ESPRIT

Machine Learning Final Project

Breast Cancer Detection and Risk Analysis

Team Members:

Mohamed Yassine Janfaoui
Ahmed Tounsi
Ahmed Rayen Aloui
Farah Boubaker
Farah Derbel
Mahdi Saoudi

Academic Year: 2025–2026

Contents

Chapter 1

Business Understanding

Breast cancer remains one of the most prevalent and life-threatening diseases among women worldwide. Early detection and accurate diagnosis are critical factors that significantly improve patient survival rates and reduce treatment costs. However, traditional diagnostic processes rely heavily on expert interpretation of medical data, which can be time-consuming and subject to human error.

This project follows the CRISP-DM methodology to design and evaluate machine learning models that assist healthcare professionals in breast cancer detection, diagnosis, and prevention. The system is designed as a decision-support tool and does not aim to replace medical expertise.

Business Objectives

- **Confirm the presence of a breast tumor:** Provide reliable support to physicians by analyzing patient data and indicating whether the data suggest the presence of a suspicious breast tumor requiring further medical examination.

- **Characterize the tumor to guide clinical decisions:** Assist clinicians by determining whether a detected tumor is benign or malignant, enabling faster and more informed diagnostic decisions.

- **Detect early risk factors in healthy patients and recommend appropriate preventive actions:** Analyze patient profiles to identify early indicators of increased breast cancer risk and support preventive healthcare strategies.

Chapter 2

Data Understanding

2.1 DSO1: Classify tumor type

2.1.1 Dataset Description

Le dataset Wisconsin Diagnostic Breast Cancer (WDBC) contient des caractéristiques dérivées d'images numérisées de masses mammaires. Il est largement utilisé dans la recherche sur la détection du cancer du sein grâce à ses données complètes et bien structurées.

2.1.2 Variables Description

Le dataset comprend 30 caractéristiques radiologiques calculées pour chaque noyau cellulaire, regroupées en trois catégories :

- **Caractéristiques moyennes** (10 features) : Moyennes des mesures de base
- **Erreurs standard** (10 features) : Écart-type des mesures
- **Valeurs maximales** (10 features) : Plus grandes valeurs observées

Les caractéristiques incluent :

- `radius` : Rayon moyen des distances du centre aux points du périmètre
- `texture` : Écart-type des valeurs de niveau de gris
- `perimeter` : Périmètre de la cellule
- `area` : Aire de la cellule
- `smoothness` : Variation locale de la longueur du rayon
- `compactness` : $\frac{perimeter^2}{area} - 1.0$

- `concavity` : Sévérité des parties concaves du contour
- `concave points` : Nombre de parties concaves du contour
- `symmetry` : Symétrie de la cellule
- `fractal dimension` : Approximation de la "côte" - 1

2.1.3 Dataset Characteristics

- **Total samples:** 569
- **Benign (B):** 357 échantillons (62.7%)
- **Malignant (M):** 212 échantillons (37.3%)
- **Features:** 30 caractéristiques numériques
- **Target:** Diagnostic (Bénin/Malin)

2.1.4 Analyse Exploratoire

- **Distribution des classes:** Déséquilibre modéré en faveur des cas bénins
- **Valeurs manquantes:** Aucune valeur manquante dans les 30 caractéristiques
- **Corrélations:** Les caractéristiques liées aux points concaves montrent la plus forte corrélation avec le diagnostic
- **Valeurs aberrantes:** Peu de valeurs aberrantes détectées via la méthode IQR

2.2 DSO2: Histopathology Image Analysis

For DSO2, we worked with a comprehensive histopathology image dataset from Kaggle focused on Invasive Ductal Carcinoma (IDC) detection. This dataset is particularly valuable as it contains actual tissue samples from breast cancer patients.

2.2.1 Dataset Description

The dataset originates from 162 whole mount slide images of Breast Cancer (BCa) specimens scanned at 40x magnification. From these slides, 277,524 patches of size 50×50 pixels were extracted:

- **Total images:** 277,524 patches
- **Class 0 (non-IDC):** 198,738 patches (71.6%)
- **Class 1 (IDC):** 78,786 patches (28.4%)

2.2.2 File Naming Convention

Each patch follows a strict naming convention that encodes important metadata:

`patientID_idx*_x*_y*_class*.png`

Example: `10253_idx5_x1351_y1101_class0.png`

- `u`: Patient ID (e.g., 10253_idx5)
- `X`: x-coordinate where the patch was cropped from the original slide
- `Y`: y-coordinate where the patch was cropped from the original slide
- `C`: Class label where 0 = non-IDC and 1 = IDC

