Predicting Pokémon Types with Clustering and Classification

University of Toronto - STA2201 - Winter 2025

Justin Zhang, Isaac Baguisa, Alex Faassen

2025-04-04

0.1 Abstract

- Report Summary
- Link GitHub?

0.2 Contributions

- Justin Zhang:
- Isaac Baguisa:
- Alex Faassen: Data pre-processing, Data exploration, Image dimension reduction

1 Introduction

Over the past two decades, Pokémon has evolved from a niche Gameboy game into a global multimedia franchise encompassing anime, trading cards, e-sports competitions, and mobile applications. For many who grew up in the late 2000s and early 2010s, the world of Pokémon was a formative part of childhood entertainment alongside cultural staples like Mario Bros. and Zelda. As Pokémon continues to captivate audiences across generations, its community-oriented game design and well-structured universe presents an intriguing opportunity for data analysis. In this project, we investigate whether a Pokémon's type—a categorical label used in-game to denote elemental or behavioral characteristics such as Fire, Water, or Grass—can be inferred from its visual appearance and numerical attributes using statistical learning techniques.

Our primary research question is: *Can clustering and classification methods uncover or predict a Pokémon's type based on its image and statstical features?*

The Pokémon type system serves not only as a game mechanic but also as a conceptual grouping based on traits like colour, strength, and theme. We aim to explore whether these groupings have an underlying statistical structure that can be detected through dimensionality reduction, clustering, and classification algorithms. This analysis will provide insight into the extent to which a Pokémon's type is reflected in its appearance and physical characteristics, or whether it is a more arbitrary design choice by game developers.

2 Data Description

Our analysis is based on two publicly available Kaggle datasets:

2.1 1. Pokémon Image Dataset

URL: Pokémon Image Dataset

The Pokémon Image Dataset consists of 809 PNG files representing unique images of Pokémon from generations 1 through 7. Each file consists of a 3-dimensional array of dimensions 120 by 120 by 4. The first 2 slices

represent a 120 by 120 grid of pixels, and the 3rd slice represents the 4 RGBA channels (Red, Blue, Green, Alpha), the colour and transparency assignments of each pixel.

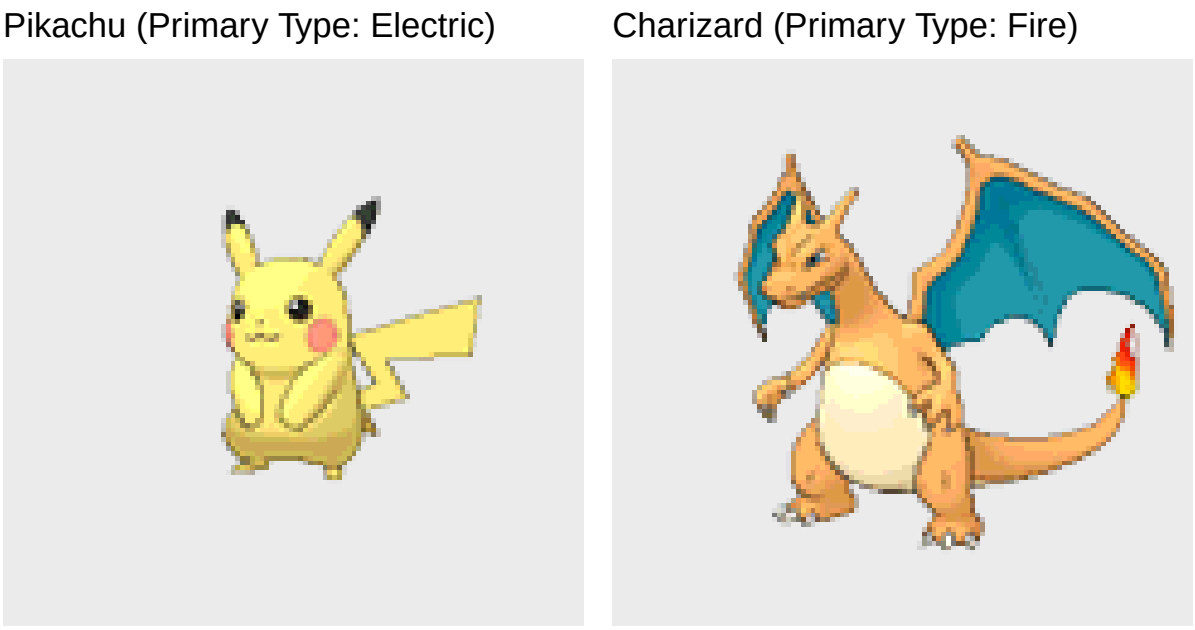

Figure 1: Images of Pikachu (Primary type: Electric) and Charizard (Primary Type: Charizard) from the Pokémon Image Dataset.

