Fall 25: CSci 4521—Applied Machine Learning

Homework 4
(Prepared by Brock Bye)

1.Binary Classification with Logistic Regression

I built a binary classifier with logistic regression using the following principles.

Dataset Balance

I checked the balance of the dataset for benign to malignant ratio as well as the numbers for each diagnostic category. An 8:2 benign to malignant ratio meant that the dataset was imbalanced. I downsampled benign and oversampled malignant.

Category	Subtype	Original	New
Benign	nv	6705	2221
	bkl	1099	871
	vasc	142	121
	df	115	87
Total Benign		8061	3300
Malignant	mel	1113	1110
	bcc	514	1110
	akiec	327	1110
Total Malignant		1954	3330

Table 1: Original vs. Resampled Dataset Counts

Data Augmentation

I performed data augmentation by introducing color jitter, horizontal flip, and rotation. To minimize overfitting, I split the dataset into train and test sets. To ensure there was no data leakage, I balanced the training set and left the test set untouched. After viewing the ROC and PR curves, I lowered the threshold from 0.5 to 0.3 in an attempt to raise recall scores.

Results

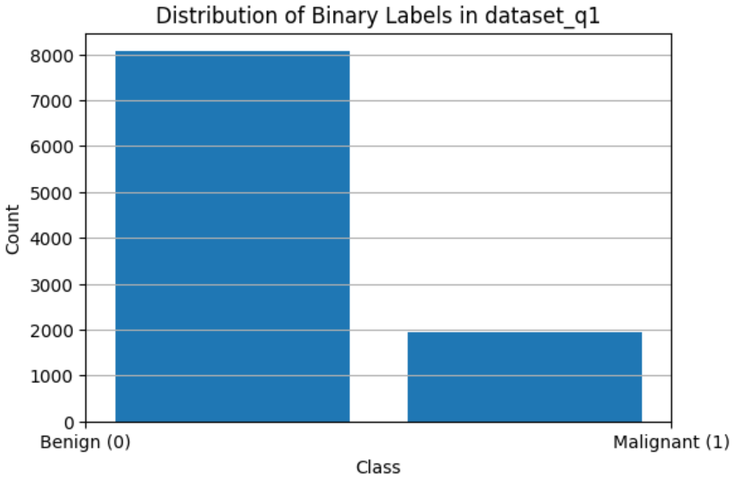

Figure 1: Benign vs Malignant

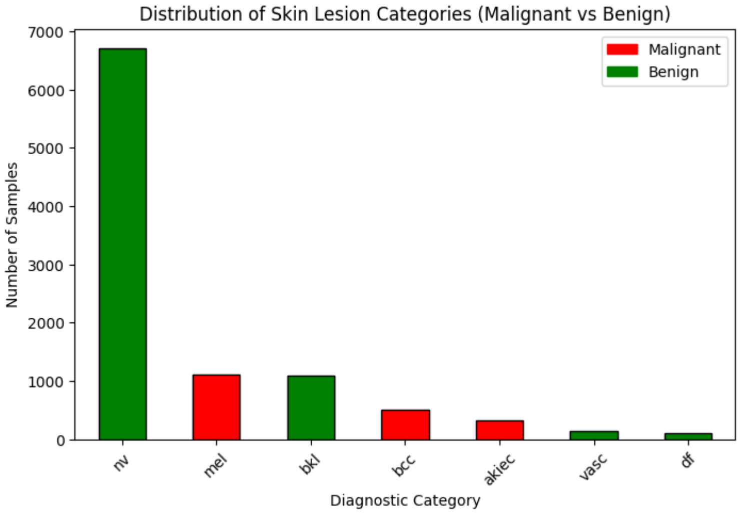

Figure 2: Across categories

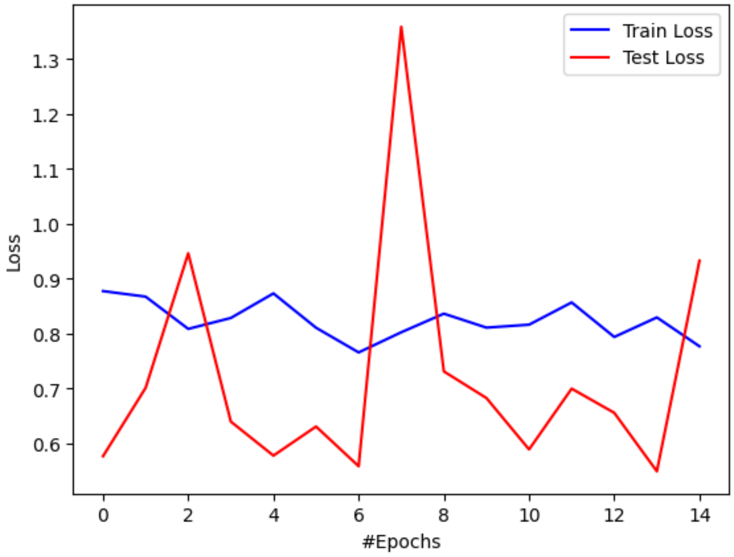

Figure 3: Loss over Epochs

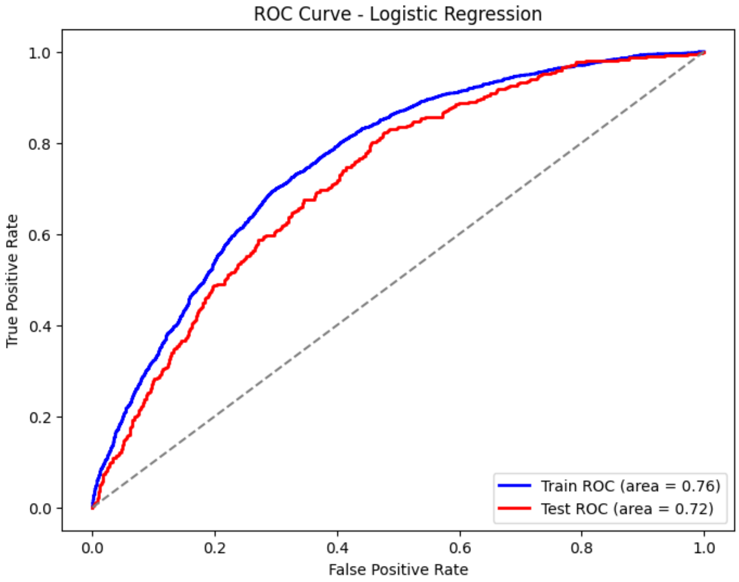

Figure 4: ROC and PR curves

Metric	Train	Test
Accuracy	0.59	0.48
Precision	0.54	0.25
