Hierarchical Clustering of Engineering Materials

1. **The Engineering Topic.** In design engineering, the problem of selecting an 'optimal' material is often imprecise and ill-defined. When navigating vast, tabulated lists of different materials it can be difficult for non-materials scientists to compare, and draw conclusions about the relationships between materials. This project attempts to use unsupervised clustering techniques to generate a hierarchical visualization of the relationships between these materials in order to assist in the design selection process.

2. **The Technical Problem of Interest.** As material datasets are inherently large and relatively high dimensional, they are perfect for machine learning techniques that highlight connections hidden deep within a feature space. Our problem involves transforming tabulated material data (roughly 50 materials in 6 dimensions) first into a linkage matrix structure that can be assessed for cluster quality and then into a visual dendrogram tree that can be used to understand material relationships and to select suitable design alternatives. The associated difficulties are:

   (a) *Data Acquisition and Preparation.* Comprehensive, open source material property databases are limited in availability and typically too large to be useful in a design selection context. Section 4 of this report expands further on the specific pre-processing steps that were taken to arrive at a reasonably sized dataset.

   (b) *Quality Visualization.* The clustering algorithm must produce a plot-friendly object that can be clearly visualized in dendrogram format. This requires the selection of appropriate cluster boundaries and visualization parameters.

   (c) *Validation.* The resulting linkage structure must be quantitatively evaluated to ensure accuracy and usefulness when applied to real design problems. This is particularly difficult in datasets of $dim > 3$ as we do not have a concrete way to visualize the initial dataset to confirm our predictions make intuitive sense.

3. **Literature Review.** The underlying algorithm assessment strategy for this analysis was adopted (and simplified) from Unsupervised clustering of materials properties using hierarchical techniques, although the specific algorithms and input datasets used are different.

   Algorithm variants and distance measures were selected based on descriptions found in Cluster Analysis for Researchers.

   The primary mathematical machinery and dendrogram plotting is performed by several functions available within SciPy's `cluster.hierarchy`.

4. **The Algorithm(s).** The algorithms performed in this report are all similar variants of unsupervised, agglomerative (start with many clusters and end with 1), hierarchical clustering techniques that organize data points into a hierarchical structure based on some measure of spatial distance. The general model strategy is laid out in figure 1. The pseudocode contained within the 'algorithm execution' blocks outlines the vital portions of `scipy.cluster.hierarchy.mst_ single_linkage` and `scipy.cluster.hierarchy.nn_chain` that my implementation utilizes.

5. **The Implementation.** See python notebook for code and the accompanying comments.

6. **Verifying the Implementation.** We can verify the correctness of my code in two main ways.

   (a) *Example 2D data* — The section entitled 'Example Data' within my code generates 3 blobs of points in 2D space and applies my implementation to the example dataset. We can then plot

