SIADS 699 Capstone Winter 2024

Settle it in the Cypher - Breakdancing Analysis

Sara Haptonstall (sarahapt@umich.edu), Sihyun Kyung (skyfront@umich.edu),
Sabrina Mahnke (smahnke@umich.edu)

Project Statement

Settle it in the Cypher (SIC) is a company that works to empower the global competitive breaking community to be able to have data-driven discourse about breaking as both sport and artform. With breakdancing entering the Olympics this year as a category, there has been more visibility for the sport. Our project aimed to conduct analysis on the breakdancing competition data that the Settle it in the Cypher team had webscraped and share our insights with the client to further their effort. The team also completed a video analysis model which predicted different types of breakdance moves which will be shared with the client for their future use. The last part of our project was to compile all the analysis that we could into a website/dashboard to showcase all of the results as well as have something the client could build off of in the future.

Methodology

The three main sections of our project are split into data analysis of the existing data used by the Settle it in the Cypher team, video analysis performed on the youtube videos of the competitions, and the dashboard/website with our results.

Data Analysis

The dataset used for the data analysis portion of the project consisted of web scraped data from the breakdancing competitions. This dataset included features such as the events, competitors, judges, and rankings for each round.

Scoring Systems Comparison

Competitive breakdancing had employed various judging systems through the years to assess participants, tailoring evaluations to reflect the diverse aspects of this dynamic dance form. Originally the winner was determined by majority vote by a panel of judges and in more recent years two different scoring systems had been used, Threefold which consist of 3 categories (Physical, Artistic and Interpretative) and Trivium which consist of 6 categories (Technique, Variety, Performativity, Musicality, Creativity, Personality). Experts in the sport theorize that the two scoring systems are representative of each other as shown in Figure 1.

Threefold	Trivium
Physical	Technique
	Variety
Artistic	Creativity
	Personality
Interpretive	Performativity
	Musicality

FIGURE1

	PC1	PC2	PC3
Creativity	0.362888	0.143561	0.713032
Musicality	0.403575	0.450374	0.269483
Performance	0.469467	0.239068	0.327888
Personality	0.409212	0.452774	0.295638
Technique	0.350112	0.704047	0.190013
Variety	0.441596	0.136793	0.433527

	Factor1	Factor2	Factor3
Creativity	0.584749	0.144549	0.356931
Musicality	0.669645	0.352689	0.148181
Performance	0.850926	0.149023	0.278467
Personality	0.728653	0.465250	0.282656
Technique	0.682145	0.707579	0.173523
Variety	0.840725	0.059744	0.537386

FIGURE 2

Considering this expert assumption that Threefold (three categories) is a dimension reduction from the six Trivium categories. We decide to utilize dimensionality reduction techniques that transform the

original Trivium (6) variables to three new variables. Specifically, We performed structure detection analysis using Principal Component Analysis (PCA), Exploratory Factor Analysis (EPA) and Confirmatory Factor Analysis (CFA) in order to test this theory.

We performed a 3 principal component (PCA) dimensionality reduction and we obtained an accumulated explained variance of 0.882. Simultaneously, we performed a 3 factor reduction (EPA) obtaining an accumulated explained variance of 0.789.

As we observe in Figure 2, the dimensionality reductions for both show a similar pattern which differs from the experts' assumptions.

Next, we performed CFA, which tests pre-defined hypotheses about the structure of the data. CFA begins with a predefined theoretical model that specifies how observed variables are expected to relate to Latent Variables (LV) that are unobserved. In our case, we are trying to explore the expert assumption (Model 1) on how the Trivium scores (observed variable) relate to a Threefold score (latent variable). Additionally, we created two more models (Model 2, Model 3) using the loadings from EPA and PCA as reference.

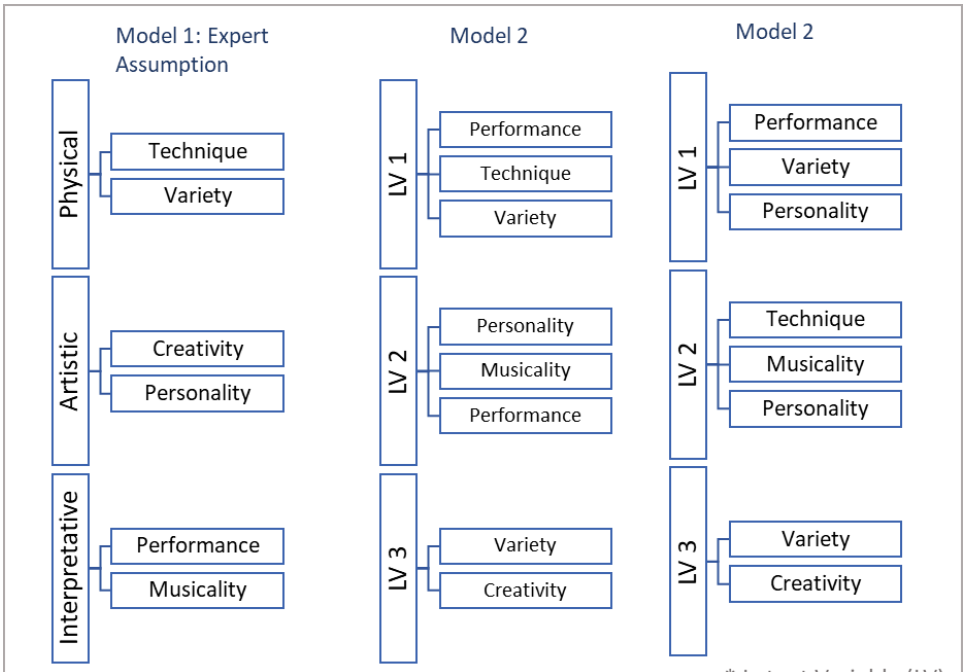


FIGURE 3: CPA Models

* Latent Variable (LV)

Upon examination of the CFA results, it became evident that Model 1 exhibited the poorest fit among the three models based on several metrics, including the highest chi-square value, highest Root Mean Square Error of Approximation (RMSEA), and lowest Comparative Fit Index (CFI), Goodness-of-Fit Index (GFI), Adjusted Goodness-of-Fit Index (AGFI), Normed Fit Index (NFI), and Tucker-Lewis Index (TLI) values. Additionally, Model 1 displayed the highest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, indicating inferior model fit compared to the other models.

In contrast, Model 2 emerged as the optimal choice, demonstrating the most favorable fit among the three models. This conclusion is supported by several key indicators, including the lowest chi-square value, lowest RMSEA, highest CFI, GFI, AGFI, NFI, and TLI values. Furthermore, Model 2 exhibited relatively lower AIC and BIC values compared to the other models.

