Business Understanding

Problem statement:

Breast cancer affects 1 in 8 women worldwide, with early detection increasing 5-year survival rates from 27% to 99%. However, current diagnostic methods face significant limitations:

24.2% of breast cancers are missed in initial screenings

Up to 50% of women receive false positives over 10 years of annual screening

15-30% of biopsy procedures are unnecessary due to benign findings

Radiologist interpretation shows 10-15% variability in diagnosis accuracy

-What we must solve:

1. Detect Breast Cancer Accurately

Tell if a breast tumor is cancerous (malignant) or not cancerous (benign)

Use data from 569 patient tests (212 cancerous, 357 non-cancerous)

2. Find the Best Computer Model

Test 6 different computer programs to see which works best

All programs must be over 90% accurate

3. Avoid Dangerous Mistakes

Don't miss real cancer (false negatives - very dangerous)

Don't scare healthy people (false positives - unnecessary worry)

4. Make It Fast and Reliable

Train programs quickly (seconds to minutes)

Work with standard medical data

Give doctors clear results to help their decisions

5. Help Doctors Save Lives

Provide a second opinion for doctors

Catch cancer early when it's easier to treat

Reduce stress for patients with faster, more accurate results

-Why This Matters Now:

With breast cancer incidence increasing by 3.1% annually and diagnostic errors costing healthcare systems $12-18 billion yearly, this AI-powered solution addresses a critical gap in women's healthcare,

potentially saving 250,000+ lives globally each year through earlier, more accurate detection and reduced diagnostic delays.

Business Objectives:

– Confirm the presence of a breast tumor:

Provide reliable support to the physician to determine whether the data indicate a suspicious tumor requiring medical intervention.

– Characterize the tumor to guide clinical decisions:

Offer clear indications on whether the tumor is benign or malignant.

– Detect early risk factors in healthy patients and recommend appropriate preventive actions.

Data science objectives:

– Classify tumor type:

Build a model to distinguish between benign and malignant tumors based on imaging and clinical features. (dataset de base + dataset simple ajoutée)

– Detect tumor presence:

Develop a predictive model to classify patient data as indicative of a tumor or not, providing a reliable alert for potential breast cancer. (dataset image)

– Identify risk factors in healthy patients:

Analyze patient data to detect early indicators of increased breast cancer risk and generate actionable preventive recommendations.

(dataset de base + dataset simple ajoutée avec plus de concentration sur les variables qui traite la maladie)

Table 1: DSO 1 – Classify Tumor Type

Model	Variables Involved	Parameters to Use / Optimize
GRU-SVM	All WDBC features	Batch Size = 128, Cell Size = 128, Dropout = 0.5, Learning Rate = 1e-3, Epochs = 3000, SVM C = 5
Linear Regression	All WDBC features	Batch Size = 128, Learning Rate = 1e-3, Epochs = 3000
MLP	All WDBC features	Batch Size = 128, Architecture = [500, 500, 500], Learning Rate = 1e-2, Epochs = 3000
Nearest Neighbor	All WDBC features	Norm = L1 or L2
Softmax Regression	All WDBC features	Batch Size = 128, Learning Rate = 1e-3, Epochs = 3000
SVM	All WDBC features	Batch Size = 128, Learning Rate = 1e-3, Epochs = 3000, SVM C = 5, Norm = L2

Table 2: DSO 2 – Detect Tumor Presence (Image Dataset)

Model	Variables Involved	Parameters to Use / Optimize
GRU-SVM	Image pixels (50x50)	Batch Size = 128, Cell Size = 128, Dropout = 0.5, Learning Rate = 1e-3, Epochs = 3000, SVM C = 5
MLP	Image pixels (50x50)	Batch Size = 128, Architecture = [500, 500, 500], Learning Rate = 1e-2, Epochs = 3000

CNN (new)	Image pixels (50x50)	Conv2D, MaxPooling, Dropout, Dense Layers, Learning Rate = 1e-3, Epochs = 100
SVM (with RBF)	Features extracted from images (e.g., HOG, SIFT)	Kernel = RBF, C = 5, Gamma = 'scale'
Softmax Regression	Features extracted from images	Batch Size = 128, Learning Rate = 1e-3, Epochs = 3000

Table 3: DSO 3 – Identify Risk Factors in Healthy Patients

Model	Variables Involved	Parameters to Use / Optimize
GRU-SVM	Clinical variables (BMI, Glucose, Insulin, etc.)	Batch Size = 128, Cell Size = 128, Dropout = 0.5, Learning Rate = 1e-3, Epochs = 3000, SVM C = 5
Linear Regression	Clinical variables (BMI, Glucose, Insulin, etc.)	Batch Size = 128, Learning Rate = 1e-3, Epochs = 3000
MLP	Clinical variables (BMI, Glucose, Insulin, etc.)	Batch Size = 128, Architecture = [500, 500, 500], Learning Rate = 1e-2, Epochs = 3000
Nearest Neighbor	Clinical variables (BMI, Glucose, Insulin, etc.)	Norm = L1 or L2
Softmax Regression	Clinical variables (BMI, Glucose, Insulin, etc.)	Batch Size = 128, Learning Rate = 1e-3, Epochs = 3000
SVM	Clinical variables (BMI, Glucose, Insulin, etc.)	Batch Size = 128, Learning Rate = 1e-3, Epochs = 3000, SVM C = 5, Norm = L2