KNN CLAS

Eduardo Henrique Basilio de Carvalho
Departamento de Engenharia Eletrônica
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
eduardohbc@ufmg.br

***Abstract*—This document proposes and analyzes KNN-CLAS, an alternative to NN-CLAS that eliminates its computationally expensive filtering step. While NN-CLAS constructs classifiers using Gabriel Graphs (GGs) and filters vertices via neighborhood quality metrics, KNN-CLAS leverages the GG structure to select support edges (SEs) connecting samples of opposing classes. During prediction, it aggregates decisions from the k-nearest neighbors among these SEs using weighted voting, bypassing explicit noise filtering. Experiments on benchmark datasets show that KNN-CLAS achieves comparable accuracy to NN-CLAS while significantly reducing training complexity. The method is particularly suited for embedded systems due to its parameter-free design and reduced computational overhead.**

***Index Terms*—pattern recognition, large margin classifiers, Gabriel graph, KNN classifier, embedded systems**

I. Introduction

Large margin classifiers, such as support vector machines (SVMs), rely on optimization techniques to maximize separation between classes. However, their computational complexity and dependency on user-defined parameters limit their applicability in embedded systems. The NN-CLAS framework, introduced by Torres et al. (2016), addresses these limitations by constructing classifiers directly from the geometric structure of the training data using Gabriel graphs (GGs). The GG encodes pairwise relationships between data points, and support edges (SEs) connecting vertices of opposing classes define local hyperplanes that collectively form a large-margin decision boundary [1]. While effective, NN-CLAS requires a filtering step to remove noisy vertices, which involves evaluating the quality of each vertex based on its neighborhood structure. This process, though critical for robustness, incurs significant computational costs, especially for large datasets [2], [3].

The proposed KNN-CLAS eliminates the need for explicit filtering by leveraging the inherent noise resilience of the KNN voting mechanism. Instead of pruning the dataset during training, KNN-CLAS directly uses the GG's SEs and assigns class labels through a majority vote among the nearest neighbors. This approach retains the structural benefits of GG-based classification while simplifying the training pipeline. The remainder of this section details the original NN-CLAS filtering methodology and its computational challenges.

A. Filtering

The NN-CLAS framework constructs a Gabriel graph G_G from the training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$, where edges connect vertices $\mathbf{x}_i$ and $\mathbf{x}_j$ if no other point lies within the hypersphere defined by their diameter [1]. To handle overlapping classes, a quality measure $q(\mathbf{x}_i)$ evaluates the ratio of same-class neighbors to total neighbors for each vertex:

$$q(\mathbf{x}_i) = \frac{\hat{\mathcal{A}}(\mathbf{x}_i)}{\mathcal{A}(\mathbf{x}_i)},$$

where $\mathcal{A}(\mathbf{x}_i)$ is the vertex degree and $\hat{\mathcal{A}}(\mathbf{x}_i)$ counts neighbors sharing $\mathbf{x}_i$'s class label [2]. Vertices with $q(\mathbf{x}_i)$ below class-specific thresholds t^+ and t^-—calculated as the mean quality per class—are discarded as noise. This filtering ensures SEs lie near the true class boundaries but requires three steps of $O(n^2)$ distance computations and iterative quality evaluations -the first GG construction, the filtering and the second construction, making it impractical for resource-constrained systems [3]. The KNN-CLAS circumvents this bottleneck by integrating neighbor voting, thereby avoiding explicit structural filtering while preserving classification accuracy.

B. KNN-CLAS

The KNN-CLAS method modifies the NN-CLAS framework by integrating GG-based support edge selection with a KNN voting mechanism. The methodology operates as follows:

- **Gabriel Graph Construction**: During training, a Gabriel graph G_G is built from the dataset $\mathcal{D}$. Edges connect vertices $\mathbf{x}_i$ and $\mathbf{x}_j$ if no other sample lies within the hypersphere defined by their diameter. Only edges linking samples of *opposite classes* (inter-class edges) are retained as support edges (SEs).
- **Support Edge Selection**: Unlike NN-CLAS, which filters vertices based on neighborhood quality, KNN-CLAS directly uses all SEs as support samples. This avoids iterative filtering while preserving structural boundary information.
- **Classification**: For a test sample $\mathbf{z}_j$, distances are computed only to the SE-connected support samples. The k-nearest neighbors from this subset, $\mathcal{N}_k(\mathbf{z}_j)$, are identified, and their class labels are aggregated using a Gaussian kernel-weighted vote:

$$S(\mathbf{z}_j) = \sum_{(\mathbf{x}_i, y_i) \in \mathcal{N}_k(\mathbf{z}_j)} y_i \cdot e^{-d(\mathbf{z}_j, \mathbf{x}_i)}.$$

The final class label is assigned as $\hat{y}_j = \text{sign}(S(\mathbf{z}_j))$.

