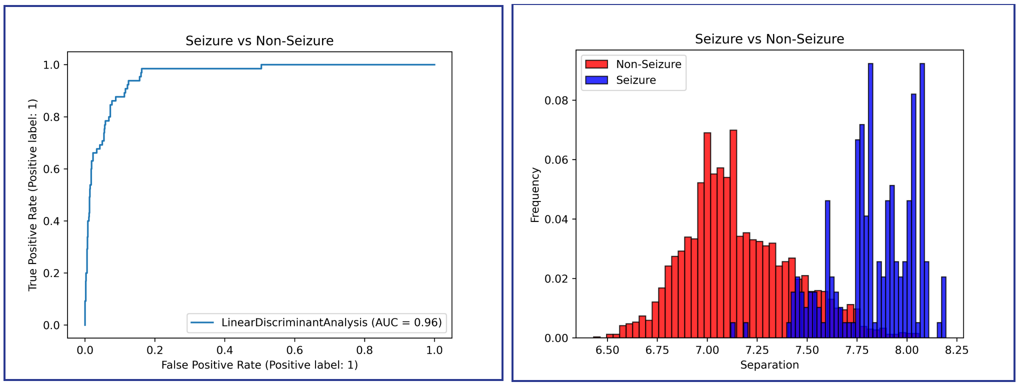


Figure 26: LDA for Seizure vs Non-Seizure for all subjects with unmounted centroids

The LDA yielded an F1 score of .83 and AUROC score of .96, lower than the scores using centroids from mounted calibration.

5.8 - Additional Feature Extraction

Additional features were extracted from the raw Pupil Core output and used for LDA. These features were: average normalized x, average normalized y, maximum, minimum and range of values for normalized x and normalized y, average distance between all eye positions, average distance between adjacent eye positions, average confidence value, standard deviation of confidence value. LDA was performed with and without inclusion of features generated from the lower resolution eye-tracking grid.

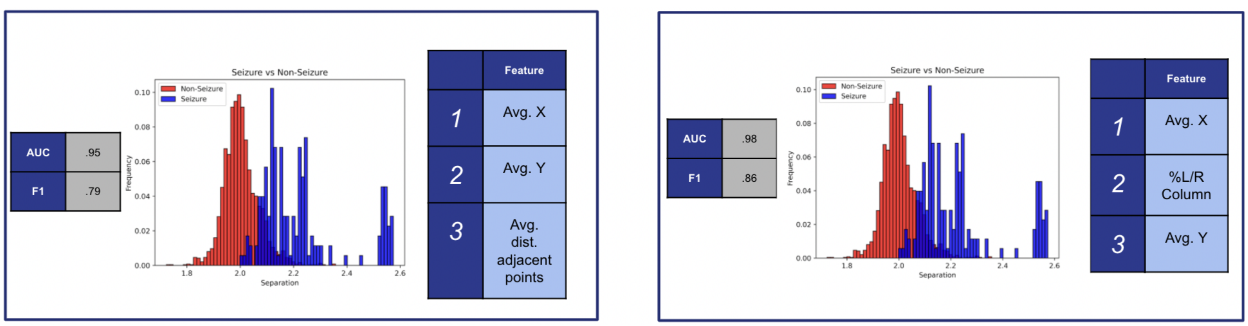


Figure 27: Additional Feature Extraction LDA Results with (right) and without (left) Low-Resolution Features

LDA shows significantly worse results when low-resolution eye-tracking features (4.3) from a 3x3 grid are not included. When included, there is no improvement in scores from only 3x3 features. This indicates that not only are low-resolution features an acceptable way to track eye position, they may be a very useful way to do so.

5.9 - Imbalanced Data

A further consideration was the imbalance in the dataset. Using 5-Fill data imputation (Section 4.2), 5.4% of all epochs were seizure epochs. Two approaches were taken to characterize the effect of balancing the dataset on LDA scores: naive undersampling, and

Synthetic Minority Oversampling Technique (SMOTE). Naive undersampling simply took a random subset of all non-seizure epochs, whereas oversampling employed SMOTE, a more sophisticated algorithm which artificially fills in gaps within the minority class (See Appendix D.3). SMOTE was implemented using the imbalanced-learn Python package (See Appendix F).

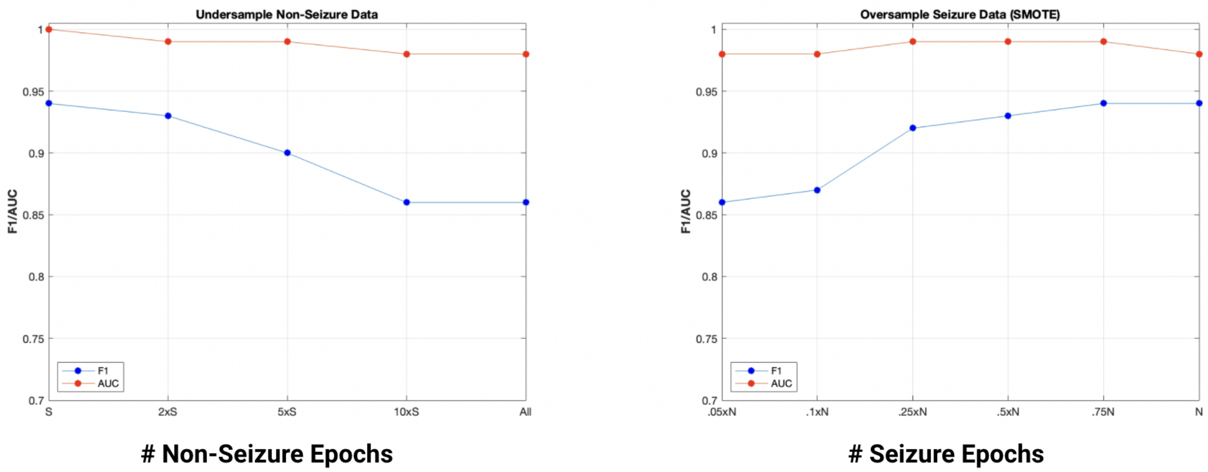


Figure 28: Effect of Undersampling (left) and Oversampling (right) on LDA scores

In the above figure, S is equivalent to total number seizure epochs, and N is equivalent to the number of non-seizure epochs. Scores improved when seizure and non-seizure epochs moved toward equalization, either by undersampling or oversampling.

6. Societal and Economic Analysis

Our project is part of a larger effort to reliably monitor seizure activity beyond the hospital. If successful, this could lead to improved diagnosis and treatment plans while limiting EMU stays and specialist consultations. This could decrease the time and cost burden of current epilepsy treatment, as well as the inherent inequity in healthcare accessibility.

We researched standard epilepsy protocols and their associated cost, from which it was evident that epileptic patients incur a larger healthcare cost burden than the typical American. The estimated costs for an epileptic patient's first year of treatment is $20,084[10] compared to an average yearly healthcare expenditure of $10,739[14] for the general population. These additional costs come from different levels of epilepsy care as detailed in Figure 29.

Level 1 Care	Level 2 Care	Level 3/4 Care
❖ Emergency room visits ❖ Epilepsy evaluation with primary care physician ❖ **Cost**: ➢ $2,400 [17]	❖ Consultation with neurologist ❖ Visits to a specialized epilepsy center ❖ **Cost**: ➢ $500 per consult ➢ Transportation Costs	❖ Stays in the EMU ❖ **Cost**: ➢ **$35-40K per EMU stay[18]** ➢ Transportation Costs ➢ Loss of work

Figure 29: Different levels of epilepsy care with associated costs

In addition to the cost of EMU stays, there are additional costs patients must incur due to transportation – which may include airfare, as there are only 170 epilepsy treatment centers nationally[10] – and loss of work – the average stay in the EMU is between 4-7 days[15]. Reducing the number of EMU stays could greatly ease the cost burden for epileptic patients. In addition, accurate, long-term seizure monitoring could expedite diagnosis and improve treatment efficacy, resulting in reductions in Level 1 and Level 2 care.

This solution could also have implications for racial disparities in healthcare. Not only is the lifetime prevalence of epilepsy – the number of patients who develop the condition over their lives – higher in Black populations, but Black and Hispanic patients are less likely to have private health insurance to cover the costs of treatment[16].

7. Discussion

The results of our research show that non-seizure behavior is highly separable from simulated seizure activity using low-resolution eye-tracking data combined with accelerometry based head-tracking data, as measured by the AUROC and F1 scores (.98 / .86) from LDA.

The next step for this project is to design and build a low-power prototype to collect eye and head tracking data for eventual incorporation into our sponsors' sensor suite. Our experimental protocol (Appendix B), methodology (Section 3 and 4) and hardware-agnostic pipeline (Appendix C) provide ways to evaluate data collected from prototypes against the "ground-truth" data from the Pupil Core device. Additionally, we have investigated the effect of eye-tracking grid resolution and error-rate in pupil position and arrived at the following recommendations.

(1) Given our sponsors' desire to reduce power-consumption and complexity of a future prototype, we recommend a **3x3 grid** for eye-tracking resolution, as this is the lowest resolution

our separability specifications are met. Should future hardware fail to match the same scores, or should future research indicate that detection of more subtle eye behaviors is necessary, we recommend increasing resolution up to a 6x6 grid to improve results.

(2) While our research indicates that an error rate within 5.5 mm still meets our separability requirements, we recommend tracking the pupil within **2.75mm**, as beyond that error range the quality of separation begins to drop off. We believe that this specification is achievable given the results of low-power research prototypes (iShadow and Battery-Free Eye Tracker)[12,13], which show the promise of using low-power NIR LED emitter and photodiode pairs to place pupil position within one millimeter.

(3) Calibrating the classification centroids without the subject's head mounted to chin rest caused separability results to decrease slightly. Future hardware development should carefully consider how best to calibrate the device and weigh the risks of fixing an epileptic patient's head to a mount.

We also believe there is room to develop head-tracking feature extraction techniques, such as more robust pre-processing and filtering to reduce noise, as well as to extract frequency domain features, such as the mean Fourier Transform and spectral centroid[5]. Though eye-tracking features alone separated data almost as well as eye and head tracking features together, we recommend investigating these features and considering other possible seizure presentations (such as head jerking seen in myoclonic seizures[22]) before omitting head-tracking sensors in future hardware design.

Finally, we believe future work should verify these results with real seizure data, and once more data is collected, consider using heavier-weight machine learning algorithms and data processing techniques to further improve results.

Due to the encouraging results for efficacy of eye and head tracking data to separate certain seizure presentations from non-seizure activity, and the manageable technical specifications identified to achieve these results, we are excited to state our full recommendation to further pursue the development of a novel low-power prototype for eye and head tracking.

References

1. Institute of Medicine (US) Committee on the Public Health Dimensions of the Epilepsies, M. J. England, C. T. Liverman, A. M. Schultz, L. M. Strawbridge, editors, “Introduction,” in *Epilepsy Across the Spectrum*, *1*st ed. Washington D.C., (Washington D.C.), US: National Academies Press 540 (US), 2012, ch. 1, sec. 2, pp. *27*.

2. M. M. Zack, R. Kobau, “National and state estimates of the numbers of adults and children with active epilepsy – United States, 2015,” MMWR Morb Mortal Wkly Rep 2017, vol. 66, no. 31, pp. 821-825, Aug, 2017. DOI: 10.15585/mmwr.mm6631a1, [Online].

