Classification and Analysis in Flow Imaging Microscopy

Tristan Pedro

TEXAS STATE UNIVERSITY

MEMBER THE TEXAS STATE UNIVERSITY SYSTEM

TABLE OF CONTENTS

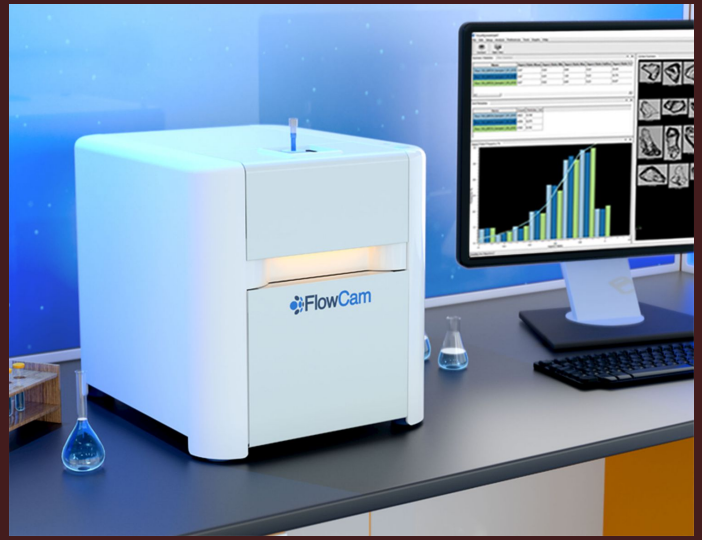
TEXAS STATE UNIVERSITY

INTRODUCTION AND BACKGROUND

TEXAS STATE UNIVERSITY

Project Overview

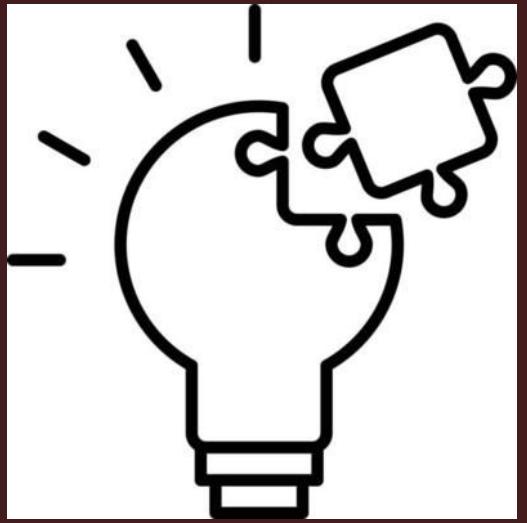
- Utilizing high-resolution FlowCam(hardware) imaging to automate algal cell classification.
- Preprocessing images to isolate individual cells and eliminate irrelevant data.
- Developing a machine learning model for accurate algal species/group classification.
- Aiming to enhance efficiency, accuracy, and understanding of algal diversity.



TEXAS STATE UNIVERSITY

Problem Statement

- Challenge in Aquatic Life Sciences:
  - Addressing the time-consuming and error-prone manual process of algal cell counting
- Opportunity with FlowCam Technology:
  - Leveraging FlowCam's detailed imaging to introduce automated analysis.
- Project Goals:
  - Creating an efficient ML model for classifying algal species.
  - Transforming algal diversity analysis in lab settings.



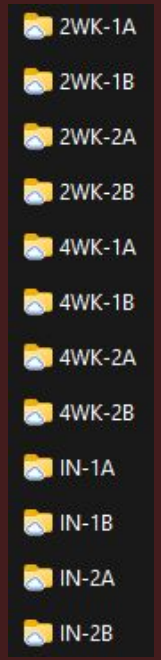
DATA COLLECTION AND PREPROCESSING

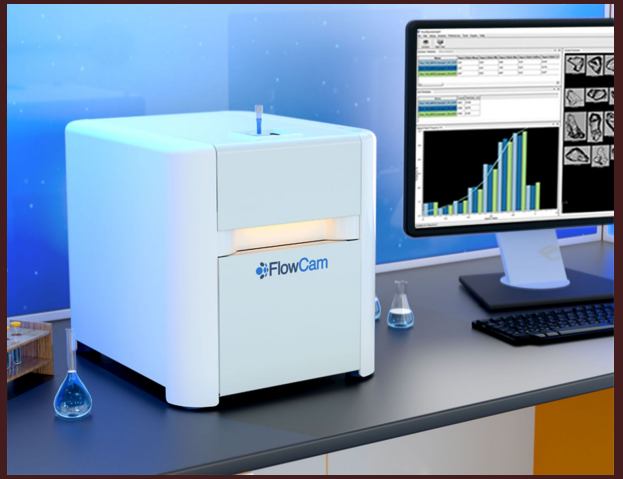
Dataset Description

Source of Data: Captured from FlowCam.

