Statistical Learning V/S Deep Learning For Breast Cancer Detection

Robustness Comparison under Noisy and Clean Conditions

The most recent statistics show that of all cancers, breast cancer is the most common killing about 9,00,000 individuals annually. It has been observed that traditional methods are very time consuming and less accurate because it depends on physical tests and nearest assumptions but on the other hand when machines predicts it using the data, the disease can be predicted early, and it can be as accurate as possible. However, in real-world healthcare datasets, data impurities such as image noise, labeling errors, and outliers can significantly reduce model accuracy and reliability.

Statistical Learning for Text data

DATASET

Tabular Dataset — Wisconsin Breast Cancer (CSV)

•Target: Diagnosis (M = malignant, B = benign) it has 569 samples and 32 numeric features and includes radius, texture, smoothness, concavity, symmetry, etc.

METHODS USED

1**. Data Preprocessing**

•Missing value handling with SimpleImputer

•Outlier detection using: IQR filtering and Isolation Forest

•Feature scaling using StandardScaler.

2.**Feature Engineering:-** Added Polynomial Features (degree = 2)

•Explores nonlinear relationships

•Logistic Regression improves slightly

•Random Forest often overfits (accuracy drops)

3.**PCA (Principal Component Analysis)**

•Reduces dimensionality, Removes noise and Improves model stability.

Results

Logistic Regression

•Baseline linear model it works well with scaled features

•Sensitive to outliers but improved using RobustScaler + PCA

Random Forest

•Non-linear tree-based ensemble and Robust to feature noise

•May overfit with too many polynomial features

Statistical Learning for Image Dataset

Data Used:

- **Dataset:** Breast Cancer MSI Multimodal Image Dataset (Kaggle)
- **Modality:** Histopathology images (1246 images)
- **Classes:** Benign vs Malignant (623 Benign vs 623 Malignant)

Noise Conditions Tested:

- **Clean:** Original images and labels
- **Pixel Noise:** Gaussian noise (o = 010) added to images before feature extraction
- **Label Noise:** 20% of training labels randomly flipped
- **Combined Noise:** Both pixel noise (o = 0.10) and 20% label flips

Training Summary:

- **Feature extractor:** Pre-trained ResNet-50 (ImageNet)
- **Input size:** 224 × 224 RGB images
- **Feature type:** Global average-pooled deep features (2048-D)
- **Classifiers:** Logistic Regression
- **Train / Test split:** 80% train, 20% test (stratified by class)

Methodology:

We extracted deep feature vectors from each image using a frozen ResNet-50 backbone. These 2048 dimensional features served as input to a Logistic Regression classifier, representing a traditional machine learning pipeline built on top of deep representations.

We evaluated how pixel noise, label noise and combined noise affect the performance compared with the clean condition. Model was evaluated using **Accuracy**, **F1 Score** and **ROC-AUC**

Results

Results for Textual dataset using Statistical Machine Learning

Logistic Regression (LR) for Malign cases	Accuracy	AUC	F1
Raw data	0.973	0.994	0.962
With Polynomial fitting	0.982	0.998	0.975
With PCA + cross validation	0.973	0.996	0.961
With Noise Filtering	0.986	0.994	0.975

Random forest (RF) for Malign cases	Accuracy	AUC	F1
Raw data	0.973	0.994	0.962
With Polynomial fitting	0.964	0.992	0.975
With PCA + cross validation	0.938	0.989	0.921
With Noise Filtering	0.934	0.986	0.869

Results for textual data using Deep Learning

Condition	Accuracy	AUC	F1
Raw Images (Baseline)	0.931	0.96	0.93
Noise Filtered Images	0.945	0.97	0.94
With Data Augmentation	0.952	0.975	0.95
With Contrast Enhanceme	0.948	0.972	0.94
ROI-Based Training	0.958	0.98	0.96

