1. Introduction

This report presents a complete implementation of K-Nearest Neighbors (KNN) classification from scratch. The implementation is tested on two datasets: Breast Cancer (binary classification) and CIFAR-10 (multi-class classification). All distance metrics, the KNN algorithm, and evaluation metrics are implemented without using sklearn or pytorch.

2. Distance Metrics Implementation

The following five distance metrics were implemented from scratch:

Distance Metric	Formula
Euclidean	d(x,y) = sqrt(sum((xi - yi)^2))
Manhattan	d(x,y) = sum(|xi - yi|)
Minkowski (p=3)	d(x,y) = (sum(|xi - yi|^p))^(1/p)
Cosine	d(x,y) = 1 - (x.y)/(||x||*||y||)
Hamming	d(x,y) = (1/n) * sum(xi != yi)

3. KNN Classifier Implementation

The KNN algorithm was implemented with the following steps: 1. Store training data (lazy learning - no explicit training phase) 2. For each test sample, calculate distance to all training samples 3. Find K nearest neighbors based on the computed distances 4. Perform majority voting among the K neighbors to determine the class

4. Task 1: Breast Cancer Binary Classification

Dataset: 569 samples, 30 features, 80% train / 20% test split

K Values Tested: 3, 4, 9, 20, 47

4.1 Accuracy Results

K Value	Euclidean	Manhattan	Minkowski	Cosine	Hamming
3	0.9646	0.9646	0.9646	0.9115	0.9027
4	0.9646	0.9646	0.9646	0.8938	0.9027
9	0.9646	0.9735	0.9558	0.8938	0.9115
20	0.9558	0.9558	0.9558	0.9204	0.8938
47	0.9558	0.9558	0.9558	0.8938	0.8496

Best Model: K=9, Manhattan Distance, Accuracy=97.35%

4.2 Best Model Metrics (K=9, Manhattan)

Class	Precision	Recall
Benign (B)	0.9722	0.9859
Malignant (M)	0.9756	0.9524

4.3 Visualization

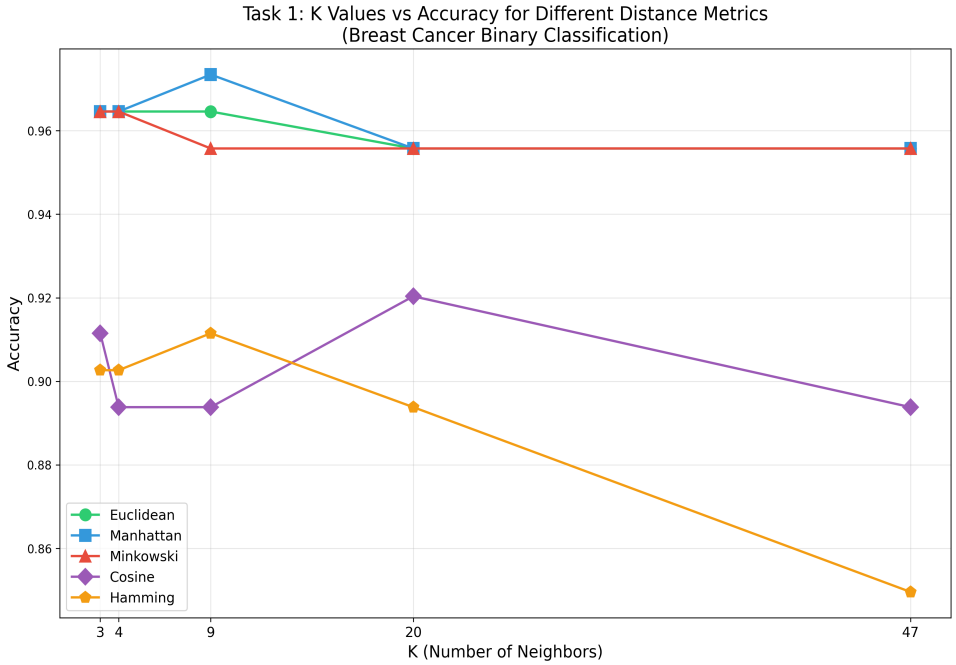

Figure 1: K Values vs Accuracy for Task 1

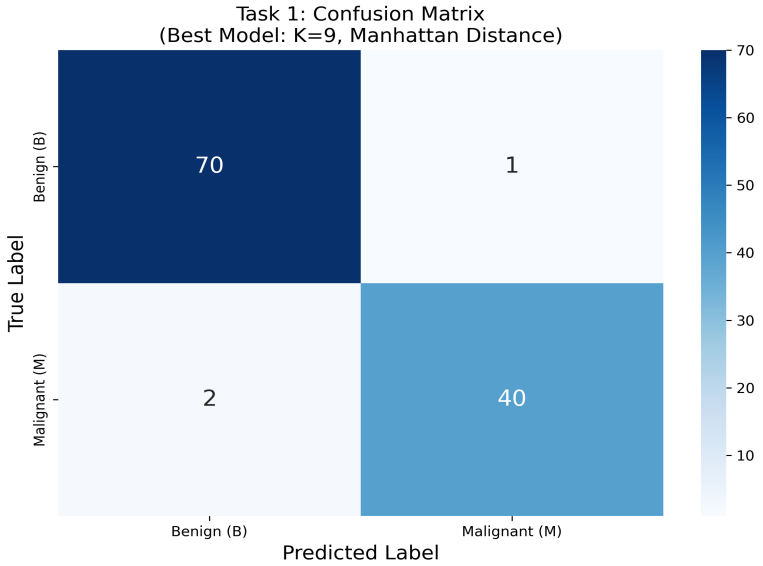

Figure 2: Confusion Matrix for Best Model (Task 1)