Diversity in Data: Wide variety of algal cell images, varying in size, shape, and color.

Volume and Organization: Four week experiment on two cultures split into two sets.





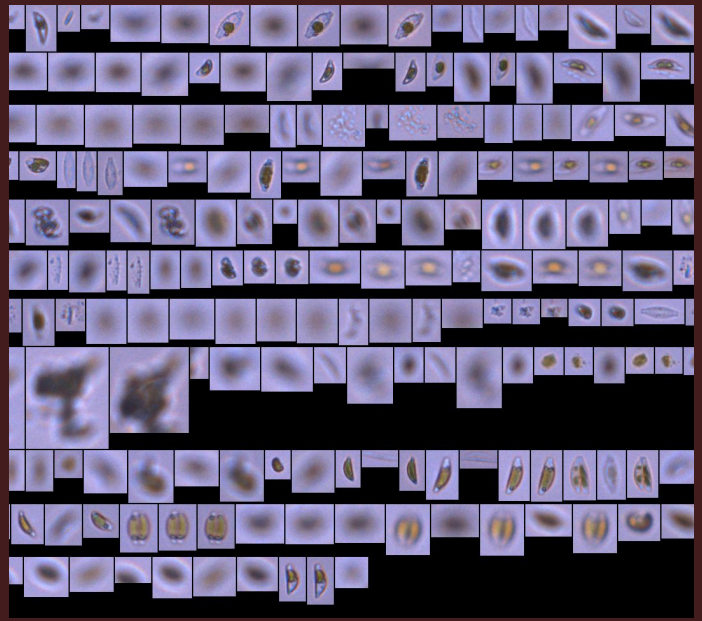
TEXAS STATE UNIVERSITY

Data Collection

High-Resolution Imaging: FlowCam provides high-resolution images of algal cells from fluid samples.

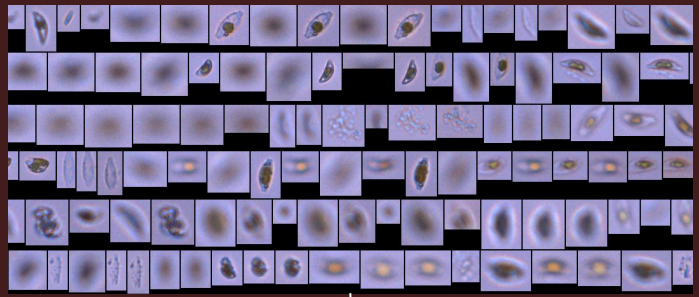
Advanced Imaging Technology: Uses optics and imaging tech to capture cells in flow.

Unique Image Format: Images presented as a series, with multiple cell images per file, accompanied by a '.lst' file containing corresponding data.

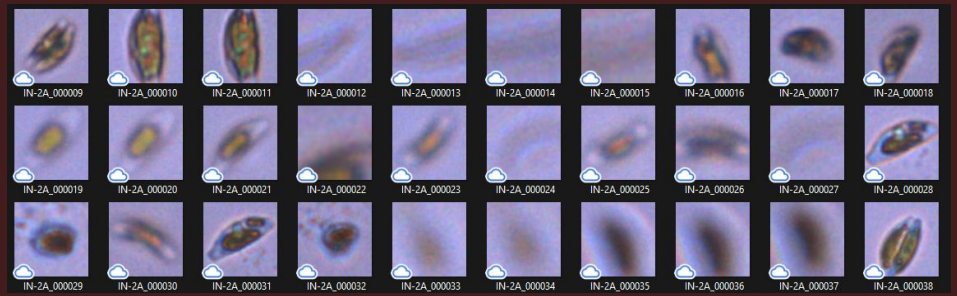


Data Preprocessing

1) Chop up grouped images into individual images
2) Filter out images that are obviously empty
3) Filter out images that are partially showing
4) Match images to similar images
    a) NOTE: Solutions will be discussed during EDA section