Recall	0.96	0.88
F1	0.69	0.40
AUC	0.75	0.72
AP	0.71	0.34

Table 2: Overall Performance Metrics

Category	Precision	Recall	F1	Counts
mel	0	0	0	228
bcc	0	0	0	28
akiec	1	0.86	0.93	226

Table 3: Performance Metrics by Malignant Category

Analysis

The model struggles to generalize. Accuracy is low even on training data and drops further on test. On test precision, only 1 in 4 predicted positives are correct, indicating a very high false positive rate. On test recall, the model is catching most positives, meaning false negatives are relatively low. On test F1, the balance between precision and recall is poor, mainly due to overpredicting positives.

The ROC curve shows the model ranks positives above negatives, but not strongly. Average precision drops sharply on test, and in general the precision-recall performance is poor in generalization. On metrics for malignant categories, the model does not perform equally well. It does poorly on *mel* and *bcc*, completely missing them, but is very strong in detecting *akiec*. The model's positive predictions are not reliable across malignant categories, as evident in the low average precision.

2. Other Binary Classifier

KNN and PCA

I built a binary classifier using KNN and PCA. PCA was applied to reduce the dimensionality while ensuring that 95% of variance was retained, making the model both fast and of solid quality. The model performed slightly better using $k = 11$ neighbors.