Although Model 3 also displayed good fit according to certain metrics, such as RMSEA and AIC/BIC values, its overall fit indices, including CFI, GFI, AGFI, NFI, and TLI, were slightly lower than those of Model 2. Consequently, while Model 3 remains a viable representation of the scoring relationship, it does not offer the optimal fit achieved by Model 2.

In summary, these findings highlight that the Model 2 created by utilizing PCA and EPA better captures the relationships between the observed and latent variables. Therefore, Model 2 is a more robust data-driven depiction of the scoring systems relations.

Source	DoF	DoF Baseline	chi2	chi2 p-value	chi2 Baseline	CFI	GFI	AGFI	NFI	TLI	RMSEA	AIC	BIC	LogLik
Model 1	6	15	1526.418	0	5296.640	0.712	0.712	0.280	0.712	0.280	0.408	27.999	107.956	1.000
Model 2	4	15	72.127	8.10E-15	5296.640	0.987	0.986	0.949	0.986	0.952	0.106	33.905	124.522	0.047
Model 3	4	15	800.602	0	5296.640	0.849	0.849	0.433	0.849	0.434	0.361	32.951	123.568	0.525

FIGURE 4: CPA Models Statistics

Breakdancer Ranking

We were informed by our client (SIC) that currently there is not an official ranking methodology that identifies top competitors in the field, therefore we decided to use the battle data to implement a ranking system that they could continue maintaining as new battles occur. We decided to experiment with Elo and Glicko2.

Elo Ranking is a very popular ranking methodology used in chess. Each player is awarded some points at the start, in this case 1500. After a battle, points are transferred from the losing player to the winning player. The amount of points transferred depends on the probability of each player winning the battle, which is calculated based on their current ratings.The K-factor controls the rate at which a player's rating changes. Higher K-factors allow more points to be transferred per battle, leading to faster changes in ratings.

Glicko2 Ranking, an enhancement of the Glicko system, assesses player strengths in battles like chess. Each player has a rating (R), rating deviation (RD), indicating the reliability, and a volatility score (σ) reflecting expected rating fluctuations. Incorporating results against others to adjust both rating and RD. The system introduces more dynamics by accounting for the uncertainty in a player's rating through RD and adjusting for unexpected battle outcomes via volatility, making it very suitable for environments with varying player activity and engagement levels.

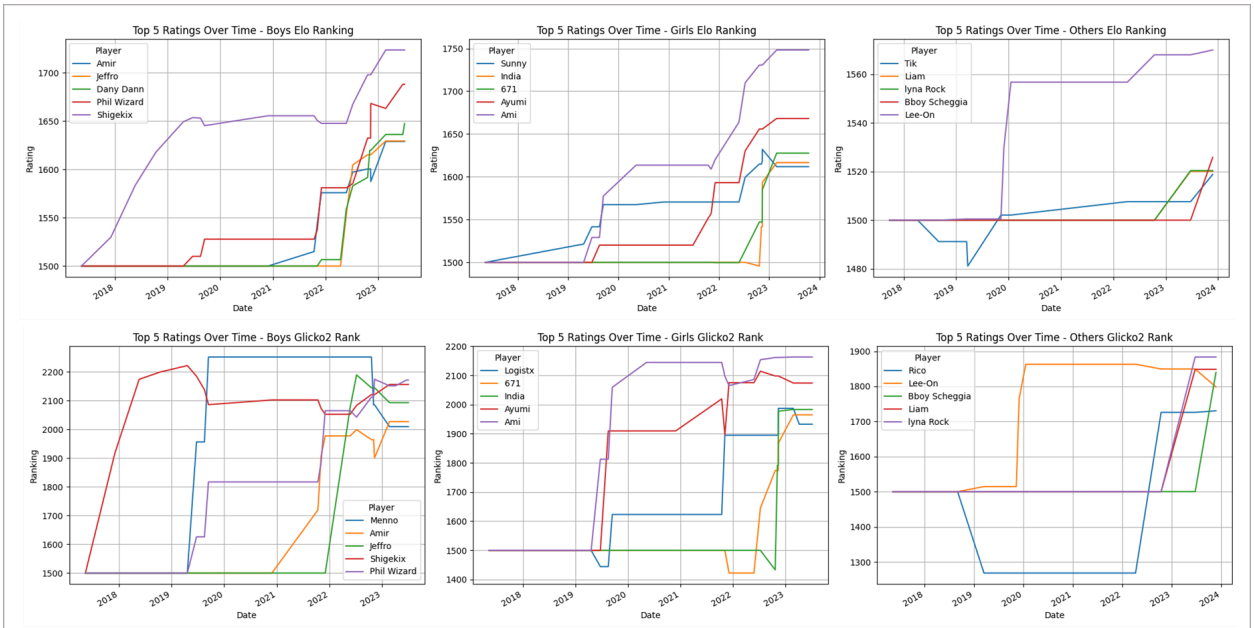


FIGURE 5: Elo and Glicko2 ranking for top 5 breakers by category

The Glicko-2 system has a more significant rating deviations than ELO, certain players experience significant shifts in rankings within short timeframes, which implies a more responsive system to recent performances. Handling of periods without change in the ELO rankings suggest inactivity, while in the Glicko-2 rankings, the RD would increase during these periods, impacting subsequent calculations, resulting in rate drops for some dancers whereas Elo rankings show a more gradual and consistent progression over time. Since Elo provides a more stable and gradual performance change over time, we suggested our client to use Elo as the initial ranking system.

Judge Bias

We analyzed the potential biases among judges in competitive breakdancing, focusing on geographic region and gender. For geographic region bias, a subset of our data from the 'WDSF BfG World Series - Kitakyushu 2023 - Japan' event was examined, comprising 83 breakers and 10 judges. Each judge's nationality was identified, revealing 10 different nationalities among the judges and 34 among the breakers. A T-test comparing judges' scores for breakers from the same versus different nationalities found a t-value of -1.035 and a p-value 0.30, which means we fail to reject the null hypothesis and there is no significant difference, indicating unbiased scoring based on nationality. But some judges' scores statistically rejected the null hypothesis which meant that there was significant difference.

Bias- Fair Judges based on T-test

Description	Fair	Bias	Total
Girls	48	10	58
Boys	14	8	22
Total	62	18	80

FIGURE 6: Accuracy on Validation by Epoch

Regarding gender bias, events were labeled as either female or male based on the event category. We performed a t-test for each judge, we performed a t-test to test that there is no difference between scores given to male or female dancers, if p-value is less than 0.05 we rejected this hypothesis and we considered that the judge is biased. Overall scores for male versus female breakers revealed bias among 18 judges, with 10 biased toward male breakers and 8 toward female breakers**.**

Judge Distraction/Fatigue

Next, we investigated the impact of judge fatigue while scoring multiple battles by analyzing the judges voting patterns. When counting the number of no interactions (represented by zero scores) in both Threefold and Trivium battles, we noticed that the percentage increased. When we look at the regression line we observed the average rate change (slope) by 0.04 (threefold), 0.01 (trivium) between scores.