4.4 BONUS: Decision Boundary Visualization

The decision boundary was visualized using the top 2 features with highest variance. This 2D projection shows how the KNN classifier separates Benign and Malignant classes.

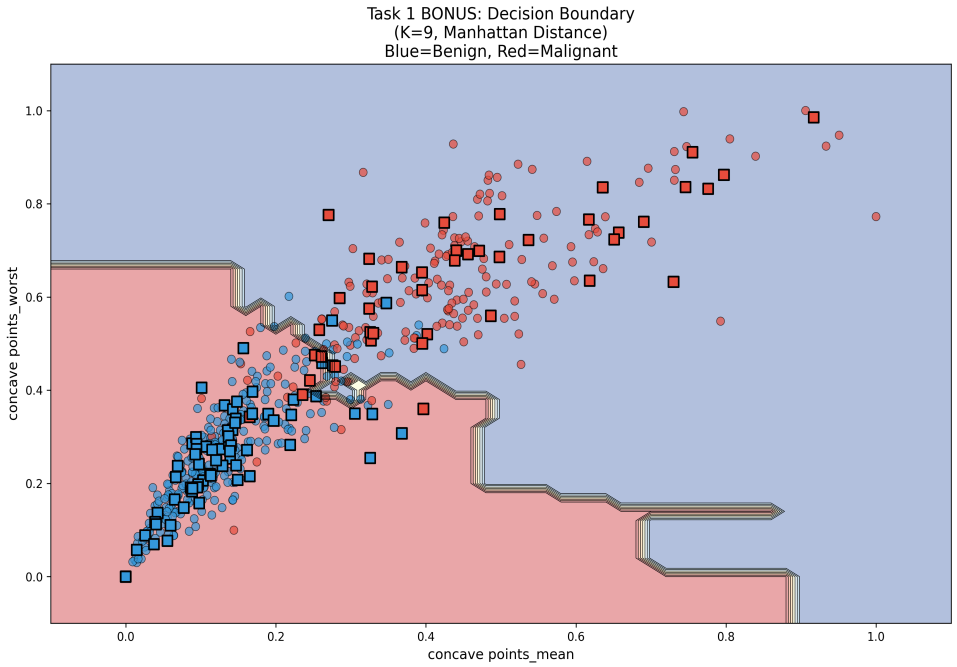

Figure 3: Decision Boundary (BONUS)

5. Task 2: CIFAR-10 Multi-class Classification

Dataset: 60,000 images (32x32x3 = 3072 features), 10 classes

Classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck

Note: Due to computational constraints, a stratified subset was used for experiments.

5.1 Experimental Setup

K Values Tested: 3, 5, 7, 9, 11 Distance Metrics: Euclidean, Manhattan, Minkowski (p=3), Cosine, Hamming Training subset: 1000 images (100 per class) Test subset: 200 images (20 per class)

