INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH AND INNOVATIVE STUDIES
ISSN: 2820-7157
www.ijsrisjournal.com

December 2025 **Volume 4** **Number 6** **06-22**

Received Date: October 20, 2025 **Accepted Date: November 11, 2025** **Published Date: December 01, 2025**

Development of a predictive model for breast cancer based on risk factors, a significant advance in improving early detection of the disease

Prof. Dr KITONDUA LUBANZADIO Richard [1] , Prof. Dr ISAKATONGA LOANIE Justin [2] , PhD Student Blaise KAPALALA KAPENDA [3] , MBUYAMBA MBUYAMBA Trésor [4] , KAPALALA BIBIANE Clarisse[5] , ETUMANGELE TSHITAKA Gabriel [6]

Pr. Dr KITONDUA LUBANZADIO Richard – National Pedagogical University – Kinshasa, DRC
Online: (+243) 999 943 126 Email:richardkitondua@gmail.com
Pr. Dr ISAKATONGA LOANIE Justin – Higher Institute of Education – Gombe – Kinshasa DRC
Online: (+243) 813 895 362 Email:issakatonga@yahoo.fr
Mr KAPALALA KAPENDA Blaise, PhD Student – Higher Institute of Education – Gombe Kinshasa
Online: (+212) 612-975163, (+243) 211459397 Email:blaisekapalala94@gmail.com
Mr Mbuyamba Mbuyamba Trésor, Higher Institute of Education – Gombe Kinshasa – DRC Online:
(+243) 851 916 057 Email:mbuyamba12@gmail.com
Ms KAPALALA BIBIANE Clarisse, Higher Institute of Medical Techniques – Kinshasa – DRC
Online: (+243) 81 08 86 930 Email: clarissekapalala3@gmail.com
Mr ETUMANGELE TSHITAKA Gabriel Higher Institute of Education – Gombe Kinshasa – DRC
Online: (+243) 817 851 599 Email: gabrieletums2015@gmail.com

Abstract

Breast cancer is one of the most common cancers among women, representing a major cause of morbidity and mortality worldwide. In the Democratic Republic of Congo, the lack of systematic screening exacerbates the difficulties in identifying high-risk patients at an early stage. This study aims to develop a predictive model based on machine learning, drawing on an analysis of breast cancer risk factors. The main objective is to facilitate early detection and improve patient care by proactively identifying those at increased risk of developing breast cancer. The methodology of this study is based on the Team Data Science Process (TDSP), which organises and optimises the various stages of predictive model development, from data collection to results evaluation. Clinical data and variables related to family history, health behaviours, and physiological characteristics were incorporated into the model. Machine learning algorithms, including Support Vector Machine (SVM), were used to build the model, offering 97% accuracy in predicting breast cancer risk. This model was then deployed in a production environment with an intuitive user interface, allowing physicians to easily analyse the results. The results showed that integrating such a tool could significantly improve early detection capabilities in settings where screening resources are limited, while reducing treatment costs associated with late diagnoses. The findings of this article pave the way for the implementation of a national screening programme based on predictive tools, thereby helping to reduce healthcare disparities and improve the prognosis for at-risk patients. Future improvements to the model could focus on integrating new biomedical data and adjusting models for more diverse clinical settings.

Keywords: model, cancer, early, factors, detection, disease, predictive, breast, significant,

1. INTRODUCTION

Breasts play an important role in femininity and in the image women have of their bodies. The biological function of the breast is to produce milk to feed a newborn baby. The structure of the breast is complex. Each breast (also called a mammary gland) is composed of fifteen to twenty compartments separated by fatty tissue, which gives the breast its familiar shape. Each of these compartments is made up of lobules and ducts. The role of the lobules is to produce milk during breastfeeding; the ducts then transport the milk to the nipple. **(Sor savoir patient, 2022)**.