2.2 2. Pokémon Stats Dataset

URL: The Complete Pokémon Dataset

This structured dataset contains 41 variables for 801 Pokémon. These features include:

- **Physical attributes:** height, weight, base experience
- **Combat statistics:** HP (Health Points), attack, defense, special attack, special defense, speed
- **Categorical indicators:** Legendary status, abilities, and generation
- **Primary and secondary type labels**

All variables are numeric or have been encoded numerically (e.g., legendary status as 0/1). There are no missing values. For this project, we **only consider primary type** labels to maintain a single-label classification structure.

Table 1: Summary Statistics for Key Features

Feature	mean	sd	min	q1	median	q3	max
hp	68.96	26.58	1.0	50.0	65.0	80.0	255.0
attack	77.86	32.16	5.0	55.0	75.0	100.0	185.0
defense	73.01	30.77	5.0	50.0	70.0	90.0	230.0
sp_attack	71.31	32.35	10.0	45.0	65.0	91.0	194.0
sp_defense	70.91	27.94	20.0	50.0	66.0	90.0	230.0
speed	66.33	28.91	5.0	45.0	65.0	85.0	180.0
height_m	1.16	1.08	0.1	0.6	1.0	1.5	14.5
weight_kg	61.38	109.35	0.1	9.0	27.3	64.8	999.9

Table 1 provides an overview of the distribution of key numeric features across Pokémon in the dataset. Several notable patterns emerge:

Combat statistics, such as `attack`, `defense`, `sp_attack`, `sp_defense`, and `speed`, show similar ranges, with means around 66–78 and standard deviations near 30. These distributions suggest balanced but varied combat capabilities, with maximum values exceeding 180 or 230, likely representing fully evolved or legendary Pokémon. `hp` follows a similar pattern with a slightly lower mean of 69 and a maximum of 255, again suggesting outliers with extreme values. In contrast, the physical attributes, `height_m` and `weight_kg`, seem to be skewed. The average height is just over 1 meter, but with a maximum of 14.5 meters. Weight displays even greater disparity, ranging from under 1 kg to nearly 1000 kg. This indicates that a few extremely large Pokémon (e.g., Wailord) heavily influence these distributions.

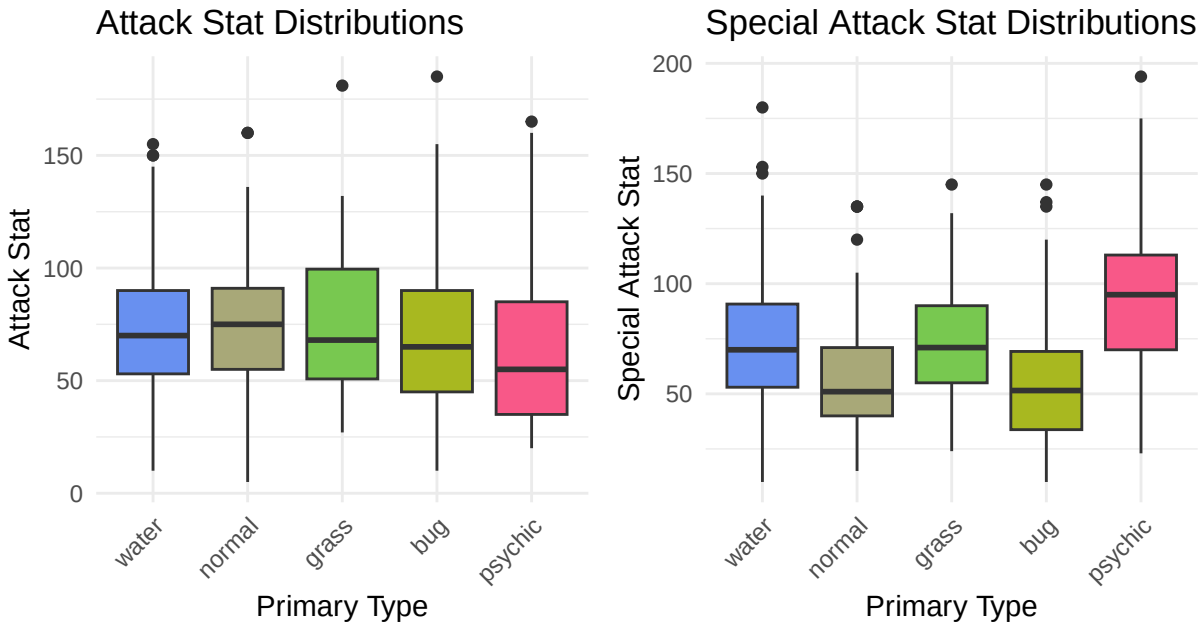

Figure 2: Frequency and attack statistic distributions of the top 5 Pokémon primary types by frequency.