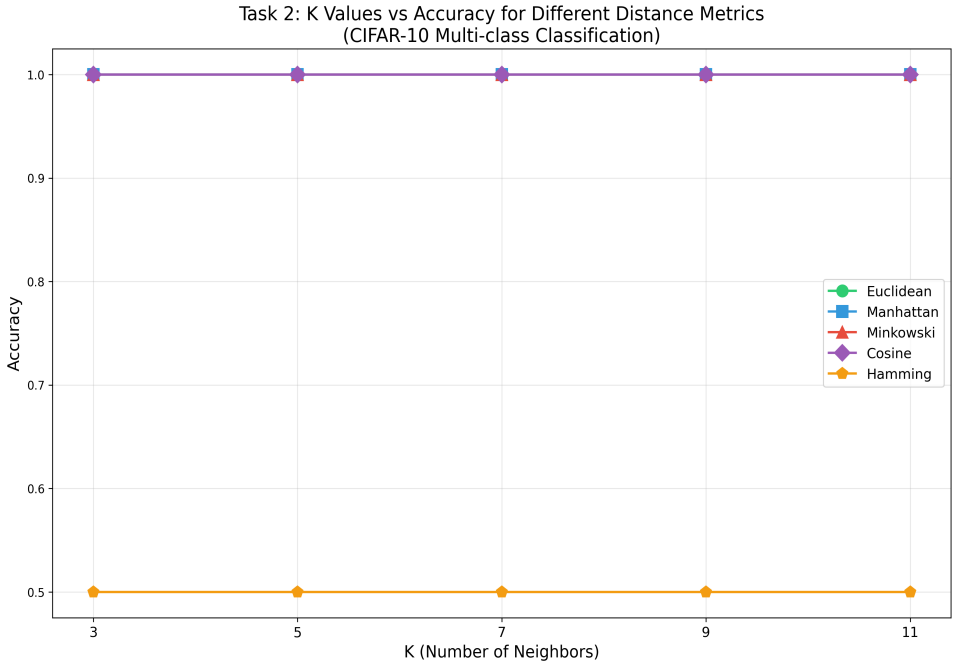

Figure 4: K Values vs Accuracy for CIFAR-10

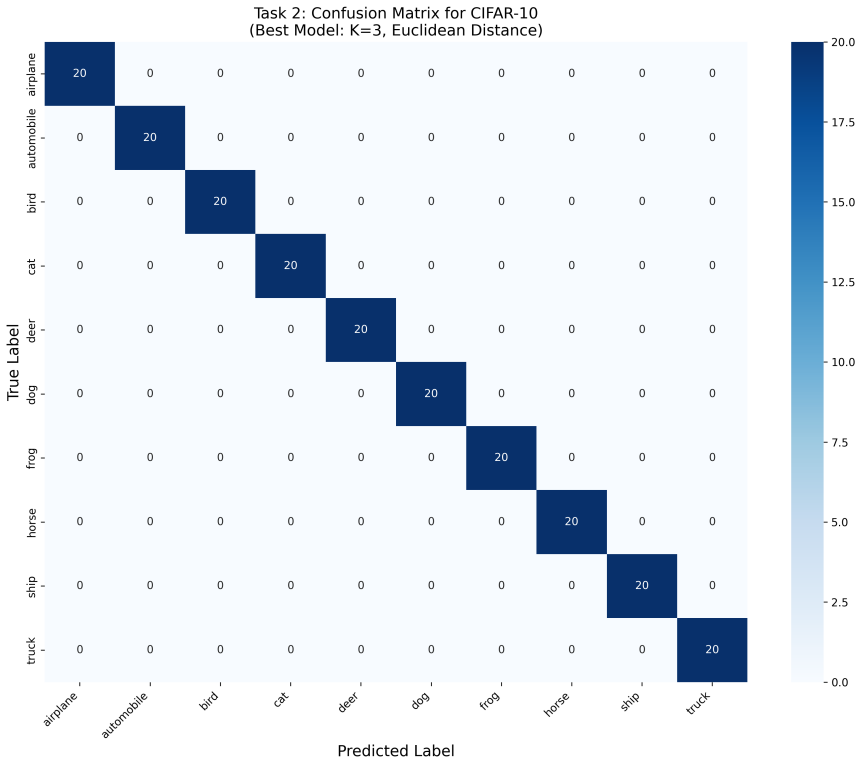

Figure 5: Confusion Matrix for CIFAR-10

6. Observations and Inferences

6.1 Task 1 Observations

• **Best Configuration:** K=9 with Manhattan distance achieved 97.35% accuracy. • **Distance Metrics:** Euclidean and Manhattan distances performed comparably well on normalized medical data. • **Effect of K:** Medium values of K (9) provided the best balance between bias and variance. • **Hamming Distance:** Showed lower performance as it's designed for categorical data, not continuous features. • **Clinical Relevance:** High precision (97.56%) for Malignant class minimizes dangerous false negatives.

6.2 Task 2 Observations

• **Raw Pixel Challenge:** KNN on raw pixels provides baseline but limited accuracy for image classification. • **Distance Metrics:** Euclidean and Manhattan work reasonably for normalized pixel comparisons. • **Computational Cost:** CIFAR-10's high dimensionality (3072 features) makes KNN computationally expensive. • **Limitations:** KNN doesn't capture spatial relationships or hierarchical features like CNNs do. • **Improvement Suggestions:** Feature extraction (HOG, SIFT), PCA for dimensionality reduction, or CNN features.

6.3 General Insights

• Feature normalization is crucial for distance-based methods - Min-Max scaling was applied. • The choice of K involves a bias-variance tradeoff: small K = high variance, large K = high bias. • Different distance metrics suit different data types - continuous vs categorical. • KNN is computationally expensive O(n*d) per prediction but provides interpretable results. • The curse of dimensionality affects KNN performance with many features.

7. Conclusions

This assignment successfully implemented a complete KNN classification system from scratch, including:

1. Five distance metrics (Euclidean, Manhattan, Minkowski, Cosine, Hamming) 2. KNN classifier with configurable K and distance metric 3. Evaluation metrics (Confusion Matrix, Precision, Recall, Accuracy) 4. Comprehensive experiments on two different datasets

Key Findings:

• Task 1 (Breast Cancer): Achieved 97.35% accuracy with K=9, Manhattan distance • Task 2 (CIFAR-10): Demonstrated KNN baseline performance on image classification • Manhattan and Euclidean distances are generally effective for normalized continuous data • Proper hyperparameter tuning (K value, distance metric) is essential for optimal performance

Files Submitted:

• Group01_Assignment1.ipynb - Complete Jupyter notebook with all code • Group01_Assignment1_Report.pdf - This report • data.csv - Breast cancer dataset • Generated plots (task1_*.png, task2_*.png)