3. Burneo J. G. (2008). The real truth behind seizure count. Epilepsy currents, 8(4), 92–93. https://doi.org/10.1111/j.1535-7511.2008.00251.x

4. C. Hoppe, A. Poepel, C. E. Elger, “Epilepsy: accuracy of patient seizure counts,” *Epilepsia*, vol. 64, no. *11*, pp. 1595-1599, DOI: 10.1001/archneur/64.11.1595, [Online].

5. A. Hamlin, E. Kobylarz, J.H. Lever, S. Taylor, L. Ray, Assessing the Feasibility of Detecting Epileptic Seizures using Non-Cerebral Sensor Data, Computers in Biology and Medicine, https://doi.org/10.1016/j.compbiomed.2021.104232

6. Elger, C. E., & Schmidt, D. (2008). Modern management of epilepsy: a practical approach. Epilepsy & Behavior, 12(4), 501-539.

7. Zhang, W., Liu, X., Zuo, L. *et al.* Ipsiversive ictal eye deviation in inferioposterior temporal lobe epilepsy—Two SEEG cases report. *BMC Neurol* 17, 38 (2017). https://doi.org/10.1186/s12883-017-0811-8

8. Eliashiv, Dawn. “Frontal Lobe Seizures.” *MedLink Neurology*, 17 June 2003.

9. Takumi Mitsuhashi, Masaki Sonoda, Hirotaka Iwaki, Kazuki Sakakura, Eishi Asano, Detection of absence seizures using a glasses-type eye tracker, Clinical Neurophysiology, Volume 132, Issue 3,2021, Pages 720-722, ISSN 1388-2457, https://doi.org/10.1016/j.clinph.2020.12.015.

10. Gumnit RJ, Labiner DM, Fountain NB, et al. (2012) Data on Specialized Epilepsy Centers: Report to the Institute of Medicine's Committee on the Public Health Dimensions of the Epilepsies. Epilepsy Across the Spectrum: Promoting Health and Understanding. https://www.ncbi.nlm.nih.gov/books/NBK100603/

11. E. Azad, D. Bajpai, R. Butler, L. Sridhar, "Final Design Review: A Device to Detect and Quantify Seizures," Thayer School of Engineering, Hanover, N.H., U.S., 2016.

12. Mayberry A, Hu P, Marlin B, Salthouse C, Ganesan D. iShadow: Design of a Wearable, Real-Time Mobile Gaze Tracker. *MobiSys*. 2014;2014:82-94. doi:10.1145/2594368.2594388

13. Li, Tianxing and Zhou, Xia, "Battery-Free Eye Tracker on Glasses" (2018). Computer Science Technical Report TR2018-861. https://digitalcommons.dartmouth.edu/cs_tr/378

14. Liang, Lan, et al. "National Inpatient Hospital Costs." *National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2017 #261*, https://www.hcup-us.ahrq.gov/reports/statbriefs/sb261-Most-Expensive-Hospital-Conditions-2017.jsp.

15. Gazzola DM, Thawani S, Agbe-Davies O, Carlson C. Epilepsy monitoring unit length of stay. Epilepsy Behav. 2016 May;58:102-5. doi: 10.1016/j.yebeh.2016.02.031. Epub 2016 Apr 9. PMID: 27064830.

16. Fantaneanu, T. A., Hurwitz, S., van Meurs, K., Llewellyn, N., O'Laughlin, K. N., & Dworetzky, B. A. (2016). Racial differences in emergency department visits for seizures. Seizure, 40, 52-56.

17. Caron, E., Wheless, C. E., Patters, A. B., & Wheless, J. W. (2015). The charges for seizures in the pediatric emergency room: a single center study. Pediatric neurology, 52(5), 517–520. https://doi.org/10.1016/j.pediatrneurol.2015.02.010

18. Yield of Monitoring in an Adult Epilepsy Monitoring Unit (P2.097). Shubhi Agrawal, Lauren Turco, Shweta Goswami, Michele Faulkner, Sanjay Singh. Neurology Apr 2015, 84 (14 Supplement) P2.097

19. ENGS 90: Eye-Tracker for Glasses-Based Computing, 2014

20. Chicco, D., Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21, 6 (2020). https://doi.org/10.1186/s12864-019-6413-7

21. Cheong Hee Park and Haesun Park. 2008. A comparison of generalized linear discriminant analysis algorithms. Pattern Recogn. 41, 3 (March, 2008), 1083–1097. DOI:https://doi.org/10.1016/j.patcog.2007.07.022

22. "Types of Seizures." *Johns Hopkins Medicine*, https://www.hopkinsmedicine.org/health/conditions-and-diseases/epilepsy/types-of-seizures.

23. "F-Score." *Wikipedia*, Wikimedia Foundation, 12 Nov. 2021,

https://en.wikipedia.org/wiki/F-score.

Appendix A: Hardware

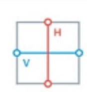
Appendix A.1: Pupil Core Technical Specifications

Pupil Core — Technical Specifications

Frame

22.75 g
W: 160
H: 51mm
D: 175 mm

Gaze Accuracy	**Accuracy** 0.60º	**Precision** 0.02º
Pupil Tracking	Dark Pupil with 3D model	
Pupil Parameters	2D Position 3D Eye model parameters	
Gaze Parameters	**2D Gaze** Normalized 2D gaze position	**3D Gaze** 3D gaze rays + 3D gaze point through binocular vergence
Pupil Diameter	Relative size in eye camera pixels, absolute size in mm through 3D eye model	
Calibration	5 point calibration. Multiple calibration methods available. See documentation	
Sampling Frequency	**Eye Camera** 200Hz @ 192x192px	**Scene Camera** 30hz@1080p 60hz@720p 120hz@480p
Latency	**Eye Camera** 8.5ms	**Processing Latency** Depending on CPU typically > 3ms
Slippage Compensation	Yes, through 3D eye model	
Recording	Pupil and gaze and user data Raw eye and world video	
Connectivity	Pupil Core headsets connect via USB to your laptop or desktop computer running Pupil Core software. Pupil Capture destop app enables data capture, recording, and	real-time data relay via WiFi or LAN. Please see network API documentation for more info.
Physical Properties	**Material** PA12 Nylon	
Scene Camera FOV	**Wide Angle Lens** 1080p H:139 V:83 720p H:99 V:53 480p H:100 V:74	**Narrow Angle Lens** 1080p H:88 V:54 720p H:63 V:37 480p H:42 V:32
Sample Recording	Download sample recording	
Desktop Software	**Pupil Capture** Real time application. Download	**Pupil Player** Post-hoc visualization and analysis. Download



Measured from the midpoint of the frame edge for both horizontal and vertical values.

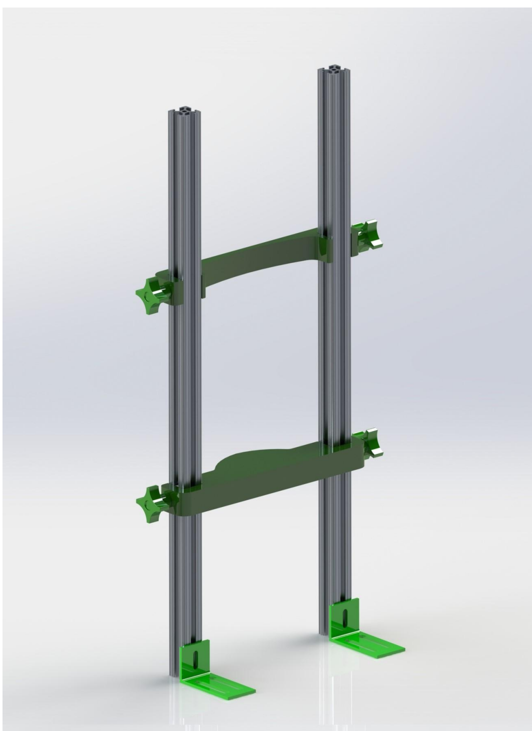
Pupil Core Technical Specifications

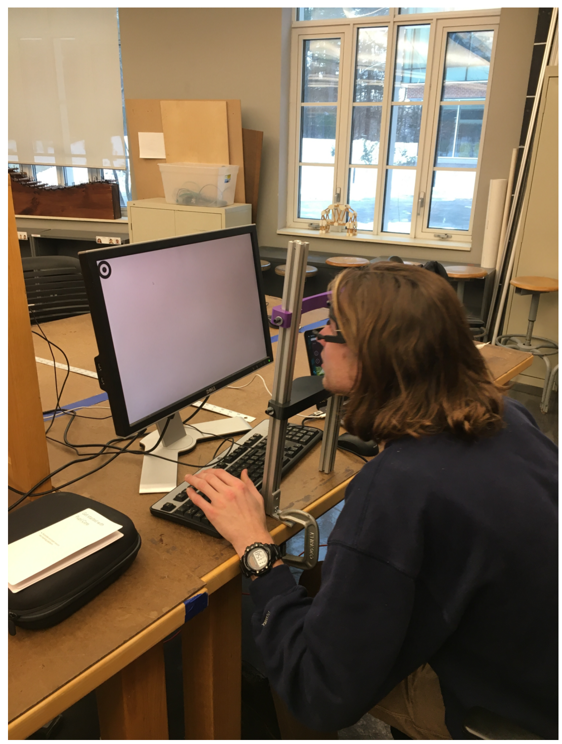
Full Pupil Core Documentation: https://docs.pupil-labs.com/core/#_1-put-on-pupil-core

Appendix A.2: GY-521 MPU-6050 Inertial Measurement Unit

Datasheet: https://invensense.tdk.com/wp-content/uploads/2015/02/MPU-6000-Datasheet1.pdf

Appendix A.3: Chin and Forehead Mount





Appendix B: Experimental Protocol & Checklist

Appendix B.1: Experimental Protocol

Eye Tracking For Seizure Detection

Experimental Protocol

Purpose

This step-by-step guide explains how to conduct an experiment to collect eye and head tracking data. The protocol is currently written using the Pupil Core hardware. However, future hardware can be substituted in, for the purpose of comparing that hardware's output to the Pupil Core's, which is assumed ground truth. When future hardware is used, remote_run.py will need to be modified to turn on and off the new hardware's data collection, since it is currently communicating with the Pupil Core hardware. Note that some steps may not be required when new hardware is substituted in. The general outline of the experiment is as follows:

1. Initial setup
2. 3x3 calibration (mounted)
3. Validation
4. 3x3 calibration (unmounted)
5. 2x2 calibration (mounted)
6. Data collection

Average total run time per subject is 48 minutes.

Initial Setup

1. Log into the computer and open up a terminal (Terminal for Mac or Command Prompt for Windows).

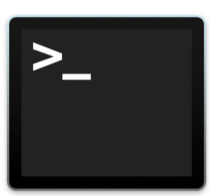


Figure 1: Terminal application icon

2. In the terminal, navigate to the directory containing the remote_run.py software. If unfamiliar with command line commands, see https://learndjango.com/tutorials/terminal-command-line-beginners
3. On the computer, create a folder in the same directory as the software to hold the experimental data. You could name this folder “experiment1”, for example. **Do not name the folder a name that already exists in the directory.**
4. Clamp the chin and forehead mount to the table in front of the computer in the middle of the screen.