In addition to summary statistics, **Figure 2** visualizes the distributions of the `attack` and `sp_attack` stats across the five most common primary types using boxplots. These five types, `water`, `normal`, `grass`, `bug`, and `psychic`, were selected based on their overall frequency in the dataset.

The boxplots reveal several key insights:

- `normal` Pokémon tend to have a slightly higher median attack, but slightly lower median special attack compared to the other types. The interquartile ranges are also both narrower, indicating a more concentrated spread of both stats within this type.
- `bug` and `psychic` types show the lowest median attack values among the five. However, `psychic` seems to make up for this with a significantly higher special attack spread, while `bug` maintains the lowest special attack stats.
- All types exhibit high-attack outliers above 150, indicating that even typically lower physically offense types can include powerful exceptions.
- The two most common types, `water` and `normal`, show narrower distributions for both stats, consistent with their broad representation across Pokémon species.

This visualization supports the idea that while some types, like `normal`, may lean toward stronger physical offense, others, like `psychic`, may rely on special attack, while some, like `bug`, may rely on `abilities` or other strategic capabilities not captured by standard combat stats alone. These findings underscore the

importance of incorporating multiple features when attempting to predict a Pokémon's type using clustering or classification models.

2.3 Pre-processing

For the images, we begin by flattening each PNG into a (120x120x4) 57600-element vector, each pixel and RGBA value representing a feature. Binding these vectors into a table identified by Pokémon name gives us a usable dataset. To integrate image features with the stats dataset, we filter both tables to common Pokémon requiring several name formatting adjustments to account for unusual characters. Finally, we match the row-order of both datasets, resulting in two easily transferable Pokémon datasets.

3 Methodology

- What methods/models were chosen and why w.r.t. research question

3.1 Image Dimension Reduction

To explore whether Pokémon types are distinguishable based on their visual features, we look to apply clustering algorithms to their images. Since we are working with high-dimensional data (801x57600), we first apply dimension reduction techniques to mitigate the curse of dimensionality. Our primary tool for image dimension reduction is Principal Component Analysis (PCA). In its own right, this method also helps us determine if type-based structure is embedded in high-dimensional image representations.

3.1.1 Principal Component Analysis (PCA)

PCA is a linear technique that identifies the directions of maximal variance in the dataset. After centering the flattened image matrix, we applied PCA and examined the principal components, loading vectors, and compression options.