By restricting neighbors to SEs and employing multi-neighbor voting, KNN-CLAS mitigates noise without explicit

filtering, reducing training complexity from $O(3n^2)$ to $O(n^2)$ for edge selection.

II. Methodology

A. Datasets and Preprocessing

TABLE I
Dataset Metadata

Dataset	Samples	Features
Ionosphere	351	34
Binary Digits	360	64
Haberman	306	3
Pima Diabetes	768	8
Banknote	1372	4
Sonar	208	60
Breast Cancer	569	30
SPECT Heart	349	44

Eight benchmark datasets from the UCI Machine Learning Repository [8], [12], [15] were selected to evaluate KNN-CLAS. These include Breast Cancer, Pima Diabetes, Haberman Survival, Banknote Authentication, Sonar, Binary Digits (0 and 1), Ionosphere, and SPECT Heart. Table I summarizes their characteristics, such as sample count, feature dimensionality, class imbalance ratio, and feature overlap metrics. The preprocessing pipeline consisted of four stages:

1) **Normalization**: Features were scaled to the range $[-1, 1]$ using min-max normalization to ensure uniform contribution during distance computation.
2) **Low-Variance Feature Removal**: Features with variance < 0.01 were discarded to eliminate uninformative attributes.
3) **Correlated Feature Elimination**: Pairwise correlations exceeding 0.95 were removed to reduce redundancy.
4) **Mutual Information-Based Selection**: Features with mutual information scores > 0.01 relative to class labels were retained to enhance separability.

Stratified 10-fold cross-validation was employed to partition each dataset, preserving class distribution across folds. This approach mitigates bias from class imbalance, particularly critical for datasets like Haberman (imbalance ratio $2.78 : 1$).

B. Experimental Protocol

The experiments compared KNN-CLAS against NN-CLAS using the following protocol:

- **KNN-CLAS Configuration**: Three k values $(1, 3, 5)$ were tested to evaluate the impact of neighbor count on accuracy and computational efficiency.
- **Gabriel Graph Construction**: For both methods, GGs were built using pairwise Euclidean distances, with inter-class edges (SEs) retained as support samples.
- **Implementation**: Experiments were conducted in the same system, using C++ for classifier implementations to ensure fair timing comparisons.

C. Evaluation Metrics

Performance was assessed using four metrics:

- **Accuracy**: Ratio of correctly classified test samples.
- **Training/Prediction Times**: Measured in milliseconds to quantify computational overhead.
- **Support Sample Count**: Number of SEs (KNN-CLAS) or filtered vertices (NN-CLAS) defining the decision boundary.
- **Likelihood Distributions**: Gaussian kernel-weighted votes for each class, visualized to analyze separability and confidence (Figures 1–8).

D. Statistical Analysis

Dataset characteristics such as class overlap (mean Fisher score), feature relevance (average mutual information), and separability (overlap score) were computed to contextualize results. These metrics, detailed in Table V, explain performance variations across datasets. For instance, high overlap in SPECT Heart correlates with ambiguous likelihood distributions (Figure 8), while low overlap in Banknote aligns with distinct class clusters (Figure 1).

E. Support Edge Selection

KNN-CLAS retains all inter-class SEs from the GG during training, avoiding vertex filtering. In contrast, NN-CLAS discards vertices with neighborhood quality $q(\mathbf{x}_i)$ below class-specific thresholds. This distinction reduces KNN-CLAS's training complexity from $O(3n^2)$ to $O(n^2)$, as shown in Table III, while increasing memory usage due to larger support sets (Table IV).