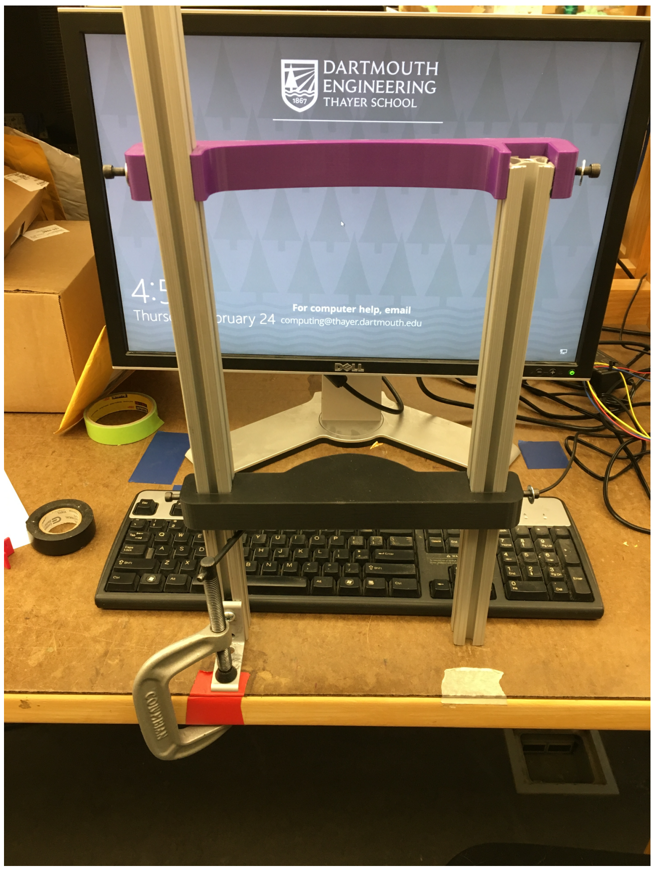


Figure 2: Chin and forehead mount clamped to table

5. Plug the Pupil Core device and Arduino into the computer. Also ensure that the power cord of the Arduino is plugged in.

Figure 3: Pupil Core plugin (top wire) and Arduino plugin (bottom wire) connected to computer. The Arduino power cord is not pictured here and should be connected to wall socket power.

6. Open the Pupil Capture software application. If necessary, toggle “Detect Eye 0” and “Detect Eye 1” to open the eye video windows on the computer.


Figure 4: Pupil Capture software icon

7. Instruct the subject to wear the Pupil Core device. Adjust the cameras such that each pupil is roughly in the center of the Pupil Core’s cameras. If necessary, adjust the height of the glasses (e.g. by placing something between the nose and nose bridge of glasses).

Figure 5: Passable eye positioning

8. Instruct the subject to look at the center of the screen while rotating the head in wide circles. This allows the Pupil Core to identify what the pupil's movement range is, which is defined by the blue circle in Figure 1.

Calibration & Validation

9. Within the experiment folder, create another directory called "subjx" where 'x' is the subject number (e.g. subj1, subj2, subj3, etc).
10. In the terminal, type the command "python3 remote_run.py -o /experiment/subjx/" where 'x' is the subject number and 'experiment' is the name of the directory you created in step 3.
11. The code will direct the subject to first undergo a 3x3 mounted calibration. Mounted means that their head will be resting on the chin mount. The subject will focus on a dot on the screen for 3 seconds at a time. Each time the subject focuses on a dot, they will press the spacebar when ready to begin data collection for that dot. After 3 seconds, the dot will change location and the process will repeat. To begin the calibration, press the spacebar. Repeat this until the collected eye positions are tightly packed around each centroid. If necessary, instruct the subject to open their eyes wider to prevent interference from eyelid

and eyelashes. Instruct the subject to keep their chin on the mount at the end of the calibration.

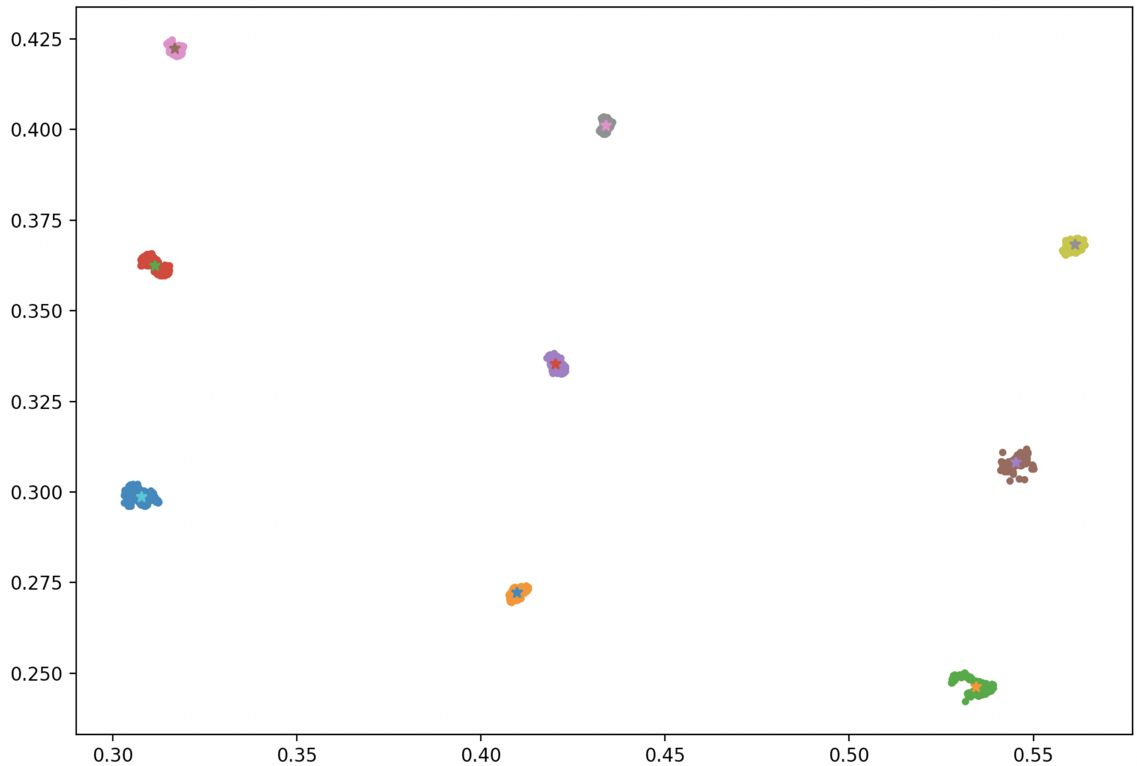


Figure 6: Example of a successful 3x3 calibration. The data points are tightly clustered.

12. After obtaining a satisfactory 3x3 calibration, instruct the subject to keep their head on the chin mount. Close the calibration graph that popped up on the screen. **Do not exit out of the full-screen calibration window.** After selecting the full-screen window, press ‘s’ to save the calibration.
13. Next, the subject completes a validation procedure by following a moving dot on the screen. When ready, press spacebar. If necessary, press ‘R’ to retry.

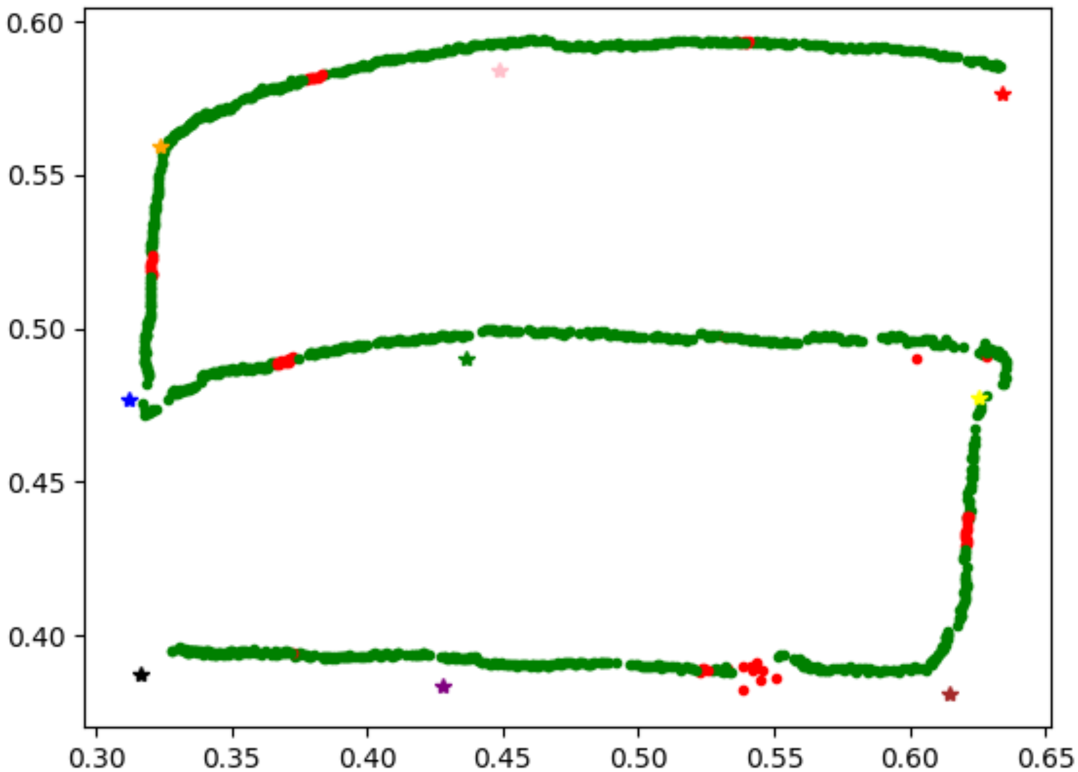


Figure 7: Example of a highly accurate validation

14. After obtaining a satisfactory validation, close the validation graph that popped up on the screen. Check the terminal and record the given accuracy score.
15. Select the main testing window again. Press spacebar to proceed.
16. Next is an unmounted 3x3 calibration, meaning the subject will not be resting their head on the chin rest. However, instruct the subject to keep their head in relatively the same position that they did during the initial calibration, just with their chin not touching the mount.
17. After obtaining a satisfactory 3x3 calibration, close the calibration graph that popped up. After selecting the main testing screen, press spacebar to proceed.
18. Next, the subject will complete a mounted 2x2 calibration, with their head resting on the chin rest. The same procedure of pressing the spacebar when focused on the dot applies. This will be the last time the subject uses the chin rest.
19. After obtaining a satisfactory 2x2 calibration, close the calibration graph that popped up. After selecting the main testing screen, press spacebar to proceed.

Data Collection

20. During the testing phase, the subject will complete four activities: using technology, conversation, eating, and simulated seizure. To begin the technology behavior, ask the subject to begin using either a phone or computer. Ask them to keep everything at or above table level. Once they begin the behavior, press ‘r’ to begin recording. Start a timer for 10 minutes.
21. After 10 minutes, press ‘r’ to stop the recording.
22. Ask the subject to begin eating something (e.g. an orange). If no food is available, ask the subject to pretend to eat something. Once they begin the behavior, press ‘r’ to begin recording. Start a timer for 10 minutes.
23. After 10 minutes, press ‘r’ to stop the recording.
24. Ask the subject to have a conversation with you. Once they begin the behavior, press ‘r’ to begin recording. Start a timer for 10 minutes.
25. After 10 minutes, press ‘r’ to stop the recording.
26. Finally, for the seizure behavior, ask the subject to turn their head all the way to the left side, and to look all the way to the left. Once they begin the behavior, press ‘r’ to begin recording. Start a timer for 1 minute.
27. After 1 minutes, press ‘r’ to stop the recording.
28. Ask the subject to turn their head all the way to the right side, and to look all the way to the right. Once they begin the behavior, press ‘r’ to begin recording. Start a timer for 1 minute.
29. After 1 minutes, press ‘r’ to stop the recording.
30. Press ‘q’ to end the experiment.
31. Navigate to the directory containing the experiment folders.
32. Rename file ‘0.csv’ as ‘technology.csv’.
33. Rename file ‘1.csv’ as ‘conversation.csv’.
34. Rename file ‘2.csv’ as ‘eating.csv’.
35. Rename file ‘3.csv’ as ‘seizure_left.csv’.
36. Rename file ‘4.csv’ as ‘seizure_right.csv’.