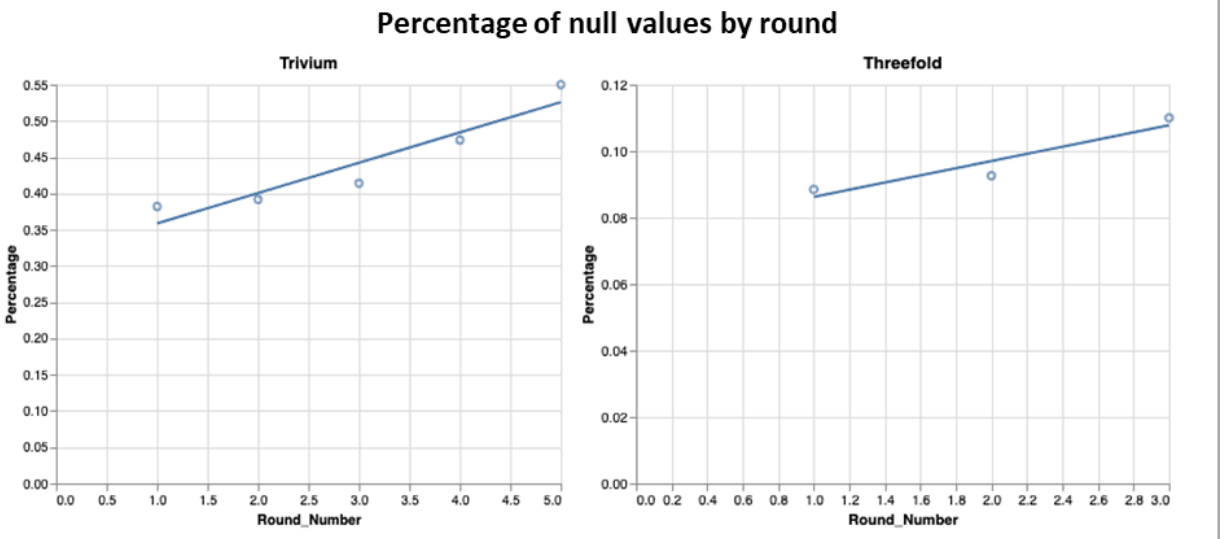


FIGURE 7: Regression Percentage of null values

We also look at the variance in scores between rounds, a Analysis of Variance (ANOVA) was conducted to examine the variance difference in the overall scores , the null hypothesis, statins that there are no differences in overall scores between rounds, was rejected based on a significant F-value (9..39) and

near-zero p-value (1.38 e-07). Subsequently multiple comparisons between rounds (Figure 8) show a decreased variance in the variance as the battle progressed, indicating signs of fatigue. Based on this it is recommended that after round 3, judges take a small break to maintain consistency on the scoring variance throughout the event. .

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
r1	r2	0.0205	0.9998	-0.2868	0.3279	False
r1	r3	-0.1899	0.919	-0.8088	0.4289	False
r1	r4	-1.868	0	-2.7804	-0.9556	True
r1	r5	-3.8789	0.1615	-8.5819	0.8241	False
r2	r3	-0.2105	0.8918	-0.839	0.418	False
r2	r4	-1.8885	0	-2.8075	-0.9695	True
r2	r5	-3.8994	0.1577	-8.6037	0.8048	False
r3	r4	-1.678	0.0002	-2.7425	-0.6136	True
r3	r5	-3.6889	0.2092	-8.4238	1.0459	False
r4	r5	-2.0109	0.7813	-6.793	2.7711	False

FIGURE 8: Mean Comparison

Score Normalization

We decided to normalize the scores in order to account for the varying scoring ranges among judges. For instance, different judges may assign scores within different ranges, such as Judge1 scoring between 1.33 and 19.67, while Judge2 scores between 1.33 and 11. Normalization standardizes the scores across judges, allowing for a more equitable comparison of performances.When we compare the normalized scores across rounds to determine if the same winner would emerge under standardized scoring conditions, we discovered that on average 18% of the rounds would have had a different winner.

Video Analysis

For video analysis, the team developed a model that aimed to predict three distinct breakdancing movements (toprock, powermove, and footwork).

Breakdancing Movement Prediction

For our dance movement video analysis we used a publicly available breakdance movements dataset (https://github.com/dmoltisanti/brace) which consists of videos where breakdancers execute several sequences. These sequences are labeled into segments based on their specific dance elements (toprock, powermove, footwork). For each labeled segment, they provide keypoints, which are specific [x,y] coordinates on a dancer's body position in a frame collected at one second intervals (additional information in the brace GitHub). Each segment has varying length from [105,(17,2)] to [1200,(17,2)] (this corresponds to [frames, (key points)].

We decided to implement several neural networks models, We started with some Long Short Term Models (LSTM) and Time Distributed models (Figure9) as baseline since this type of model is specifically helpful when dealing with image recognition, object detection and spatial patterns. These types of models are powerful for analyzing video clips, specifically for movement recognition, since they can extract temporal and spatial patterns within our data. Unfortunately, our Model1 and Model 1A with maximum length padding have undesirable accuracies of 0.361 and 0.362. Model1B with padding equal to the 90th

FIGURE 9: Baseline models

Model Description	Accuracy
Model1: Baseline LSTM, padding: max length	0.361
Model1A: Baseline Time Distributed, padding: max length	0.362
Model 1B: Baseline LSTM, padding 90th, frontal trucating	0.715

percentile segment length and truncating the clips and the beginning has a better performance. Therefore for all further experiments we will use this padding structure to fit our key points data.

In an attempt to increase the accuracy, we decided to augment our data by calculating several different distances and angles during each frame. Our hope is to be able to find significant differences based on the breakdancer position while performing these dance moves. As you can see in Figure 10, the three movements have a distinct structure. For example, toprock movements are generally more upright and involve a variety of steps and turns at the rhythm of the music. Footwork involves movements at a ground level using the hands for support and maneuvering the legs in sweeping circular motions close to the floor, while powermoves are more acrobatic and involve spins, flips and rotations such as headspins or windmills.

FIGURE 10: Breakdance three class movements

https://www.wikihow.com/Toprock

https://www.youtube.com/watch?app=desktop&v=xaw1tsYv4Mc

https://thekidshouldseethis.com/post/59010190235

We decided to calculate the leg angle, hip-knee-ankle angle, head to floor distance, and wrists to floor distance. Toprock has the highest mean leg angle, indicating a more upright posture compared to the others. Powermoves show a lower leg angle, suggesting bent legs during moves like spins. Footwork, while still lower than toprock, has a moderate leg angle, indicating a semi-crouched position close to the floor while the dancer hips and knees are almost parallel to the floor, typical for moves that involve ground-level maneuvers.. After calculating these angles we are able to observe differences based on each distinct dance movement.