2.2.3 Key Statistics and Observations

- **Class Distribution:** Significant class imbalance with approximately 71.6% non-IDC and 28.4% IDC cases.
- **Data Integrity:** All 277,524 PNG files were validated against the naming convention with zero invalid filenames found.
- **Patient-Wise Splitting:** The data was split at the patient level (not image level) to prevent data leakage, ensuring that images from the same patient do not appear in both training and testing sets.
- **Visual Patterns:** Initial visual inspection revealed meaningful texture and structural patterns that distinguish IDC from non-IDC tissue.

2.3 DSO3: Identify risk factors in healthy patients

This DSO focuses on the analysis of clinical and biochemical data to identify early risk factors associated with breast cancer in patients who may not yet show detectable tumors.

2.3.1 Dataset Description

The dataset used for DSO3 contains clinical measurements collected from both healthy individuals and breast cancer patients. All variables are numerical and correspond to metabolic, hormonal, and inflammatory indicators.

2.3.2 Variables Description

The dataset includes the following features:

- **Age**: Age of the patient (years)
- **BMI**: Body Mass Index
- **Glucose**: Blood glucose concentration
- **Insulin**: Insulin level
- **HOMA-IR**: Insulin resistance indicator
- **Leptin**: Hormone related to fat metabolism
- **Adiponectin**: Hormone regulating glucose and lipid metabolism
- **Resistin**: Inflammatory marker
- **MCP-1**: Monocyte Chemoattractant Protein-1
- **Classification**: Target variable (0 = Healthy, 1 = Patient)

2.3.3 Dataset Characteristics

- **Total samples:** 4000
- **Healthy controls:** 1784
- **Patients:** 2216
- **Missing values:** None detected
- **Data distribution:** Slight class imbalance

Initial exploratory analysis revealed that several metabolic indicators (Glucose, BMI, Insulin) show noticeable differences between healthy and patient groups.

Chapter 3

Data Preparation

3.1 DSO1: Classify tumor type

3.1.1 Data Cleaning

Le dataset nécessitait un nettoyage minimal :

- Suppression de la colonne `Unnamed: 32` entièrement vide
- Conversion de la variable cible : M=1 (Malin), B=0 (Bénin)
- Vérification de la cohérence des 30 caractéristiques

3.1.2 Feature-Target Separation

Séparation des données en features (X) et target (y) :

- **Features:** 30 caractéristiques radiologiques
- **Target:** Variable binaire (0: Bénin, 1: Malin)
- **ID:** La colonne d'identifiant a été supprimée

3.1.3 Train-Test Split

Division des données selon la méthodologie de l'article :

- **Training set:** 70% (398 échantillons)
- **Testing set:** 30% (171 échantillons)
- **Stratification:** Préservation de la distribution des classes
- **Random state:** 42 pour la reproductibilité

3.1.4 Feature Standardization

Normalisation des caractéristiques pour améliorer la performance des modèles :

- **Méthode:** Standardisation Z-score
- **Formule:** $z = \frac{x-\mu}{\sigma}$
- **Implémentation:** StandardScaler de scikit-learn
- **Note:** Les paramètres sont ajustés sur le training set seulement

3.2 DSO2: Histopathology Image Analysis

For the image dataset, we implemented a comprehensive preprocessing pipeline tailored for deep learning models:

3.2.1 Data Loading and Preprocessing

- **Image Loading:** Images were loaded using TensorFlow's optimized I/O operations.
- **Resizing:** All images were resized to 50×50 pixels to maintain consistency with the original patch size.
- **Normalization:** Pixel values were scaled to the range [0, 1] by converting to float32 and dividing by 255.
- **Data Augmentation:** Applied random transformations (rotations, flips) to increase dataset variability and prevent overfitting.

3.2.2 Patient-Wise Data Splitting

To ensure realistic evaluation and prevent data leakage:

- **Patient Identification:** Extracted unique patient IDs from filenames.
- **Stratified Splitting:** Split patients (not individual images) into training (70%), validation (10%), and testing (20%) sets.
- **Class Balance Preservation:** Maintained similar class distributions across all splits.