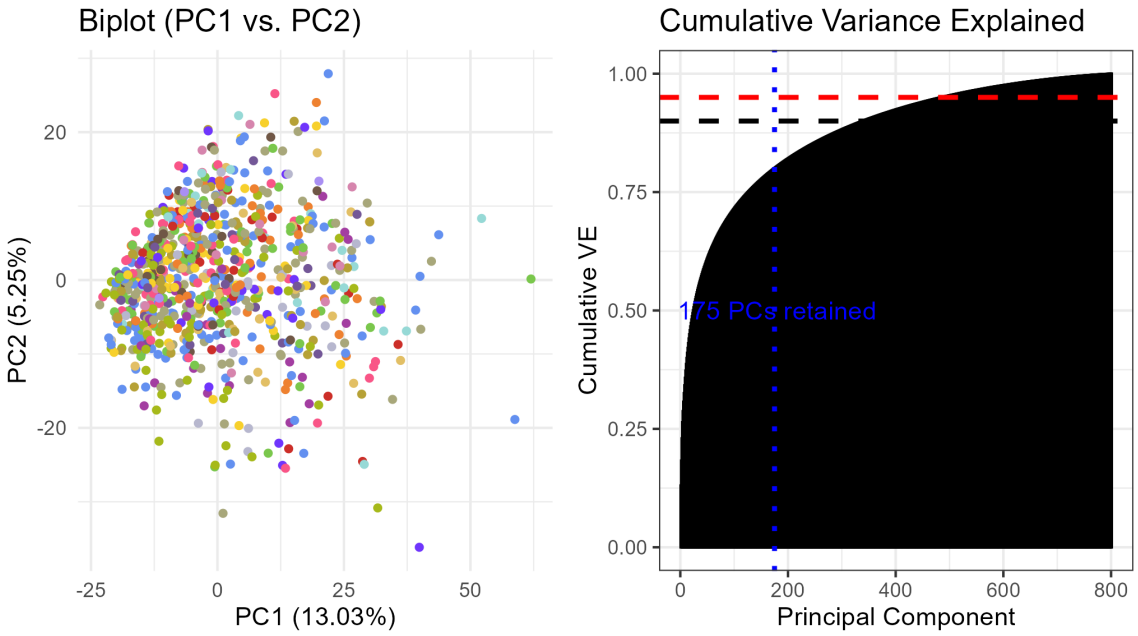

Figure 3: Left: Cumulative Variance Explained (VE) by principal components (PC). The blue and black dashed lines represent 80% and 95% of VE respectively. Right: Biplot for the first 2 PCs.

Figure 3 above summarizes the results of applying PCA to the flattened Pokémon image data.

On the left, the biplot visualizes the Pokémon with respect to the first 2 principal components (PCs), **PC1 (13.03%)** and **PC2 (5.25%)**. Each point corresponds to a Pokémon, colour-coded by its primary type. While its possible that there are weak groupings, all types seem to be heavily overlapping. This is somewhat unsurprising given the first 2 PCs only represent roughly 18% of the variance. Further, **Appendix A** examines the first 2 **loading vectors** which seem to correspond to Pokémon area (i.e. height-width stretch) and height-width ratio respectively, neither of which seem particularly promising on their own for type clustering. These findings motivate the use of non-linear methods, such as **UMAP** or **t-SNE**, which are better suited for uncovering type-based structure in complex image data.

On the right, the **cumulative variance explained** plot shows that the first few PCs capture a substantial portion of the variance, but with diminishing returns beyond the first 100 components. To retain at least **80% of the total variance**, we selected the first **175 PCs** (indicated by the blue vertical line). This threshold strikes a balance between compression and information preservation with a slightly greater emphasis on compression, and these 175 components are used in downstream clustering and classification steps. Below, **Figure 4** and **Appendix B** provide visual examples of the degree of compression applied before applying clustering algorithms.

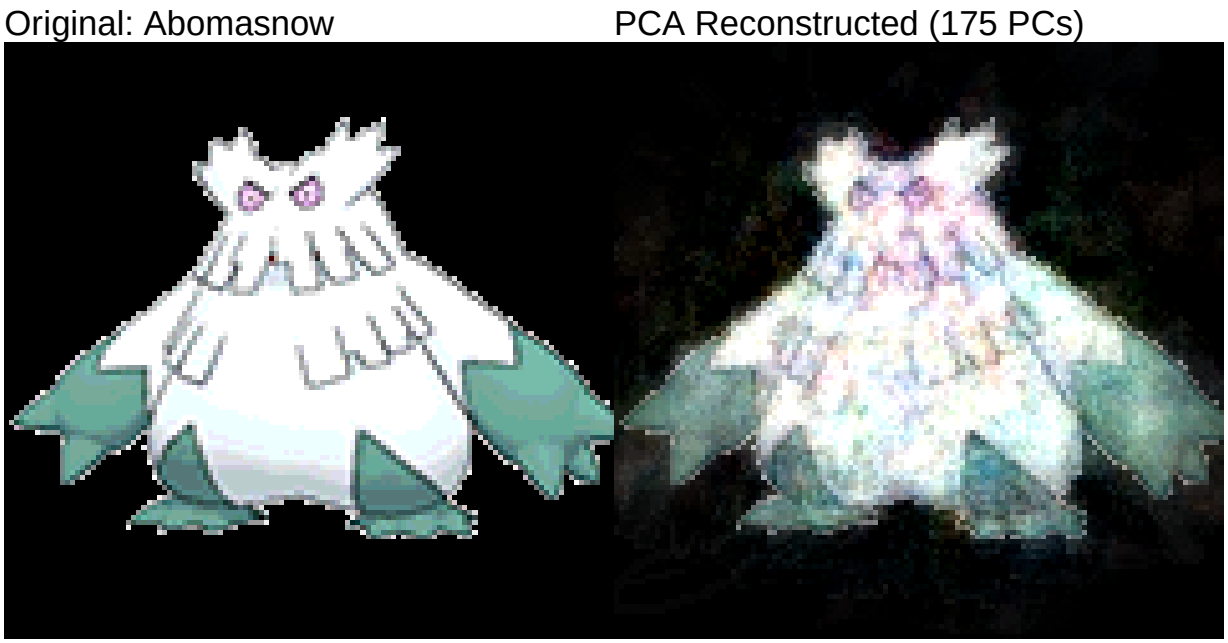

Figure 4: Images of the Pokémon Abomasnow. Left: Original. Right: After image compression with 175 PCs, capturing 80% of the variance explained.