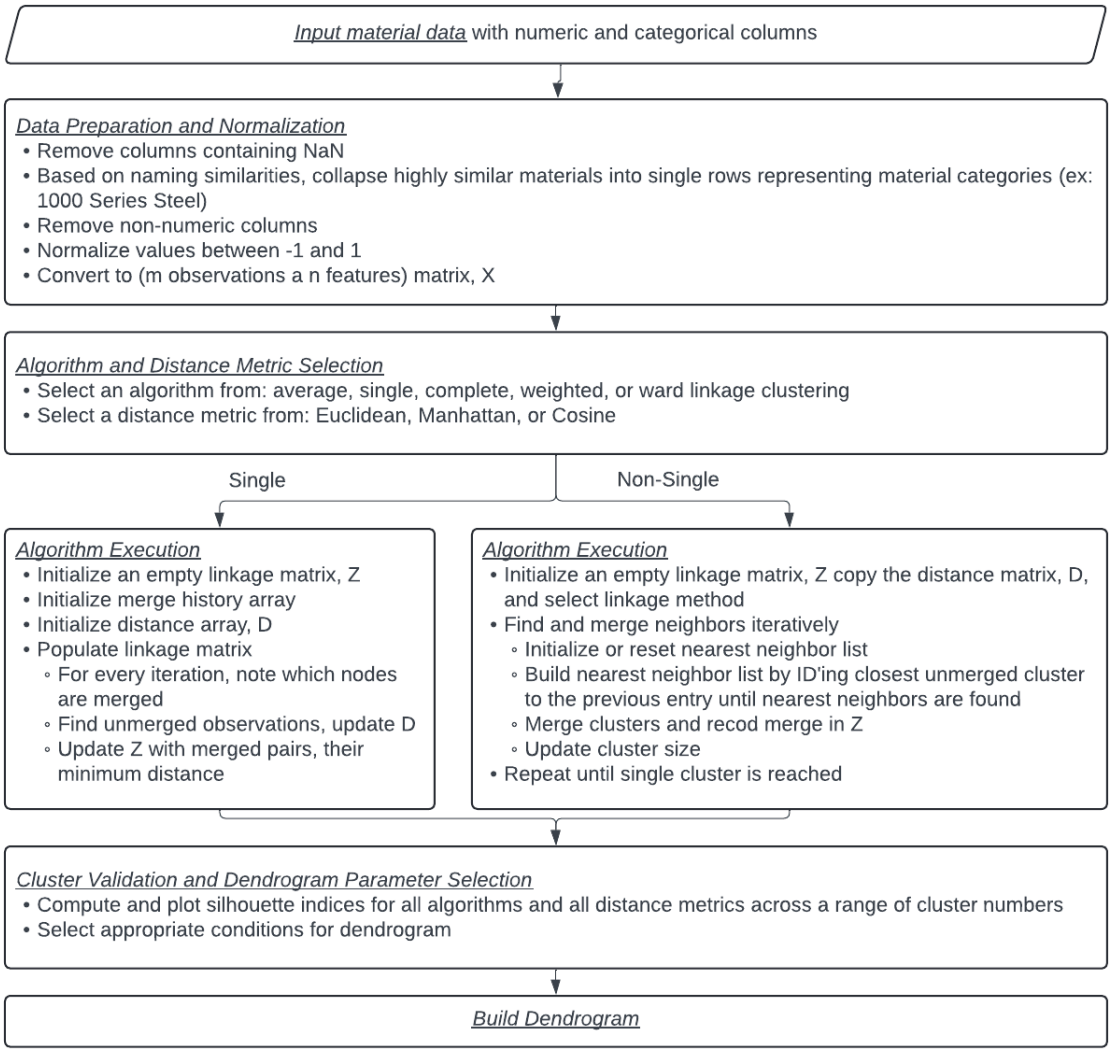

Figure 1: High Level Algorithm Overview

both the original data and the resulting dendrogram to ensure our algorithm is behaving as expected. We can see from the resulting plots that the dendrogram matches the 2D scatter plot and correctly captures the 3 distinct blobs in the data.

(b) *Qualitative comparison with known material categories.* — From the dendrogram, we can compare the overall grouping and highlighted clusters with the material category labels present in the original dataset. NOTE: These labels were not considered during the clustering process.

7. **The Datasets, Problem Settings, and Parameters.** The dataset used was an open source tabulated collection of around 1500 rows of materials in 15 dimensions. In order to avoid an over-crowded dendrogram, these rows were condensed into 50 rows representing more general material categories. The resulting dataframe was around 50 rows, with each row containing numeric values for yield strength, tensile strength, Young's modulus, shear modulus, and density.

   Using silhouette indices as the scoring method, trials were ran to explore all possible combinations of the following problem settings and parameters.

   (a) *Choice of algorithm* — average, single, complete, weighted, or ward linkage hierarchical

clustering.

(b) *Choice of distance metric* — Euclidean, Manhattan, or cosine

(c) *Number of Clusters* — All algorithms will be tested with all distance metrics for a range of clusters from 3-12.

Sobh et. al., found the average linkage algorithm using euclidean distance to generally produce higher silhouette indices as compared to other configurations, so we'll take this configuration as our baseline.