Appendix B.2: Experimental Checklist

Eye Tracking for Seizure Detection
Experimental Checklist

Purpose

The following is a checklist of materials and software required to complete the experimental protocol and analyze the data. New hardware can be substituted in for the Pupil Core.

Checklist

Hardware:

- ❖ Eye tracking device
- ❖ GY-521 MPU-6050 Inertial Measurement Unit (IMU)
- ❖ Arduino Uno Rev3
- ❖ 4 sufficiently long cables from IMU to Arduino (~2-3 ft)
- ❖ Chin & Forehead mount
- ❖ C-clamp to secure mount
- ❖ Computer, monitor (~22” was used in experiment), keyboard, mouse

Software:

- ❖ Remote_run.py
- ❖ Process_raw.py
- ❖ Analysis.py
- ❖ Figure_gen.py
- ❖ graphics.py

Appendix C: GitHub README

GitHub Repository: https://github.com/MatterBaby/ENGS90_Software

ENGS90: Eye Tracking for Seizure Detection

This repo includes all software for Thayer School ENGS90 Group 592: Eye Tracking which allows the user to:

1. Calibrate and collect eye and head tracking data from the Pupil Core eye-tracker and Arduino compatible IMU (***remote_run.py***)
2. Process the hardware-specific output into a hardware-agnostic form, and run analysis using hardware specific features (***process_raw.py***)
3. Use a hardware-agnostic .csv file to analyze collected data (***analysis.py***)

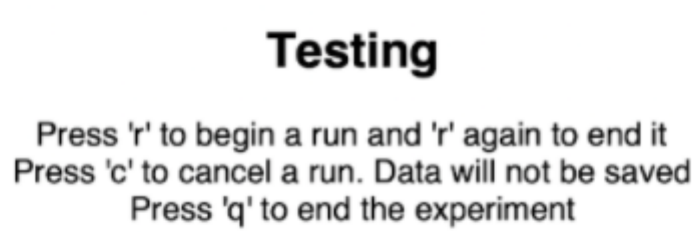
   Python 3+ required. Neccessary packages for each script noted in the code documentation

remote_run.py

Calibrates Pupil Core and collects synchronized data from Pupil Core and Arduino IMU. Writes out centroid coordinates and raw experimental data to a designated directory. The user is able to control the process via a simple GUI.

Useage:

```
python remote_run.py -o [output directory] -c [optional: path to csv of calibrated centroids] -v [optional: run validation protocol, defaults to False] -p [optional: specifies Arduino port, defaults to COM4] -n [optional: NxN grid number, defaults to 3 (3x3 grid)] ``
```



- Pressing 'r' begins and ends an experimental run, saving the data to the directory with the `-o` argument.
- Pressing 'c' will cancel the current run -- no data will be saved.
- Pressing 'q' will end the experiment and the GUI will close
- While data is being recorded, a bold " **RECORDING** " will show up beneath the text shown above

Detailed experimental protocol for calibrating the Pupil Core and collecting data is contained in the Appendix of the Final Report

process_raw.py

Processes the raw output files written by remote_run.py into a hardware-agnostic .csv file of the form:

Eye position classification, Acceleration_X, Acceleration_Y, Acceleration_Z, Gyroscope_X, Gyroscope_Y, Gyroscope_Z

Useage:

```
python process_raw.py -e [experiment directory] -c [centroid file]
```

Relies on the experimental data packaged in a directory with sub-directories subj1,...,subjn corresponding to each subject in the experiment. Each subdirectory contains .csv files named, 'technology', 'eating', 'conversation', 'seizure_right' and seizure_left', along with a set of centroid files, each calibrated in different ways, with suffixes standarded across each subject subdirectory. The user can designate which of these files to use when classifing eye position with the `-c` argument.

Additionally, performs analysis requiring specific output from the Pupil Core (x, y, and confidence values) such as extracting additional features for LDA analysis.

analysis.py

Loads in the experimental data from the hardware-agnostic files from an experimental directory as specified above. Each behavior .csv once converted to a hardware-agnostic format will contain a "data_" prefix, which analysis searches for while loading data.

Useage:

```
python analysis.py -e [experiment directory] -o [output directory]
```

Feature Extraction

Using loaded .csv files, the data is chunked into ~5 second long "epochs" of 130 samples each. The following features are extracted from each epoch:

Eye Tracking

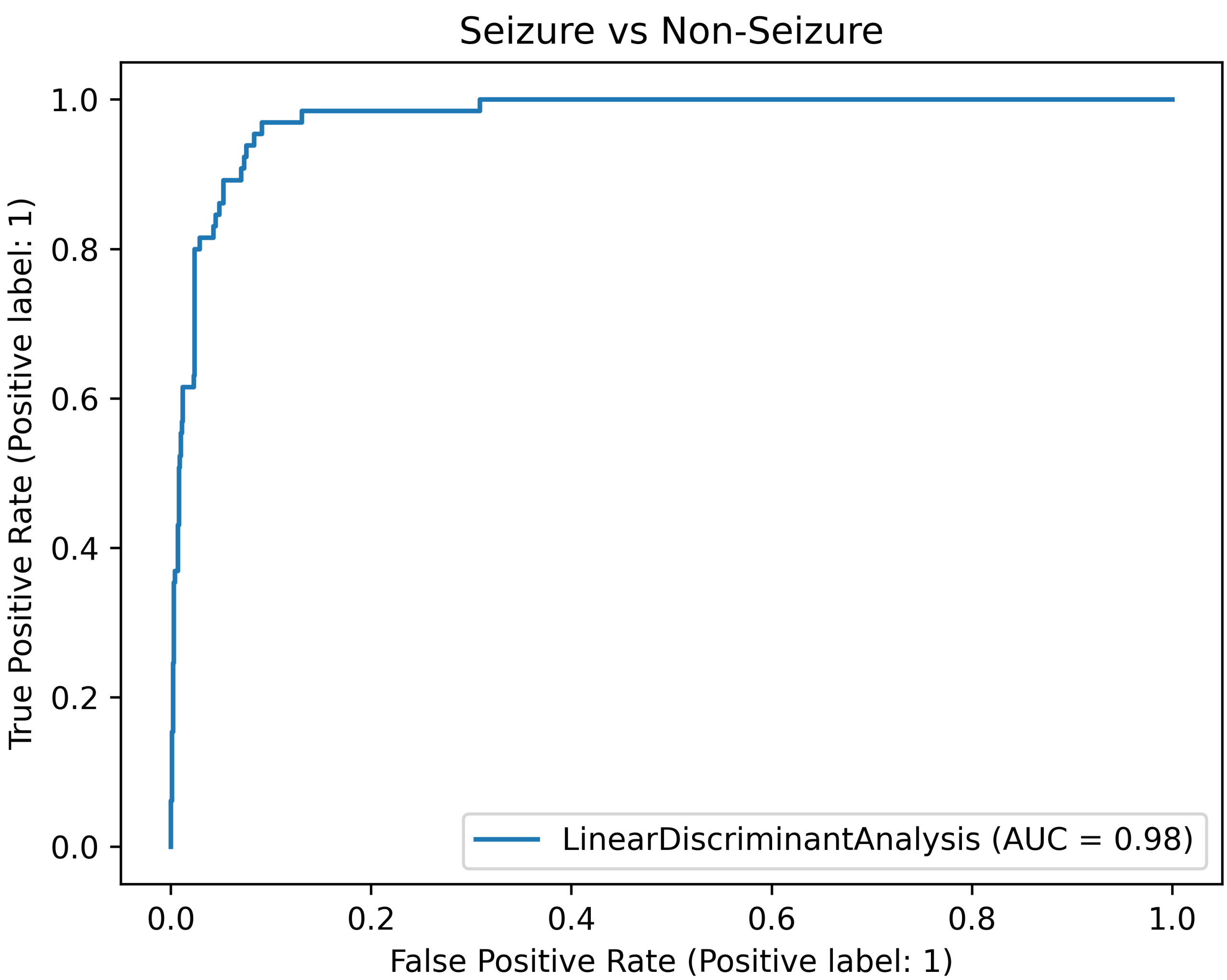
- Average classification value
- Number of unique classifications
- Number of changes between classifications from one sample to the next
- Number of low-confidence classifcations
- Percentage of samples in the top and bottom edges of the grid (for 3x3 grid and up)
- Percentage of samples in left and right edges of the grid (for 3x3 grid and up)

Head Tracking

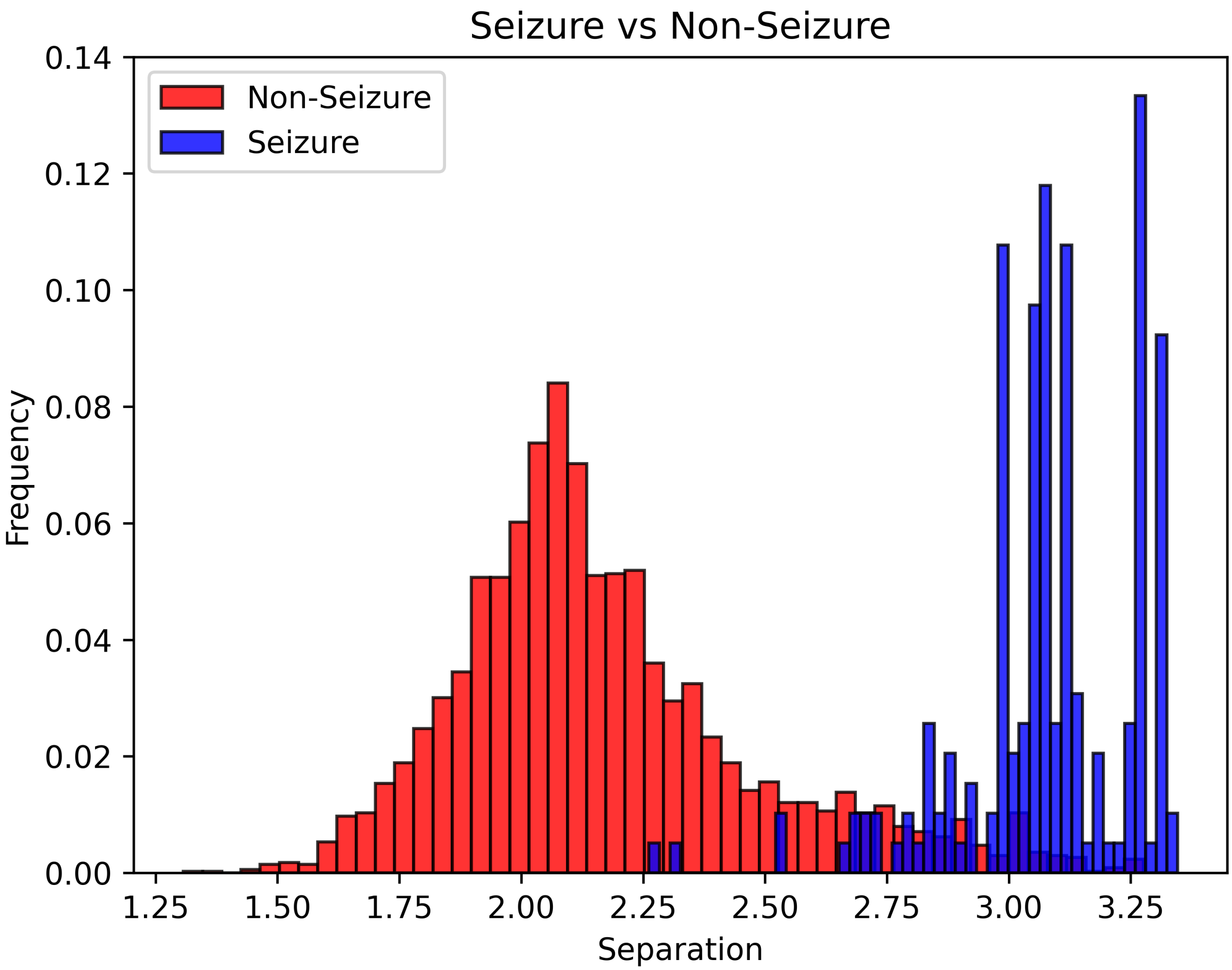
- Average magnitude of accelerometer readout
- Average magnitude of gyroscope readout
- Standard deviation of accelerometer readout
- Standard deviation of gyroscope readout
- Maximum magnitude of accelerometer readout
- Maximum magnitude of gyroscope readout
- Minimum magnitude of accelerometer readout
- Minimum magnitude of gyroscope readout
- Range of magnitude of accelerometer readout
- Range of magnitude of gyroscope readout