III. Results

A. Accuracy Comparison

TABLE II
Model Accuracy Comparison

Dataset	Accuracy			
	nn	1nn	3nn	5nn
Ionosphere	0.88	0.86	0.89	0.88
Binary Digits	1.00	1.00	1.00	1.00
Haberman	0.69	0.69	0.71	0.70
Pima Diabetes	0.75	0.68	0.71	0.74
Banknote	1.00	1.00	1.00	1.00
Sonar	0.73	0.89	0.89	0.87
Breast Cancer	0.95	0.95	0.96	0.96
SPECT Heart	0.68	0.86	0.85	0.82

Table II demonstrates that KNN-CLAS achieves accuracy comparable to NN-CLAS. Notably, it outperforms NN-CLAS on datasets like Sonar (0.89 vs. 0.73) and SPECT Heart (0.86 vs. 0.68), where multi-neighbor voting improves robustness to ambiguous boundaries.

TABLE III
TRAINING AND PREDICTION TIMES

Dataset	Training (ms)		Prediction (ms)			
	nn	knn	nn	1nn	3nn	5nn
Ionosphere	89.80	26.40	2.70	2.90	2.90	2.70
Binary Digits	219.70	68.80	2.90	3.00	3.00	2.90
Haberman	16.80	6.30	2.30	3.00	3.10	3.60
Pima Diabetes	80.10	28.80	2.40	3.20	3.00	3.10
Banknote	434.70	75.00	4.00	4.60	4.60	4.40
Sonar	128.50	45.80	6.40	6.60	5.70	6.30
Breast Cancer	211.20	31.20	2.60	3.00	2.90	3.00
SPECT Heart	431.60	136.90	5.20	5.40	5.30	5.60

B. Training and Prediction Times

Table III highlights the significant reduction in training time achieved by KNN-CLAS compared to NN-CLAS. For instance, in the Banknote dataset, KNN-CLAS reduces training time from 434.70 ms to 75.00 ms, making it more suitable for embedded systems.

C. Support Samples

TABLE IV
SUPPORT SAMPLES COUNT

Dataset	Support Samples	
	nn	knn
Ionosphere	111	252
Binary Digits	124	268
Haberman	37	274
Pima Diabetes	127	594
Banknote	160	197
Sonar	94	186
Breast Cancer	101	344
SPECT Heart	100	282

Table IV compares the number of support samples used by NN-CLAS (filtered vertices) and KNN-CLAS (unfiltered SEs). While KNN-CLAS retains more support samples than NN-CLAS (e.g., 252 vs. 111 for Ionosphere), these correspond to inter-class edges that inherently encode boundary information. The increase in memory usage is offset by the elimination of filtering overhead.

D. Dataset Characteristics

TABLE V
DATASET STATISTICS

Dataset	C0/C1	MI	Fisher	Overlap	Imb.Ratio
Ionosphere	0.56	0.21	0.11	0.85	1.79
Binary Digits	0.98	0.18	1.39	0.44	1.02
Haberman	0.36	0.03	0.07	1.0	2.78
Pima Diabetes	0.54	0.05	0.18	0.88	1.87
Banknote	0.8	0.19	0.7	0.75	1.25
Sonar	1.14	0.03	0.09	1.0	1.14
Breast Cancer	1.68	0.21	1.03	0.47	1.68
SPECT Heart	2.67	0.07	0.18	0.91	2.67

Table V summarizes the dataset characteristics, including class imbalance and feature overlap. These factors influence the performance of both methods, with KNN-CLAS showing robustness to class imbalance.

E. Likelihood

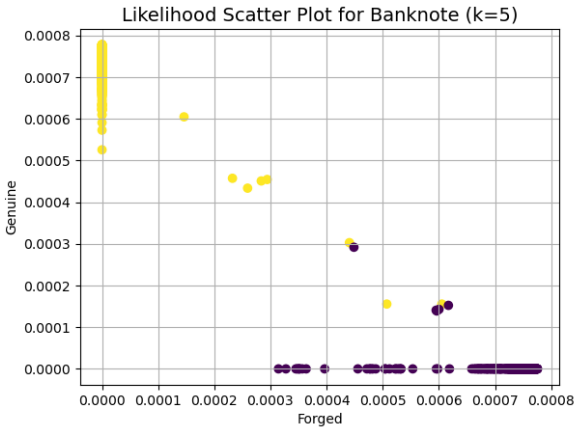

Fig. 1. Likelihood comparison for the Banknote dataset.