3.2.3 TensorFlow Data Pipeline

Implemented an efficient data pipeline using TensorFlow's `tf.data` API:

- **Parallel Processing:** Used `.map()` for parallel image loading and preprocessing.
- **Shuffling:** Applied shuffling with a buffer size of 5000 to ensure random ordering each epoch.
- **Batching:** Created batches of 64 images for efficient GPU utilization.
- **Prefetching:** Used `.prefetch(1)` to overlap data preprocessing and model execution.

3.3 DSO3: Identify risk factors in healthy patients

3.3.1 Data Cleaning

The dataset required minimal cleaning. Column names were standardized, and all variables were verified to be numeric. No missing or duplicated values were detected.

3.3.2 Feature Engineering

To capture clinically relevant interactions between variables, additional features were created:

- **BMI_Glucose** = BMI × Glucose
- **HOMA_Leptin** = HOMA-IR × Leptin
- **Insulin_Resistin** = Insulin × Resistin
- **Adiponectin_Leptin_ratio** = Adiponectin / Leptin

These features reflect known biological relationships between metabolism, inflammation, and cancer risk.

3.3.3 Feature Scaling

All numerical features were standardized using the **StandardScaler** to ensure:

- Equal contribution of variables
- Improved convergence for SVM and logistic models
- Better performance for distance-based classifiers

3.3.4 Handling Class Imbalance

To address class imbalance, resampling techniques were applied:

- SMOTE
- SMOTETomek

The SMOTETomek method was selected for final training as it produces a balanced dataset while reducing noise near class boundaries.

Chapter 4

Modeling

4.1 DSO1: Classify tumor type

Six algorithmes d'apprentissage automatique ont été implémentés et comparés pour la classification binaire des tumeurs mammaires. Tous les modèles ont été entraînés avec les hyperparamètres exacts spécifiés dans l'article de référence.

4.1.1 Modèle 1: Régression Linéaire

Rationnel: La régression linéaire pour la classification binaire utilise une fonction de seuil (0.5) pour prédire les classes. Bien que simple, elle peut fournir une baseline pour les modèles plus complexes.

Paramètres utilisés:

- Taille de batch: 128
- Taux d'apprentissage: 0.001
- Nombre d'epochs: 3000
- Fonction de perte: MSE (Equation 9 de l'article)
- Optimiseur: SGD

Architecture:

- Couche linéaire unique: $y = WX + b$
- Fonction de seuil: $\hat{y} = 1_{(y \geq 0.5)}$
- 31 paramètres entraînables (30 poids + 1 biais)

4.1.2 Modèle 2: Classification par Plus Proche Voisin (k=1)

Rationnel: L'algorithme des k-plus proches voisins (avec k=1) est une méthode géométrique non paramétrique qui classe les points basés sur la similarité avec les échantillons d'entraînement.

Métriques de distance évaluées:

- **Manhattan (L1):** $d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$ (Equation 12)
- **Euclidienne (L2):** $d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$ (Equation 13)

Caractéristiques:

- Aucun entraînement requis
- Calcul géométrique direct
- Sensible à l'échelle des features (d'où l'importance de la standardisation)

4.1.3 Modèle 3: Régression Softmax

Rationnel: La régression softmax étend la régression logistique aux problèmes multi-classes. Pour la classification binaire, elle est équivalente à la régression logistique mais formulée dans un cadre multi-classes.

Paramètres utilisés:

- Taille de batch: 128
- Taux d'apprentissage: 0.001
- Nombre d'epochs: 3000
- Fonction de perte: Entropie croisée (Equation 15)
- Optimiseur: SGD

Architecture:

- Couche linéaire suivie de softmax: $P(y = j|x) = \frac{e^{w_j^T x}}{\sum_{k=1}^{K} e^{w_k^T x}}$
- 62 paramètres entraînables (30×2 poids + 2 biais)
- Encodage one-hot pour les labels

4.1.4 Modèle 4: Machine à Vecteurs de Support (SVM)

Rationnel: Les SVM cherchent l'hyperplan optimal qui maximise la marge entre les classes. La version L2-SVM utilisée implémente une pénalité quadratique pour les erreurs.

Paramètres utilisés:

- Taille de batch: 128
- Taux d'apprentissage: 0.001
- Nombre d'epochs: 3000
- Paramètre C: 5 (force de régularisation)
- Norme: L2 (Equation 20)
- Optimiseur: Adam

Fonction de perte:

$$L = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b))^2$$

4.1.5 Modèle 5: Perceptron Multicouche (MLP)

Rationnel: Le MLP est un réseau de neurones feed-forward capable d'apprendre des relations non-linéaires complexes entre les features et la target.