Linear Discriminant Analysis

The extracted features are then run through a Linear Discriminant Analysis. 5-fold cross validation averages area under the receiver operating characteristic curve (ROC AUC score) and F1 macro score to evaluate the classifier. The following figures are produced:



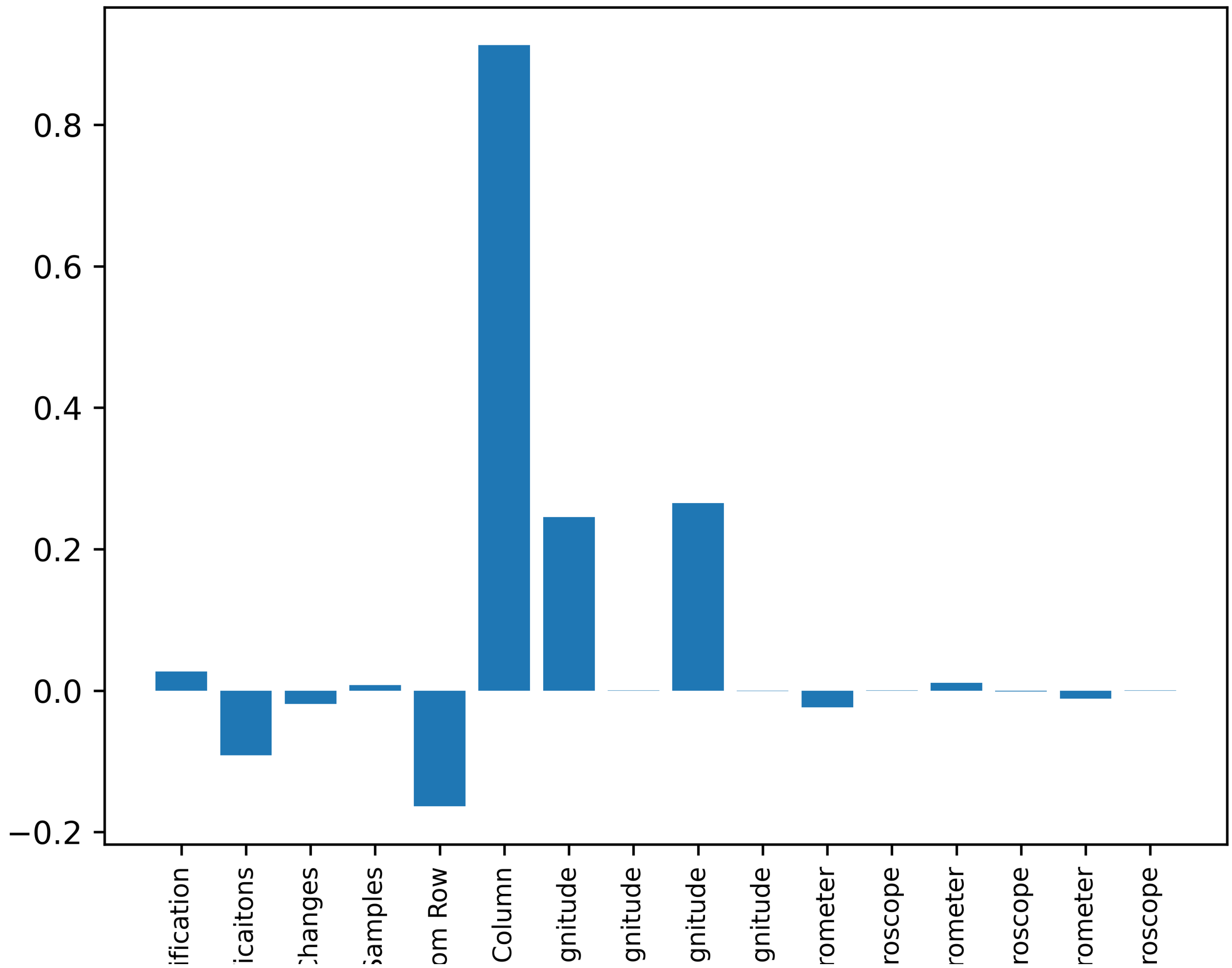
The Receiver Operating Characteristic



LDA Visualization Frequency Histogram

Feature	Normalized Coefficient
% Left/Right Column	0.9124131811886481
Std Accelerometer Magnitude	0.2648102778582806
Average Accelerometer Magnitude	0.24527142683230263
% Top/Bottom Row	-0.16383343604541575
# Unique Eye Classificaitons	-0.09173234588405464
Average Eye Classification	0.02696155150928424
Maximum Accelerometer	-0.023973341438551253
# Classifcation Changes	-0.018706615754867765
Range Accelerometer	-0.01158304614473444
Minimum Accelerometer	0.010747854851590294
# Low-Confidence Samples	0.007738464978454073
Minimium Gyroscope	-0.0013136120379477964
Std Gyroscope Magnitude	-0.0006100464007446332
Average Gyroscope Magnitude	0.0003970863271147222
Range Gyroscope	5.670620370080295e-05
Maximum Gyroscope	5.499685962724558e-05

Ranked Feature Table



Feature Bar Chart

experiment

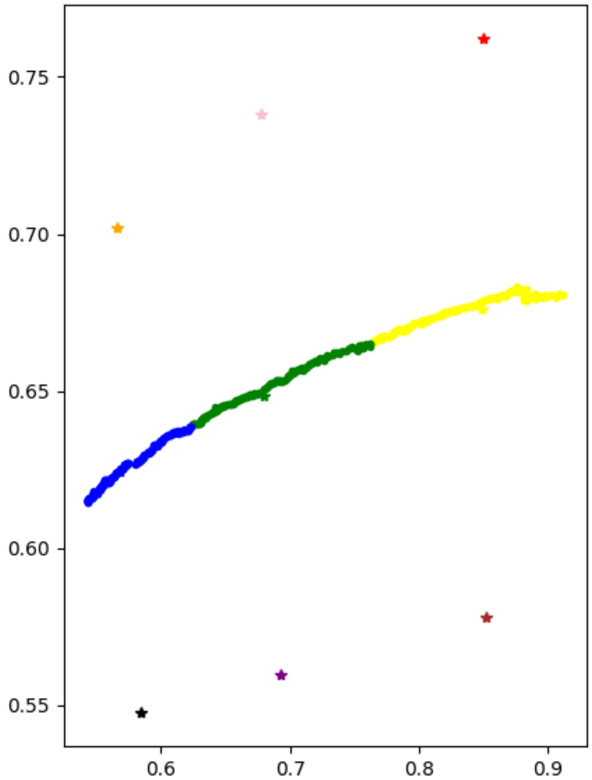
Directory containing all the raw recorded data for each of our subjects, designated by behavior.

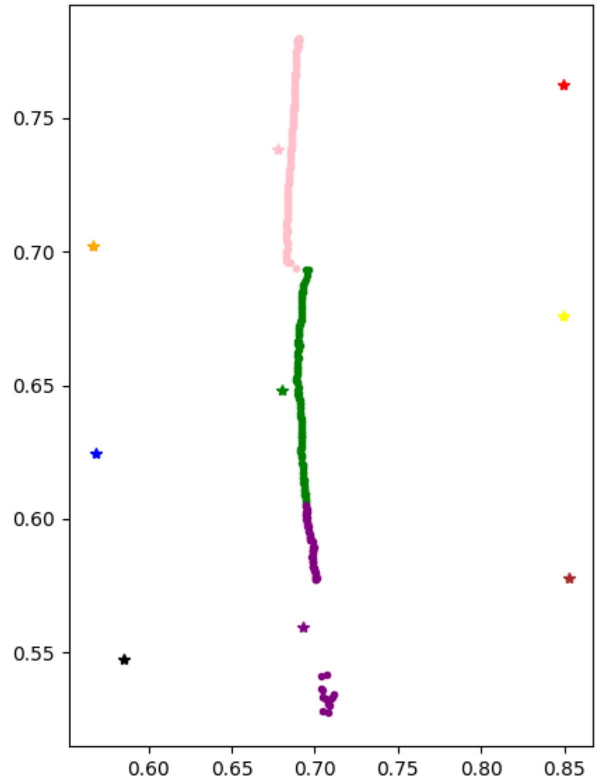
figure_gen.py

Contains code for generating the supplemental figures used in the Final Report.