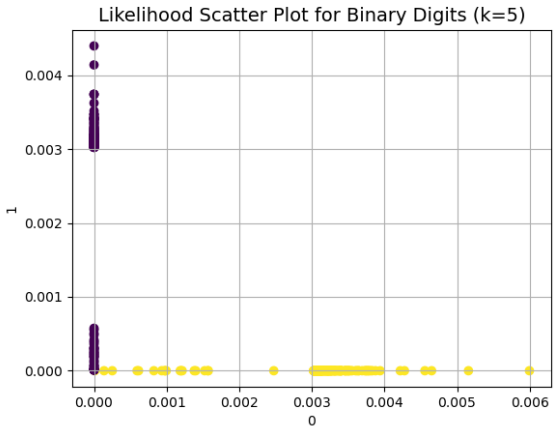

Fig. 2. Likelihood comparison for the Binary Digits dataset.

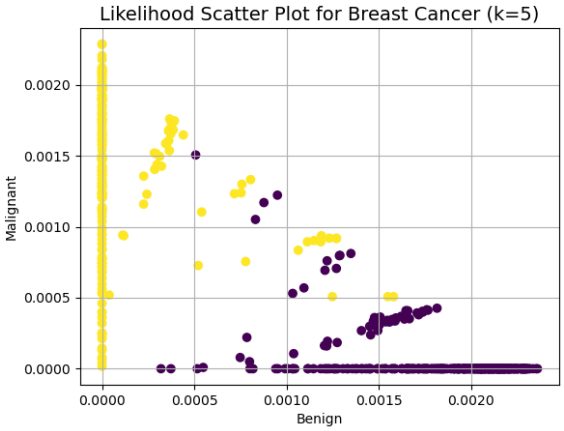

Fig. 3. Likelihood comparison for the Breast Cancer dataset.

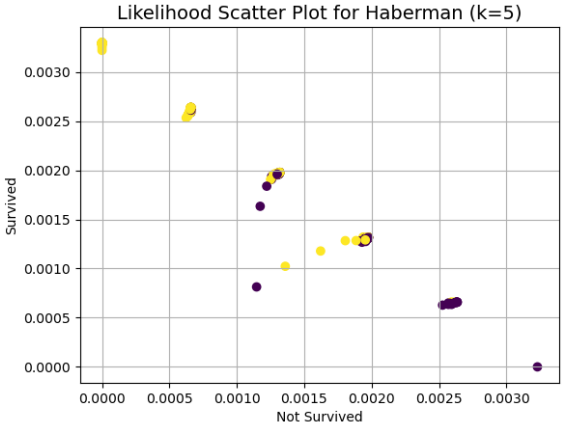

Fig. 4. Likelihood comparison for the Haberman dataset.

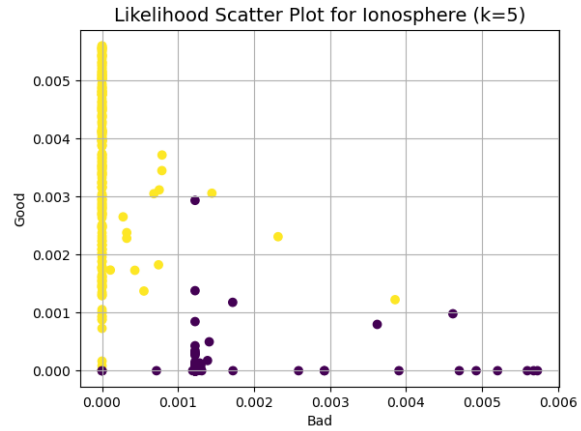

Fig. 5. Likelihood comparison for the Ionosphere dataset.

