Genetic Syndrome Classification Using Image Embeddings

Walter Monteiro

This report describes the implementation of a machine learning pipeline for classifying genetic syndromes using 320-dimensional image embeddings produced by a pre-trained model. The pipeline includes data preprocessing, exploratory analysis, 2D visualization via t-SNE, K-Nearest Neighbors (KNN) classification with Euclidean and Cosine distances, and performance evaluation using 10-fold stratified cross-validation. Evaluation metrics include macro F1-score, macro one-vs-rest AUC, top-k accuracy, and micro-average ROC curves averaged across folds. Results demonstrate strong discriminative performance for both distance metrics, with consistent improvements using Cosine similarity.

1 Introduction

Genetic syndrome recognition from medical images can be supported by representation learning models that map images into embedding vectors. Given embeddings extracted from a pre-trained model, the objective of this work is to analyze embedding distributions and build a robust baseline classifier to predict syndrome identifiers associated with each image embedding.

This report presents: (i) dataset preprocessing and exploratory analysis, (ii) visualization of embedding structure using t-SNE, (iii) KNN-based classification using different distance metrics, and (iv) cross-validated evaluation using manually implemented metrics and ROC analysis.

2 Dataset and Preprocessing

2.1 Dataset Structure

The dataset is provided as a hierarchical dictionary where each embedding is indexed by syndrome ID, subject ID, and image ID. Each embedding is a 320-dimensional vector.

2.2 Flattening and Integrity Checks

The hierarchical structure was flattened into:

- Feature matrix $X \in R^{N \times 320}$
- Label vector $y \in \{0, \ldots, C-1\}^N$, where C is the number of syndrome classes

Integrity checks were applied to ensure:

- All embeddings have dimensionality 320
- No missing values and no non-finite values (NaN/Inf)
- Consistent mapping between syndrome identifiers and numeric class labels

2.3 Exploratory Data Analysis (EDA)

Basic descriptive statistics were computed, including number of syndromes, total images, and per-class image counts. The dataset exhibits moderate class imbalance.

Insert EDA Summary Table

Table 1: Dataset summary statistics.

Statistic	Value
Number of syndromes (C)	**10**
Total images (N)	**1116**
Embedding dimension	**320**
Images per syndrome (min)	**64**
Images per syndrome (max)	**210**
Images per syndrome (mean)	**111.6**
Images per syndrome (median)	**93.5**

Insert Class Distribution Table (Optional)

Table 2: Images per syndrome (sorted by frequency).

Syndrome ID	Number of Images
300000034	210
300000080	198
100192430	136
300000007	115
300000082	98
100610443	89
300000018	74
100180860	67
100610883	65
700018215	64

3 Embedding Visualization (t-SNE)

To inspect the embedding geometry and potential separability between classes, t-SNE was applied to project 320-dimensional embeddings into 2D space.

