Machine Learning

Final Assignment

Classifying Cell Dynamics in *Drosophila* Development

Group 10:

Vincent Chan, Van Grando, Wasif Khan, Vasavdutta Purohit, Christopher Ringwood, Tim Rogalsky

August 12, 2024

Objective:

This project supports the research of cancer biologist Dr. Nicolas Malagon, Assistant Professor of Biology at Canadian Mennonite University, who investigates the developmental cell biology of *Drosophila melanogaster*, commonly known as the fruit fly. Our primary objective is to classify cell size oscillation patterns in normal development. This will provide a reference point for future comparisons to cancerous cell growth, with the long-term aim of early cancer detection using machine learning.

The sex comb, a male-specific group of bristles located on the foreleg of a fruit fly, is essential for courtship behaviors (Malagon et al., 2014). During development, tissues in the foreleg change in area, to facilitate the rotation of the sex comb into its functional position. Cells in the distal (lower) region of the foreleg expand while those in the proximal (upper) region contract. These overall trends are marked by complex oscillations in size, an evolutionary mechanism thought to eliminate unfit cells. Understanding these normal dynamics is crucial for establishing a baseline for comparison with pathological conditions like cancer (Malagon et al., 2018).

Machine learning is indispensable for this classification problem due to the complexity and subtlety of the oscillations. Traditional analytical methods struggle to capture the detailed spatial and temporal relationships in the data. By using unsupervised learning methods, such as clustering, we can group cells based on specific cellular parameters. This approach will help uncover hidden patterns and enhance our ability to analyze healthy developmental processes at a finer scale.

To summarize, our objectives are:

1. **Classify Cell Oscillation Patterns**: Apply machine learning to classify cell size oscillation patterns during fruit fly leg development, identifying both spatial and temporal patterns in cell size changes.

2. **Identify Key Features**: Determine the key features that contribute to the classification of cell dynamics.

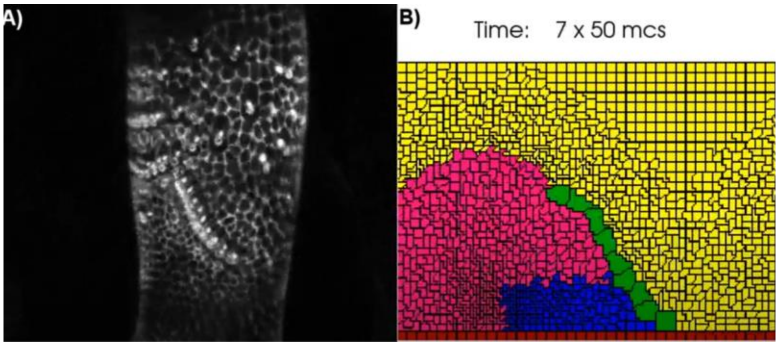


Figure 1. Image from Movie 1, showing sex comb and surrounding cells. A) Time-lapse movie. B) Simulation. Our study focuses on the development of the yellow epithelial cells proximal to the green rotating comb cells (Malagon, 2018).

Data Preparation:

The dataset for this project was provided by Dr. Malagon, in Excel format. It had been generated through advanced live imaging microscopy using ImageJ software, from three time-lapse movies, each documenting cell development in a different fruit fly. The full time-lapse of Movie 1 is available in Malagon's post on The Node (2018). For a screenshot, see Figure 1.

Cells were measured at 20-minute intervals in the proximal region of the foreleg, adjacent to the sex comb. The dataset consists of measurements from 131 cells, with time steps ranging from 39 to 48. Area and Delta (the change in area between time points) were provided. Our analysis focused primarily on Delta, as recommended by Dr. Malagon.

Missing values were encountered, indicating instances where accurate data could not be captured. These occurred exclusively at the heads and tails of time series, in consecutive blocks. Figure 2 shows the distribution of NaNs (Not a Number).

