Project 3 : Breast Cancer Detection

Orlane BOCHARD, Pauline DUMAS, Simon LABRACHERIE, Emeline LAVAUX

1) Preprocessing

Scaling the images to the [0, 1] range (scaling step), then z-score **normalization** using the training set's mean and standard deviation.

Additional tweaks: **CLAHE** method (Contrast Limited Adaptive Histogram Equalization) to improve local contrast and better highlight the contours of the masses.
Median blur applied to reduce noise while preserving important edges.

Data augmentation step: makes the model more robust to real-world variations in medical imaging by increasing the variability of the training set.
Geometrical: horizontal flip, rotation $\pm 10°$, slight zoom.
Photometric: brightness $\times$ 0.8–1.2, contrast $\times$ 0.9–1.3, speckle noise with $\sigma = 0.10$.

Skipped step: Resizing since BreastMNIST images are already in a uniform 128×128 format. Downscaling (e.g. 64×64) would reduce image quality and blur important textural and structural details that are crucial for detecting malignant patterns. Keeping the original resolution preserves diagnostic detail critical for breast cancer classification.

2) Feature Engineering

Handcrafted radiomic features were designed to capture tumor heterogeneity at multiple scales by combining statistical, structural, and morphological information. Gray levels were first reduced from 255 to 32 to stabilize the **GLCM** computation.
From this matrix, six descriptors are extracted — **contrast**, **homogeneity**, **energy**, **correlation**, **ASM**, and **dissimilarity** — averaged over several distances and angles. Local micro-texture was captured with **LBP** (P=8, R=1), summarized by the dominant pattern and histogram entropy. **Canny edge density** quantified contour diversity, while global **intensity statistics** (mean, std, median, 75th percentile, entropy) summarized the brightness distribution. Additional descriptors included a **Gabor filter bank** (4 angles × 3 frequencies) for directional textures, and **Hu moments** (7 on Otsu mask + 7 on Canny edges) for shape invariance. Several **composite ratios and products** (e.g., contrast/homogeneity, edge×entropy) were added to link texture, contour, and intensity cues. Feature selection combined **Pearson correlation**, **ANOVA**, **Mutual Information**, and **RFE** with Logistic Regression.
The final set of features, built as the union of top features, has much better results: **AUC = 0.823**, **Accuracy = 0.808**, and **F1 = 0.872**, confirming the gain from multi-scale and shape descriptors.

Table 1: Comparison before and after extended feature engineering.

Model	**AUC**	**Accuracy**	**F1-score**
Baseline (GLCM + LBP + Canny + Intensity)	0.784	0.724	0.790
Extended (multi-scale + Gabor + Hu + combos)	**0.823**	**0.808**	**0.872**

3) Modeling

For the model of our choice, we used DenseNet121 because it benefits from ImageNet pretraining. Although ImageNet is non-medical, its learned visual features (edges, textures, shapes) transfer effectively to medical images after fine-tuning.

Table 2: Summary of key models (AUROC, Accuracy, F1). Best values in bold.

Model (TEST)	**Threshold**	**AUC**	**Precision (+)**	**Recall (+)**	**F1**
Logistic Regression @0.5	0.500	0.821	0.889	0.772	0.826
Logistic Regression @F1(VAL)	0.410	0.821	0.870	0.825	0.847
Calibrated SVM @0.5	0.500	0.829	0.841	**0.930**	**0.883**
Calibrated SVM @F1(VAL)	0.532	0.829	0.839	0.912	0.874
CNN DenseNet121 (128×128) @0.5	0.500	0.908	0.879	**0.956**	0.916
CNN DenseNet121 (128×128) @F1(VAL)	**0.653**	**0.908**	**0.916**	**0.956**	**0.936**

4) Evaluation

Table 3: Model comparison on the test set

Model (TEST)	Threshold	AUC	ACC	F1
LR @0.5	0.500	0.823	0.750	0.817
LR @F1(VAL)	0.318	0.823	0.808	0.872
SVM (cal.) @0.5	0.500	0.829	0.821	**0.883**
SVM (cal.) @F1(VAL)	0.532	0.829	0.808	0.874
CNN DenseNet121 (128×128) @0.5	0.500	0.913	0.865	0.910
CNN DenseNet121 (128×128) @F1(VAL)	0.653	0.913	0.865	0.908
CNN DenseNet121 (224×224) @0.5	0.500	**0.937**	**0.891**	**0.928**
CNN DenseNet121 (224×224) @F1(VAL)	0.878	0.937	0.878	0.912

Our best-performing model is DenseNet121 trained at 224×224 resolution. It achieved the highest AUROC (0.94) and F1-score (0.93), with a recall of 0.96, meaning it correctly identifies almost all malignant cases. Thanks to transfer learning from ImageNet and fine-tuning on our dataset, it captures fine mammographic details and provides the best balance between high sensitivity and reliable precision, which is ideal for breast cancer screening.

5) Reflection

Figure 1: Whitch metric should matter most ?

Accuracy	Precision	Recall
Not relevant	May be important but not the priority	Should **matter the most**
Imbalanced dataset: fewer malignant cases than benign ones	In medical detection, we prefer to have some false positives (which can be checked by further tests) rather than missing a real cancer case	In breast cancer detection, the goal is to detect all malignant cases. Missing a positive one has huge consequences on health (delay treatment and harm the patient)

We should maximize recall while keeping an acceptable level of precision.

Table 4: Comparison of key aspects across different error types

Error Type	Consequences	Who is harmed?
False Positive	A benign case is incorrectly classified as malignant: This may lead to unnecessary treatments, additional tests, and increased stress or anxiety for the patient. It also consumes hospital resources and increases operational costs.	**Primarily the patient** (psychological and emotional impact) Hospital (financial and resource-related impact)
False Negative	A malignant case is incorrectly classified as benign: This can delay treatment, allow the disease to progress, and reduce the patient's chances of recovery or survival. The hospital may also face ethical and legal consequences.	**Primarily the patient** (risk to health and life) Hospital (ethical, legal, and reputational consequences)

False negatives are much more dangerous than false positives, as they directly threaten the patient's life.

- If the model is used for **automatic detection**, recall remains critical since a false negative could be completely unnoticed and threaten the patient's life. The model must be highly sensitive and detect all malignant cases, even at the expense of more false positives.
- If the model is used as an **assisting tool for doctors**, precision becomes more important, it should balance recall and precision to effectively assist medical experts. Missing a false negative is less critical because the final decision is made by a human.

Conclusion: Since we are dealing with cancer detection, such models should be used **as decision-support tools** rather than fully replacing radiologists.
[1]

[1] we used AI/tool assistance for the code and Hugging Face website