8. **The Results.** Shown below are 3 bar charts of the silhouette indices for all algorithm variants corresponding to the 3 distance metrics that were assessed.

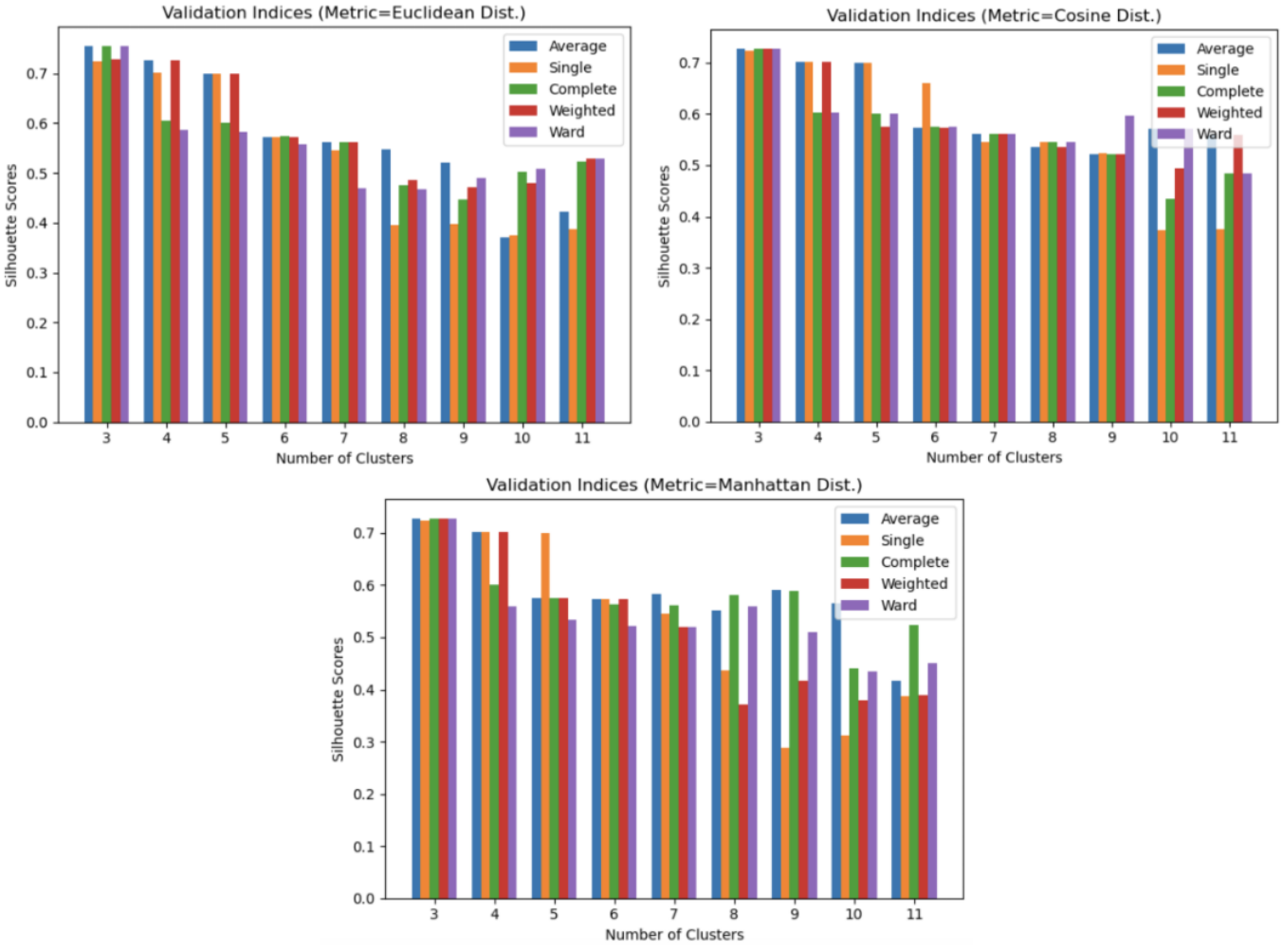

Figure 2: Validation indices for all algorithm configurations.