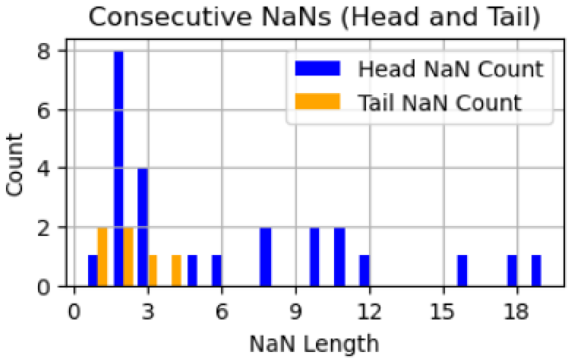


Figure 2. Distribution of consecutive missing values (NaNs) at the start (Head) and end (Tail) of time series for cell measurements.

To address missing data, we explored two strategies: imputation and non-imputation. One imputation technique involved dropping cells with more than 5 NaNs then filling the remaining gaps using a rolling average with window size of 15 to 17 time steps. Additionally, we experimented with imputing missing values using zeros and random values.

Recognizing that any form of imputation could introduce bias, our second approach avoided it altogether. We dropped time series with excessive NaNs, using various relative frequency thresholds. Then to compare cells we employed feature engineering or Dynamic Time Warping (DTW) (Berndt and Clifford, 1994) which can compare time series patterns even in the presence of missing data.

To assess whether additional data transformation was necessary, we examined the distributions of both Area and Delta. The Area distribution was right-skewed, but applying a logarithmic transformation yielded a distribution that was approximately normal. The Delta distributions were already roughly normal, centered around zero (as shown in Figure 3), which aligns with the oscillatory patterns typically observed in cell development. Many outliers were present in both datasets, some of them extreme. Although we experimented with further normalizing the data, this did not significantly impact the classification results.

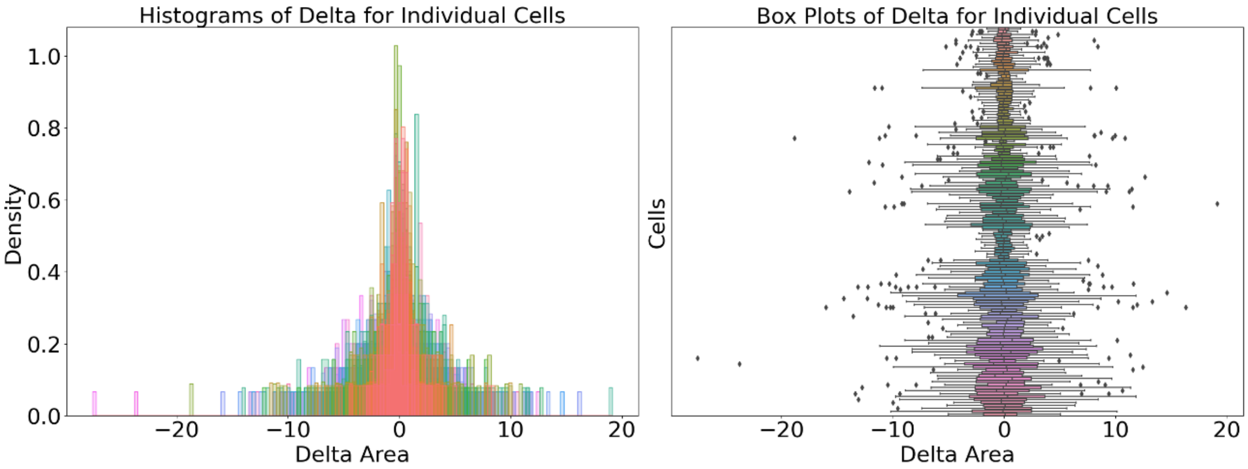


Figure 3. Distributions of Delta Area for Individual Cells—Left: Histograms showing approximately normal distributions with outliers. Right: Box plots highlighting the spread and variation across individual cells.

Model Design:

We used four machine learning models to classify the time series data. K-Means was chosen for its simplicity and efficiency, although it is more effective for well-separated, spherical groups than for oscillatory time series such as ours. The elbow method helped determine the optimal number of clusters by analyzing the sum of squared distances as a function of cluster count.