INDIVIDUAL IMAGES IN DIRECTORY



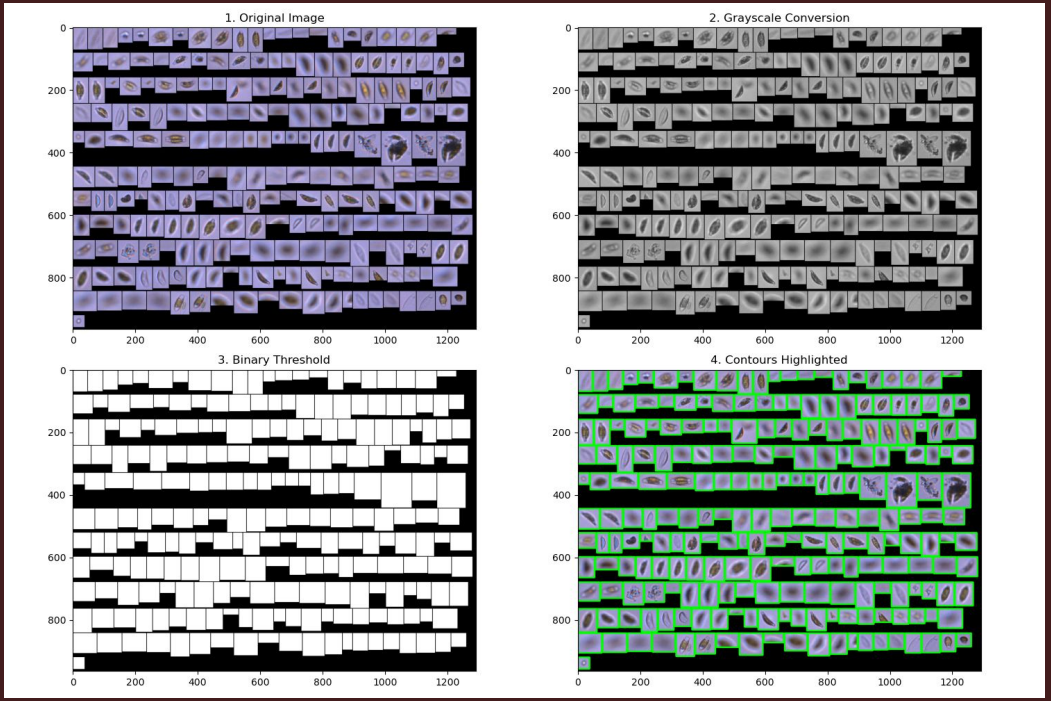
TEXAS STATE UNIVERSITY

EXPLORATORY DATA ANALYSIS

TEXAS STATE UNIVERSITY

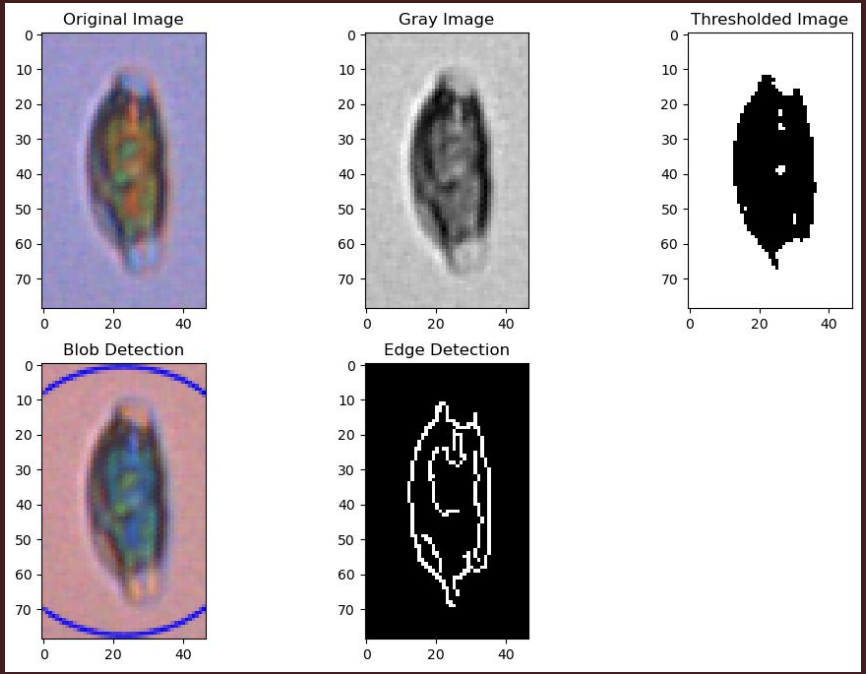
Preprocessing Solution cont.