Feature Dimensions

- Original feature dimension: 49152
- Reduced feature dimension: 69

Performance Metrics

Metric	Train	Test
Accuracy	0.77	0.76
Precision	0.84	0.34
Recall	0.64	0.26
F1	0.73	0.30
AUC	0.87	0.67
AP	0.84	0.30

Table 4: KNN + PCA Performance Metrics

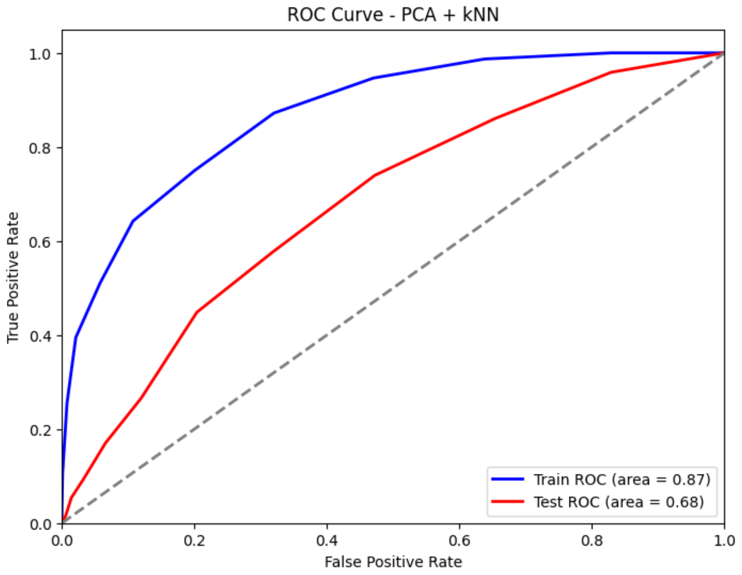

Figure 5: ROC curve

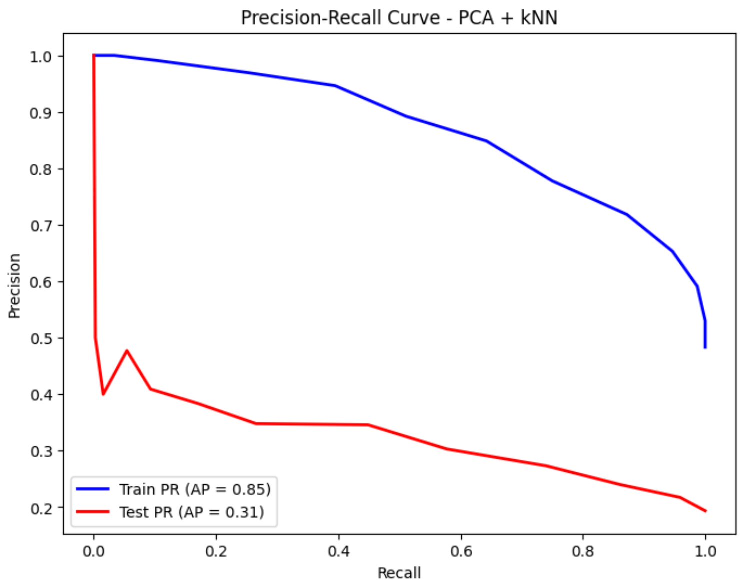

Figure 6: PR curve

Analysis

The train-test accuracy is fairly stable, suggesting the model is not severely overfitting. However, precision drops to 0.34, meaning more than half of the predicted positives are false alarms on the test set. Recall drops from to 0.30, indicating the model misses over half of the true positives. The balance between precision and recall is poor on the test F1 score.

The ROC AUC is decent, showing the model can rank positives above negatives reasonably well. However, the average precision (AP) drops sharply to 0.30, indicating the classifier struggles to maintain precision at useful recall levels.

SVM and PCA

I also built a binary classifier using SVM and PCA. The PCA implementation was the same as with the KNN model.