Hierarchical clustering was selected for its ability to capture nested clusters, allowing us to explore cluster relationships at multiple levels. The dendrogram visually represented the hierarchy, guiding our choice of cluster cut points.

Gaussian Mixture Models (GMMs) were used for their flexibility in modeling clusters of varying shapes and sizes, especially when clusters could be approximated by Gaussian distributions. We used the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) to determine the optimal number of components.

Finally, DBSCAN was utilized to identify clusters of varying shapes and densities. DBSCAN effectively manages outliers and noise, which was beneficial for our dataset. We experimented with different epsilon values and minimum samples to fine-tune the model.

We employed two metrics for classification: Euclidean and DTW. The Euclidean metric was used for both imputed time series and engineered features, providing a straightforward measure of similarity. For time series, however, it assumes perfect alignment between time steps, which limits its effectiveness for irregularly oscillating biological data.

To address these challenges, we used the DTW metric, for direct comparison of non-imputed time series. DTW aligns time series in a nonlinear fashion, minimizing the distance between corresponding points even if the series vary in length or experience temporal distortions. Unlike the Euclidean metric, DTW effectively captures similarities between time series with different lengths or missing data.

Model Evaluation:

Given the unsupervised nature of this classification task, we adopted multiple strategies to explore the data comprehensively. The following sections outline our various approaches and assess how well they revealed patterns in the data.

Clustering of Non-Imputed Data with Dynamic Time Warping

To compare the Delta series directly, without imputation, we applied four clustering models using the DTW metric.Time Series K-Means and DBSCAN failed to identify meaningful clusters, over a range of parameters. Hierarchical Clustering and GMM, however, identified two highly distinct clusters, with one cluster subdividing into two smaller clusters (Figure 4). The result was robust - clusters were virtually the same, for both models, whether or not longer series were truncated, and for a range of NaN thresholds.

