Look Over Here!

Aya Hourani
Software Engineering
Rutgers University
New Brunswick, NJ, USA
aah173@scarletmail.rutgers.edu

[1] ***Abstract*—This paper explores the correlation between gaze data and game performance using a dataset featuring 16 Atari games. Leveraging feature engineering and machine learning, the study aims to discern differences in gaze patterns between skilled and less-skilled players. Building upon prior research on eye tracking in gaming, the dataset includes diverse games with varying difficulty levels. The experimentation section details the dataset's structure, feature engineering techniques applied, and the final data chart. The evaluation section introduces the cumulative distribution function to select games based on trials and performance. Six games from different difficulty levels are chosen for gaze data extraction, and the process involves selecting trials based on performance. The gaze positions extracted from the original dataset are organized for each trial, with labels indicating "good" or "bad" trials. The Gaussian Mixture Model is then employed for unsupervised classification, predicting the probability of classification for each instance based on gaze data. This research provides insights into the intricate relationship between gaze patterns and gaming performance, contributing to our understanding of player behavior in diverse gaming environments.**

***Index Terms*—Machine Learning, Feature Engineering, Gaussian Mixture Model, Cumulative Distribution Function, Gaze Positions**

I. INTRODUCTION

The video game industry has made an immense impact on our society on a global scale. In fact, video gaming reached new heights during the breakout of the pandemic, becoming one of the mainstream ways people are spending their time. One of the contributing factors leading to this is how accessible video games have become to the average consumer. Wide ranges of gaming consoles, as well as portable gaming devices and mobile games, make it incredibly easy for a consumer to engage in gaming. More recent features like streaming allow consumers to connect with other players globally, as well as profit off creating gaming content while engaging with the audience live; This is currently being done using streaming platforms like Twitch. It is easy to assume that given this information, video gaming is targeting a wide range of audiences, from toddlers to adults. In 2021, more than eight in ten internet users from ages sixteen to forty-four played video games on any device [1]. From a business perspective, companies producing these games and consoles have even more of an incentive to continue releasing games, as they're generating over $206 billion in revenue [1].

A. Problem Statement

In this paper, we explore the correlation between gaze data and game performance. What relationships can we observe? Is there a difference between the gaze patterns belonging to a good player and a bad player? Using a dataset that includes 16 games, all with varying levels of difficulty, we will investigate these questions using feature engineering and machine learning.

B. Related Work

Similar research has been done investigating the difference in gaze control between esports expert players and lower skill players using a popular real-time strategy game, StarCraft [2]. Players had to maintain a certain distance from the monitor and performed tasks for 3 minutes while their gaze movement was recorded by an eye tracker. Continuous research involving eye tracking and video games has been conducted to study human behavior and performance. Some studies show that expert performance doesn't correlate to general intelligence, meaning expertise in one activity doesn't guarantee expertise in a different activity [3]. Using eye tracking, we can observe how a player's eyes behave and where they look on the screen and gauge their performance potential.

II. EXPERIMENT

This section will describe the dataset used and methodology used to extract the relevant fields and reconstruct the dataset.

A. Data Description

The dataset used to conduct this experiment comes from the "Atari-HEAD: Atari Human Eye-Tracking and Demonstration Dataset" [4]. This dataset features 175 trials of gameplay from 4 different subjects and 16 different Atari games. The data was recorded using an EyeLink 1000 eye tracker set at 1000 Hz, and the games were recorded in a frame-by-frame manner. Movements that were tracked include the human keystroke action, reaction time to perform the action, the gaze positions, and immediate reward returned by the environment. These movements were used to calculate the average validation error of each trial for each game and subject. Each subject played a game for 15 minutes, and a 15-minute break was given between each trial. The relevant parameters within the dataset that will be explored include the game, trial number, subject ID, average validation error per trial, and best score achieved per trial.