FIGURE 11: Key point angles

Keypoints Mean Values			
Angle Distance Description	Toprock	Powermove	Footwork
Leg Angle (degrees)	130.42	97.55	110.16
Hip-Knee-Floor Angle (degrees)	88.36	22.88	1.93
Head to Floor distance (pixels)	0.971	0.761	0.508
Wrists to Floor Distance (pixels)	0.65	0.335	0.285

The descending order of mean head to floor distances—from toprock to footwork—indicates the level of elevation of the dancer's head from the floor. Toprock, being performed upright, has the greatest distance. Powermoves involve the head being closer to the floor, especially in moves like headspins. Footwork, which happens near or on the floor, has the shortest distance, signifying that the dancer's head is close to the ground.

FIGURE 12: Better Models

Model Description	Accuracy
Model 1B: Baseline LSTM.	0.715
Model 2: LSTM w/ sequence and frame	0.708
Model 3: LSTM w/Angles & Distances	0.789

Similar to the head-to-floor distance, the wrists' proximity to the floor decreases from toprock to footwork. In toprock, the wrists are not usually used for support, hence the greater distance. In powermoves and footwork , the wrists are closer to the floor as they often provide support and

balance. (Additional figures showing these angle distances can be found in the appendix.)

Considering the difference in angles and distances between the three movements we decided to incorporate that along with data about the image sequence and segment length to train two additional models, Model 2: LSTM with Keypoints and sequence count and segment length, We used a LSTM to process key points and a dense layer to process sequence count and segment length, the output from both are concatenated and passed through a dense layer, a dropout layer then a final dense layer and Model 3: LSTM with angles and distances; this model extracts sequential features (key points) using LSTM, then processes angles and distances using a dense layer and after that it concatenates the weights from both; this is then processes using a dense layer and a dropout layer to prevent overfitting . These models performed much better with a 0.708 and 0.789 accuracy on our test data.

Now that we have three models that have an accuracy above 0.708 on test data (Figure 12), we decided to perform hyperparameter tuning in order to maximize the improvement in our models. Hyperparameter tuning is a critical step when developing artificial neural network models, because parameters such as learning rate of the optimizer layers can significantly impact performance. We attempt to maintain a balance between overfitting and underfitting our best models in order to ensure that the final model generalizes well on new data. We experimented with a wide range of parameters to enhance accuracy and reliability of our models. We tuned our best models over a max on 10 epochs, 1 trial, a validations split of 0.20 for the data in order to maximize for accuracy

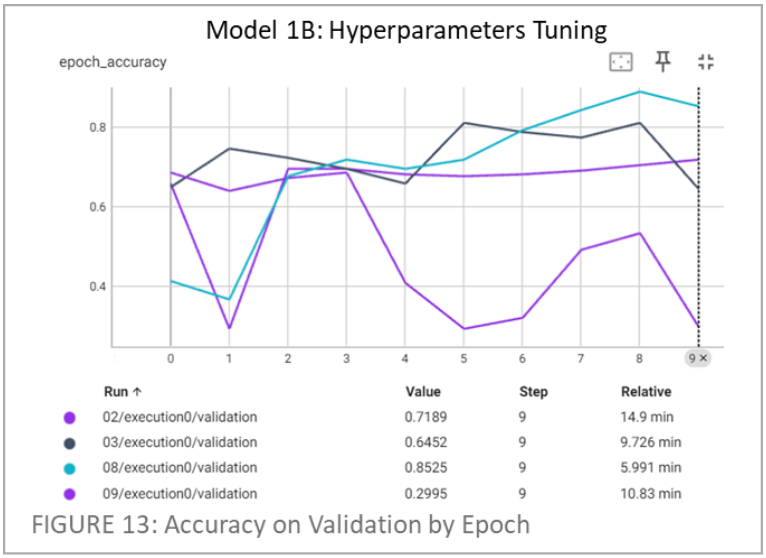


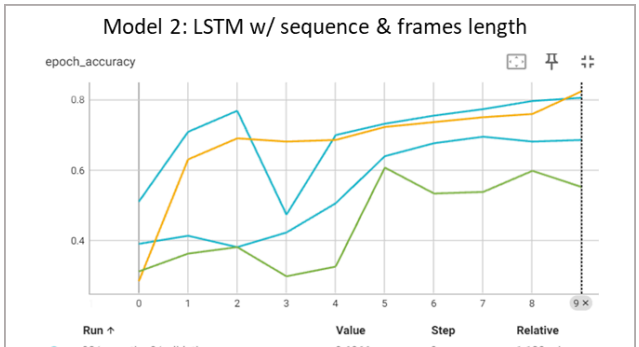
FIGURE 13: Accuracy on Validation by Epoch

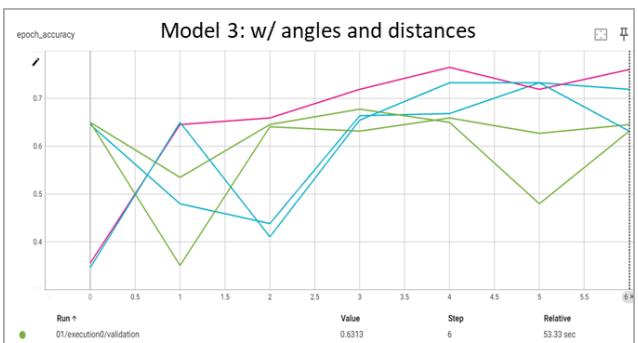
While tuning Model 1B, we noticed great variability on accuracy across epochs with the highest accuracy on the validation of 0.834. After selecting the best model parameter from epoch 8, we achieved an accuracy of 0.859 in our test set, a 0.144 improvement from our baseline model.

Model 2 exhibits slightly less volatility during the training epochs. The training time was 90% faster, accuracy in the validations set reached 0.825 and in the test set reached 0.827, which is a 0.159 improvement from the baseline model that includes sequences and frames.

For Model 3, the training progress shows a moderate level of variability in validation accuracy across epochs, the highest accuracy 0.760, in the validation data and 0.745 in the test data. This model performs worse than the initial model and the convergence is not clear.

In conclusion, the hyperparameter tuning process has notably transformed Model 1B, initially this model suffered from inefficiencies in learning from the complex temporal-spatial data patterns inherent in breakdancing video analysis. However, through hyperparameter optimization, we have fine-tuned critical aspects such as the number of LSTM units and the dense layer configuration, alongside introducing a dropout rate to combat overfitting. This highlights the vital role of hyperparameter tuning in machine learning, since it enables a model with less additional data to perform better than more complex counterparts.