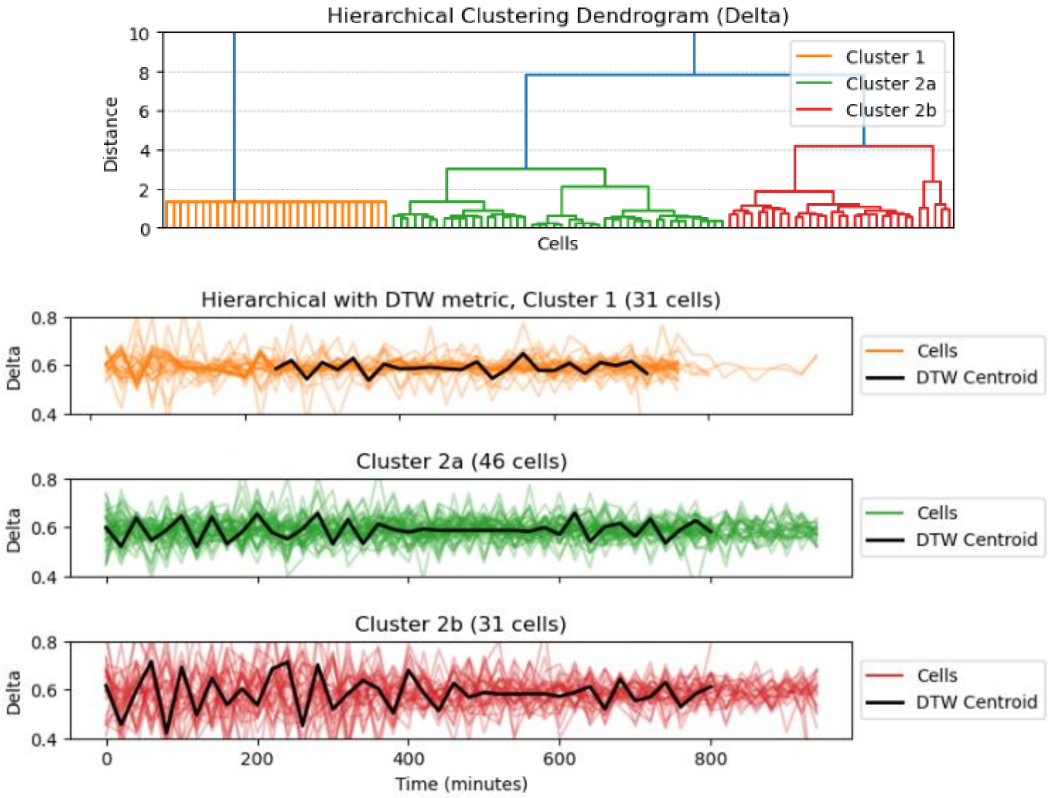


Figure 4. Hierarchical Clustering for non-imputed Delta time series, using the DTW metric. A) Dendrogram. B) Clusters at threshold distances 10 and 6.

Significantly, Cluster 1 was composed primarily of cells from Movie 1, suggesting a strong spatial or temporal pattern specific to this movie that differentiates it from the others. Cluster centroids, calculated with DTW barycenter averaging (DBA), are shown in Figure 4B. Note that these centroids are shorter than some time series because DBA can only be performed where data exists across all series within a cluster.

Cluster 1 shows the least variation, suggesting stable cell size oscillation and uniform developmental processes, likely linked to Movie 1. Cluster 2a has a broader range of Delta values, indicating more variable cell behavior, especially in earlier stages. The early stages of Cluster 2b have higher magnitude and a different temporal pattern compared to 2a, suggesting significant developmental changes before stabilizing.

When DTW analysis was applied to imputed data, clustering results were less distinct across models, and cells from Movie 1 were more dispersed. Imputation may have introduced noise, blurring the patterns observed seen in non-imputed data. This highlights the importance of preserving original time series data to accurately capture biological phenomena.

Clustering with Rolling Statistics

Rolling Statistics functions were applied to the pre-processed data to extract the mean, median and standard deviation of the cell area and the delta of area across all three datasets as well as the final concatenated dataframe. The results from this were saved in separate dataframes for each individual movie as well as in the combined dataframe. The rolling statistics window was set as 3 while calculating the final values.

K-Means clustering was applied to the rolling mean datasets for both area and delta series, initially for each movie individually (Mov1, Mov2, and Mov3), and then on a combined dataset. The optimal number of clusters was determined using the elbow plot and inertia metric. Cluster labels were plotted for each movie, and the silhouette score was used to assess the clustering performance. However, the clustering results were suboptimal, as indicated by a low silhouette score (< 0.5), suggesting poor cluster separation. Consequently, K-Means may not be the most suitable method for this study.

Next, hierarchical clustering was performed on the rolling mean statistics combined dataset of both area and delta series. The optimal number of clusters were determined by looking at the dendrogram (delta - Figure 5) and cutting it at a certain height to identify the number of clusters that exist at that level of similarity. The delta series dendrogram was cut at a little above 20 to get five optimal clusters. The image (Figure 5B) displays five line plots, each corresponding to a different cluster identified from hierarchical clustering for delta series. The two cells in Cluster 1 exhibit unique and irregular behavior, suggesting they may be outliers or represent a rare state. Clusters 2, 3, and 4 show varying trends of activity or growth, with Cluster 2 cells recovering and stabilizing, while Clusters 3 and 4 show a decline. Cluster 5's stability indicates that the majority of cells maintain a steady state throughout the observed period.