[1]

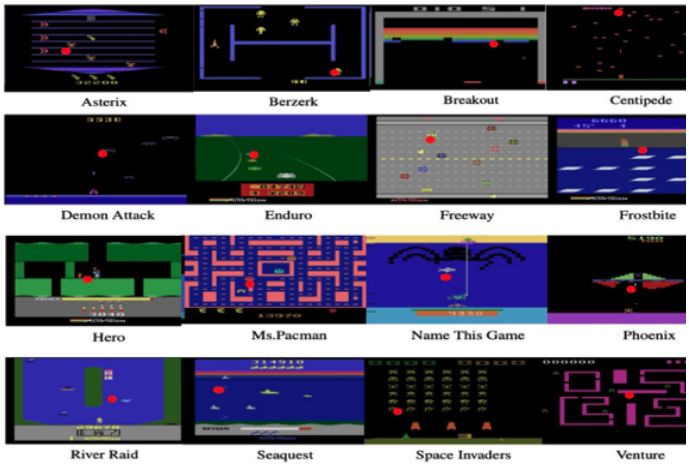

Figure 1: The 16 Atari video games included in the dataset

	Game	TrialNumber	SubjID	AverageValError	BestScore
0	Asterix	160.0	R	0.45	95500.0
1	Asterix	162.0	R	0.41	84500.0
2	Asterix	163.0	R	0.54	99000.0
3	Asterix	213.0	R	0.40	88000.0
4	Asterix	243.0	R	0.25	106500.0

Figure 2: Structure of the original dataset

Figure 2 displays the first 5 entries of the original dataset before applying any feature engineering techniques. The validation error and best score were recorded for each subject, trial, and corresponding game.

B. Feature Engineering

Before applying any algorithmic techniques to the dataset, modifications were made to produce a finalized chart of the data. The first step was to calculate the mean validation error per game. This meant taking the average validation errors in all the trials for a given game and finding the total mean error. This process was also done for the best scores achieved, all the scores for a given game were used to output a total best score; what we're left with are the mean validation errors and mean best scores for each game. To determine which game performed the best, the top worldwide scores for each game had to be calculated. Due to each game being scaled on a different point system, simply looking at the mean best scores and choosing the highest value would be inaccurate. For example, a game with a score of over 100,000 points doesn't necessarily outperform a game with 30 points. To combat this issue, the top 10 scores ever recorded for each game were averaged and used to compute the performance in terms of percentage [5]. The equation used to calculate these values is as shown:

$$Performance\ (\%)\ per\ game\ =\ \frac{Mean\ Best\ Score}{Mean\ Top\ Score}\ \times\ 100\%$$

Now that all the values have been finalized, the final data chart will be updated to contain the pre existing fields, as well as the mean top scores and scaled performances.

C. Final Data Chart

Figure 3 below illustrates the finalized chart containing the relevant fields and values obtained from applying the described feature engineering techniques.

	Game	Mean Validation Error	Mean Best Score	Mean Top Score	Difficulty(%)
0	Asterix	0.402000	171300.000000	335500.000000	48.940000
1	Berzerk	0.378000	5730.000000	70610.555600	91.890000
2	Breakout	0.418300	435.083300	634.500000	31.430000
3	Centipede	0.443200	66770.545500	124056.300000	46.180000
4	DemonAttack	0.512900	6810.000000	18920.000000	64.010000
5	Enduro	0.333600	404.500000	1741.090000	76.770000
6	Freeway	0.483800	31.125000	31.571400	1.410000
7	Frostbite	0.330000	25074.000000	327237.000000	92.340000
8	Hero	0.460000	44657.500000	129870.833300	65.610000
9	Mspacman	0.416500	32797.411800	85855.600000	61.800000
10	NameThisGame	0.364000	7266.000000	25220.000000	71.190000
11	Phoenix	0.472900	30041.428600	79658.750000	62.290000
12	Riverraid	0.402300	17840.000000	182715.000000	90.240000
13	Seaquest	0.332900	111764.285700	194390.000000	42.510000
14	SpaceInvaders	0.454000	1791.000000	44547.000000	95.980000
15	Venture	0.408500	8546.153800	104400.000000	91.810000

Figure 3: Finalized Dataset

The highlighted column represents the top scores that were obtained outside the original dataset. The performance values calculated were used to obtain the difficulty percentage of each game.

$$Difficulty\ (\%)\ =\ 100\ -\ Performance\ (\%)$$

From this chart, we are able to identify the games with the best/worst performances based on their level of difficulty.