3.1.2 UMAP

To uncover non-linear structure in the image data, we applied **Uniform Manifold Approximation and Projection (UMAP)**. Unlike PCA, which is a linear method, UMAP is designed to preserve both local and global relationships in high-dimensional spaces, making it particularly effective for image-based clustering. **Figure 5** displays the 2D UMAP projection of the 801 Pokémon image vectors, coloured by their primary type.

Once again, with the large numper of types, it is very hard to visually identify groupings. Upon close inspection, there does seem to be slightly more grouping between certain types, such as `dark`, `fairy`, and `poison`. However, for the most part, most types appear to be largely overlapping. Though types aren't

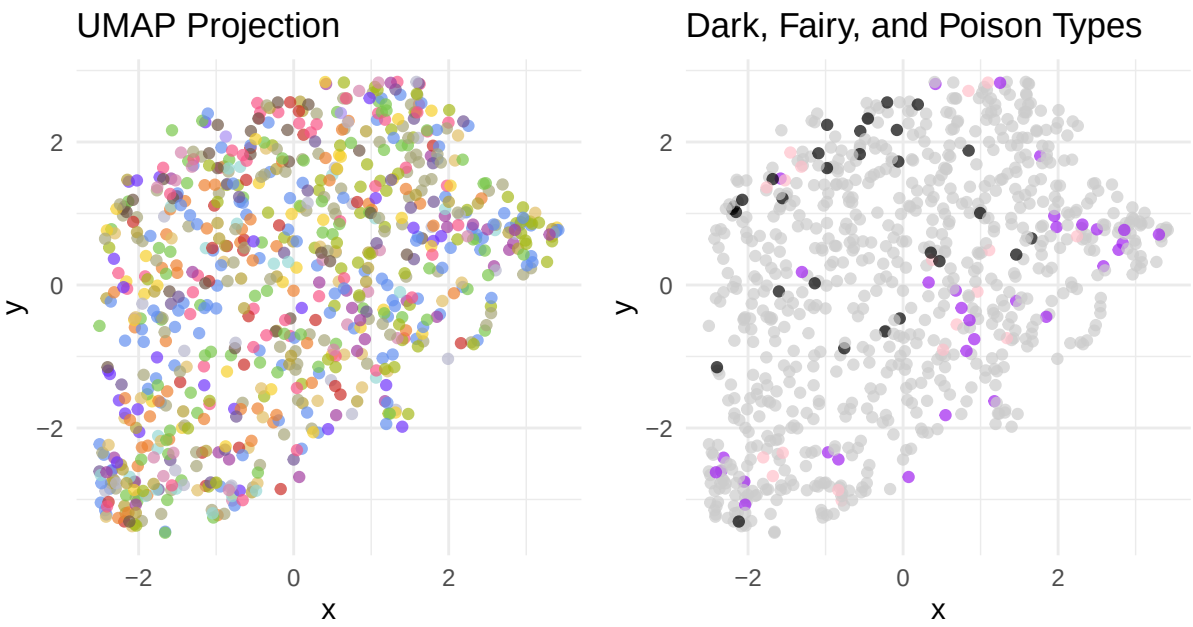

Figure 5: 2D UMAP projection of the 801 Pokémon image vectors. Left: Coloured by their primary type. Right: All type colours set to grey, save Dark, Fairy, and Poison types.

particularly distinct, there does seem to be some improvement over the PCA biplot. Overall, UMAP provides stronger visual evidence that type-specific patterns may exist in the high-dimensional image data.

3.2 Unsupervised Model

- e.g. Weighted K-means

3.3 Supervised Model

- e.g. Log reg, KNN, LDA

4 Results

- Summarize findings (include tables/plots)
- Interpret results w.r.t. research question

5 Discussion

- Discuss model performance
- Limitations of methodology
  - Potential sources of bias
- Challenges encountered
- Recommendations for overcoming these + improvements for future work

Notes: - Dual-typing might be causing some difficulties - adding complexity. - Perhaps other approaches, such as deep learning, might do better.

6 References

- Use `knitcitations`? - manual easier?

