Abstract
We present a novel per-patient evaluation of the Synthetic Minority Oversampling Technique (SMOTE) on structured, class-imbalanced medical data. Instead of using aggregate accuracy metrics alone, we validate each generated synthetic sample's quality by comparing it with test-set instances via the Hassanat distance. We calculate a per-patient SMOTE error rate by measuring how often synthetic samples generated for a given minority-class patient lie closer to majority-class instances in the test set. This allows us to identify which patients are "safe" for SMOTE augmentation and which ones are "risky." Our approach is inspired by the validation framework proposed in "Stop Oversampling for Class Imbalance Learning: A Review" by Tarawneh et al. (2022), where similarity-based validation replaces blind reliance on accuracy metrics.

1. Introduction
Medical datasets often exhibit class imbalance, making it difficult for traditional classifiers to detect rare but critical positive cases. SMOTE is commonly applied to address this imbalance, yet it assumes that the synthesized instances always lie within the minority class region, which can be problematic in practice. We seek to verify whether SMOTE-generated examples truly reflect the minority class per patient, a concern raised in recent work [Tarawneh et al., 2022; Hassanat et al., 2023].

2. Methodology
We implemented a Hassanat-distance-based SMOTE validation system with the following steps:

- **Environment Setup and Libraries Used:**
  - pandas for data manipulation.
  - numpy for numerical computations.
  - matplotlib.pyplot for plotting histograms and bar charts.
  - sklearn.model_selection.train_test_split for stratified train-test splitting.
  - sklearn.neighbors.NearestNeighbors for k-NN lookup in SMOTE generation.
- **Data Splitting:** The dataset is split 60/40 (train/test) with stratification on the class label.
- **Minority Detection:** We isolate each minority-class patient in the training set.
- **SMOTE Generation:** For each minority patient, we generate k = 5 synthetic points using the classic SMOTE approach — linear interpolation between the patient and a randomly selected nearest minority neighbor.
- **Distance Metric:** We use Hassanat distance to compare synthetic samples with all instances in the test set. The distance is bounded, robust to scale, and handles both negative and positive values. The function is defined as follows:

$1 - (1 + \min(p_i, q_i)) / (1 + \max(p_i, q_i))$ if both ≥ 0
$1 - (1 + \min + |\min|) / (1 + \max + |\min|)$ if $\min < 0$

- **Error Rate Computation:** Each synthetic sample is matched to its nearest test instance. If that nearest neighbor belongs to the majority class, the sample is considered erroneous. We compute the mean error per patient to get the "SMOTE error rate."

3. Code Walkthrough

- hassanat_dist(p, q): Custom distance function that applies the Hassanat formula feature-wise and returns the total distance.
- train_test_split(...): Performs stratified data split.
- NearestNeighbors(n_neighbors=k).fit(X_min): Finds k nearest neighbors for each minority patient.
- synth = X_min[i] + δ * (X_min[j] - X_min[i]): Generates a new synthetic point between a patient and one of its neighbors.
- dists = [hassanat_dist(synth, t) for t in X_test]: Calculates similarity scores between synthetic and test points.
- records.append(...): Stores error as 1 if the synthetic's nearest neighbor is majority.
- groupby('patient_id')['error'].mean(): Computes average error rate per patient.

4. Results and Visualizations

- Histogram: Distribution of SMOTE error rates across patients.
- Horizontal Bar Chart: Ranking of patients based on their SMOTE error rate.

5. Interpretation

- High-error patients (e.g., SB-003, SB-033, SB-071) are "risky"—SMOTE tends to generate misleading examples.
- Low-error patients (e.g., SB-011, SB-112, SB-001) are "safe"—synthetic data generated near true minority instances.

6. Paper Inspiration and Theoretical Foundation The implementation and methodology were inspired by:

- **Tarawneh, A.S. et al. (2022)**: Introduced the concept of validating oversampling quality by calculating the proportion of synthetic samples whose nearest neighbors belong to the opposite class. This paper also introduced the oversampling error formula:
  Error=CMSS\text{Error} = \frac{C_M}{S_S} Where CMC_M = count of synthetic samples closest to majority, and SSS_S = total synthetic samples.
- **Hassanat Distance**: Used as the similarity metric due to its robustness to scale and mixed-value features, as proposed in:
  - *The Jeopardy of Learning from Over-Sampled Class-Imbalanced Medical Datasets* (Hassanat et al., 2023)

7. Conclusion
This per-patient audit framework offers a more interpretable and trustworthy way to assess the effectiveness of SMOTE in medical applications. By using Hassanat distance and analyzing error rates individually, practitioners can identify which synthetic data points are likely helpful and which might mislead the classifier. This promotes better sampling decisions and safer model deployment.

References

- Ahmad S. Tarawneh et al. (2022). Stop Oversampling for Class Imbalance Learning: A Review. IEEE Access. DOI: 10.1109/ACCESS.2022.3169512

- Hassanat et al. (2023). The Jeopardy of Learning from Over-Sampled Class-Imbalanced Medical Datasets. IEEE ISCC 2023. DOI: 10.1109/ISCC58397.2023.10218211

- Yang Zhao et al. (2018). A Framework of Rebalancing Imbalanced Healthcare Data. Journal of Healthcare Engineering. DOI: 10.1155/2018/6275435