Dashboard/Website

Using python library Streamlit, the team built a dashboard/website to showcase the different types of analysis performed by utilizing the sidebar function. This allowed for the website to be split into different pages labeled as Breaker Rankings, Breaker Match Simulation, Judge Profiles, and Our Data. Screenshots of each page can be seen in the appendix. The home page gives an overview of the Settle it in the Cypher team and their mission as well as lists of current projects, potential future projects, and GitHub repos either written or referenced by the team.

The Breaker Rankings page shown in appendix 4 as screenshot 1 shows the comparison of Elo rankings to Glicko rankings between the breakers over time. Each plot also includes tooltips showing the date and exact ranking value when you hover over the plots. Users will be able to select the group they would like to view (girls, boys, other) as well as how many of the top breakers they would like to view on the plots (with a maximum of ten).

The Breaker Match Simulation page (screenshot 2) takes the Elo model and predicts what the outcome of a match between two breakers would be with a probability. It prompts the user, like the Breaker Rankings page, to select which group they would like to see the results for. The website also allows for two different types of matches, a random selection of two breakers or for the user to choose the two competitors that they would like to see the outcome of.

The Judges Profiles page (screenshot 3) is a visual representation of the statistical distribution of judge's scores in each category. The website allows users to select from a list of judges and a table will appear showing the analysis of the chosen judge. This is composed of all of the data from events that the Settle it in the Cypher team has web scraped going back to 2019.

Using the Streamlit cloud, the team was able to link to the GitHub repository so that there is a public version of the website available. The website can be found at https://teambreakers.streamlit.app.

Evaluation Strategy

Overall, after meeting with the client, we were able to meet the requirements given to us as well as introduce other topics of interest that the client had not previously considered after analyzing the data.

While we are very satisfied with our findings in regards to the scoring system comparison, we are not surprised to find that these categories are closely related due to the artistic form of this sport. During our evaluation of ranking systems we noticed an accuracy of 0.57 average for both Elo and Glicko2, which is low considering a binary classification with equal probability. While our client was able to confirm that the individuals listed as "top 5 rank" were elite competitors in the field, additional feedback from the industry will be required and possible testing of other ranking methods that can achieve better results if there is an intention to use the ranking as a predictor of winners during battles.

We meticulously examine judges' scoring behavior in order to test several hypotheses regardless of fairness and impartiality. When working with the judging bias, the team was aware of the ethical dilemma that was determining whether or not a judge was showing bias. Since we don't know the actual thoughts and feelings of the judges, we needed to clarify that these results are purely objective based on the data and didn't reflect the actual feelings of the judges. For that reason, the team elected to leave the judge analysis off of the website so that it was not publicly available information that could be construed incorrectly. Moreover, we scrutinize distractions that may impact accuracy while scoring and through statistics modeling we were able to identify instances where fatigue or distractions impacted the overall outcome of the batter, we hope this will be valuable to our client to better understand external factors that may impact battle outcomes.

The breakdance movement classification that we developed demonstrated a significant capability to differentiate among various dance moves. After rigorous hyperparameter tuning, the model achieved a notable increase in accuracy. This model's success marks a step forward in applying machine learning to artistic sports, offering valuable insights and analytics to the breakdancing community.

Following our mentor's advice we performed several experiments using (VGG16) to try to determine the durations of rounds in videos, unfortunately after testing we were not able to develop a model that could separate the rounds' duration in the future we would like to expand upon that as well.

Statement of Work

Sara Haptonstall	Lead for Competitor Analysis, Analyst for Video Analysis
Sihyun Kyung	Lead for Judge Analysis, Analyst for Video Analysis
Sabrina Mahnke	Lead for Website Development, Analyst for Video Analysis

References

Dmoltisanti. (n.d.). *Dmoltisanti/brace: Brace: The breakdancing competition dataset for Dance Motion Synthesis*. GitHub. https://github.com/dmoltisanti/brace

Get started with tensorboard : tensorflow. TensorFlow. (n.d.). https://www.tensorflow.org/tensorboard/get_started

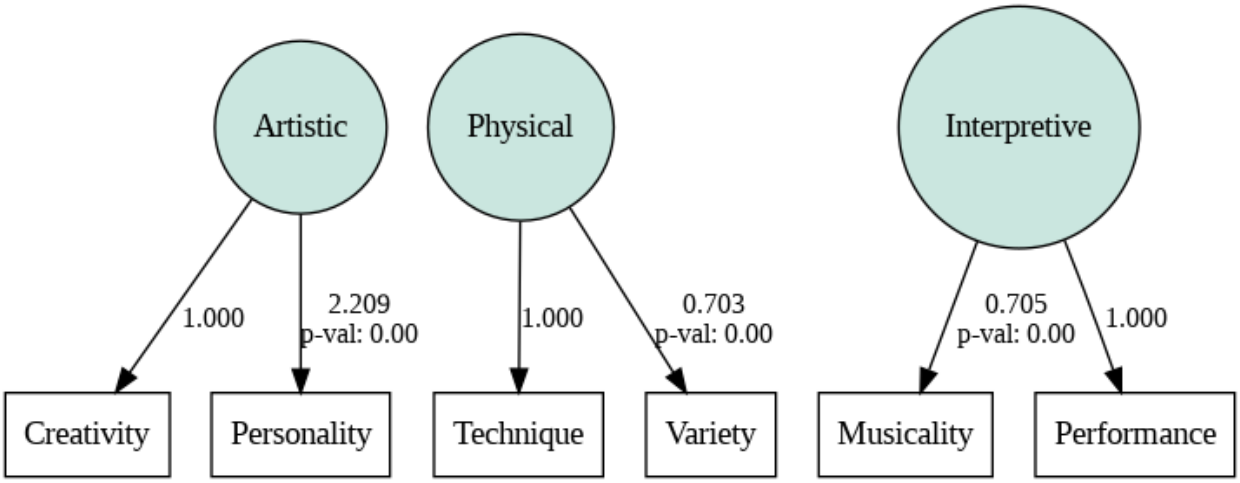
Team, K. (n.d.). *Keras Documentation: Getting started with Kerastuner*. https://keras.io/guides/keras_tuner/getting_started/

Wikimedia Foundation. (2024, January 21). *Confirmatory factor analysis*. Wikipedia. https://en.wikipedia.org/wiki/Confirmatory_factor_analysis#:~:text=In%20statistics%2C%20confirmat