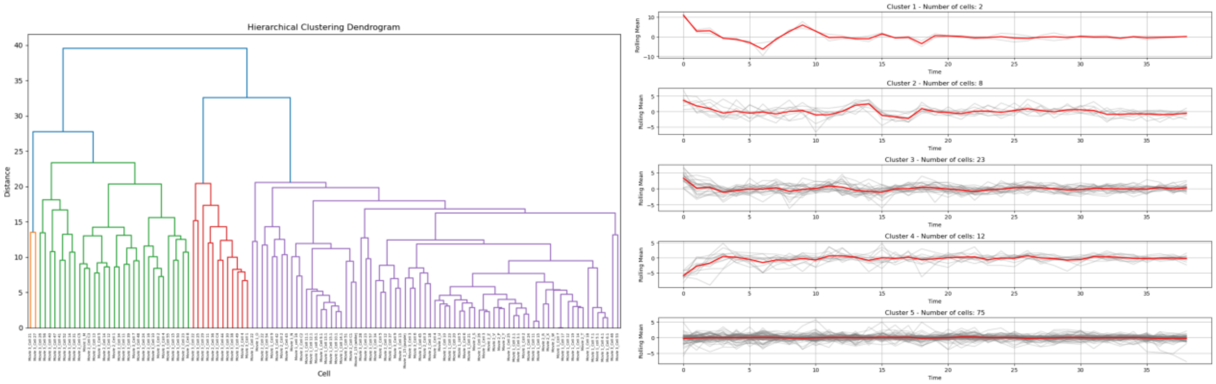


Figure 5. A)Hierarchical clustering dendrogram for delta series & B) cluster label line plots

Peak-Trough Analysis

The peak and troughs in Figure 6 represent the delta in the cell area over time. The preliminary findings of the three slides show slight differences in cellular behaviour. While both Movie 2 and Movie 3 show stable and slower oscillations, Movie 1 displayed slightly higher variability and cycling. These drastic oscillations could suggest that cells may be going through abnormal growth patterns similar to that of a cancerous state. However, the 20 minute intervals used in the data could skew the minute changes not being recorded and would require further testing parameters or data to better determine patterns associated with early cancerous states. While the changes are small, it does provide insight into possible phases of growth and contractions on a cellular level. The initial data does imply that there are varying differences even in healthy cells but further data could help decipher drastic oscillations between fit and unfit cells.

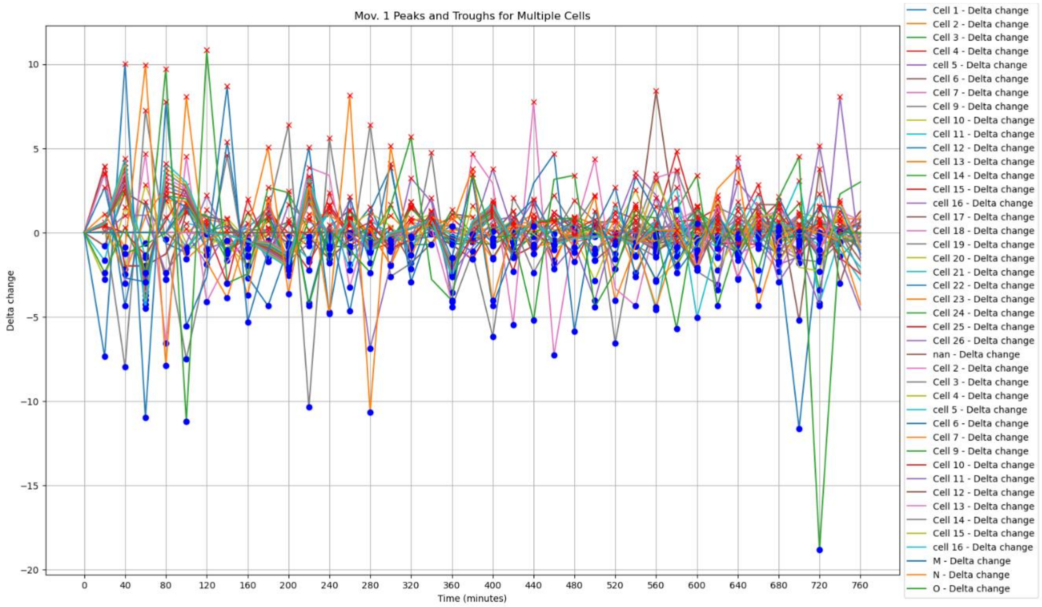
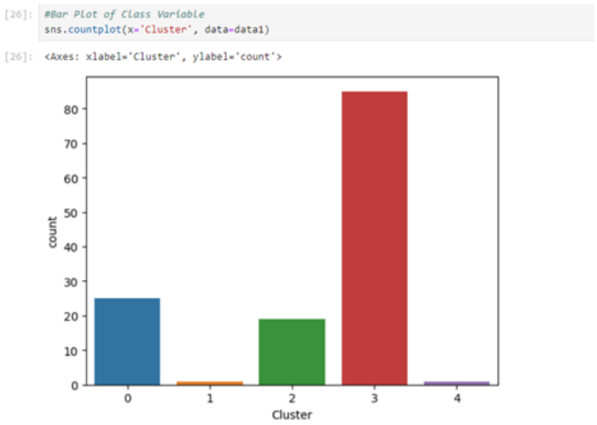