Architecture:

- **Couche d'entrée:** 30 neurones
- **Couches cachées:** 3 couches de 500 neurones chacune
- **Fonction d'activation:** ReLU
- **Couche de sortie:** 2 neurones avec softmax
- **Dropout:** Non utilisé (selon spécifications de l'article)

Paramètres utilisés:

- Taille de batch: 128
- Taux d'apprentissage: 0.01
- Nombre d'epochs: 3000
- Fonction de perte: Entropie croisée
- Optimiseur: SGD

4.1.6 Modèle 6: GRU-SVM

Rationnel: L'hybride GRU-SVM combine les capacités d'apprentissage séquentiel des GRU avec la puissance de classification des SVM. Bien que le dataset ne soit pas séquentiel, les GRU peuvent apprendre des représentations hiérarchiques des features.

Architecture:

- **Couche GRU:** 128 cellules
- **Dropout:** 0.5 (training seulement)
- **Couche SVM:** Couche dense linéaire
- **Paramètre C:** 5

Paramètres utilisés:

- Taille de batch: 128
- Taux d'apprentissage: 0.001
- Nombre d'epochs: 3000
- Optimiseur: Adam

4.2 DSO2: Histopathology Image Analysis

For DSO2, we developed and compared two deep learning architectures for IDC detection from histopathology images:

4.2.1 Model 1: Custom Convolutional Neural Network (CNN)

We designed a custom CNN architecture specifically tailored for the 50×50 histopathology images:

- **Architecture Rationale:** CNNs are ideal for image data as they automatically learn hierarchical spatial patterns (edges → textures → tissue structures) through convolutional layers with weight sharing.
- **Layer Structure:**
  1. Conv2D (32 filters, 3×3) + ReLU + MaxPooling2D
  2. Conv2D (64 filters, 3×3) + ReLU + MaxPooling2D
  3. Conv2D (128 filters, 3×3) + ReLU + MaxPooling2D

    4. Flatten layer
    5. Dense layer (128 units) + ReLU + Dropout (0.4)
    6. Output layer (1 unit) + Sigmoid activation

- **Parameters:** 683,329 trainable parameters
- **Training:** 10 epochs with Adam optimizer, binary cross-entropy loss

4.2.2 Model 2: Transfer Learning with ResNet50

To leverage pre-learned features and potentially improve performance:

- **Base Model:** ResNet50 pre-trained on ImageNet (frozen weights)
- **Custom Head:**
    1. Global Average Pooling 2D
    2. Dense layer (128 units) + ReLU + Dropout (0.4)
    3. Output layer (1 unit) + Sigmoid activation
- **Parameters:** 23,850,113 total parameters (262,401 trainable)
- **Training:** 5 epochs with Adam optimizer, binary cross-entropy loss

4.2.3 Model Selection Rationale

- The custom CNN was designed for efficiency and interpretability with the specific image size.
- ResNet50 was chosen for transfer learning to benefit from features learned on large-scale image datasets.
- Both models use sigmoid activation for binary classification (IDC vs. non-IDC).

4.3 DSO3: Identify risk factors in healthy patients

Several machine learning models were evaluated to identify patients at higher risk of breast cancer based on clinical variables. The selected models cover different learning paradigms, including tree-based, margin-based, and ensemble approaches, allowing a comprehensive comparison under identical experimental conditions.

4.3.1 Model 1: Random Forest

Model Rationale: Random Forest is an ensemble learning method based on multiple decision trees. It is particularly well suited for structured medical data, as it can capture non-linear relationships and interactions between clinical variables while remaining robust to noise.

Parameters Used:

- Number of trees (*n_estimators*): 100
- Class weight: balanced
- Maximum depth: unconstrained
- Random state: 42

Strengths:

- Handles heterogeneous clinical features effectively
- Reduces overfitting through bootstrap aggregation
- Provides feature importance for medical interpretability

—

4.3.2 Model 2: Support Vector Machine (SVM)

Model Rationale: Support Vector Machines aim to find an optimal separating hyperplane that maximizes the margin between classes. SVMs are effective for complex, non-linear decision boundaries when combined with kernel functions.

Parameters Used:

- Kernel: Radial Basis Function (RBF)
- Regularization parameter: default
- Class weight: balanced
- Probability estimation: enabled

Strengths:

- Strong theoretical foundations
- Effective in high-dimensional feature spaces
- Suitable for non-linear separations

—

4.3.3 Model 3: Gradient Boosting

Model Rationale: Gradient Boosting builds a strong classifier by sequentially combining weak learners, where each new model focuses on correcting the errors of the previous ones.