Cancer is a disease of the cell. The cell is the basic unit of life. There are more than two hundred different types of cells in the body. All of them have a specific role: muscle cells, nerve cells, bone cells, etc. However, when a cell becomes cancerous, it loses its ability to repair itself. It then begins to multiply and eventually forms a mass called **a malignant tumour**.

Malignant tumour cells tend to leave their original tissue and invade neighbouring tissues; this is known as **invasive cancer**. Some tumours remain in their original tissue without invading neighbouring tissues. This is known as **cancer in situ ("remaining in place")**. **(Sor savoir patient, 2022).** Cancer cells tend to migrate to other organs or parts of the body, where they develop new tumours called metastases. In this case, the cancer is said to be **metastatic**. **(Alessandro Furlan, 2007). Breast cancer is a malignant tumour** that develops **in the breast**. There are different types of breast cancer. The most common (95%) develop from cells in the ducts (ductal cancer) and lobules (lobular cancer). These are called adenocarcinomas.

Scientists are trying to find out why it occurs. Despite advances that have led to a better understanding of the mechanisms of cancer development, **the causes of breast cancer are currently unknown**. However, studies have identified certain **risk factors** that promote breast cancer, such as age, genetics, family history, etc. This is why breast cancer screening is widespread in some countries. Throughout the country, women between the ages of 50 and 74 are invited to have a mammogram every two years, which is covered 100% by the National Health Service (with no upfront costs).

2. Purpose of this article

2.1 Main objective

The main objective of this article is to develop a machine learning model that can predict the risk of developing breast cancer in women, using breast cancer risk factors, in a context where there is no national systematic screening programme.

2.2 Specific objectives

- **Analyse risk factors**: Identify and analyse the main risk factors associated with breast cancer, taking into account their interaction and their impact on the probability of developing the disease.
- **Design and train the predictive model**: Use machine learning algorithms to design a predictive model, then train it with the collected data to optimise its accuracy and robustness.
- **Test the clinical application of the model**: Evaluate the feasibility and effectiveness of integrating the predictive model into clinical practice, ensuring that it can help healthcare professionals identify high-risk women and recommend appropriate screening tests.
- **Explore the prospects for implementing a national screening programme**: Use the results of this study to formulate recommendations and lay the groundwork for a future national systematic breast cancer screening programme.

3. Issues related to breast cancer

Breast cancer is a disease characterised by the abnormal and uncontrolled proliferation of breast cells that form a malignant tumour. It is one of the most common cancers in women worldwide and a major cause of female mortality. However, it can also affect men, although to a much lesser extent (approximately 1% of cases).

The incidence of breast cancer varies from region to region, but it remains a serious public health problem. Its impact is medical, psychological, economic and social, as it affects not only the patient, but also her family and society as a whole. Thanks to advances in medicine (early detection, targeted treatments, awareness), survival rates have improved, but many challenges remain. **Some of the problems associated with breast cancer include:**

4. Public health issues

Breast cancer is a major global public health issue, as it is the most common cancer in women and the leading cause of cancer death in women, with millions of new cases and hundreds of thousands of deaths each year. The scale of the problem:

1. ***A global disease***: Breast cancer affects virtually all women and men (in less than 1% of cases) from puberty onwards, with incidence peaking with age.
2. ***Alarming statistics***: In 2022, an estimated 2.3 million women were diagnosed with breast cancer and 670,000 died from the disease.
3. ***Inequalities***: Diagnosis and mortality rates vary considerably between countries, particularly in terms of access to screening and care.

— **Firstly, the lack of a systematic screening programme**: Unlike some countries where regular breast cancer screening programmes are in place, our country does not have such a programme. This leaves a large proportion of the female population without systematic monitoring, increasing the risk of late diagnosis.

— **Secondly, the limitations of traditional methods**: Although mammograms and breast MRIs are effective tools for breast cancer screening, they are often only used when there are already symptoms or specific reasons to suspect cancer. This reactive approach can lead to late diagnoses, when treatment is less effective.