Figure 6. Peak and Trough in Movie 1

Clustering of Zero-Imputed Data with Euclidean Metric

The notebook performs an essential raw data inspection and analysis series to series using the Euclidean method where all three movie datasets are concatenated, addressing missing values by filling NaNs with zeros. The notebook also outlines the selection of machine learning models i.e. K-means, Dendrogram, and DBSCAN.



```
[26]: #Bar Plot of Class Variable
      sns.countplot(x='Cluster', data=data1)

[26]: <Axes: xlabel='Cluster', ylabel='count'>
```

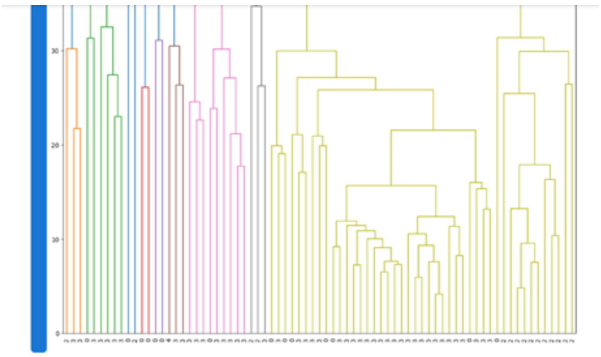


Figure 7. Dendrogram and K-means

The Dendrogram and K-means clustering chart above reveal at least five distinct clusters. If the dendrogram is cut at a height of 20, several distinct clusters would emerge, each containing similar data points. The hierarchical merges show that some clusters only merge at higher levels, indicating their dissimilarity from others. For example, clusters merging above a height of 40 are significantly less similar to other clusters. The height of the branches in the dendrogram reflects the dissimilarity measure; higher branches indicate greater dissimilarity between clusters. Notably, one cluster appears to be dominant, please refer to Cell_Size.ipynb for more details.

Clustering of Randomly Imputed Data with Feature Engineering

The next step involved data imputation in the cell datasets using a random imputation method, where missing values were replaced with randomly selected values from the same row, ensuring consistent data distribution while ignoring zeros. This method, effective for randomly distributed missing data, preserves the dataset's inherent characteristics. Feature engineering followed, with new features created using descriptive statistics (e.g., mean, median, standard deviation) and autocorrelation, adding insights into data central tendency, variability, distribution shape, and temporal dependencies. Also Monvie number was included as a feature. Scatter and density plots, alongside PairGrid visualizations, revealed clean cluster separations, though some overlap remains, suggesting effective but not perfect clustering potential.