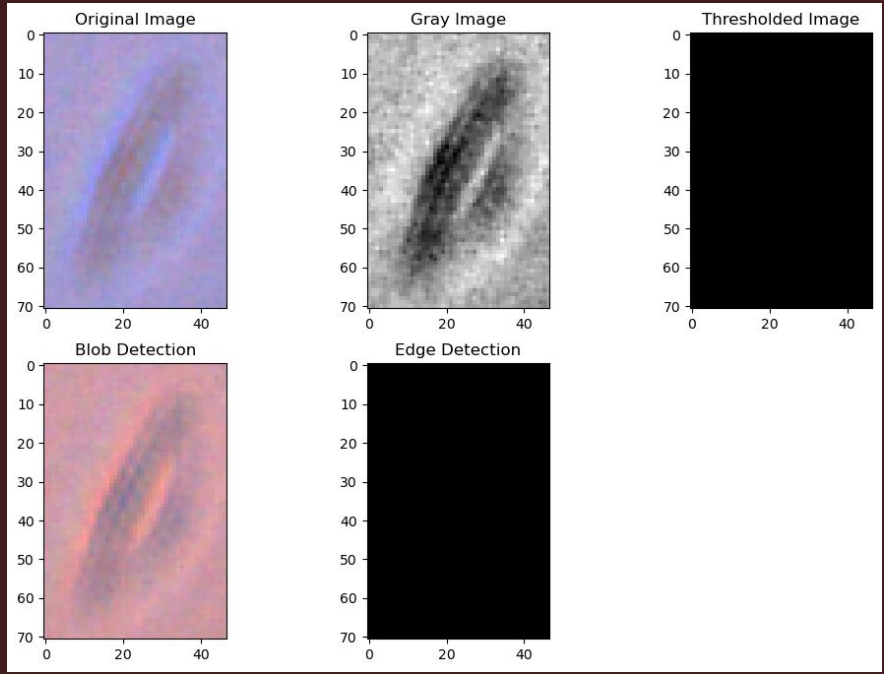
- **Contours** are continues lines or curves that bound or cover the full boundary of objects in an image



TEXAS STATE UNIVERSITY

Preprocessing Solution cont.

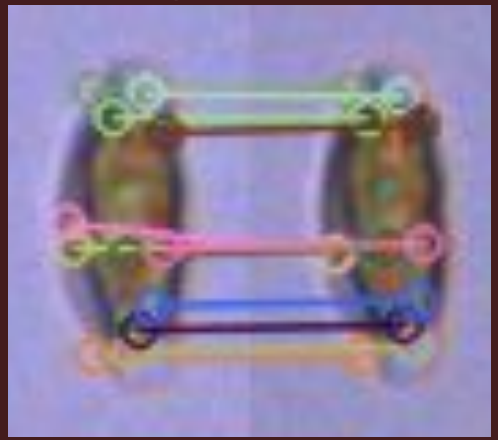




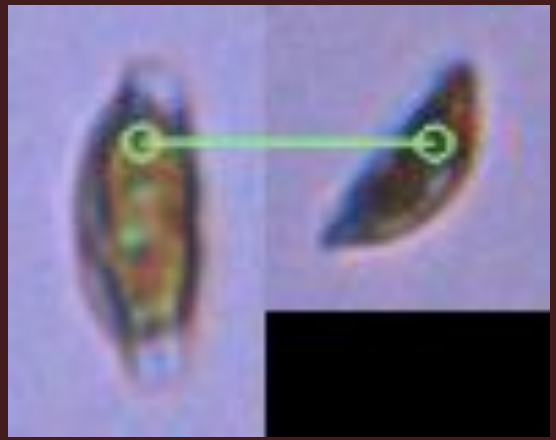
Preprocessing Solution cont.