— **Tercio regarding the underutilisation of available data**: Breast cancer risk ctors, including medical history, genetic information, and lifestyle factors, are not fully exploited for proactive breast cancer prediction. There is an opportunity to use this data to identify high-risk women before symptoms appear.

5. Scope of the study

The scope of this article is defined along three main axes: thematic, temporal, and spatial. These axes allow us to situate the limits and scope of the research in the context of developing a predictive model for breast cancer based on its risk factors.

5.1 Thematic scope

This article focuses on the ***development of a predictive model for breast cancer using breast cancer risk factors***. It explores the application of artificial intelligence and machine learning algorithms in modelling the risks associated with breast cancer. Thematic aspects include the identification and analysis of risk factors (genetic, clinical, and lifestyle-related), the design of a predictive model, and its potential integration into clinical practice to improve early cancer detection.

Our research is limited to the analysis of available data and the application of supervised learning algorithms to predict risks, without addressing therapeutic or cancer treatment aspects.

5.2 Time frame

This research covers the period from November 2024 to September 2025, during which the data used for modelling was collected and analysed. This period was chosen because of the accessibility and relevance of recent data that incorporates recent advances in the field of artificial intelligence applied to medicine.

5.3 Spatial scope

For this research, we chose the city of Kinshasa as the location for our research, with a particular focus on the LK Hospital healthcare facility and the local population that benefits from Kinshasa. The predictive model is designed to be adaptable and applicable in various clinical contexts, but it also specifically targets a female population without uniform access to screening programmes.

6. Methodology

This document presents a detailed analysis of the methodologies used to develop a predictive model for breast cancer based on risk factors, as well as recent significant advances that improve the early detection of this disease. The objective is to provide a structured overview that is useful in the context of an academic thesis, research project or clinical development.

6.1. Methodologies for developing a predictive model

a) Classical statistical approaches

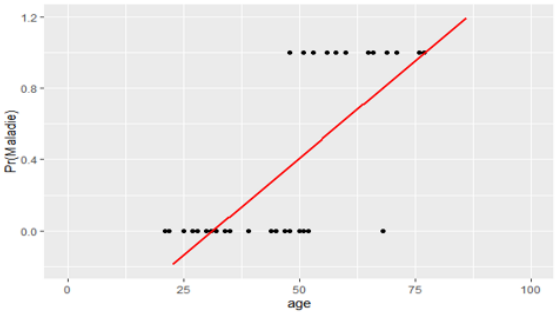

Figure 1: Logistic regression

• **Logistic regression**: used to estimate the probability o developing breast cancer based on risk factor (age, famil history, breast density, hormonal status). It offers high interpretability but has difficulty capturing complex relationships. Examples: Gail and Tyrer-Cuzick model.