III. EVALUATE

In this section, we discuss the methodology for selecting games using the cumulative distribution function, procedure for extracting the actual gaze data from each game, and

A. Cumulative Distribution Function

The cumulative distribution function (CDF) of a random variable X, evaluated at x, will give the probability that X will take a value less than or equal to x. The definition is as follows:

$$F_X(x)\ =\ P(X\ \leq\ x)$$

For the scope of this experiment, we will apply CDF to all 16 games and determine which games to select based on the following criteria:

- Number of trials: Each game was played for a different number of trials. Games with more trials will result in more accurate results.
- Performance/Difficulty: To allow for a more robust analysis, games from varying difficulty levels (easy, medium, and hard) were selected.

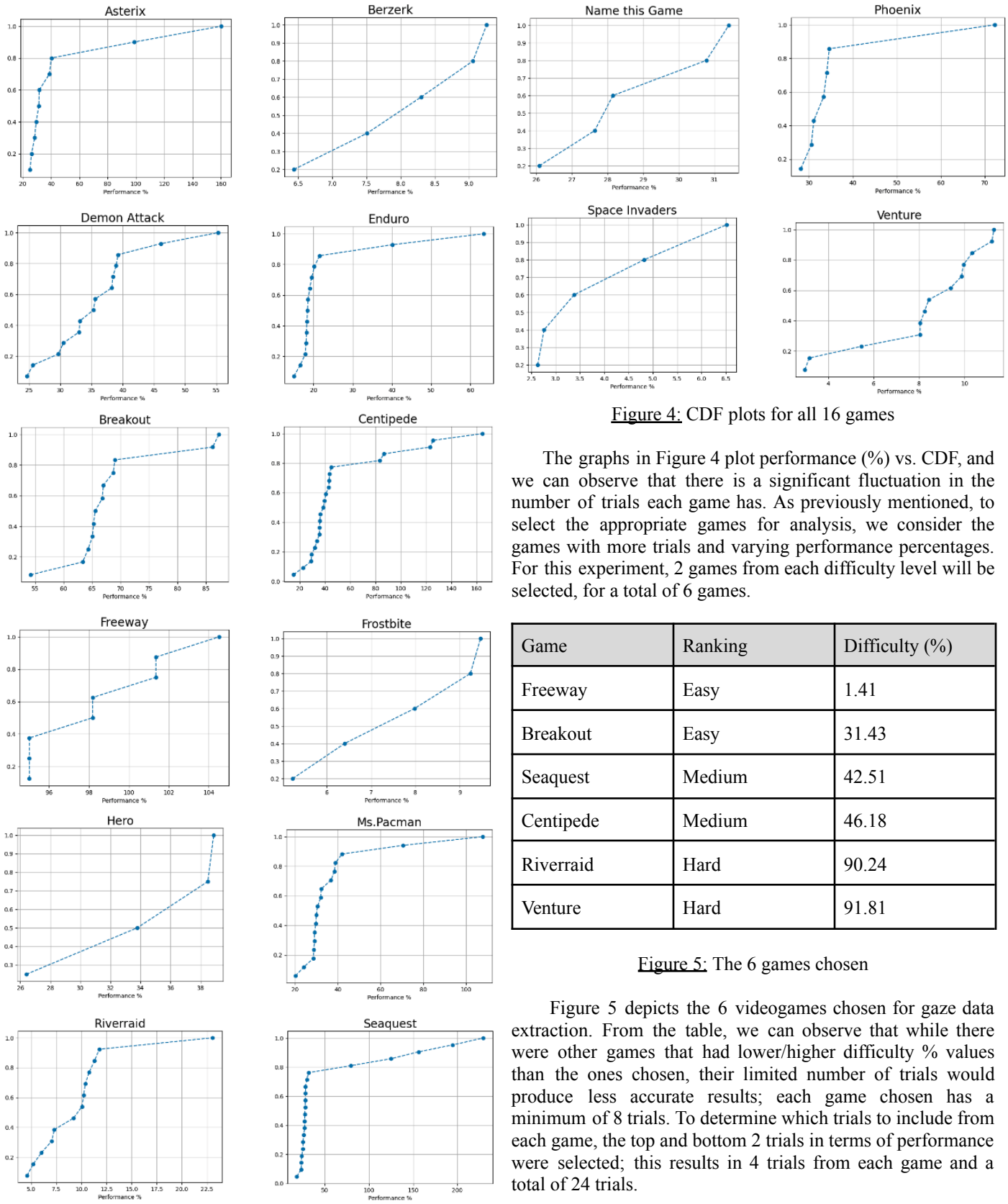

Figure 4: CDF plots for all 16 games

The graphs in Figure 4 plot performance (%) vs. CDF, and we can observe that there is a significant fluctuation in the number of trials each game has. As previously mentioned, to select the appropriate games for analysis, we consider the games with more trials and varying performance percentages. For this experiment, 2 games from each difficulty level will be selected, for a total of 6 games.

Game	Ranking	Difficulty (%)
Freeway	Easy	1.41
Breakout	Easy	31.43
Seaquest	Medium	42.51
Centipede	Medium	46.18
Riverraid	Hard	90.24
Venture	Hard	91.81

Figure 5: The 6 games chosen

Figure 5 depicts the 6 videogames chosen for gaze data extraction. From the table, we can observe that while there were other games that had lower/higher difficulty % values than the ones chosen, their limited number of trials would produce less accurate results; each game chosen has a minimum of 8 trials. To determine which trials to include from each game, the top and bottom 2 trials in terms of performance were selected; this results in 4 trials from each game and a total of 24 trials.