Parameters Used:

- Number of estimators: 100
- Learning rate: default
- Maximum depth: default

Strengths:

- Capable of modeling complex non-linear patterns
- Good bias-variance trade-off
- Frequently used in structured data problems

—

4.3.4 Model 4: XGBoost

Model Rationale: XGBoost is an optimized gradient boosting algorithm designed for efficiency and scalability. It is widely used in medical and tabular data competitions due to its strong predictive performance.

Parameters Used:

- Number of estimators: 100
- Learning rate: 0.1
- Evaluation metric: log-loss
- Random state: 42

Strengths:

- High predictive power
- Efficient handling of non-linear relationships
- Built-in regularization to reduce overfitting

—

4.3.5 Ensemble Learning

Voting Classifier

Description: A soft voting classifier combining Random Forest, Gradient Boosting, and XGBoost. The final prediction is based on the average predicted probabilities of the individual models.

Rationale: Soft voting ensembles reduce model variance and leverage complementary decision boundaries from different classifiers.

—

Stacking Classifier

Description: A stacking ensemble using Random Forest and XGBoost as base learners, with Logistic Regression as the meta-classifier.

Rationale: Stacking allows the meta-learner to optimally combine predictions from multiple models, potentially improving generalization.

—

4.3.6 Training Strategy

Prior to modeling, the dataset was balanced using SMOTETomek to address class imbalance. The balanced dataset was then split using stratified sampling:

- **Training set:** 75%
- **Test set:** 25%

All models were trained using medically reasonable hyperparameters to preserve interpretability and avoid overfitting.

Chapter 5

Evaluation

5.1 DSO1: Classify tumor type

5.1.1 Métriques d'Évaluation

Tous les modèles ont été évalués avec les mêmes métriques pour permettre une comparaison équitable :

- **Accuracy:** Pourcentage total de prédictions correctes
- **TPR (Sensibilité):** Capacité à détecter les vrais positifs (cancers malins)
- **TNR (Spécificité):** Capacité à identifier les vrais négatifs (cancers bénins)
- **FPR:** Taux de faux positifs
- **FNR:** Taux de faux négatifs
- **Matrice de confusion:** Analyse détaillée des types d'erreurs

5.1.2 Résultats Comparatifs

Algorithme	**Accuracy**	**TPR**	**TNR**	**Comparaison Article**
Régression Linéaire	96.49%	90.62%	100.00%	+0.40%
Plus Proche Voisin (L1)	95.91%	93.75%	97.20%	+2.34%
Plus Proche Voisin (L2)	94.15%	90.62%	96.26%	-0.58%
Régression Softmax	98.83%	96.87%	100.00%	+1.17%
SVM (L2)	98.25%	95.31%	100.00%	+2.15%
MLP	96.49%	92.19%	99.07%	-2.55%
GRU-SVM	96.49%	90.62%	100.00%	+2.74%

Table 5.1: Performance des algorithmes sur le dataset WDBC

5.1.3 Analyse des Performances

Meilleurs Performants

- **Régression Softmax:** Meilleure accuracy globale (98.83%) avec un excellent équilibre TPR/TNR
- **SVM:** Performance très proche de Softmax (98.25%) avec une spécificité parfaite
- **GRU-SVM:** Amélioration significative par rapport aux résultats de l'article (+2.74%)

Performances Modérées

- **Régression Linéaire:** Performance solide malgré sa simplicité, dépassant les résultats de l'article
- **Plus Proche Voisin:** Bonne performance, particulièrement avec la distance de Manhattan

Analyse des Erreurs

- **Faux Négatifs:** Critiques en oncologie, minimisés par Softmax et SVM
- **Faux Positifs:** Moins critiques mais génèrent de l'anxiété et des tests supplémentaires
- **MLP:** Sous-performance relative possiblement due au surapprentissage malgré l'architecture profonde

5.1.4 Comparaison avec l'Article de Référence

Dans l'ensemble, 5 des 6 algorithmes ont dépassé les performances rapportées dans l'article de référence :

- **Améliorations:** +0.40% à +2.74% selon l'algorithme
- **Déclin:** Seul le MLP a performé moins bien (-2.55%)
- **Consistance:** Les tendances relatives entre algorithmes sont globalement préservées

