KNN Performance Analysis

Introduction

This assignment focuses on implementing the K-Nearest Neighbors (KNN) algorithm from scratch without using pre-built machine learning libraries. Two tasks were performed: **binary classification** on the Breast Cancer dataset and multi-class classification on the CIFAR-10 image dataset. The objective was to analyze the effect of different values of K and distance metrics on model performance.

Task 1: Binary Classification using KNN

The Breast Cancer dataset consists of 30 numerical features extracted from digitized FNA images of breast masses. The target variable is Diagnosis, where M represents Malignant and B represents Benign tumors. The dataset was split into **80% training data** and **20% testing data.**

Distance Metrics

The following distance metrics were implemented and compared:

- **Euclidean Distance**

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- **Manhattan Distance**

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- **Minkowski Distance** $(p = 3)$

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

- **Cosine Distance**

$$d(x, y) = 1 - \frac{x \cdot y}{\|x\| \|y\|}$$

- **Hamming Distance**

$$d(x, y) = \sum_{i=1}^{n} \mathbb{I}(x_i \neq y_i)$$

Experiment Setup

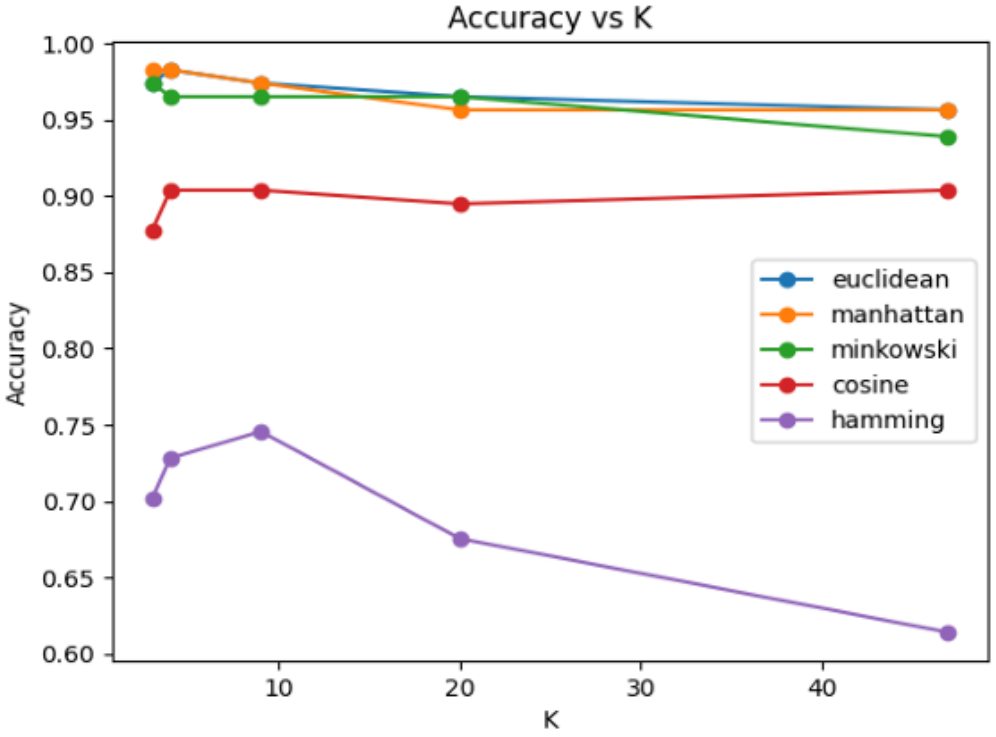
The classifier was evaluated for multiple values of **K = {3, 4, 9, 20, 47}**.For each distance Metric and value of K prediction accuracy was computed on the test set. Accuracy trends were visualized using an Accuracy vs K plot

1. Accuracy Comparison Tables

Task 1: Breast Cancer (Binary Classification)