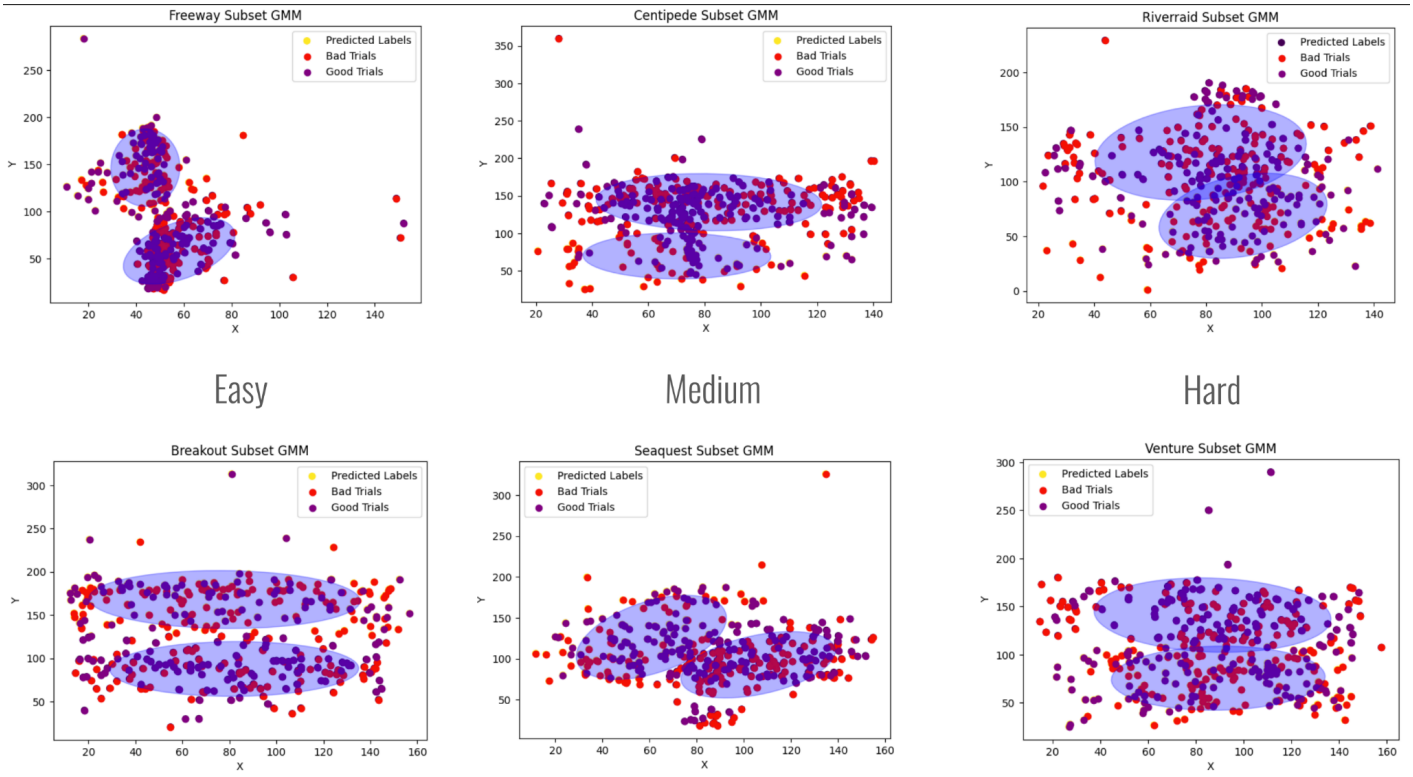

Figure 7: GMM subplots for each game

B. Gaze Extraction and Processing

The gaze positions for each trial come from the original Atari dataset; gaze positions, as well frame by frame gameplay for every trial for every game are provided. The gaze positions are composed of (x, y) coordinate pairs, each pair referring to each frame in a given trial. Each trial contains approximately 400-450 thousand coordinate pairs, and after combining all gaze positions from the 4 selected trials we have a total of ~1.5 million coordinate pairs for each game. To organize this data, each game is a list of tuples, each tuple containing the x and y coordinates, as well as a label describing whether the pair comes from a "good" or "bad" trial. This is important for keeping track of the data points when using this data to generate the Gaussian Mixture Model. The data structure for each trial is as follows:

game_trial = $[(x_0, y_0, "label"), (x_1, y_1, "label"), (x_2, y_2, "label")..... (x_n, y_n, "label")]$, where n = # of pairs in a trial

C. Gaussian Mixture Model

The Gaussian Mixture Model (GMM) is a probabilistic model that assumes the data points come from a limited set of Gaussian distributions with uncertain variables. We use the mean and covariance matrix to characterize each individual Gaussian distribution. GMM has many applications, including density estimation, clustering, and image segmentation. For clustering, we can use GMM to group together data points that come from the same Gaussian distribution. The formula for calculating the Gaussian distribution is as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \text{ where}$$

$\mu = mean\ of\ x$

$\sigma = standard\ deviation\ of\ x$