- Feature matching finds corresponding parts in different images
- Utilized Scale-Invariant Feature Transform algorithm
- SIFT algorithm detects and matches key points invariant to scale and rotation.
- High number of consistent key points between images indicates a match.
- Effective for comparing and selecting the best image within a group.

Group of same cells

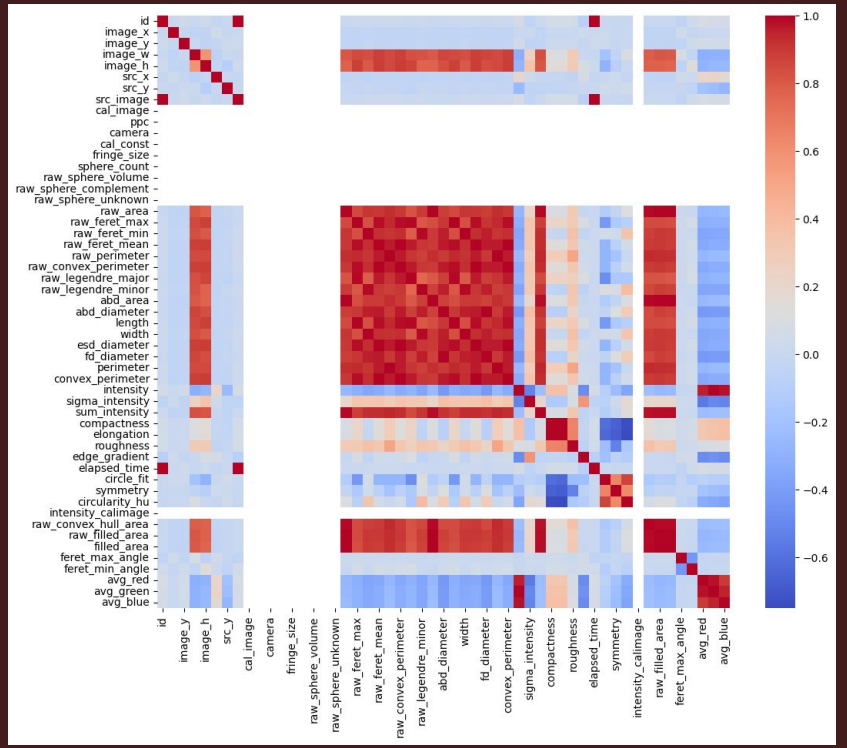


Different Grouped Cells



Correlation & Preprocessing Solution cont.

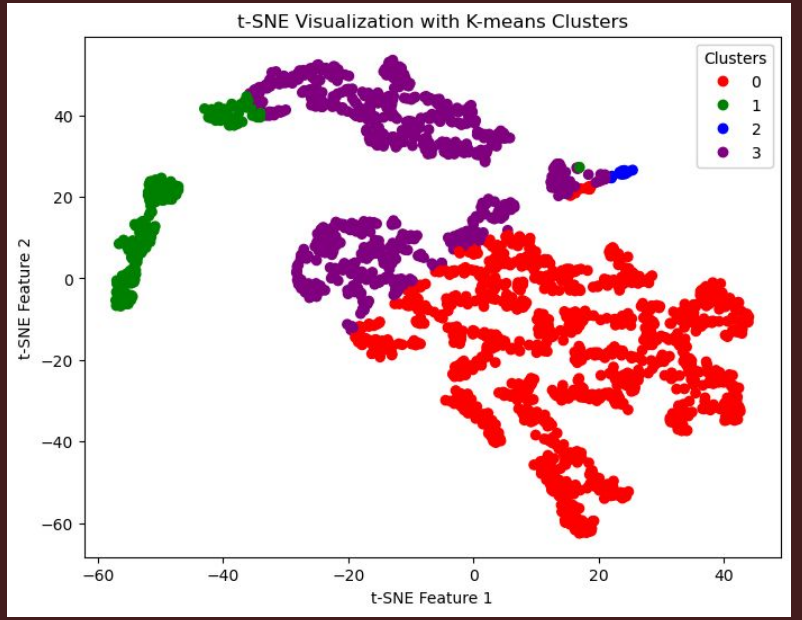
- Geometric Correlations:
  - Indicate cell presence through image dimensions and area metrics.
  - 'id' with 'elapsed_time' OR 'image_width' with 'raw_area'.
- Non-Geometric Correlations:
  - Validate cell characteristics via intensity and texture analysis.
  - Insightful: 'width' with 'sum_intensity' (cell density/pigmentation).
  - 'convex_perimeter' with 'sum_intensity' (shape complexity and internal properties).
- Applying to our Machine Learning Model:
  - Leverage these correlations for automated cell detection and analysis.
  - Utilize patterns in both geometric and non-geometric features.