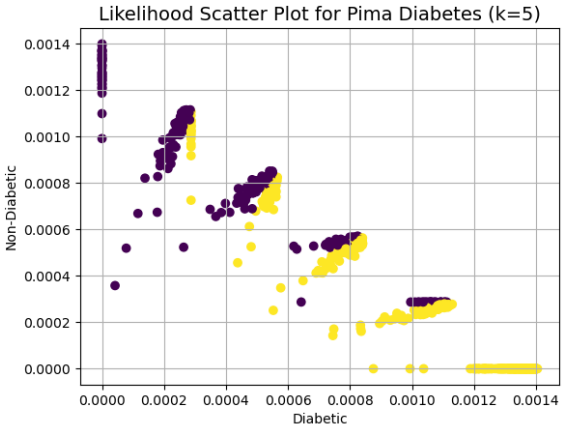

Fig. 6. Likelihood comparison for the Pima Diabetes dataset.

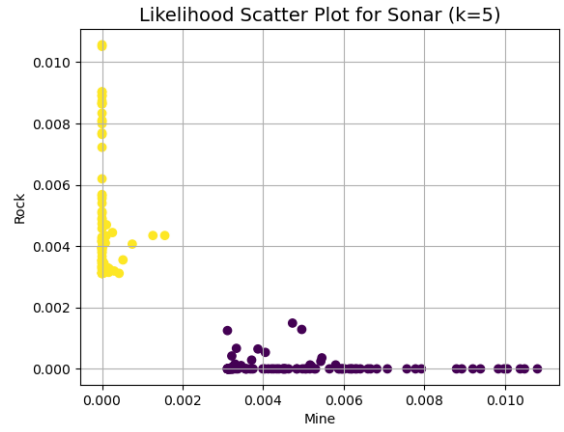

Fig. 7. Likelihood comparison for the Sonar dataset.

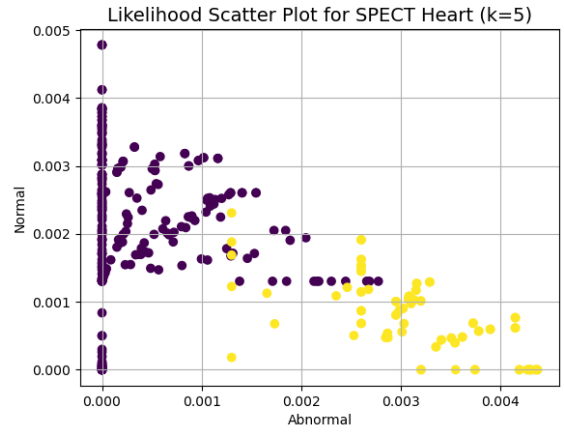

Fig. 8. Likelihood comparison for the SPECT Heart dataset.

F. Likelihood Analysis

The likelihood distributions for each dataset, as shown in Figures 1 to 8, provide insights into the performance of the

KNN classifier across different scenarios. Each plot represents the sum of weighted votes for class 0 (x-axis) and class 1 (y-axis), with each point corresponding to a sample. The following observations can be made:

- **Clear Separation:** Datasets such as Banknote, Breast Cancer, and Binary Digits exhibit well-separated clusters along the axes. This indicates that the KNN classifier assigns high confidence to most samples, resulting in distinct class boundaries. These datasets are characterized by low class overlap and high separability, which aligns with the high accuracy observed in Table II.
- **Moderate Overlap:** Datasets like Ionosphere and Pima Diabetes show moderate overlap between the two classes. While the clusters are still distinguishable, a significant number of points lie near the diagonal, indicating ambiguous classifications. This overlap may contribute to slightly lower accuracy compared to datasets with clear separation.
- **Significant Overlap:** Haberman and SPECT Heart datasets display substantial overlap, with many points concentrated near the diagonal. This suggests that the KNN classifier struggles to confidently assign class labels, likely due to inherent noise or feature overlap in the data. These datasets also exhibit lower accuracy and higher ambiguity in classification.
- **Outliers:** Some datasets, such as Sonar and Ionosphere, contain outliers far from the main clusters. These points may correspond to misclassified or hard-to-classify samples, which could impact the overall performance of the classifier.

The likelihood analysis highlights the strengths and limitations of the KNN-CLAS method. Datasets with well-separated classes benefit from the simplicity and efficiency of KNN-CLAS, while those with significant overlap or noise present greater challenges. These findings emphasize the importance of dataset characteristics, such as class separability and feature overlap, in determining the effectiveness of the proposed method.