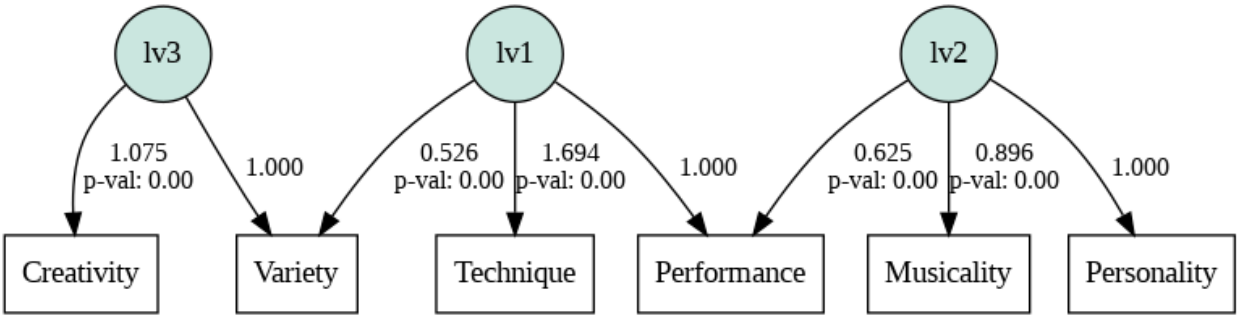
YouTube. (2020, September 10). *157 - What is tensorboard and how to launch it in a browser?*. YouTube. https://www.youtube.com/watch?v=PG4XGqUeYnM

Appendix 1: Confirmatory Factor Analysis Models Weights

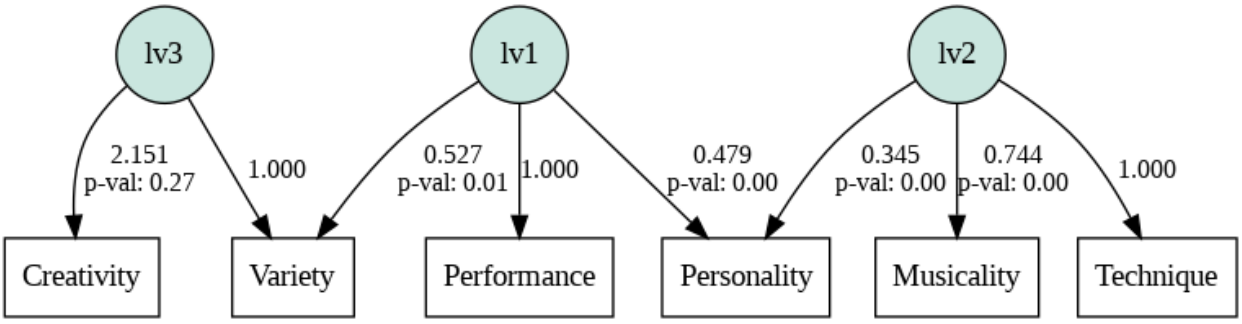
Model 1- Expert Assumptions



Model 2



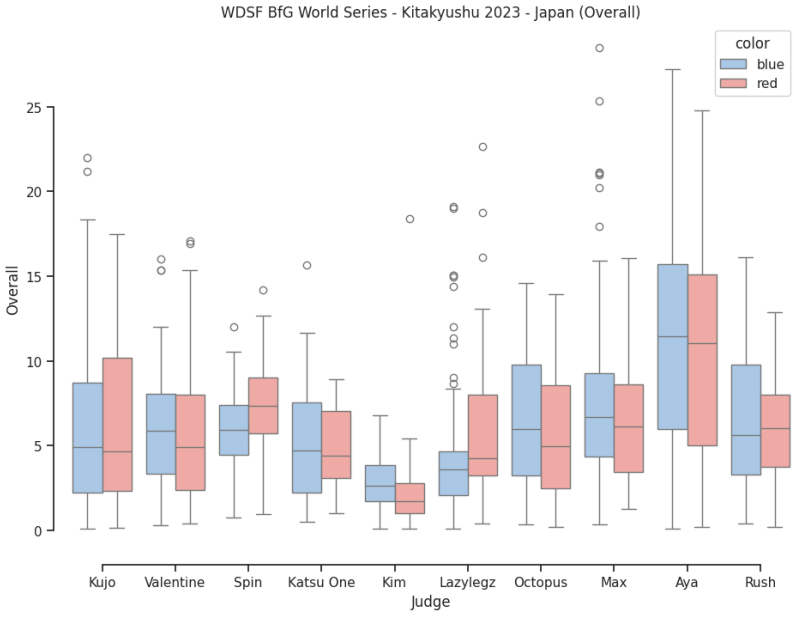
Model 3

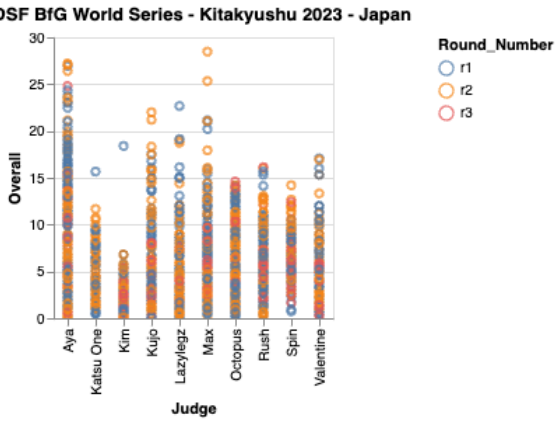


Appendix 2: Judge Analysis

Judge Profiles

We created individual judge profiles that include statistics as well as a time series analysis. Here are some examples of our visualizations that we used to explore the scores.

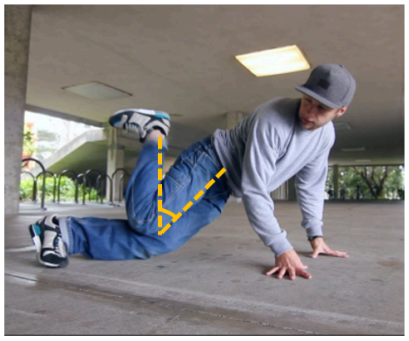




Appendix 3: Angle and Distance

Here are graphic representation on how the angles and distances were defined

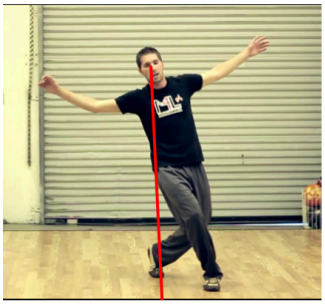
Hip-Knee-Angle Angle:



Hip-Knee-Floor

Head Distance to the floor



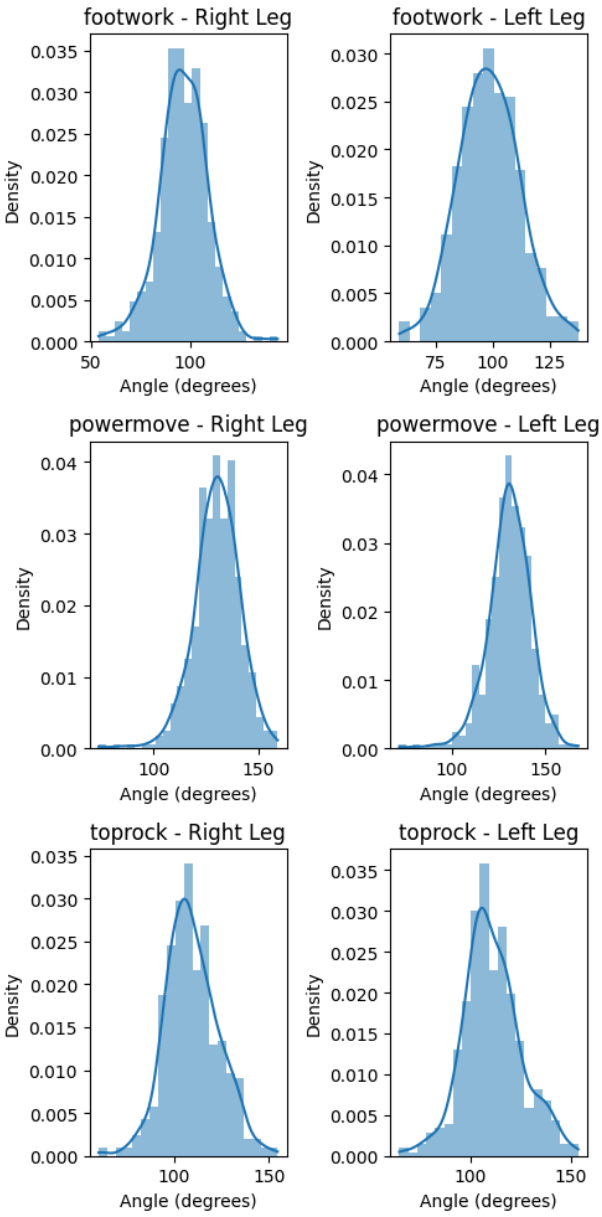


Wrists distance to the floor

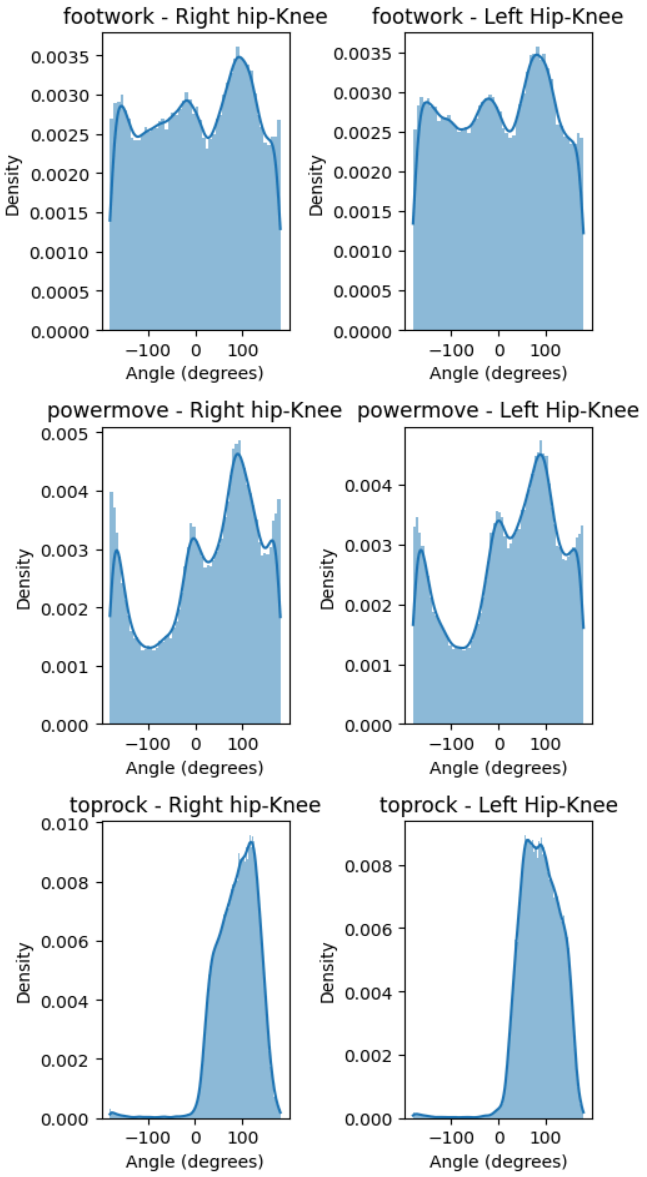


Means for angle and distances

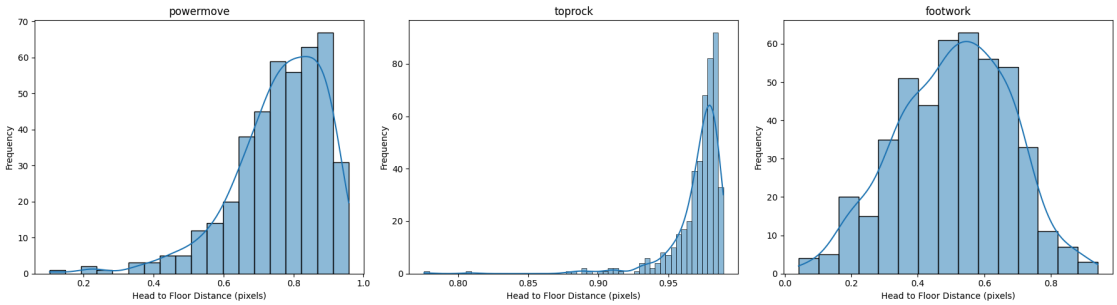
Hip-Knee-Angle Angle



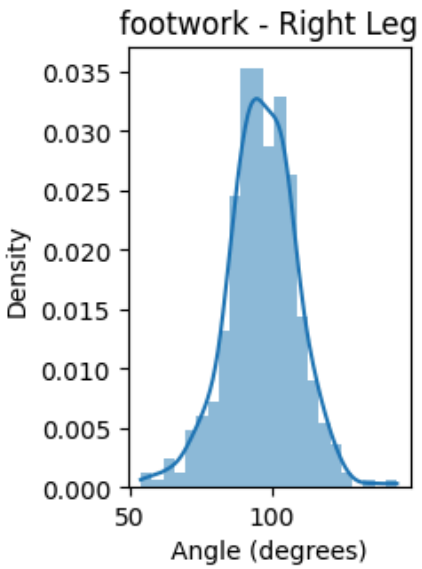
Hip-Knee-Floor

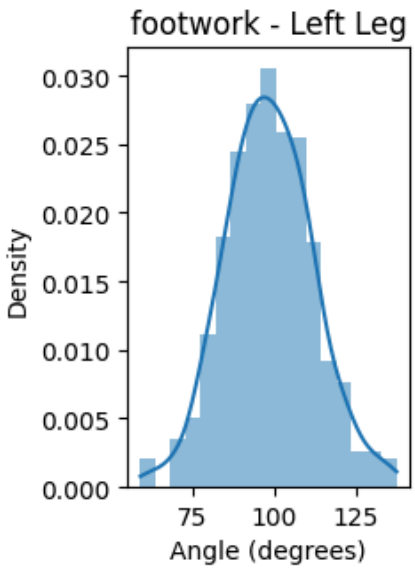


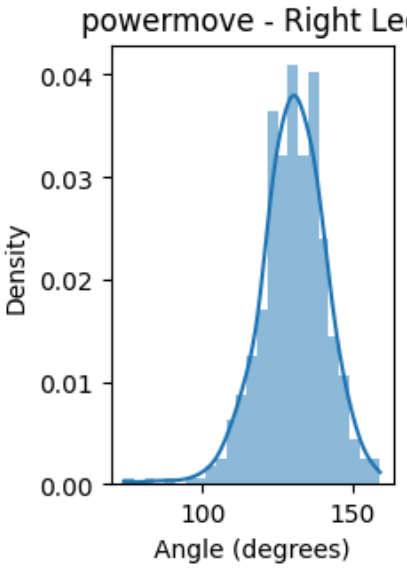
Head to floor distance

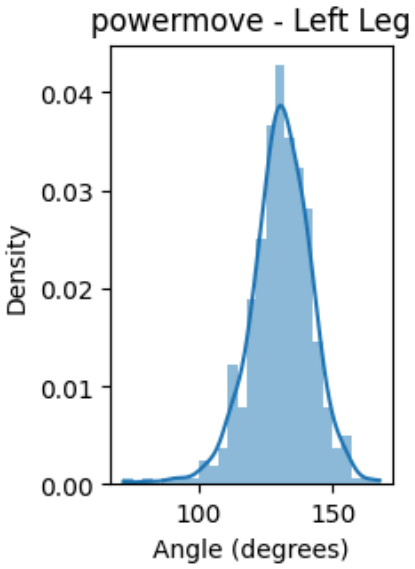


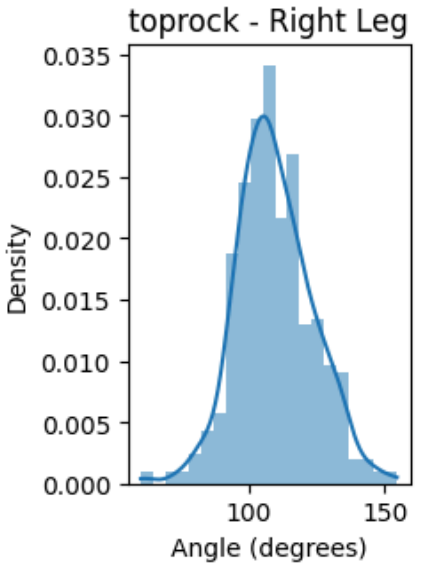
Wrists to floor distance

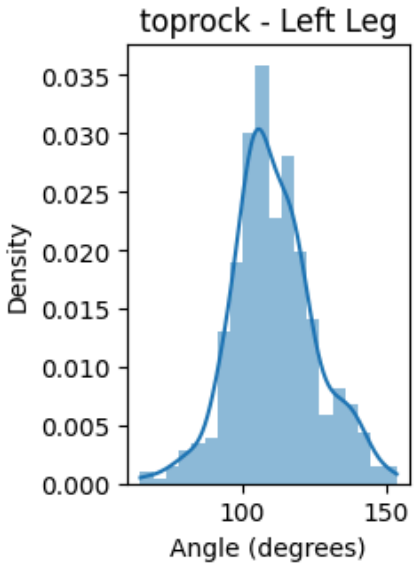

footwork - Right Leg
Density
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
50
100
Angle (degrees)


footwork - Left Leg
Density
0.030
0.025
0.020
0.015
0.010
0.005
0.000
75
100
125
Angle (degrees)


powermove - Right Leg
Density
0.04
0.03
0.02
0.01
0.00
100
150
Angle (degrees)


powermove - Left Leg
Density
0.04
0.03
0.02
0.01
0.00
100
150
Angle (degrees)