TEXAS STATE UNIVERSITY

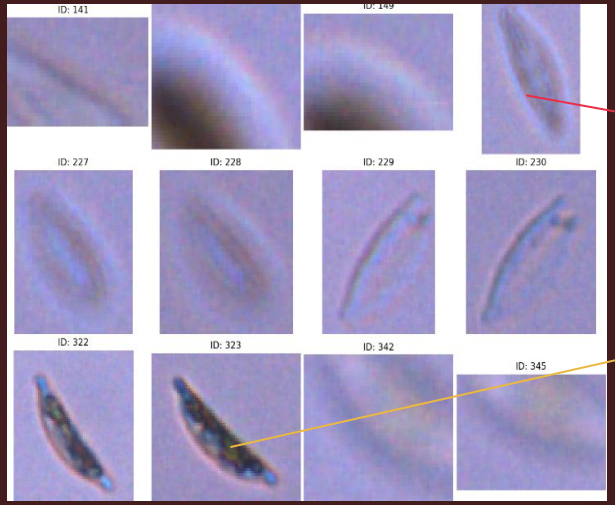
Clustering

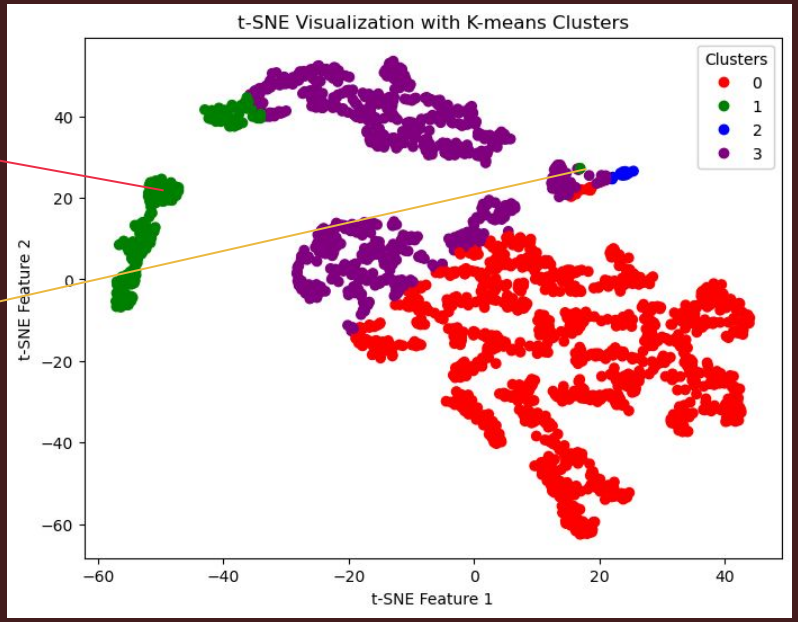
- Aims to uncover patterns and relationships in the data
- Utilized for the project to identify grouping in the algal cell data
- Utilized t-SNE, a clustering technique for dimensionality reduction. Resulting in the structure of the figure on the right
- Took t-SNE groupings, and applied the K-Means clustering algorithm to group clusters based on a given K (4 on right)



TEXAS STATE UNIVERSITY

Clustering cont.