3.1 t-SNE Plot

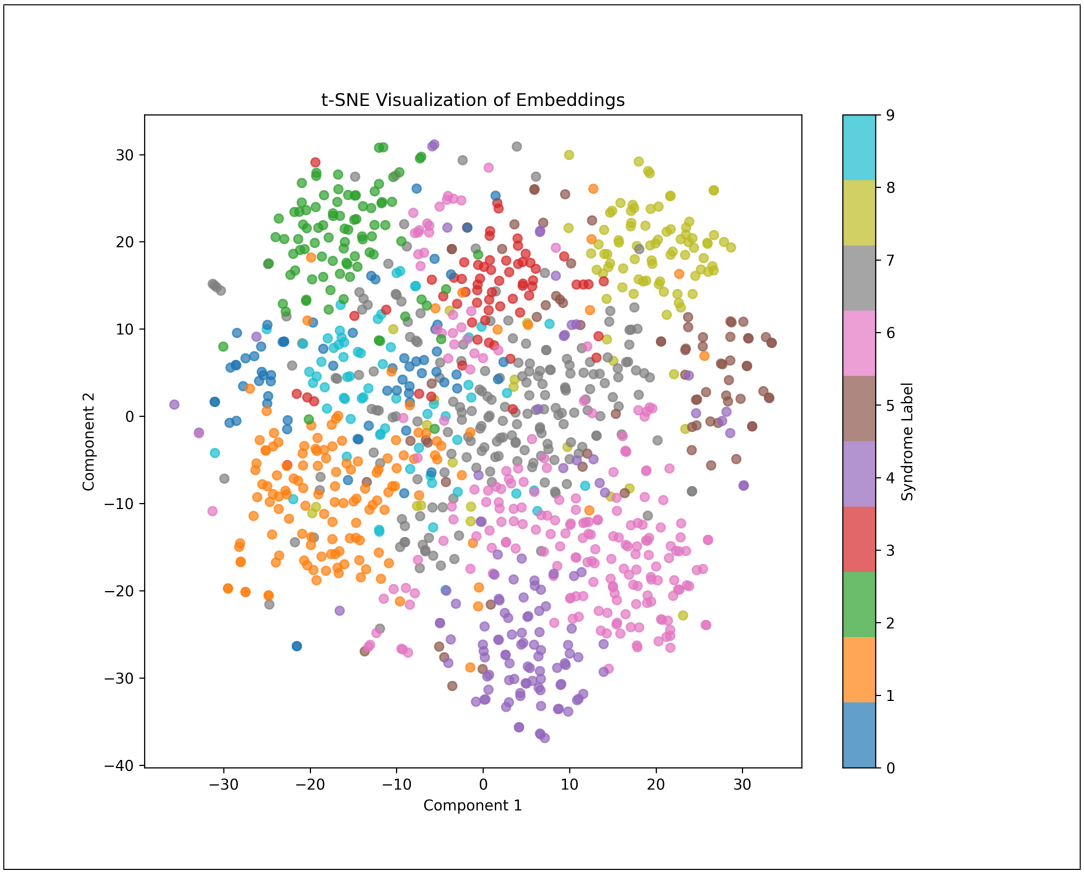

Figure 1: t-SNE projection of embeddings into 2D space, colored by syndrome class label.

3.2 Interpretation

The t-SNE visualization indicates partially separable clusters: some syndromes form compact groups, while several regions show overlap across classes. This suggests that classification performance may vary by syndrome and that neighborhood-based methods may be sensitive to the chosen similarity metric.

4 Classification Methodology

4.1 Model Choice: K-Nearest Neighbors

A K-Nearest Neighbors (KNN) classifier was used as a strong non-parametric baseline for embedding classification. KNN is well-suited for embedding spaces where similarity measures are meaningful.

4.2 Distance Metrics

Two distance metrics were evaluated:

- **Euclidean distance**, which measures absolute distance in the embedding space

- **Cosine distance**, which measures angular similarity between vectors and often aligns well with embedding representations

4.3 Hyperparameter Search

The number of neighbors was selected from:

$$k \in \{1, 2, \ldots, 15\}$$

The optimal k was chosen using cross-validation, maximizing macro one-vs-rest AUC.

4.4 Cross-Validation Protocol

Performance was estimated via **10-fold stratified cross-validation**. For each fold, the model was trained on 9 folds and evaluated on the held-out fold. Results were averaged across folds to obtain robust estimates.

5 Evaluation Metrics

The following metrics were computed:

- **Macro F1-score**: averages class-wise F1 equally across classes
- **Macro one-vs-rest AUC**: averages AUC across classes under one-vs-rest decomposition
- **Top-k accuracy** ($k \in \{1, 3, 5\}$): measures whether the true class appears among the top-k predicted classes
- **Micro-average ROC curve**: computed by flattening one-hot labels and predicted probabilities across classes

AUC and ROC computations were implemented manually using ranking-based formulations and explicit threshold sweeps, rather than relying on library metric functions.

6 Results

6.1 Best Hyperparameters

Both distance metrics achieved their best performance with $k = 15$ under the selected selection criterion.

6.2 Performance Summary (Best Model per Metric)

Table 3: Best KNN configuration per distance metric (10-fold cross-validation averages).

Metric	Best k	AUC (macro OVR)	F1 (macro)	Top-1	Top-3
Euclidean	15	0.9551	0.7370	0.7616	0.9166
Cosine	15	0.9657	0.7685	0.7992	0.9435

Note: Top-5 accuracy for Euclidean and Cosine were 0.9695 and 0.9749, respectively.