7 Appendix

7.1 A: Loading Vectors

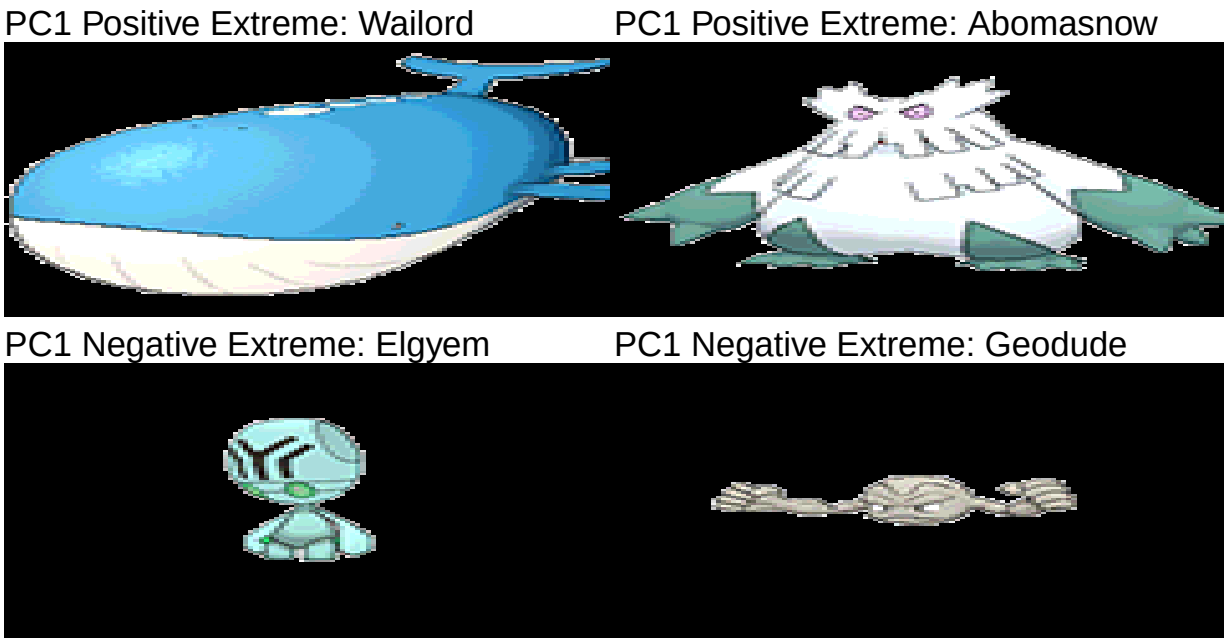

Figure 6: PCA PC1 extreme value Pokémon. Seems to reflect area.