From the above bar charts, we can see that all 5 algorithms perform similarly well under the assumption of 3 clusters, although average and weighted also seem to perform well at 4 clusters. A cluster quantity of 4 more closely aligns with the number of standard categorical labels present in the original prepared dataset. (These labels were not used during clustering)

In terms of distance metrics, euclidean distance consistently performs best across any comparison of same-cluster-quantity and same-method trials.

Given the results of my silhoutte index analysis and some prior knowledge of material categories specific to this particular dataset, we can select our optimal algorithm configuration as:

Method: Average Hierarchical Clustering,
Distance Metric: Euclidean Distance,
Number of Clusters: 4

The resulting optimal dendrogram is plotted below in figure 3. and a larger figure is included within the appendix of this document.

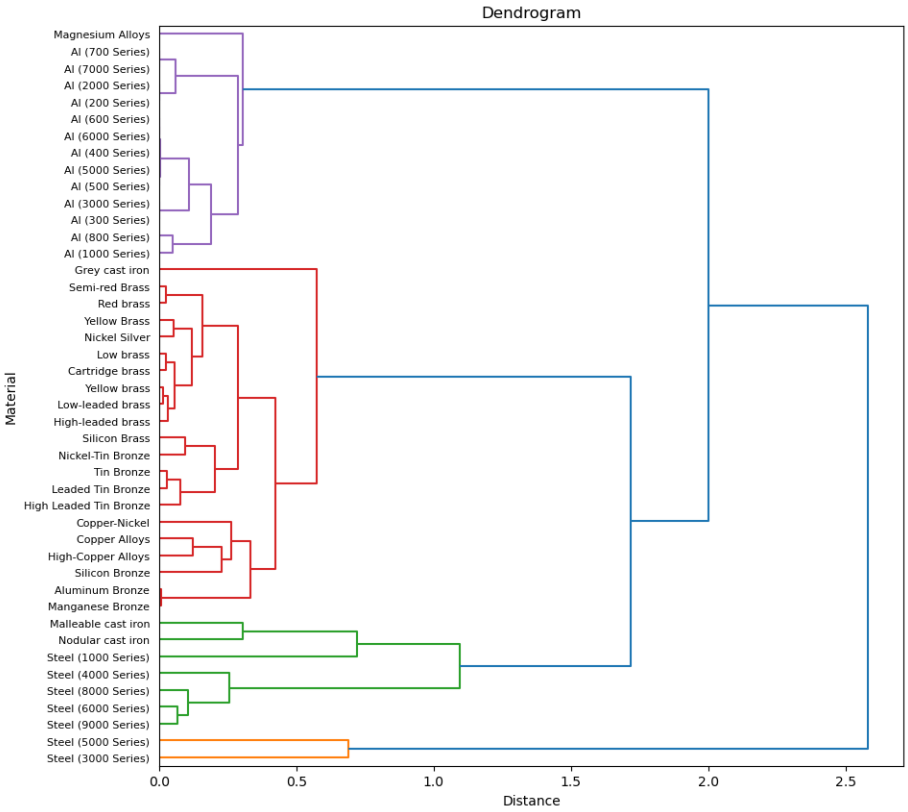

Figure 3: Optimal Dendrogram.

From a practicality standpoint, the resulting dendrogram is intuitive and easy to use. The generated clusters match well with basic material science principles and more complex relationships between seemingly dissimilar materials. One such relationship that the algorithm captured was the mechanical behavior similarities between dissimilar metals that share an alloying ingredient. Revealing such connections was the original objective of this project.

9. **The Limitations.** One of the primary limitations of this approach lies in the variability that exists within any of the 50 material categories that the original data was condensed into. For a more comprehensive and useful tool, the dendrogram should be generated from all 1500 material variants and should be searchable/navigable so that the benefits of the visualization are still able to be enjoyed.

   Additionally, this approach lacks a deterministic way to decide the optimal number of clusters, as several cluster numbers typically result in the same (or similar) silhouette score. A more robust scoring metric is needed to address this issue.