trace_line

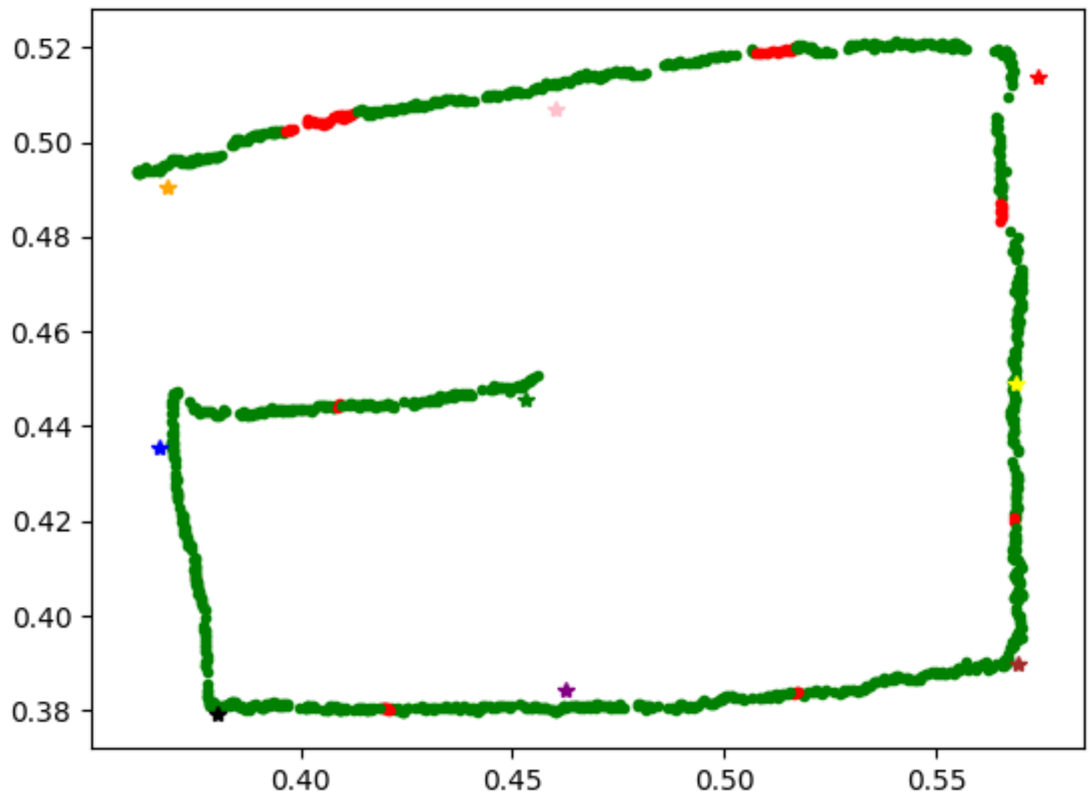
Run via the `-l` argument from `remote_run`. Graphs eye position data colored by classification value with colored centroids for reference


0.75
0.70
0.65
0.60
0.55
0.6
0.7
0.8
0.9


0.75
0.70
0.65
0.60
0.55
0.60
0.65
0.70
0.75
0.80
0.85

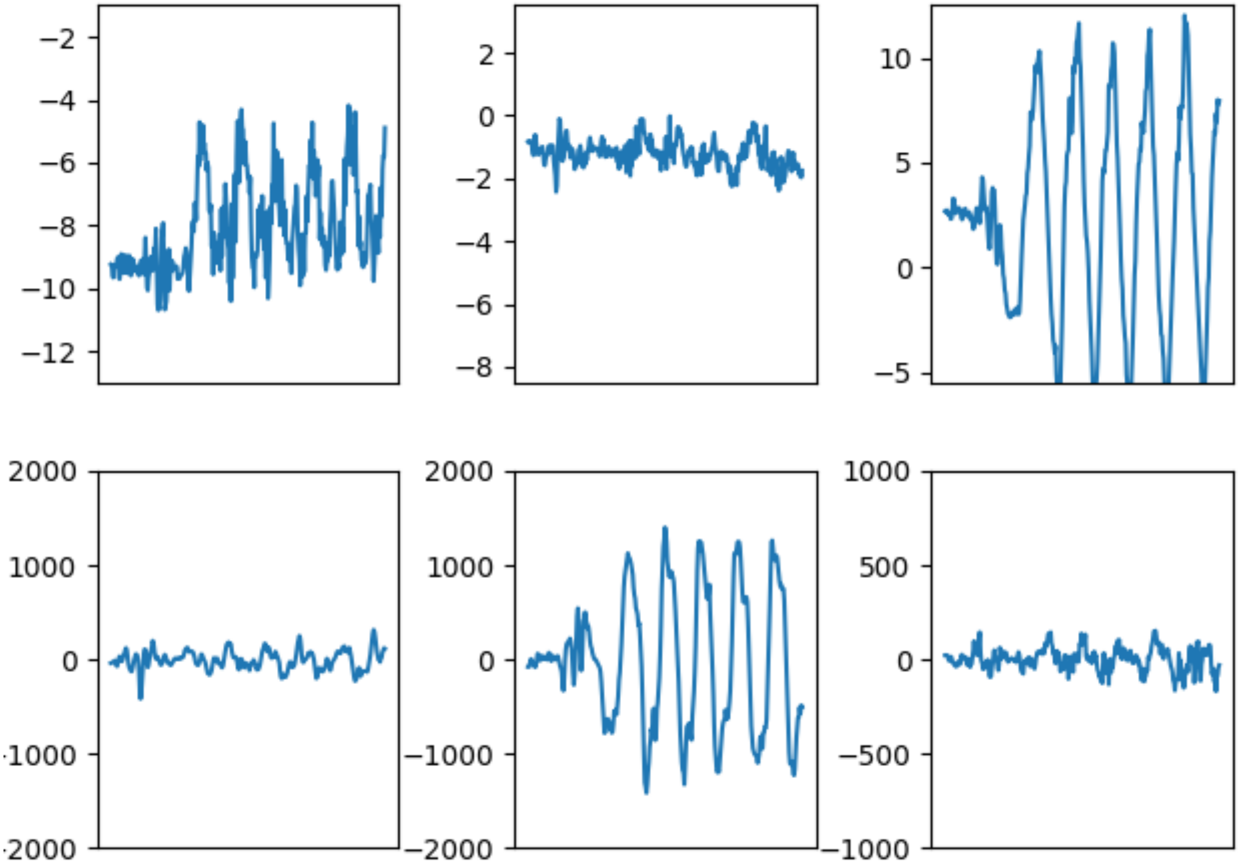
validate

Run as default during the calibration process to benchmark the accuracy of the calibrated centroids



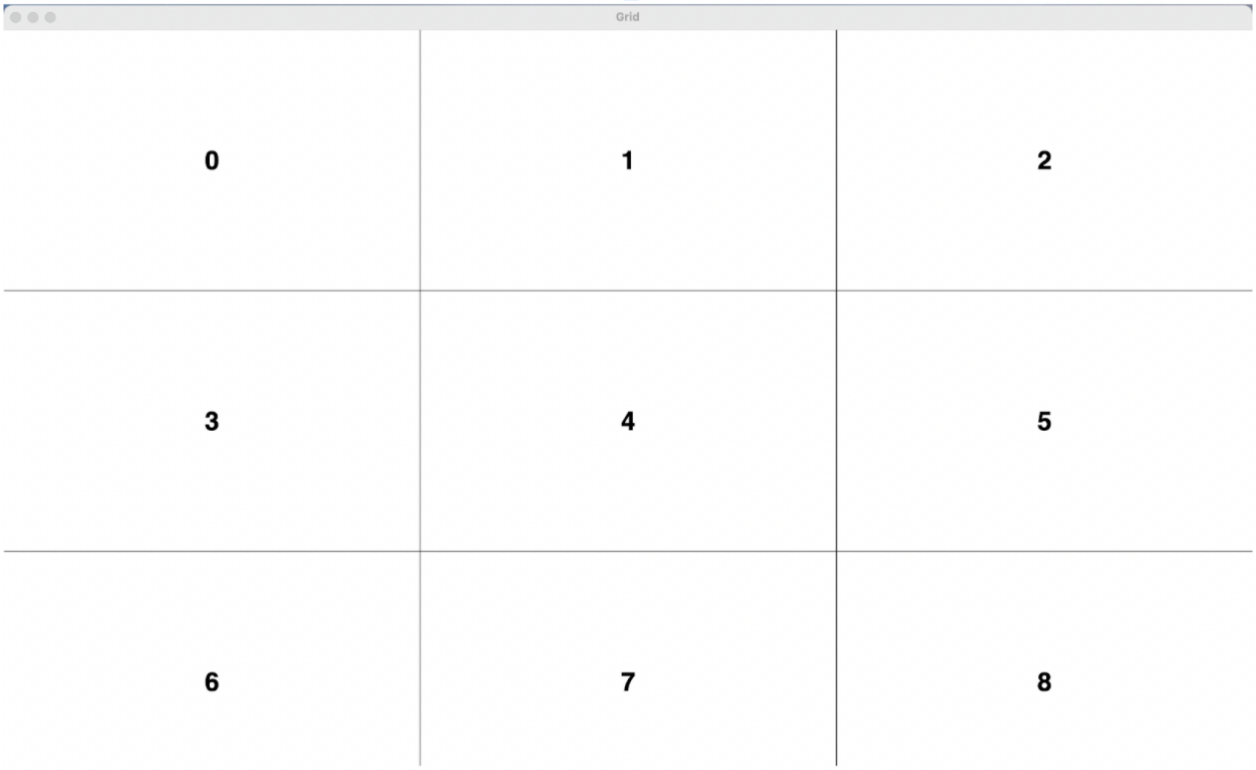
plot_accel

Generates a figure of gyroscope and accelerometer readout. xlim and ylim values may need to be manually adjusted

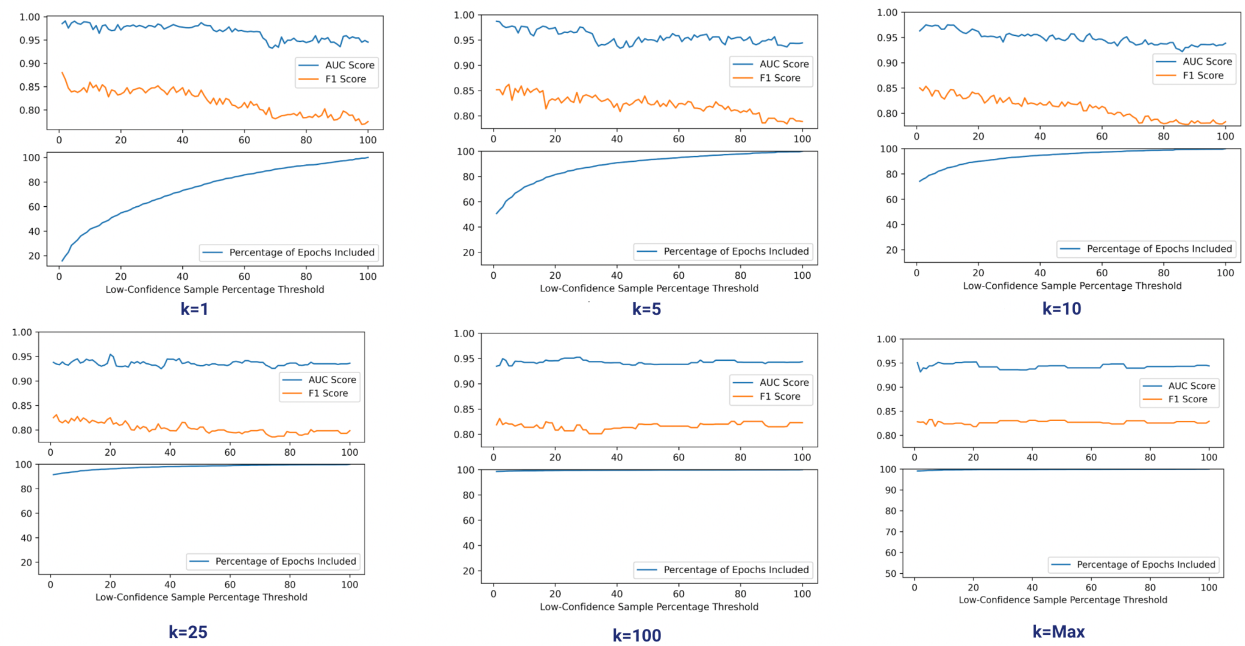


grid

Generates a reference grid labeled with classification values.

