Analysis Report – KNN Model

October 1, 2025

1 Executive Summary

This report presents the results of an analysis using the **KNN (K-Nearest Neighbors)** algorithm with vector representations (*embeddings*) of the samples. The objective is to evaluate the model's potential to correctly identify different syndromes, analyzing both technical performance and practical applicability.

Overall, the results were very positive:

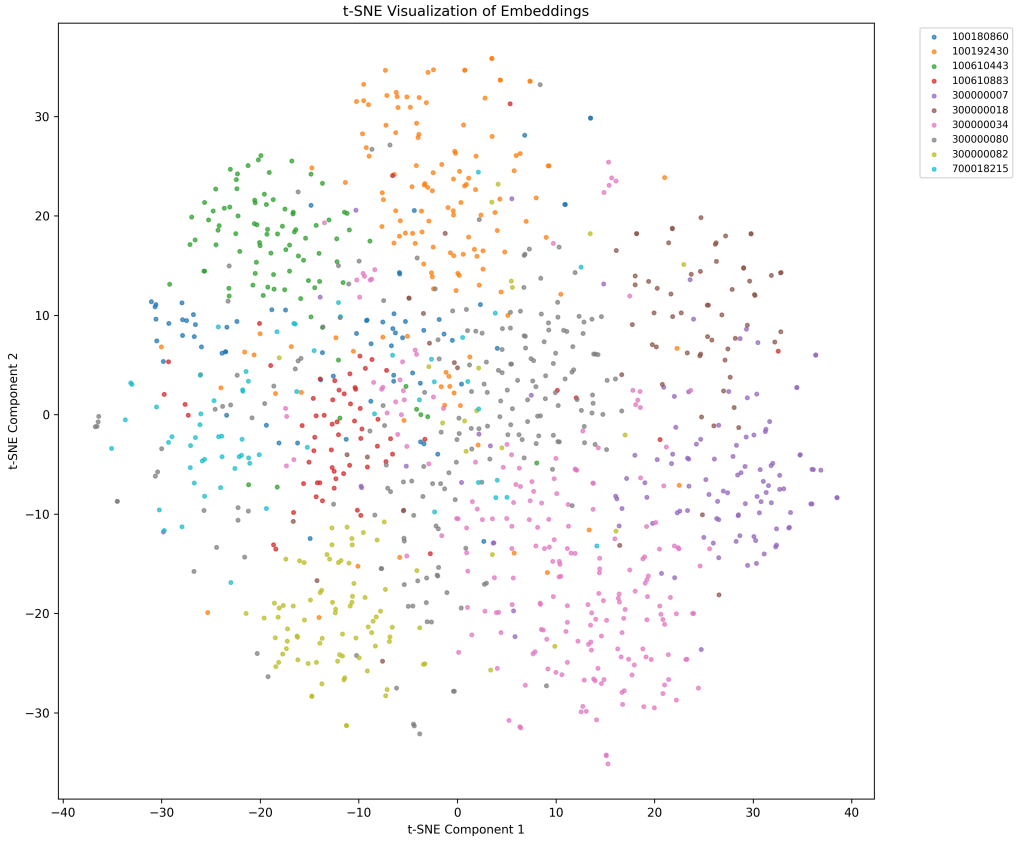
- The model achieved a **Macro AUC of 0.964**, indicating an excellent ability to distinguish between classes.
- The **cosine** distance metric outperformed the Euclidean metric.
- The optimal number of neighbors was $k = 15$, providing robustness against noise.
- The **Top-3 accuracy = 94.4%** is especially promising, as it ensures that in 94% of cases, the correct answer is among the top three suggestions.

These points reinforce that the model has high potential for use in practical scenarios, especially in supporting specialists.

2 Visualizations and Findings

The embeddings used allowed for good separation between the syndromes. This means that, in the vector space, patients with the same diagnosis tend to be close to each other, facilitating correct identification by the KNN algorithm.