GMM is an unsupervised classification algorithm which builds upon K-means instructions to predict the probability of classification for each instance. For this research, we use GMM to compare the gaze data of both "good" and "bad" trials; each GMM contains 2 good and bad trials, for a total of 4 trials for each game. The objective is to observe the strength of the difference in gaze data depending on the level of difficulty for each game, as well as the game itself. Will games with a "hard" label have a more distinct pattern than "easy" games? After initial GMM plots for each of the games, we conclude that it is difficult to observe any pattern due to the large amount of data being plotted, as shown in Figure 6.

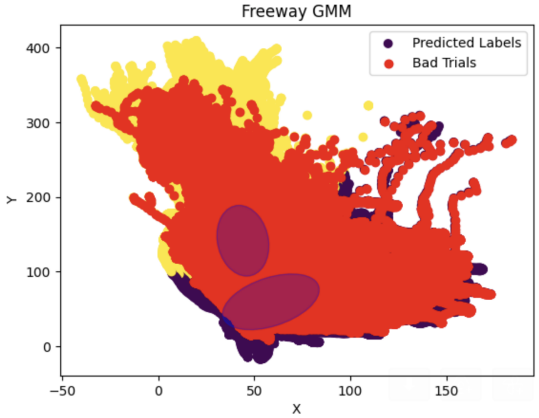

Figure 6: Initial GMM for Freeway

To mitigate this issue, we generate subplots by selecting a subset of the original dataset. This is done using random sampling, where the sample size is calculated first given an infinite population, and then adjusting it to the required population. The formula for calculating the sample size for an infinite population is as follows:

$$S = Z^2 \times \frac{p(1-p)}{M^2}, \text{ where}$$

$S = Sample\ size\ for\ infinite\ population$
$Z = Z\ score\ given\ confidence\ level$
$p = population\ proportion\ (assumed\ to\ be\ 50\% = 0.5)$
$M = Margin\ of\ error$

For this dataset, we assume confidence level to be 95% and let Z=1.96, and M=0.05 to account for 5% margin of error. Using these values, we get a value of 384.16 for the sample size. To adjust this for the required population, we use this formula:

$$\text{Adjusted Sample Size} = \frac{S}{1+\frac{(S-1)}{population}}$$

Each dataset has approximately ~1.5 million data points, resulting in the same sample size of 384. Using this value, we generate GMM subplots of each game using random sampling, as shown in Figure 7.