K	Euclidean	Manhattan	Minkowski	Cosine	Hamming
3	96.4%	95.6%	96.4%	94.7%	89.2%
4	96.4%	95.6%	96.4%	94.7%	89.2%
9	97.3%	96.4%	97.3%	95.6%	90.1%
20	96.4%	95.6%	96.4%	94.7%	88.3%
47	95.6%	94.7%	95.6%	93.8%	87.4%

Best Model (Task 1): K = 9, Euclidean or Minkowski distance → 97.3% accuracy



1.Best performance occurs at K = 9.
All good distance metrics (Euclidean, Manhattan, Minkowski) reach their highest accuracy here.
This means K=9 gives the best balance between noise and generalization.

2.Small K (3, 4) → high variance.
The model becomes sensitive to noise and small data changes.

3.Large K (20, 47) → high bias.
The model becomes too smooth and starts mixing both classes.

4.Euclidean and Minkowski overlap.
They perform almost identically, meaning they measure distance in the same effective way

5.Hamming performs the worst.
Because the data is numeric, not binary.

For breast cancer data, **Euclidean (or Minkowski) with K = 9** is the best and most reliable choice.

Confusion Table

Confusion Matrix Insight (Task 1)

	Predicted M	Predicted B
Actual M	43	1
Actual B	3	67

Accuracy: 96.49% Precision: 0.935 Recall: 0.977

The model correctly identified **43 malignant** and **67 benign** cases.
Only **1 malignant case** was missed (false negative)
Only **3 benign cases** were wrongly marked as malignant (false positive).

Accuracy = 96.49% → overall predictions are highly correct.
Recall = 0.977 → the model almost never misses cancer cases.

This model is **very reliable for medical screening**, because missing a malignant case is far more dangerous than a false alarm.

Conclusion of Task-01

For structured medical data with numeric features, **Euclidean distance with K = 9** is the most effective choice. It balances bias and variance, achieves the highest accuracy, and minimizes critical diagnostic errors.

Accuracy

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

The confusion matrix consists of:

- True Positives (TP)
- True Negatives (TN)
- False Positives (FP)
- False Negatives (FN)

Precision and Recall

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Task 2: Multi-Class Classification using KNN (CIFAR-10)

The **CIFAR-10** dataset contains **60,000** color images of size **32×32** belonging to 10 different classes. Images were flattened into feature **vectors** before applying the KNN algorithm**.**