2.1 t-SNE Visualization of Embeddings



The graph above shows the distribution of embeddings in a two-dimensional space (via t-SNE). We can observe that:

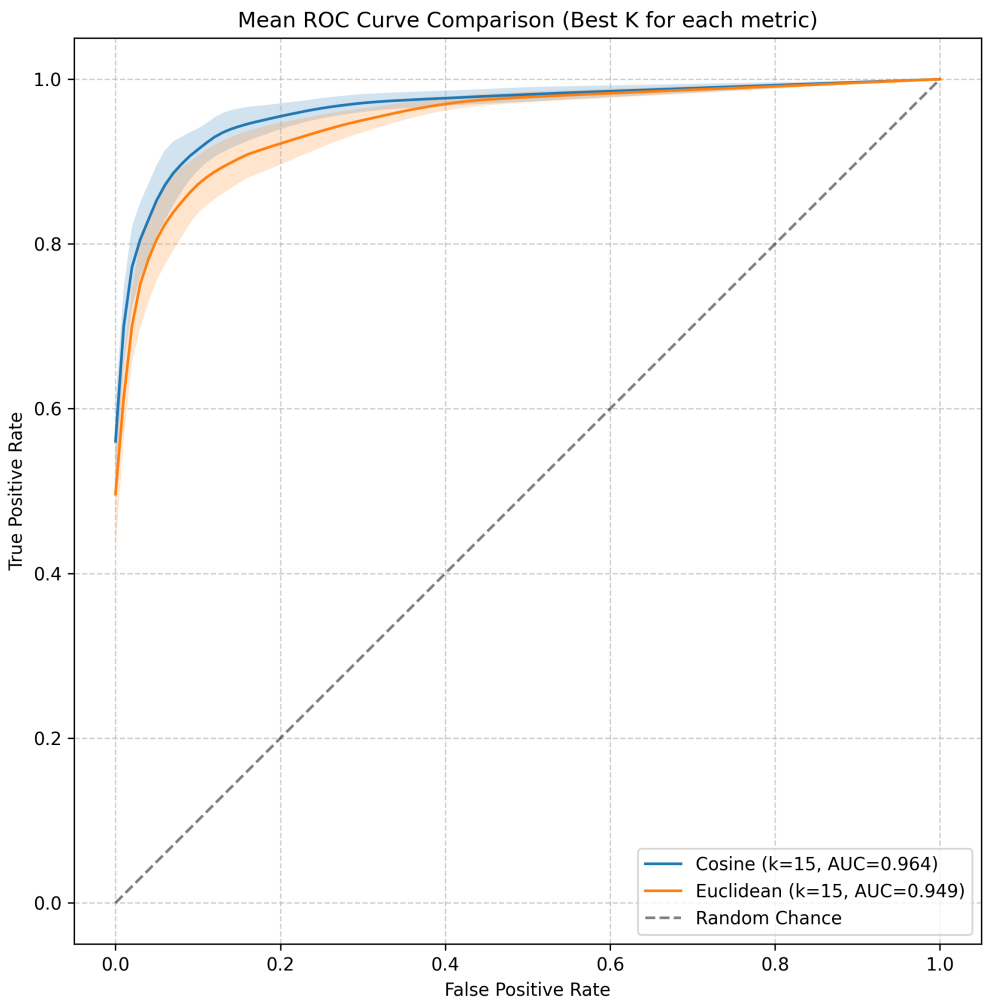
- The samples tend to cluster according to their corresponding syndrome.
- Although there is some overlap between classes, the groups are mostly well-defined.
- This confirms that the embeddings carry sufficient information for discriminating between the syndromes.

2.2 ROC Curves – Comparison of Distance Metrics

Two proximity metrics were compared:

- **Euclidean Distance** – considers the total magnitude between vectors.
- **Cosine Distance** – considers the angle between vectors (direction).

The results showed that **cosine distance was clearly superior** (AUC 0.964 vs. 0.949). This indicates that, for this problem, the direction of the vector is more relevant than its magnitude, which is common in high-dimensional embeddings.