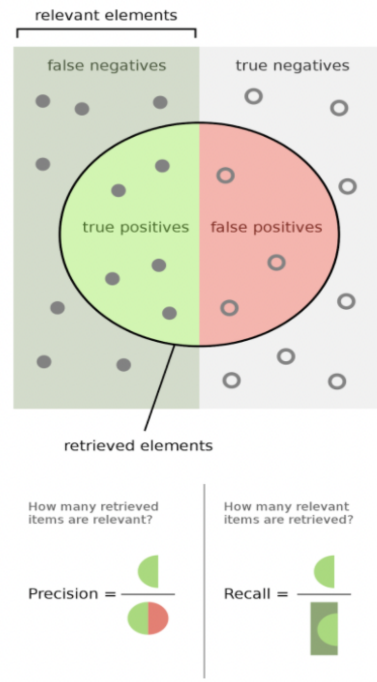

Grid
0
1
2
3
4
5
6
7
8

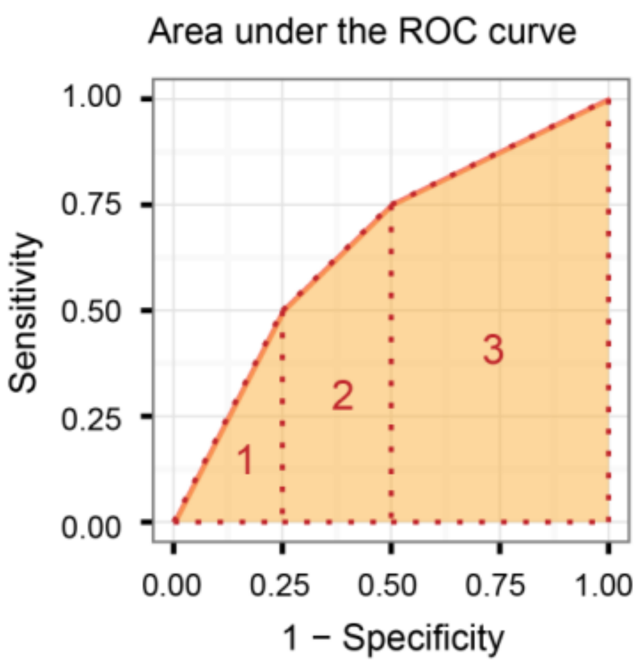
Appendix D: Data Processing Supplementary Figures



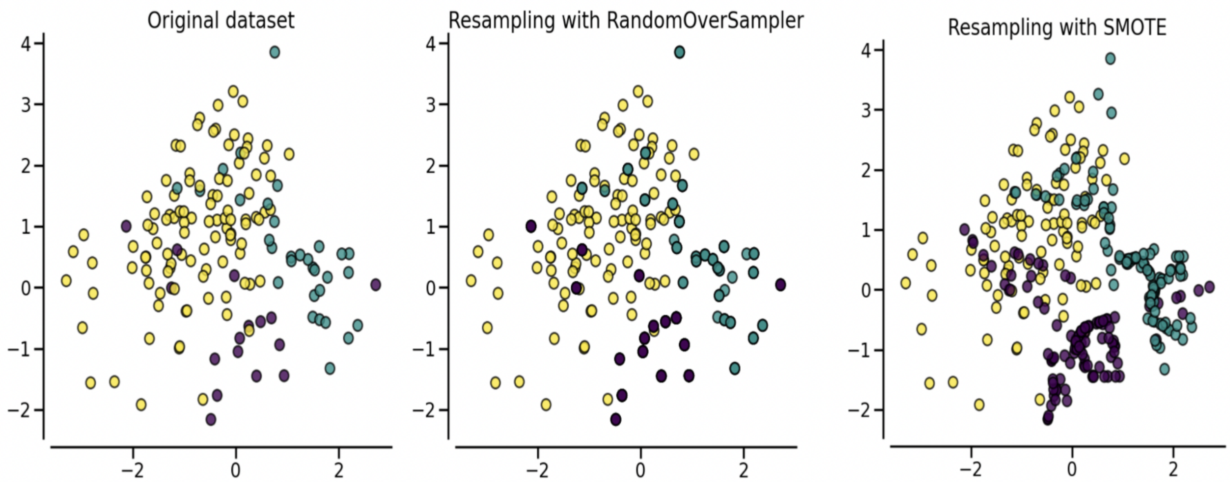
Appendix D.1: k-Fill Algorithm for Data Imputation for all k

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$





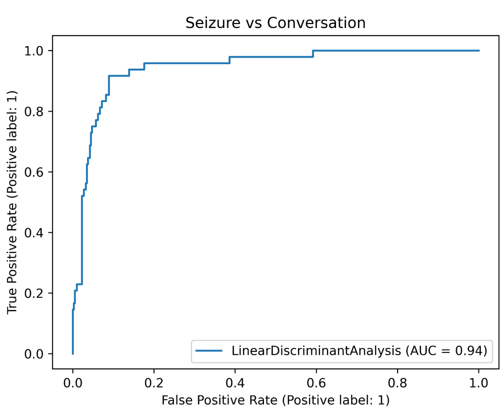
Appendix D.2: F1 and AUROC Scores[23]

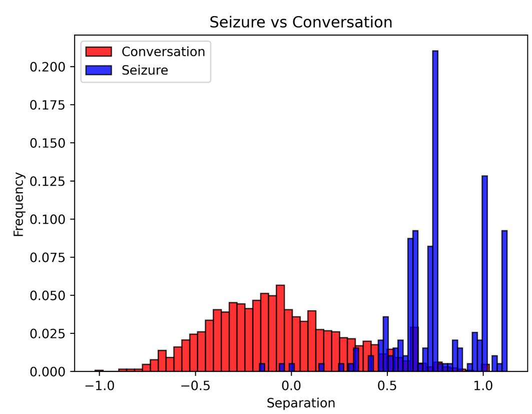


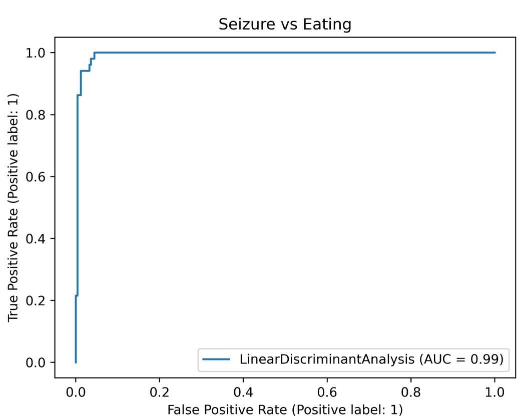
Appendix D.3: SMOTE (https://imbalanced-learn.org/stable/over_sampling.html)

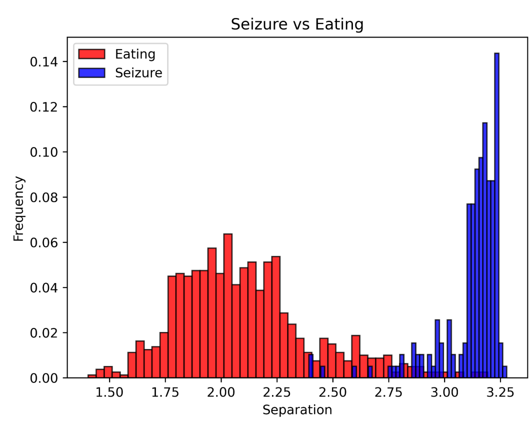
Appendix E: Supplementary Results Figures

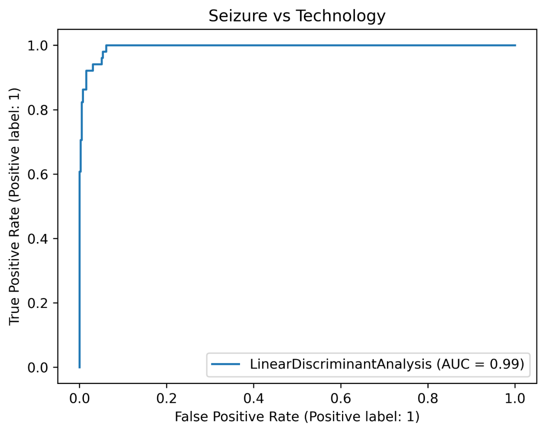
Appendix E.1: Full Experiment Results for All Subject Data

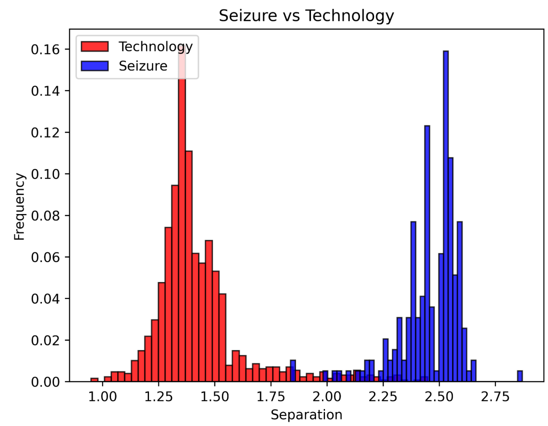


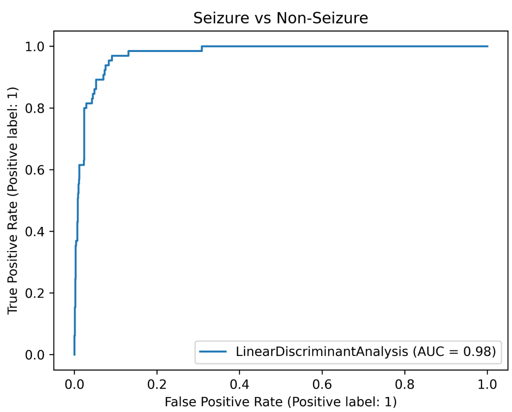


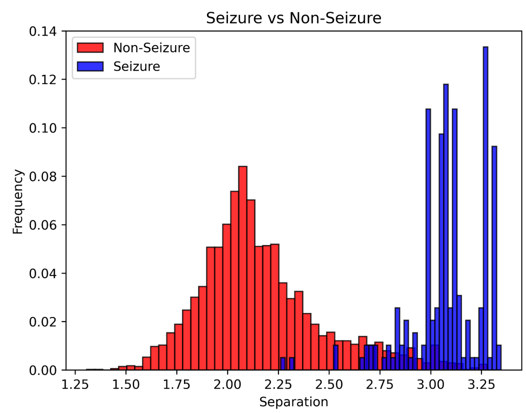




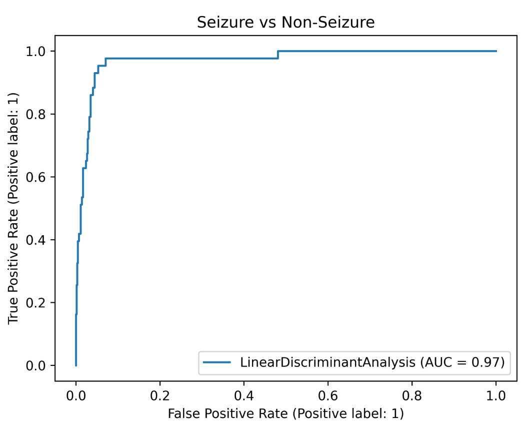


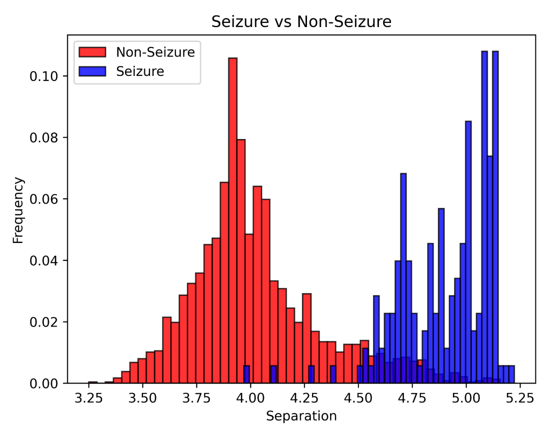






Appendix E.2: Baseline Results (without Data Imputation) with a 25% Failed Sample Threshold