For Image Dataset

Results for Statistical Machine Learning

Condition	Accuracy	F1	AUC
Clean	0.82	0.80	0.88
Pixel	0.78	0.75	0.84
Label	0.68	0.66	0.76
Combined	0.65	0.63	0.73

Results for Deep Learning

Condition	Accuracy	F1	AUC
Clean	0.89	0.88	0.98
Pixel	0.86	0.85	0.96
Label Noise	0.74	0.77	0.87
Combined	0.81	0.81	0.90

ResNet-18

Condition	Accuracy	F1	AUC
Clean	0.93	0.93	0.99
Pixel Noise	0.89	0.90	0.97
Label Noise	0.79	0.77	0.88
Combined	0.79	0.80	0.89

DenseNet-121

Condition	Accuracy	F1	AUC
Clean	0.65	0.74	0.83
Pixel	0.58	0.37	0.83
Label Noise	0.58	0.49	0.72
Combined	0.62	0.64	0.73

VGG-16

Conclusion

The project successfully explored breast cancer detection using both statistical machine learning and deep learning methodologies across image and text datasets.

Deep Learning for Image Data: Deep learning models, specifically Convolutional Neural Networks (CNNs), proved superior for image-based breast cancer detection.

Statistical Machine Learning for Text Data: Conversely, traditional statistical learning models, particularly Logistic Regression, outperformed Random Forest and deep learning models for text-based data analysis.

Deep Learning for Text Data

Data Used:

- **Batch Size:** 32
- **Epochs Trained:** 50
- **Total Images Used:** 2,000+ (Train/Val/Test split: 80/10/10)
- **Optimizer:** Adam
- **Learning Rate:** 1e-3 (0.001)
- **Loss Function:** Binary Cross Entropy

Preprocessing Conditions Tested:

o Clean Images o Noise Filtered Images

CLAHE Contrast-Enhanced Images o Data-Augmented Images o ROI-based Patches

A single, consistent CNN architecture (Conv - RelU —MaxPool → Dense - Softmax) was trained and evaluated across all conditions to measure robustness.

Performance was measured using Accuracy, F1-Score, and AUC-ROC.

Dataset: CBIS-DDSM Full Mammogram Dataset

Modalities: Mammogram Images (Calcification + Mass Views)

Results:

The CNN is highly reliable for mammogram analysis, achieving 93-96% accuracy depending on condition.

Data augmentation and ROI-based cropping provide the strongest boost in performance.

Contrast enhancement (CLAHE) and noise filtering also improve robustness

Deep Learning for Image Dataset

Training Summary:

- **Batch:** 15
- **Epoch:** 16
- **Optimizer:** Adam
- **Learning Rate:** 3e-4

Data Used:

- **Dataset:** Breast Cancer MSI Multimodal Image Dataset
- **Modalities:** Histopathology Images (1246 Images)
- **Classes:** Benign Vs Malignant (623 Benign vs 623 Malignant)
- **Noise Condition Tested:**
  - Clean, Pixel Noise(σ = 0.10), Label Noise (20% Flipped), Combined Noise

We trained three deep learning models (ResNet-18, VGG-16, DenseNet121) on histopathology images to evaluate how pixel noise, label noise and combined noise affect classification performance. Models were tested under multiple random seeds and evaluated using Accuracy, F1-score and ROC-AUC

RESULTS:

- ResNet-18 is the **most consistent ad noise robust** model.
  - Smaller performance drops under noise
  - Stronger F1 recovery under combined noise
  - More stable across seeds
- DenseNet121 performs **best on clean data** but is not the most noise robust.
  - DenseNet121 achieves the highest clean accuracy and AUC, but its performance drops sharply under label noise, especially when comapred to ResNet-18
- VGG-16 has the **worst performance** in all cases across all the seeds.

STEVENS INSTITUTE OF TECHNOLOGY

Group Name:- Ai-Sian Rising

Group Member:- Rishi Patel, Ishbat Mahmud, Rishika Pilli, Malvi Patel

ECE Department, Advisor:- Prof. Hao Wang,