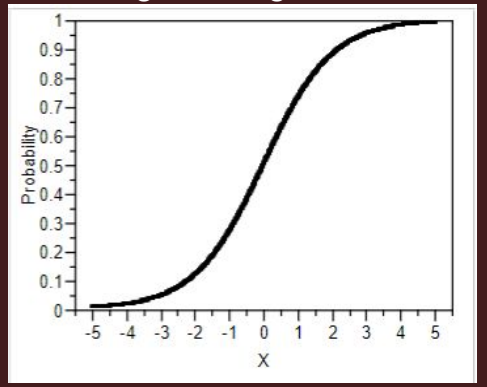
TEXAS STATE UNIVERSITY

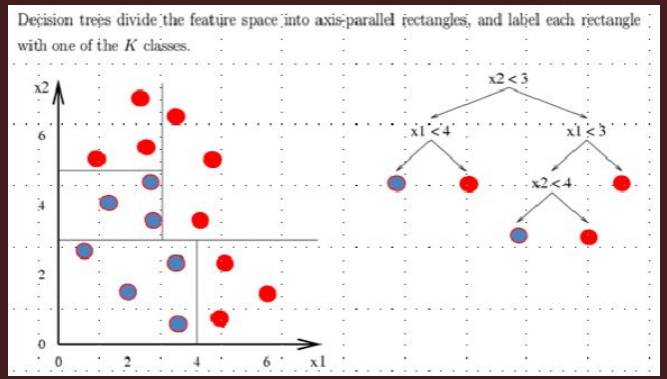
MACHINE LEARNING & CONCLUSION

Baseline Model

- Purpose: Sets minimum performance standards, paves the way for more complex models
- Model Choice: Logistic regression or Decision Trees
- Benchmarking: Assesses basic model efficiency using metrics that we can compare to later models

Logistic Regression



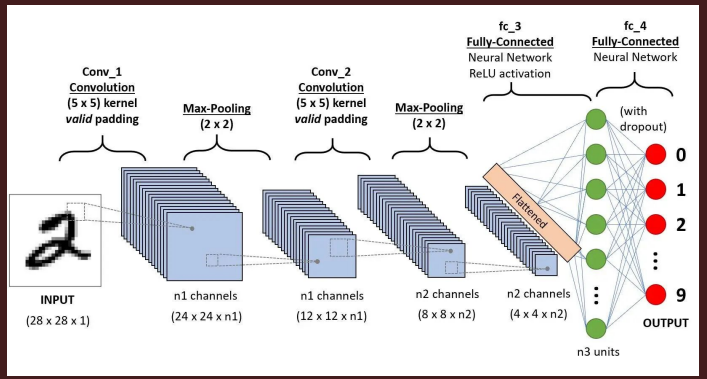


TEXAS STATE UNIVERSITY

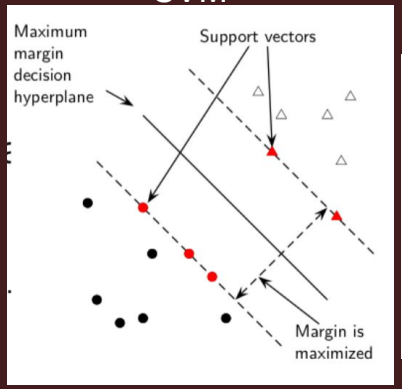
Possible Model Selections

- **Convolutional Neural Networks (CNNs)**
  - Ideal for image classification, capturing spatial hierarchies in image data.
  - Learns features automatically from image data, enhancing adaptability.
- **Support Vector Machines (SVM)**
  - Effective in high-dimensional spaces like image pixels.
  - Versatile with various kernel functions for linear and non-linear data.
- **Random Forests**
  - Robust ensemble method for classification.
  - Handles numerous features and helps in understanding key distinctive elements.

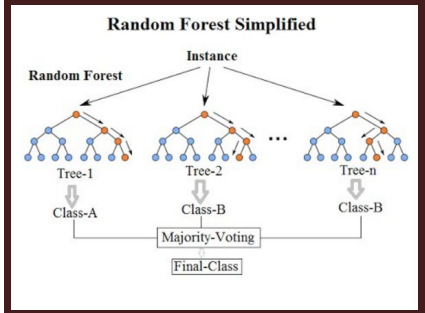
CNN



SVM



Random Forest



TEXAS STATE UNIVERSITY

What I Am Currently Working On

- Developing a method to group similar images, based on features or characteristics (exploring using SIFT or data)
- Implementing a system to select the most representative image from each group, ensuring dataset quality
- Finalizing a dataset with one distinct image per group, representing different algal classes
- Training the model on this curated dataset to accurately classify cells into various algal categories

QUESTIONS??

TEXAS STATE UNIVERSITY