Performance Metrics

Metric	Train	Test
Accuracy	0.84	0.78
Precision	0.89	0.42
Recall	0.77	0.34
F1	0.83	0.37
AUC	0.91	0.73
AP	0.92	0.38

Table 5: SVM + PCA Performance Metrics

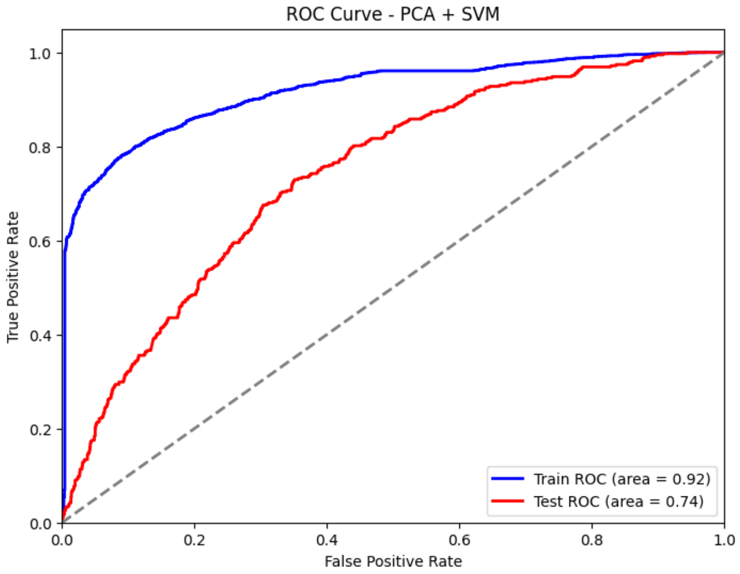

Figure 7: ROC curve

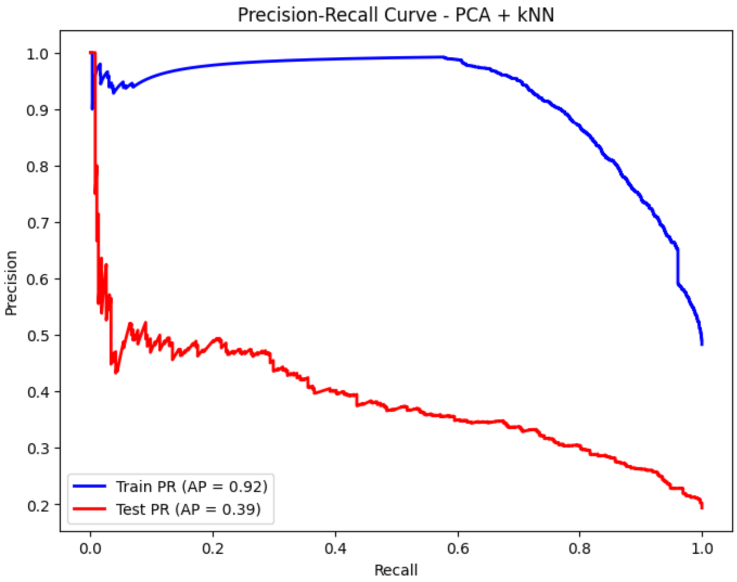

Figure 8: PR curve

Analysis

The train-test accuracy is solid, showing little evidence of severe overfitting. However, precision drops significantly, with nearly half of the predicted positives being false alarms. Recall also decreases, with about 40% of true positives missed. This results in a poor balance between precision and recall.

The AUC is strong, indicating the model can rank positives above negatives reasonably well. However, the average precision is weak, dropping from 0.91 to 0.54, which suggests that positive predictions are unreliable.

Model Comparison

Model	Accuracy	Precision	Recall	F1	AUC	AP
Logistic Regression	0.48	0.25	0.88	0.40	0.72	0.34
KNN + PCA	0.76	0.34	0.26	0.30	0.67	0.30
SVM + PCA	0.78	0.42	0.34	0.37	0.73	0.38

Table 6: Comparison of Models Across Metrics

Analysis

SVM + PCA performs best, then KNN + PCA, then Log Regression. Across the columns, SVM model has higher scores, meaning the best overall generalization. I think that this honestly makes sense. I think that Logistic Regression is designed to be good at discerning classes that are linearly separable. Skin lesions are highly non-linear in structure, which makes sense why log regression struggles, it's basically as about as good as randomly guessing.

3. Specific Diagnosis Classification

4. Visual Analysis