7.2 B: Gardevoir Image Compression

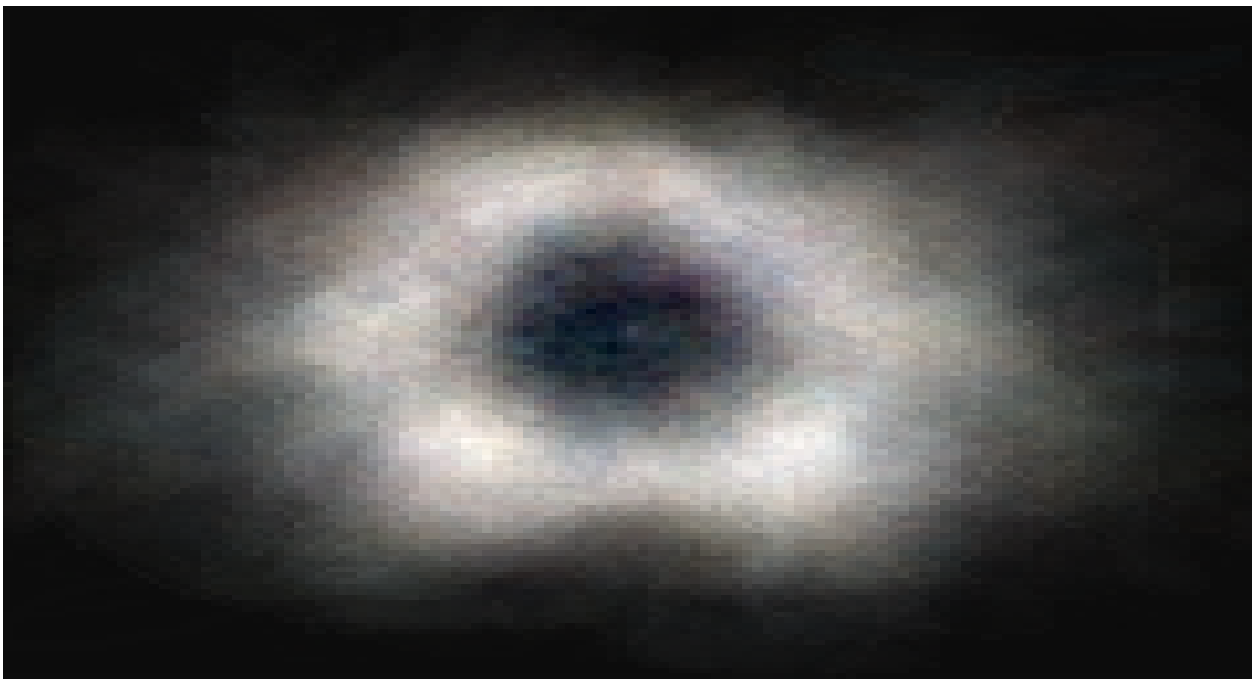

Figure 7: PCA 1st loading vector.

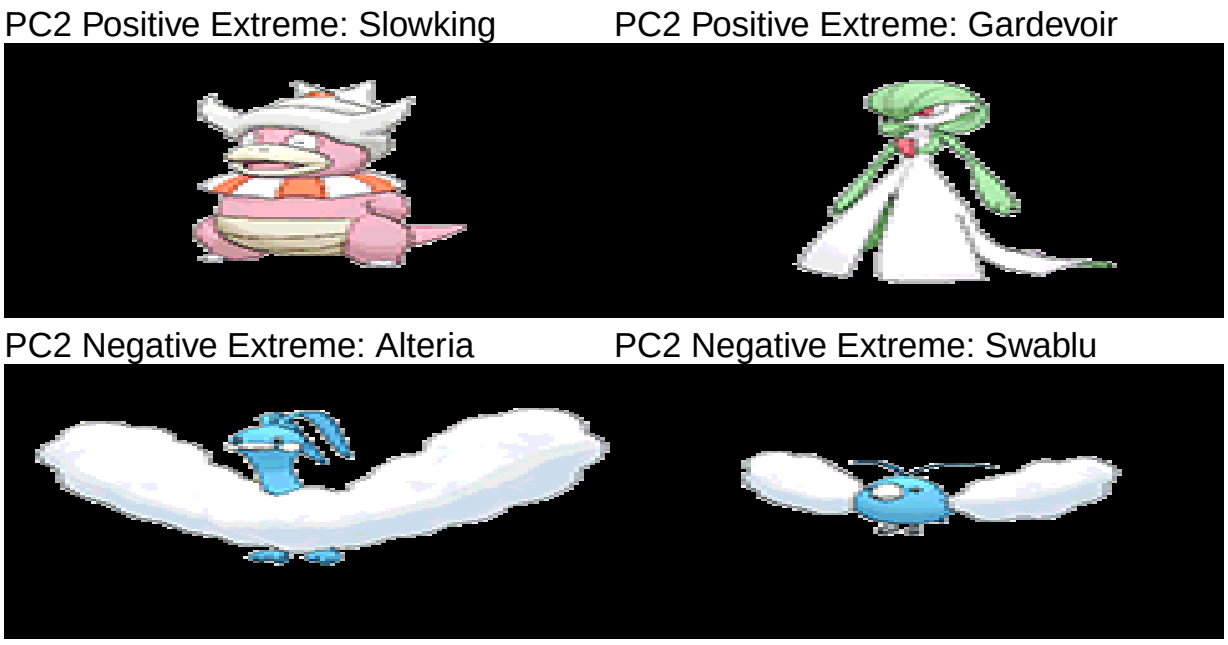

Figure 8: PCA PC2 extreme value Pokémon. Seems to reflect height-width ratio.