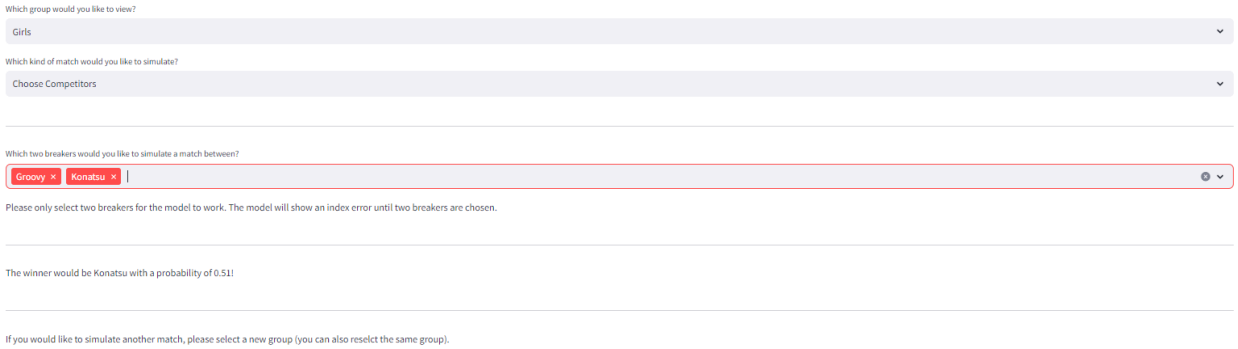

toprock - Right Leg
Density
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
100
150
Angle (degrees)


toprock - Left Leg
Density
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
100
150
Angle (degrees)

Appendix 4: Website Screenshots

Screenshot 1: Breaker Rankings

Screenshot 2: Breaker Match Simulations



Screenshot 3: Judge Profiles

Judge Profiles

Which judge would you like to view?

Kujo

You selected: Kujo

	Overall	Technique_Physical	Variety_Artistic	Performance_Interpretive	Musicality	Creativity	Personality
count	636	636	636	636	495	495	495
mean	6.89	3.3759	1.484	2.6284	1.3557	2.5325	1.4426
std	5.2781	2.245	1.6332	1.9544	0.9648	1.8268	1.0525
min	0.07	0	0	0	0	0	0
25%	2.715	1.8	0	1.4	0.67	1.4	0.67
50%	5.6	3	1.33	2.2	1.2	2.2	1.33
75%	9.885	4.6	2.13	3.8	1.87	3.4	2.13
max	32.33	11.67	11	10	4.93	10.2	4.8

This analysis is performed on all events included in the dataset. These events go back to 2019.