AUROC: .97, F1: .849

Appendix E.3: Drop-One Subject Analysis (AUROC / F1)

Dropped Subject #	*Technology*	*Eating*	*Conversation*	**Non-Seizure**
1	1.0/.93	.98/.92	.98/.80	.98/.81
2	1.0/.95	1.0/.92	.96/.84	.98/.83
3	1.0/.94	1.0/.94	.97/.86	.98/.856
4	1.0/.95	1.0/.95	.96/.82	.98/.83
5	1.0/.94	1.0/.95	.96/.82	.97/.85
6	1.0/.93	1.0/.94	.97/.87	.99/.85
7	1.0/.94	.99/.94	.97/.85	.99/.88
8	.96/1.0	.99/.93	.97/.85	.98/.85
All (no drop)	.99/.93	.99/.93	.95/.85	.98/.862

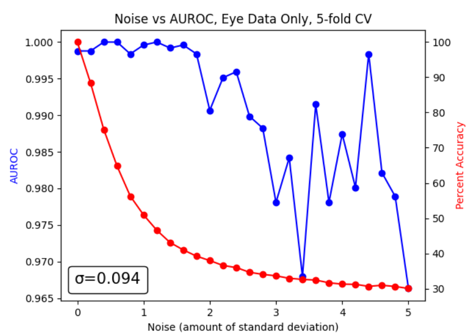
Appendix F: Python Packages

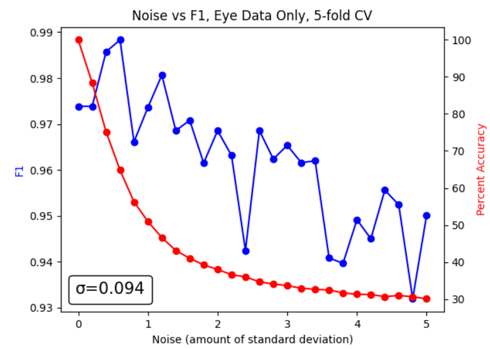
Library	Documentation	Use
imbalance-learn	https://imbalanced-learn.org/stable/	Implement SMOTE Oversampling
matplotlib	https://matplotlib.org/	Figure generation
numpy	https://numpy.org/	Mathematical functions
pandas	https://pandas.pydata.org/	Data analysis and manipulation
pyplr	https://pypi.org/project/pyplr/	Communication with Pupil Core
pySerial	https://pyserial.readthedocs.io/en/latest/pyserial.html	Communication with Arduino IMU
scipy	https://scipy.org/	Distance calculations
sckit-learn	https://scikit-learn.org/stable/	LDA, k-Fold Cross Validation, F1 and AUROC Scores, Binary Classifiers
tkinter	https://docs.python.org/3/library/tkinter.html	Graphical User Interface for experiment data collection

Appendix F : Used Python packages with links to full documentation

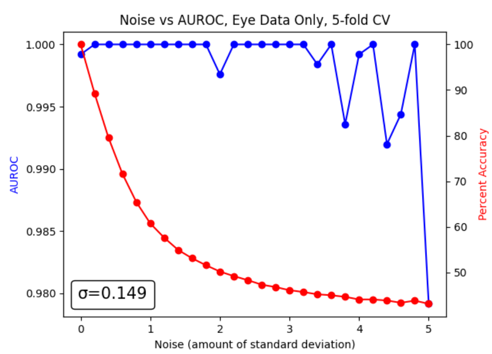
Appendix G: Noise

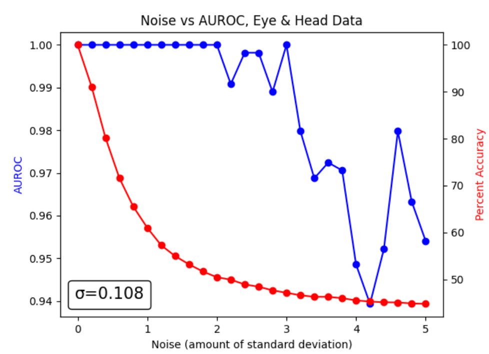
Appendix G.1: Subject 1 individual data, F1 and AUROC vs noise



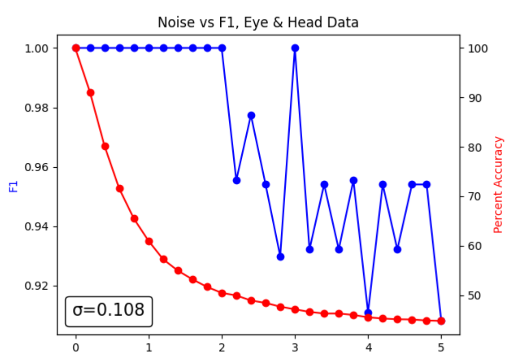


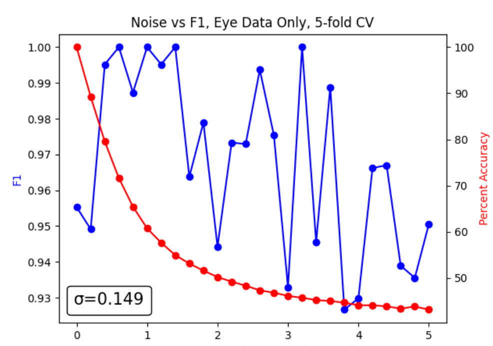
Appendix G.2: Subject 2 individual data, F1 and AUROC vs noise



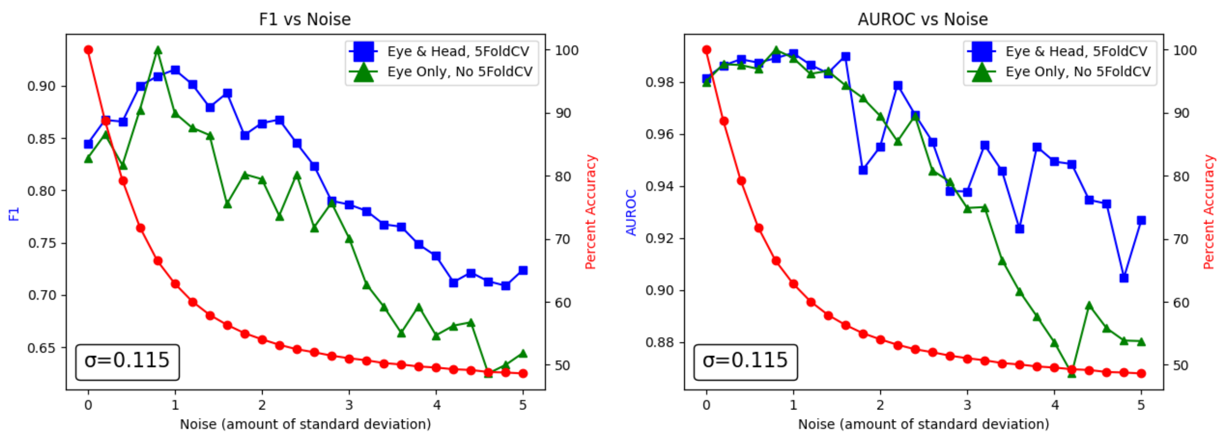


Appendix G.3: Subject 3 individual data, F1 and AUROC vs noise





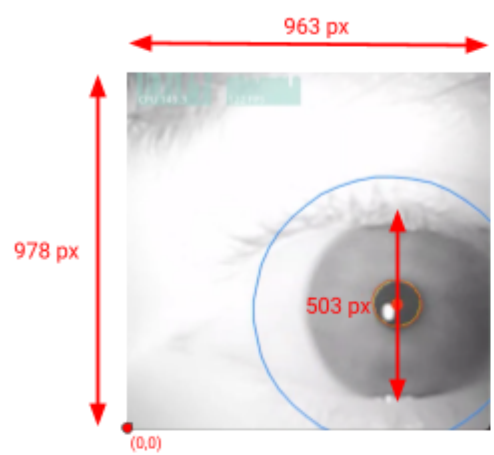
Appendix G.4: Verifying usefulness of eye and head features and 5-fold cross validation



Appendix H: Normalized Coords to Millimeters

Appendix H.1: Deriving equations to convert from normalized coordinates to millimeters

- Pupil Core
  - Origin bottom left
  - X = x pixel position / image width
  - Y = y pixel position / image height
  - Ranges from [0, 1] for X and Y
- Average pupil is 12.5 mm in diameter



(http://wordpress.artificialeyeclinic.com/eye-anatomy/iris-limbus-and-sclera/)

503 px = 12.5 mm

1px ≈ 0.025 mm

Normalized -> cm conversion:

X_{mm}= (X*963 px)*0.025mm/px

Y_{mm}= (Y*978 px)*0.025mm/px

Appendix H.2: Calculating recommended accuracy

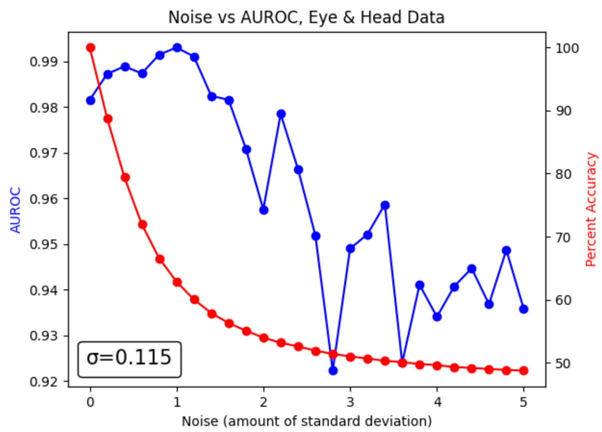
- F1 > 0.85, AUROC > 0.95 at 1 standard deviation of noise

σ = 0.115 normalized units

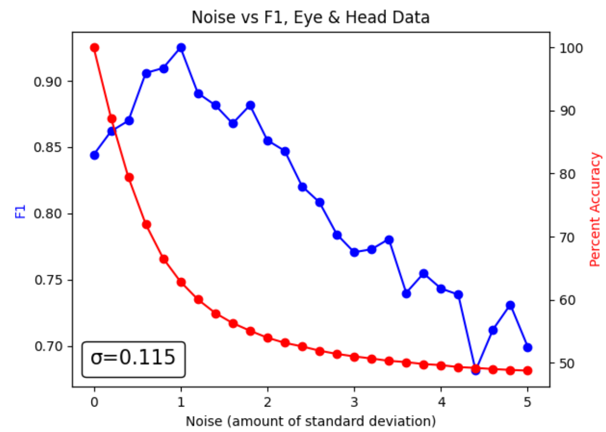
Converting from normalized coordinates to mm:

0.115*(970 px)*(0.025mm/px)

σ_{mm} ≈ 2.75 mm → Recommend hardware track center of pupil accurately within 2.75 mm



Noise vs AUROC, Eye & Head Data
AUROC
Percent Accuracy
σ=0.115
Noise (amount of standard deviation)



Noise vs F1, Eye & Head Data
F1
Percent Accuracy
σ=0.115
Noise (amount of standard deviation)

CPU 149.3
122 FPS
2.75mm