6.3 ROC Curve Comparison

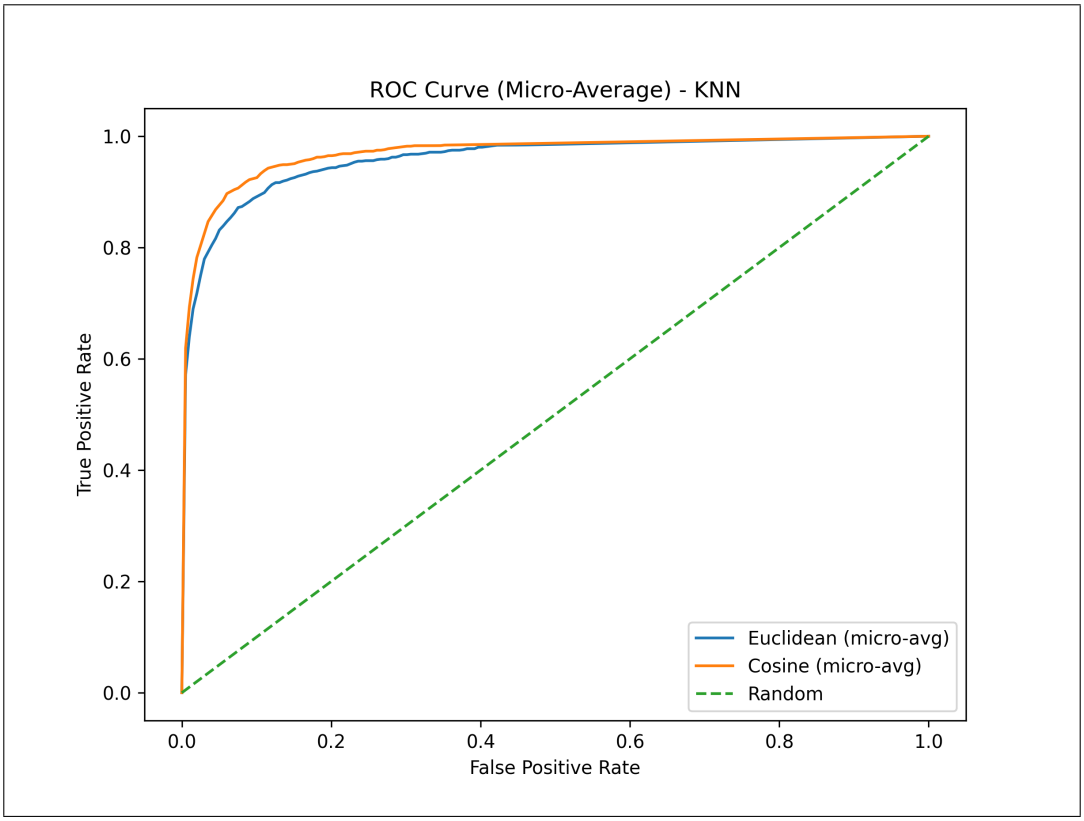

Figure 2: Micro-average ROC curves for KNN using Euclidean and Cosine distances, averaged across cross-validation folds.

6.4 Discussion of Results

Both models substantially outperform the random baseline and achieve high discriminative capability. The Cosine metric consistently outperforms Euclidean across AUC, F1, and top-k metrics, suggesting that angular similarity captures more relevant structure in this embedding space.

Performance improvements with larger k indicate that smoothing decisions over a broader neighborhood helps reduce sensitivity to local noise and overlapping classes, which is consistent with the partial class overlap observed in the t-SNE visualization.

7 Insights

1. **Distance metric selection is critical.** Cosine similarity provided consistently stronger performance, aligning with common behavior in high-dimensional embedding spaces where vector direction carries more discriminative information than magnitude.
2. **High Top-3/Top-5 accuracy indicates strong ranking quality.** Even when Top-1 predictions fail, the true class often appears among the top candidates, which can be valuable in decision-support workflows involving human review.
3. **Visualization supports observed errors.** Overlapping clusters in t-SNE provide a

qualitative explanation for misclassifications that persist even under high AUC.

8 Challenges and Mitigations

- **High-dimensional space (320D):** mitigated using t-SNE for visualization and neighborhood-based modeling for classification.
- **Class imbalance:** handled via stratified cross-validation to preserve label proportions in each fold.
- **Robust evaluation:** ensured by averaging metrics across 10 folds and computing ROC curves using micro-averaging.