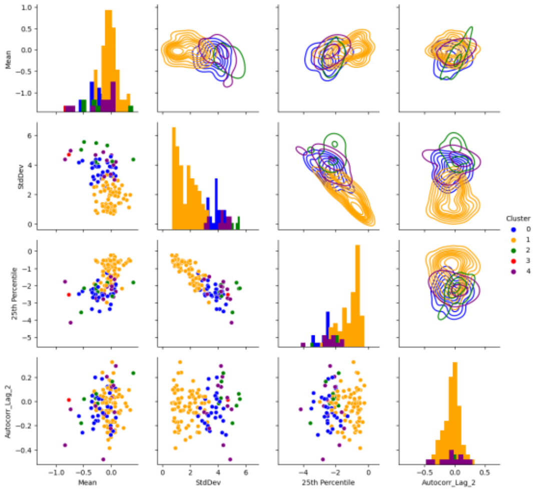


Figure 8. DBSCAN

DBSCAN was applied to the imputed dataset with engineered features, clustering data points based on density. The key parameters are Epsilon (eps), the maximum distance between neighbors, and Minimum Samples (min_samples), the minimum points to form a cluster. To determine the optimal eps, a plot of the nearest neighbors' distances is analyzed for a "knee" point using the NearestNeighbors object. Refer to 1-2024.08.11-Cell_Size.ipynb for details.

After that we create a DBSCAN object with the parameters found and visualize the clusters. Each point is colored according to its actual label. For comparison, each instance is drawn with a marker according to the label found by the clustering algorithm.

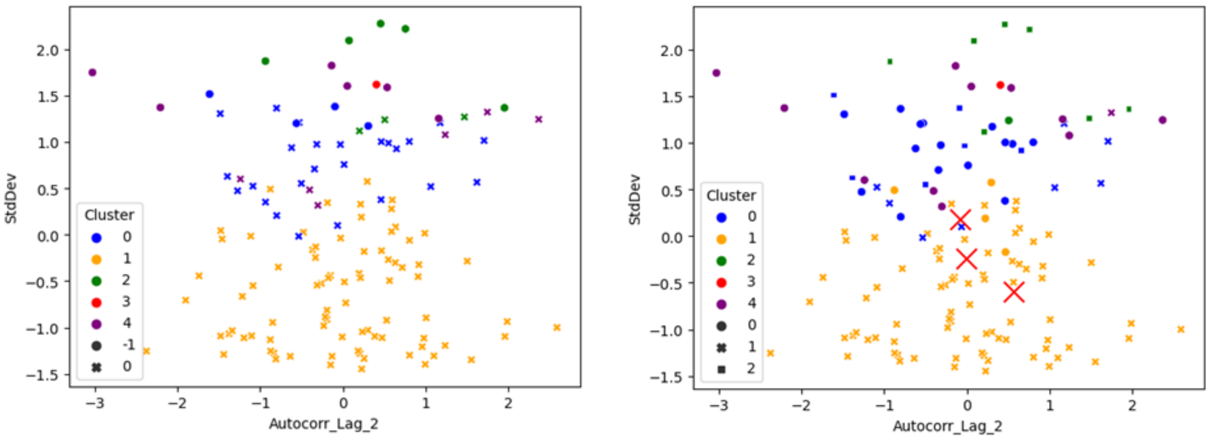


Figure 9. Scatter plots

The scatter plots compare K-means and DBSCAN clustering, highlighting three clusters with K-means and two prominent clusters with DBSCAN. This analysis provided insights into the dataset's structure, revealing natural groupings and potential for classification. These methods help understand patterns in cell area changes, facilitating classification into distinct clusters. Notably, one distinct cluster is associated with Movie #1, indicating its uniqueness compared to Movie #2 and Movie #3.

Clustering of Rolling Average Imputed Data with STL Decomposition

In this approach, feature-based modeling was conducted using Seasonal-Trend Decomposition Using LOESS (STL). Since STL Decomposition requires values at all time steps, missing data were imputed using a rolling average and cells with excessive missing values were removed. Then, using a period of 3, we extracted trend, seasonal, and residual components from Area and Delta as engineered features.

K-Means clustering with the Euclidean metric was applied to both Area and Delta trends.Cells were color-coded according to their assigned cluster labels for visualization. Elbow curve analysis for Area showed a clear elbow at k=2. This is confirmed in Figure 10 which shows the pronounced and separate clusters. For Delta it was more ambiguous, and we used k=3 as an approximation.