Appendix

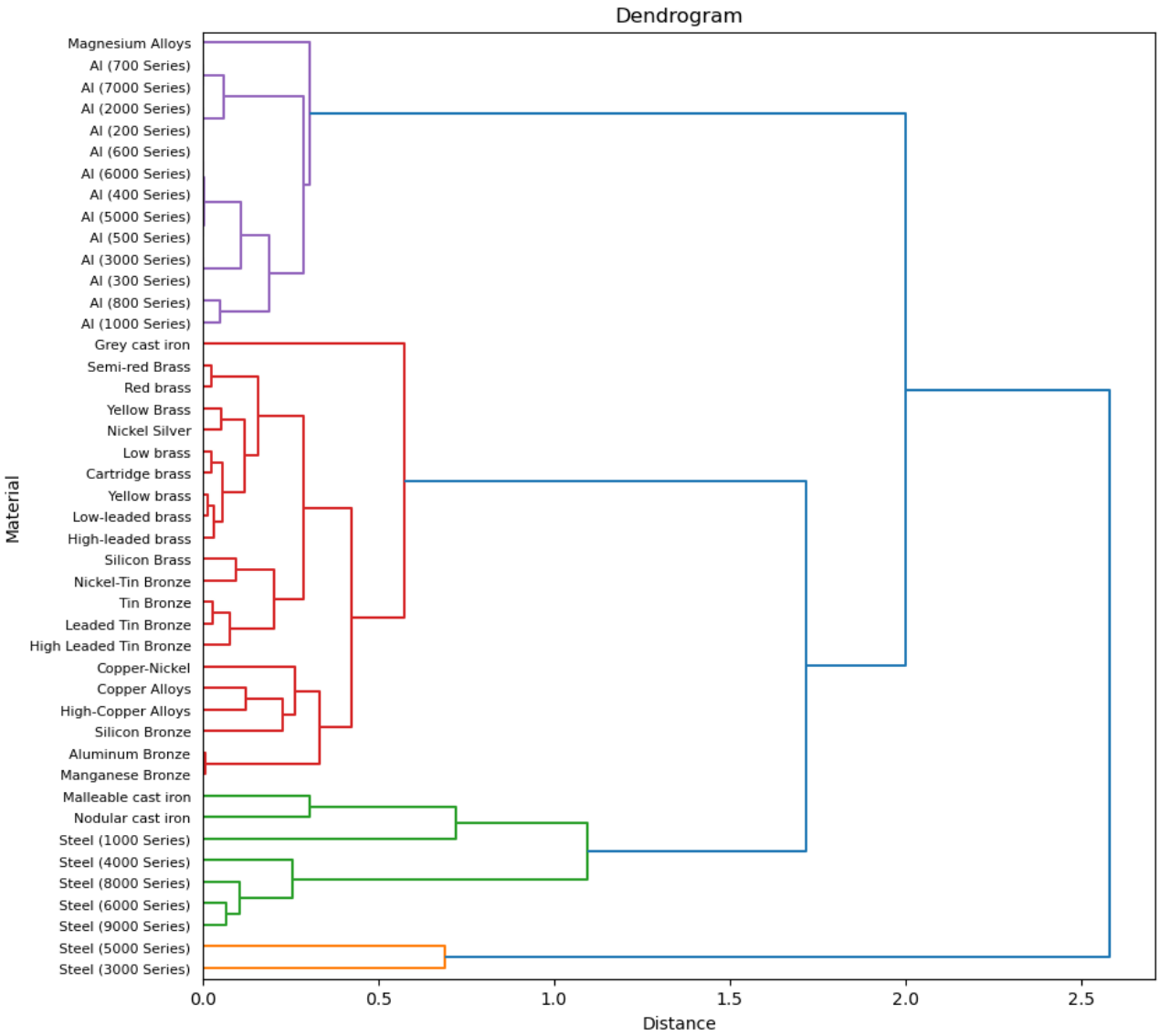

Figure 4: Optimal Dendrogram (Larger).