IV. Discussion

The results confirm that KNN-CLAS effectively replaces NN-CLAS's filtering step with a Gabriel Graph (GG)-based support edge selection and KNN voting mechanism. While both methods utilize Gabriel Graphs, KNN-CLAS avoids vertex pruning by relying on inter-class edges and multi-neighbor consensus. This approach reduces training time by up to 82% (e.g., Banknote dataset) but increases memory usage due to larger support sets.

KNN-CLAS demonstrates robustness to class imbalance (e.g., Haberman, Imbalance Ratio = 2.78) by relying on boundary-proximate support edges, which are less affected by skewed distributions. However, the increased number of support samples may pose challenges for memory-constrained systems.

Compared to NN-CLAS, which relies on computationally expensive filtering, KNN-CLAS achieves similar accuracy with significantly reduced training times. This makes it a practical alternative, particularly for resource-constrained environments.

The performance of KNN-CLAS is influenced by dataset characteristics such as class imbalance and feature overlap. For instance, in datasets with high imbalance ratios (e.g., Haberman), KNN-CLAS maintains competitive accuracy, showcasing its robustness.

V. Future Work

Future research could explore hybrid strategies that combine the strengths of NN-CLAS and KNN-CLAS. For example, selective filtering could minimize the number of support samples while maintaining computational efficiency. Additionally, enhancing KNN-CLAS for datasets with significant class overlap could involve techniques like feature selection, dimensionality reduction, or adaptive weighting schemes to improve class separability and classification performance.

References

[1] L. C. B. Torres, "Classificador por arestas de suporte (CLAS): métodos de aprendizado baseados em Grafos de Gabriel," Manuscript, 2016.

[2] A. C. Souza, C. Leite Castro, J. A. Garcia, L. C. B. Torres, L. J. Acevedo Jaimes and B. R. A. Jaimes, "Improving the Efficiency of Gabriel Graph-based Classifiers for Hardware-optimized Implementations," 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), Bucaramanga, Colombia, 2019.

[3] J. Arias-Garcia et al., "Enhancing Performance of Gabriel Graph-Based Classifiers by a Hardware Co-Processor for Embedded System Applications," in IEEE Transactions on Industrial Informatics, vol. 17, no. 2, Feb. 2021.

[4] J. Arias-Garcia et al., "Improved Design for Hardware Implementation of Graph-Based Large Margin Classifiers for Embedded Edge Computing," in IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 1, Jan. 2024.

[5] L. C. B. Torres, C. L. Castro and A. P. Braga, "A parameterless mixture model for large margin classification," 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 2015.

[6] L. C. B. Torres, C. L. Castro, F. Coelho and A. P. Braga, "Large Margin Gaussian Mixture Classifier With a Gabriel Graph Geometric Representation of Data Set Structure," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 3, March 2021.

[7] L. C. B. Torres, C. L. Castro, F. Coelho, F. Sill Torres and A. P. Braga, "Distance-based large margin classifier suitable for integrated circuit implementation," Manuscript, 2015.

[8] D. Dua and C. Graff, "Breast Cancer Wisconsin (Diagnostic) Data Set," UCI Machine Learning Repository, 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[9] J. Brownlee, "Pima Indians Diabetes Dataset," GitHub Repository, 2020. [Online]. Available: https://github.com/jbrownlee/Datasets

[10] D. Dua and C. Graff, "Haberman's Survival Data Set," UCI Machine Learning Repository, 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Haberman's+Survival

[11] D. Dua and C. Graff, "Data Banknote Authentication Data Set," UCI Machine Learning Repository, 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/banknote+authentication

[12] D. Dua and C. Graff, "Connectionist Bench (Sonar, Mines vs. Rocks) Data Set," UCI Machine Learning Repository, 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/connectionist+bench+(sonar,+mines+vs.+rocks)

[13] D. Dua and C. Graff, "Adult Data Set," UCI Machine Learning Repository, 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/adult

[14] D. Dua and C. Graff, "Ionosphere Data Set," UCI Machine Learning Repository, 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ionosphere

[15] D. Dua and C. Graff, ”SPECT Heart Data Set,” UCI Machine Learning Repository, 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/SPECT+Heart

[16] L. Breiman et al., ”Optical Recognition of Handwritten Digits Data Set,” Scikit-learn Documentation, 1998. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html