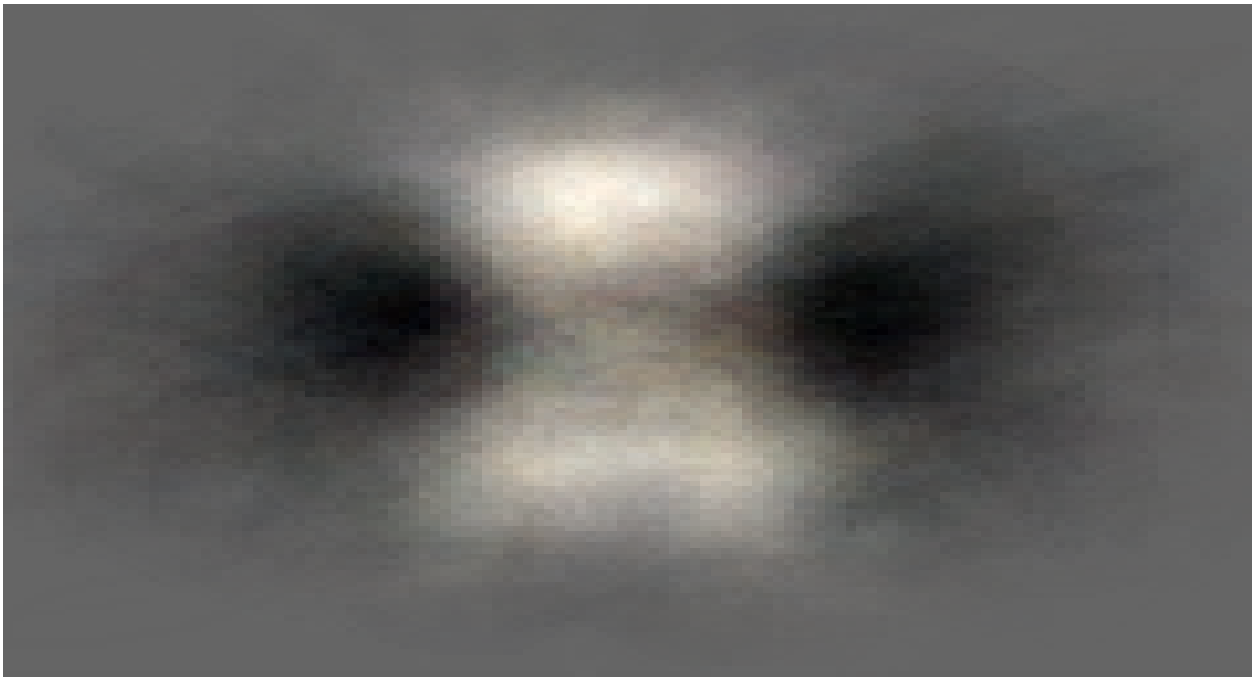

Figure 9: PCA 2nd loading vector.

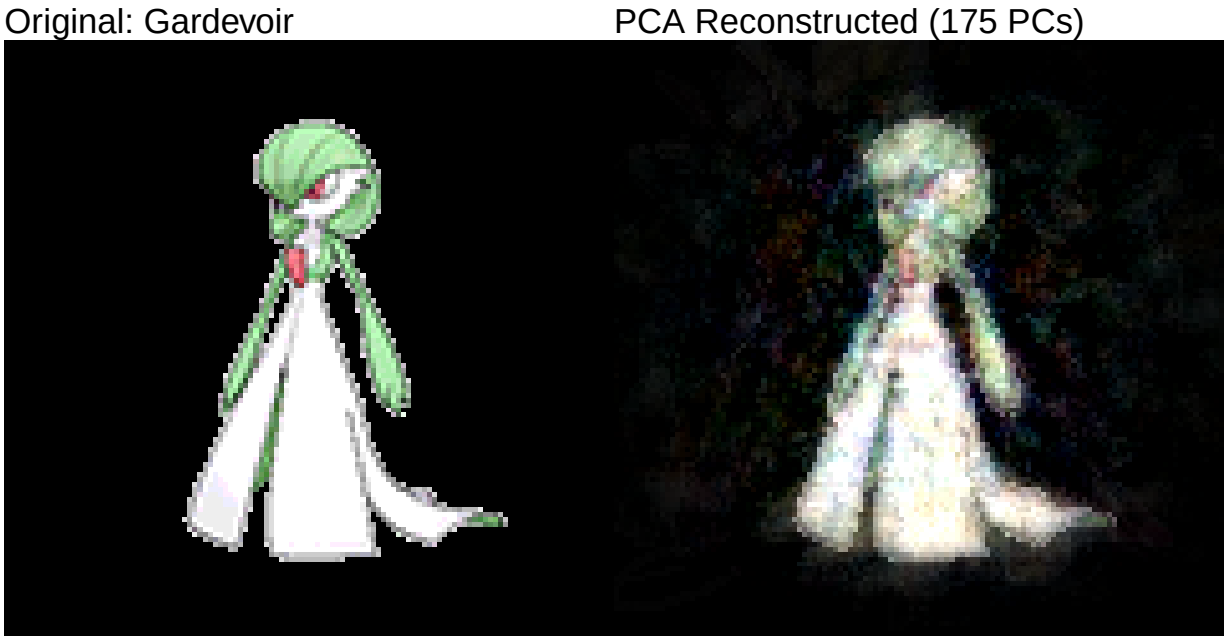

Figure 10: Images of the Pokémon Gardevoir. Left: Original. Right: After image compression with 175 PCs, capturing 80% of the variance explained.