D. Silhouette Score

To study the strength of the clusters formed in the GMM plots, we calculate 10 silhouette scores corresponding to 10 randomly sampled subsets of a game, and calculate the average silhouette score. A higher silhouette score would indicate well-clustered data, thereby suggesting a strong correlation between the gaze data and performance. A box plot was then generated to show the distribution of the silhouette scores for each game to compare, as shown in Figure 8.

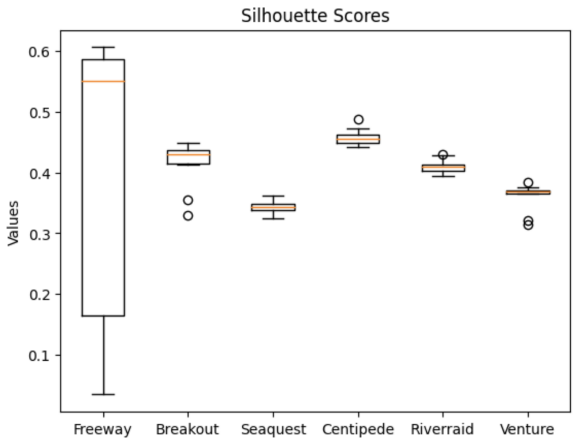

Figure 8: Box plot of silhouette scores for all games

IIII. RESULTS

We can draw a few conclusions from the generated GMM plots and box plot, the most notable being that there is a relationship between the gaze data and performance. The range of silhouette scores indicate moderate clustering, suggesting that there is a correlation between the gaze positions and game performance. An interesting finding is that higher silhouette scores were observed in “easy” games, meaning the clusters were more defined. Given that the scores were in the moderate range suggest that more featured learning needs to be done to provide more refined findings. Gaze data is certainly a contributing factor, however it is not the defining factor.

V. CONCLUSION

In conclusion, the video game industry's global impact surged during the pandemic, propelled by widespread accessibility. With revenues exceeding $206 billion, gaming targets a broad audience. This paper explores the correlation between gaze data and game performance, using the "Atari-HEAD" dataset, feature engineering, and Gaussian Mixture Models. The dataset, featuring 175 trials across 16 Atari games, underwent thorough feature engineering, resulting in a finalized chart. The evaluation phase employed the Cumulative Distribution Function (CDF) for game selection based on trials and performance and difficulty. Gaze extraction, GMM application, and Silhouette Scores provided insights. Results show a moderate correlation between gaze data and game performance. Notably, "easy" games exhibit clearer clusters. While gaze data contributes, it isn't the sole factor in determining game performance. This exploration enhances understanding of the interplay between human gaze behavior and gaming proficiency.

Acknowledgment

I would like to thank Prof. Jorge Ortiz for advising me throughout this research and introducing me to various machine learning techniques.

References

[1] Clement, J. “Topic: Video Gaming Worldwide.” Statista, 14 Nov. 2022, https://www.statista.com/topics/1680/gaming/#topicOverview.
[2] Jeong I, Nakagawa K, Osu R, Kanosue K (2022) Difference in gaze control ability between low and high skill players of a real-time strategy game in esports. PLoS ONE 17(3): e0265526. https://doi.org/10.1371/journal.pone.0265526
[3] Frank, Anders. “Eye Tracking Analytics Hits Esports.” Tobii, www.tobii.com/blog/from-f1-racing-to-league-of-legends. Accessed 26 Nov. 2023.
[4] Zhang, Ruohan, Zhuode Liu, Luxin Zhang, Jake A. Whritner, Karl S. Muller, Mary M. Hayhoe, and Dana H. Ballard. "AGIL: Learning attention from human for visuomotor tasks." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 663-679. 2018.
[5] Atari 2600 High Scores, https://www.jvgs.net/2600/top50.htm.