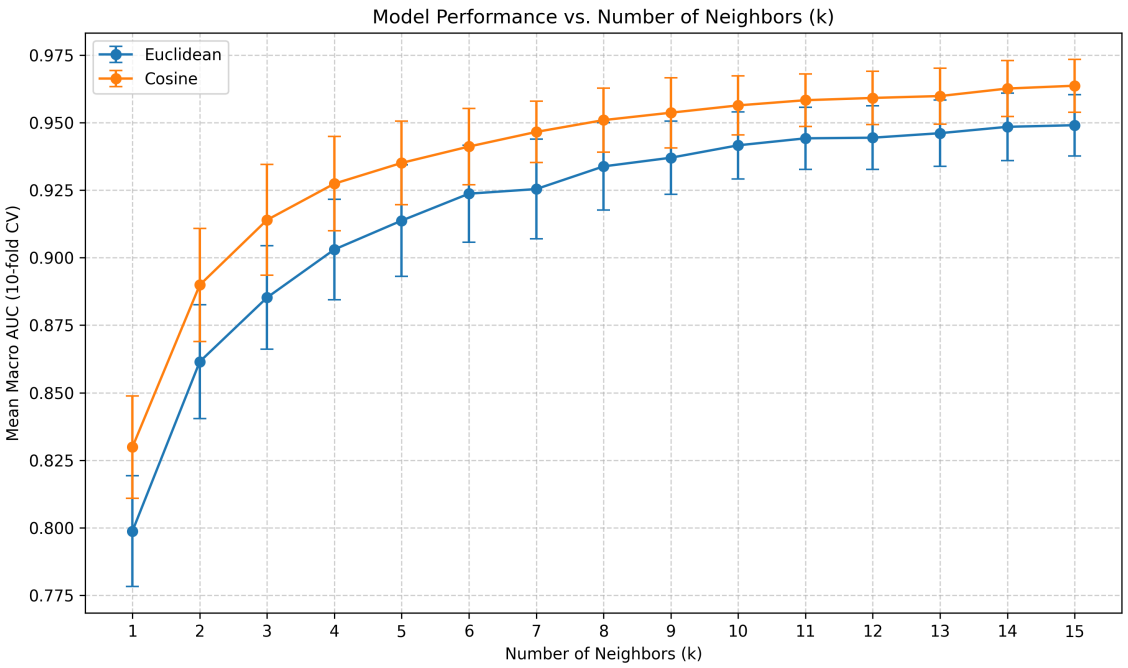
The ROC curve compares the performance between **cosine** and **Euclidean** distance, both with $k = 15$:

- The curve for cosine distance (blue) is consistently above the Euclidean one (orange).
- The AUC for cosine was **0.964**, compared to **0.949** for Euclidean.
- The difference, although numerically small, is consistent across the entire range of False Positive Rates.

Conclusion: Cosine distance is the best metric for this problem, confirming that the direction of the vectors is more relevant than their magnitude.

2.3 Performance as a Function of k

The best performance was achieved with $k = 15$ **neighbors**. Very low values of k make the model more sensitive to noise, while higher values smooth the decision boundary.



The evolution of the mean AUC as a function of the number of neighbors shows that:

- For low values of k, the model is unstable and more susceptible to noise.
- As k increases, both cosine and Euclidean metrics show improved performance.
- The cosine metric maintains an advantage across all values of k.
- Performance stabilizes from $k \approx 12$, reaching its peak at $k = 15$.

Conclusion: Using $k = 15$ neighbors is the optimal balance between stability and precision.

3 Comparison Table

Table 1: Mean performance for different values of k and distance metrics.

Distance	k Range	Mean AUC	Mean F1	Mean Top-3
Euclidean	1–15	0.7988 – 0.9490	0.5806 – 0.7273	0.6218 – 0.7597
Cosine	1–15	0.8299 – 0.9636	0.6613 – 0.7631	0.7008 – 0.8055

4 Overall Conclusion

The results obtained demonstrate that:

- The KNN model with **cosine distance and** $k = 15$ showed the best overall performance.
- The **high AUC (0.964)** confirms the excellent separability of the classes.

- The **Top-3 accuracy (94.4%)** reinforces its potential for use in real-world scenarios, acting as a recommendation system for specialists.

In summary, the KNN model with **cosine distance metric** and $k = 15$ presented:

- Excellent class distinction capability (AUC).
- Good robustness against noise.
- Great practical applicability through the Top-3 metric.

The Top-3 accuracy reached **94.4%**. In practice, this means that in almost all cases, the correct syndrome will be among the top three suggested by the model. This type of metric is extremely valuable in **clinical or research applications**, as it allows the system to function as a **recommendation assistant**, supporting specialists by reducing the search space.