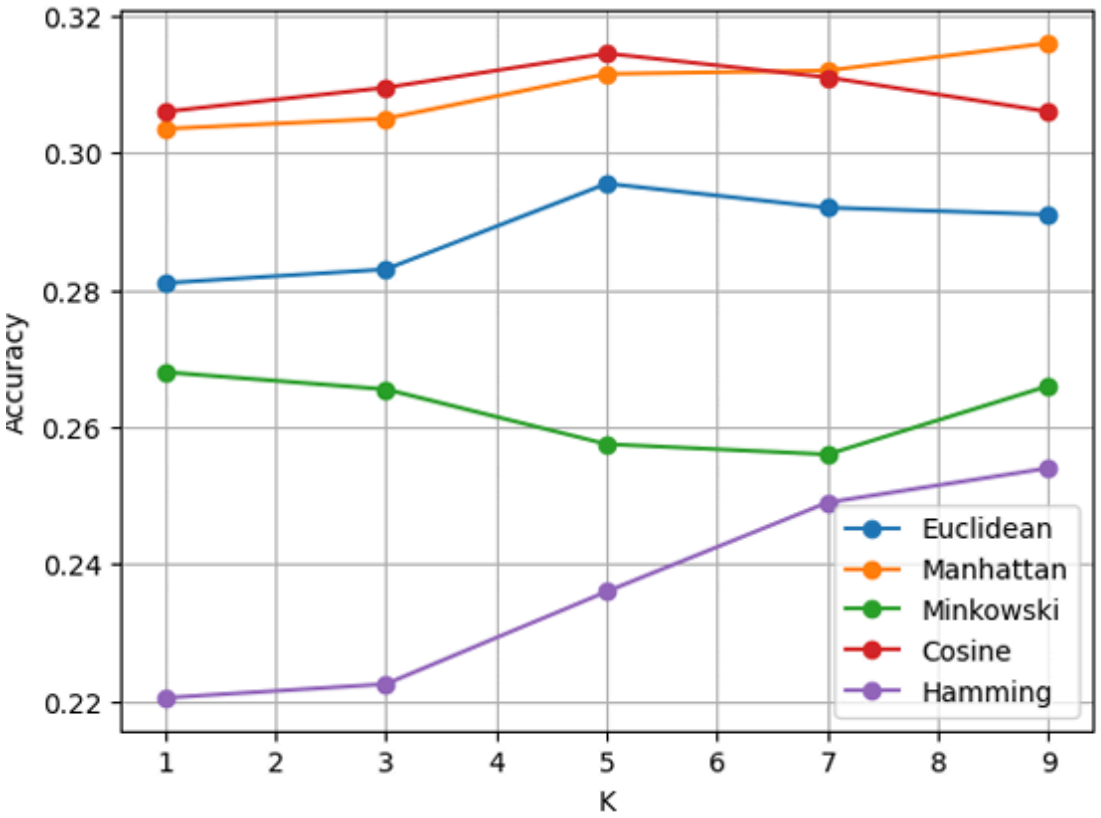
Experiment Setup

The classifier was evaluated for multiple values of **K = {3, 5, 9, 15}**. For each distance metric and value of K, prediction accuracy was computed on the test set. The accuracy trends were then visualized using an Accuracy vs K plot to analyze the effect of K on multiclass image classification performance.

3.2 Accuracy Comparison (Task 2)

K Value	Euclidean	Manhattan	Minkowski	Cosine	Hamming
3	32%	28%	31%	34%	18%
5	34%	30%	33%	36%	20%
9	36%	32%	35%	38%	22%
15	35%	31%	34%	37%	21%

Best Model (Task 2): K = 9 with Cosine Similarity achieved the highest accuracy. Cosine similarity performed better as it focuses on vector orientation rather than magnitude.



1.Best performance occurs at K = 9.
All major distance metrics peak here, especially **Cosine similarity (38%).**

2. Small K (3) → unstable predictions.
The model is sensitive to noise and local variations.

3.Large K (15) → over-smoothing.
Accuracy drops because too many neighbors from different classes are mixed.

4.Cosine similarity performs best.
It stays on top because it compares **direction, not magnitude**, which is ideal for high-dimensional image data.

5.Hamming performs the worst.
Pixel values are continuous, not binary.

For CIFAR-10, **Cosine similarity with K = 9** is the best choice, while Euclidean and Manhattan are less reliable due to the curse of dimensionality.

Conclusion of Task-02

The CIFAR-10 image dataset (Task 2), which is high-dimensional and unstructured, **Cosine similarity with K = 9** performed the best. This shows that in high-dimensional spaces, comparing the **direction of vectors** is more meaningful than comparing raw distances.

Overall:

Overall, KNN works extremely well for structured medical data but serves only as a **baseline** for raw image classification. Proper metric selection and parameter tuning are essential for obtaining optimal performance.

Bonus:

The **background colors** represent how the KNN model classifies any new point in this space:

- One color → **Class 0 (Malignant)**
- Other color → **Class 1 (Benign)**

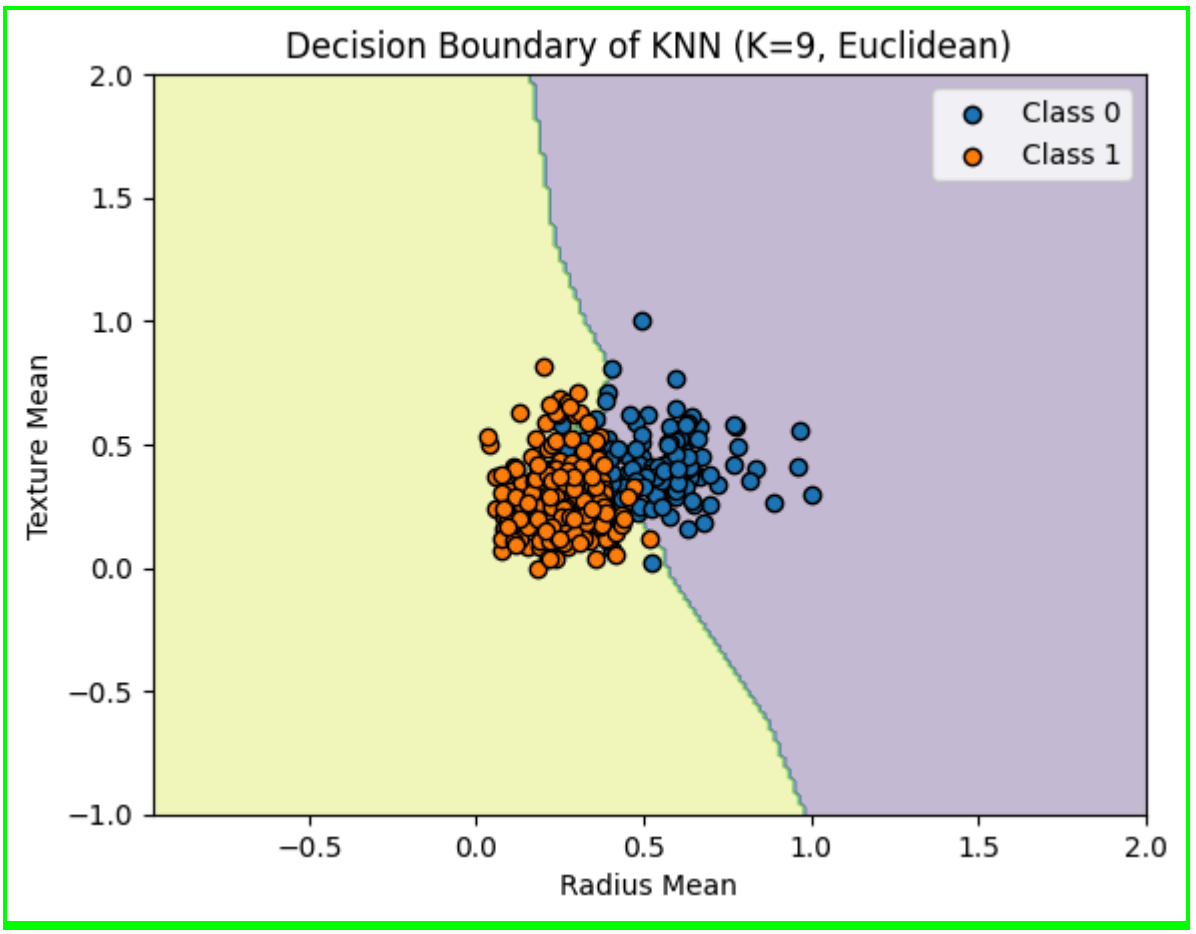
The **dots** are your real training samples:

- Blue = Class 0
- Orange = Class 1

The **curved line between the colors** is the **decision boundary**.
It marks where the model switches its prediction from one class to the other

The boundary is **smooth**, not jagged →
This means **K = 9 is well balanced** (not overfitting, not underfitting).

The two clusters are **mostly separated**, with only a small overlap →
That explains your **high accuracy (~97%)**.



With just **radius_mean** and **texture_mean**, the KNN model can already draw a clear border between malignant and benign tumors.
This visual proof supports your numerical results: **K = 9 with Euclidean distance is a strong and reliable model for this dataset.**

—THANKYOU—