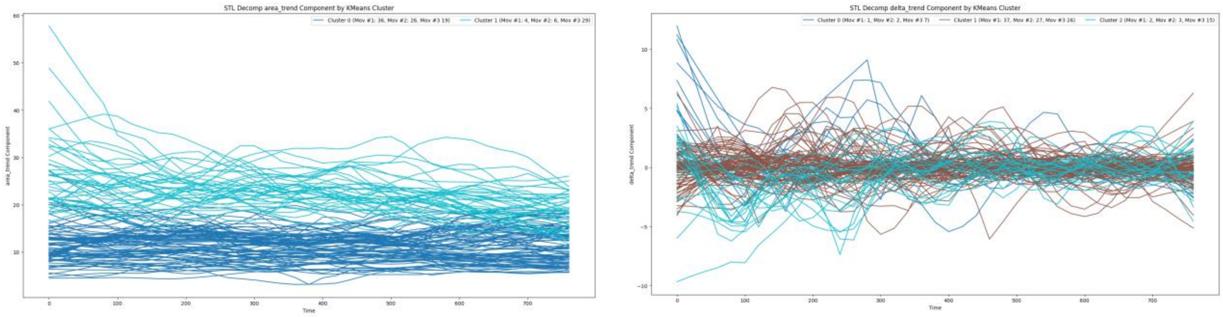


Figure 10. STL Trend clusters using K-Means. A) Area (k=2). B) Delta (k=3).

In the Delta trend, Cluster 1 emerged as the dominant group, containing 90 cells with relatively equal representation across all three movies. In Area Trend, Cluster 0 was the largest group with 81 cells, again showing a balanced contribution from each movie. Notably, for the Area trend Cluster 1 is predominantly composed of cells from Movie 3, suggesting that unique experimental or biological factors may have influenced this particular specimen. Finally, as has been observed with other approaches, the cells from Movie 1 were grouped together in both cases.

Conclusion:

In this study, we aimed to classify cell size oscillation patterns during fruit fly leg development using various machine learning models. Robust results were obtained with non-imputed data, which provided distinct and reliable clusters across several models. By contrast, there was much less overlap between clusters formed with imputed data, suggesting that imputation may have introduced enough noise to make it difficult to identify clear patterns. Despite this, one consistent finding is that a variety of models and approaches grouped cells from the same movie together, suggesting that developmental factors unique to each specimen may significantly influence cell Area and Delta over time.

Although we did not create a definitive model for identifying healthy cell development, our analysis highlighted the importance of preserving original time series data and the effectiveness of unsupervised learning in uncovering meaningful patterns. Future work could involve exploring additional features and testing on larger datasets, to improve classification accuracy and robustness.

References

Berndt, Donald J., and James Clifford. 'Using Dynamic Time Warping to Find Patterns in Time Series.' *AAA1-94 Workshop on Knowledge Discovery in Databases*, vol. 10, no. 16, pp. 359-370. 1994. https://cdn.aaai.org/Workshops/1994/WS-94-03/WS94-03-031.pdf.

Malagon, J. 2018. “Sex Combs in Motion - the Node.” The Node. November 14, 2018. https://thenode.biologists.com/sex-combs-in-motion-using-computer-simulations-and-mathematical-modeling-to-study-the-evolution-of-morphogenesis/research/.

Malagon, J., Ahuja, A., Sivapatham, G., Hung, J., Lee, J., Muñoz, S., Atallah, J., Singh, R., and Larsen, E. (2014). Evolution of Drosophila sex comb length illustrates the inextricable interplay between selection and variation. *PNAS*. September 30, vol. 111 no. 39, pp. E4103–E4109. https://doi.org/10.1073/pnas.1322342111.

Malagon, J., Ho, E., Ahuja, A., Singh, R., Larsen, E. (2018). Rotation of sex comb in Drosophila melanogaster requires precise and coordinated spatio-temporal dynamics from forces generated by epithelial cells. *PLOS Computational Biology*. Vol. 14 no. 10 pp. E1006455. https://doi.org/10.1371/journal.pcbi.1006455.