5.1.5 Implications Cliniques

- **Softmax et SVM:** Recommandés pour leur haute précision et faible taux de faux négatifs
- **Interprétabilité:** La régression linéaire offre la meilleure interprétabilité clinique
- **Temps de Calcul:** Les modèles simples (régression, k-NN) sont préférables pour le déploiement en temps réel

5.2 DSO2: Histopathology Image Analysis

5.2.1 Evaluation Metrics

Both models were evaluated using comprehensive metrics suitable for medical image analysis:

- **Accuracy:** Overall correctness of predictions
- **Loss:** Binary cross-entropy loss during training and testing
- **Confusion Matrix:** Detailed breakdown of true positives, false positives, true negatives, false negatives
- **ROC-AUC:** Area under the Receiver Operating Characteristic curve

5.2.2 Model Performance Results

Custom CNN Model:

- **Test Accuracy:** 85.69%
- **Test Loss:** 0.3918
- **Training Time:** 500 seconds per epoch

ResNet50 Transfer Learning Model:

- **Test Accuracy:** 85.78%
- **Test Loss:** 0.3516
- **Training Time:** 700 seconds per epoch (slower due to larger architecture)

5.2.3 Comparative Analysis

- **Performance:** Both models achieved similar accuracy (85.7-85.8%), with ResNet50 showing slightly better loss reduction.
- **Efficiency:** The custom CNN trained faster (10 epochs in 5000s vs. ResNet50's 5 epochs in 3500s).
- **Interpretability:** The custom CNN is more interpretable with fewer parameters and simpler architecture.
- **Clinical Relevance:** High recall is crucial for cancer detection to minimize false negatives (missed IDC cases).

5.2.4 Key Insights

- Patient-wise splitting prevented optimistic bias that could occur with random image splitting.
- Class imbalance was addressed through appropriate loss functions and evaluation metrics.
- Both architectures demonstrated capability in learning discriminative features from histopathology images.

5.3 DSO3: Identify risk factors in healthy patients

5.3.1 Evaluation Metrics

Given the medical application, evaluation focused on both predictive performance and clinical relevance:

- Accuracy
- Precision
- Recall (Sensitivity)
- F1-score
- ROC-AUC
- Confusion Matrix

Recall was prioritized to minimize false negatives, which correspond to missed high-risk patients.

5.3.2 Model Performance Comparison

Among all tested models, the Random Forest classifier achieved the best overall performance:

- **Accuracy:** 63.9%
- **ROC-AUC:** 0.691
- **Sensitivity (Recall):** 0.633
- **Specificity:** 0.646

Other models such as SVM and XGBoost showed competitive accuracy but lower ROC-AUC.

5.3.3 ROC Curve Analysis

ROC curve analysis confirmed that Random Forest provides the strongest discrimination capability between healthy individuals and patients compared to other models.

5.3.4 Clinical Interpretation

Feature importance analysis revealed that metabolic indicators such as Glucose, BMI, HOMA-IR, and Insulin are key contributors to breast cancer risk. These findings are consistent with established medical research linking metabolic syndrome to cancer development.

Chapter 6

Deployment

To make the developed models accessible and usable in a real-world context, a web-based deployment was implemented.

A web application was developed using the **Flask** framework for the backend and **HTML, CSS, and JavaScript** for the frontend. The application allows users to input patient data through a graphical interface, which is then processed by the trained machine learning model.

The backend handles data validation, preprocessing, and model inference, while the frontend provides clear and intuitive feedback regarding tumor presence predictions. This deployment approach demonstrates how machine learning models can be integrated into practical clinical decision-support systems.

For the text-based models (DSO1 and DSO3), the deployment would include:

- Forms with data verification
- Real-time preprocessing (data preparation pipeline)
- Model confidence scores

For the image-based model (DSO2), the deployment would include:

- Image upload functionality for histopathology patches
- Real-time preprocessing (resizing, normalization)
- Model confidence scores

Sources

- Scikit-learn Documentation:
  https://scikit-learn.org

- TensorFlow Documentation:
  https://www.tensorflow.org

- World Health Organization (WHO):
  Breast Cancer Fact Sheets

- Histopathologic images:
  https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images

- Clinical data:
  https://www.kaggle.com/code/yasserhessein/breast-cancer-coimbra-classification-with-eda-ml/input

- WDBC:
  https://www.kaggle.com/code/nancyalaswad90/analysis-breast-cancer-prediction-dataset/input