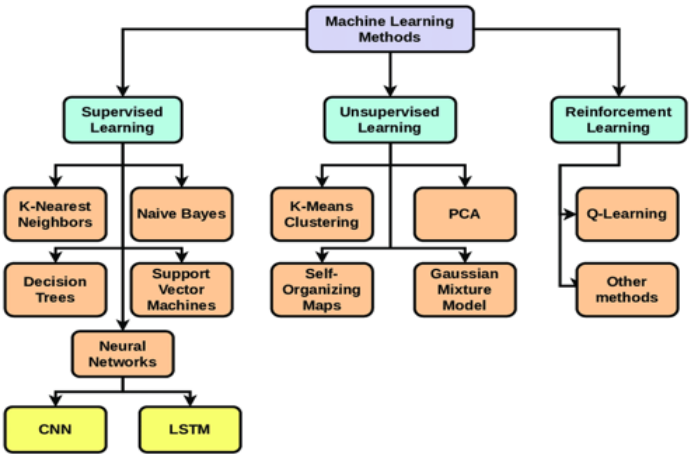

Figure 3: Machine learning methods

• **Cox models (survival analysis)**: used to estimate the risk of occurrence over time, useful for monitoring progression.

```
library(parameters)
parameters(fit, exponentiate = TRUE)

Parameter    | Coefficient |       SE |         95% CI |     z |     p
-------------------------------------------------------------------------
rx [Lev]     |        0.95 |     0.11 | [0.76, 1.18]   | -0.49 | 0.623
rx [Lev+5FU] |        0.69 |     0.08 | [0.54, 0.87]   | -3.11 | 0.002
age          |        1.00 | 4.17e-03 | [1.00, 1.01]   |  0.83 | 0.405
sex          |        0.98 |     0.09 | [0.81, 1.18]   | -0.25 | 0.799
perfor       |        1.18 |     0.31 | [0.70, 1.97]   |  0.63 | 0.532
surg         |        1.22 |     0.13 | [0.99, 1.50]   |  1.86 | 0.062

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed
  using a Wald z-distribution approximation.
```

Figure 2: Cox models (survival analysis)

b) Machine learning methods

— **Decision trees and Random Forest**: capture complex interactions between factors. Random Forest reduces overfitting and provides measures of variable importance;

— **Gradient Boosting Machines (XGBoost, LightGBM)**: highly effective for tabular data, more accurate but less interpretable;

— **SVM (Support Vector Machines)**: effective for separating high- and low-risk patients, performs well on small datasets;

— **K-Nearest Neighbours (KNN)**: classification based on similarity, limited in cases of excessive dimensionality.

c) Deep learning

— **Deep neural networks (DNN)**: allow clinical, biological and genetic data to be integrated simultaneously;

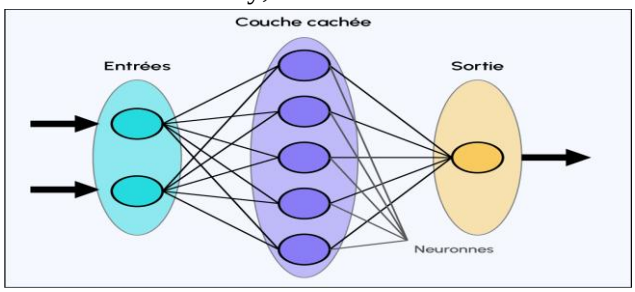

Figure 4: Deep neural networks (DNN)

— **Convolutional Neural Networks (CNN)**: used for medical image analysis (mammography, MRI, ultrasound), identify masses and microcalcifications;

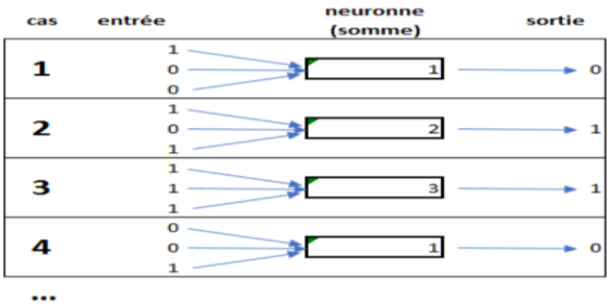

Figure 5: Convolutional Neural Networks (CNN)

— **Multimodal models**: combine imaging, clinical and genetic data, producing more robust and personalised predictions.

Figure 6: Multimodal models

d) Radiomics

A method that involves extracting quantitative characteristics from medical images (shape, texture, intensity). This data is then integrated into machine learning models to identify signatures correlated with tumour risk or aggressiveness.

e) Validation and evaluation

— **Internal validation**: separation into training, validation and test data;
— **External validation**: application to another cohort to verify robustness;
— Metrics: AUC-ROC, sensitivity, specificity, positive and negative predictive values;
— **Calibration**: verification that the estimated probabilities correspond to the actual risk.

6.2. Significant advances in early detection

a) Artificial intelligence applied to mammography

AI algorithms help radiologists detect abnormalities. The result is improved sensitivity and reduced variability between readers. Examples: AI that automatically marks suspicious areas.

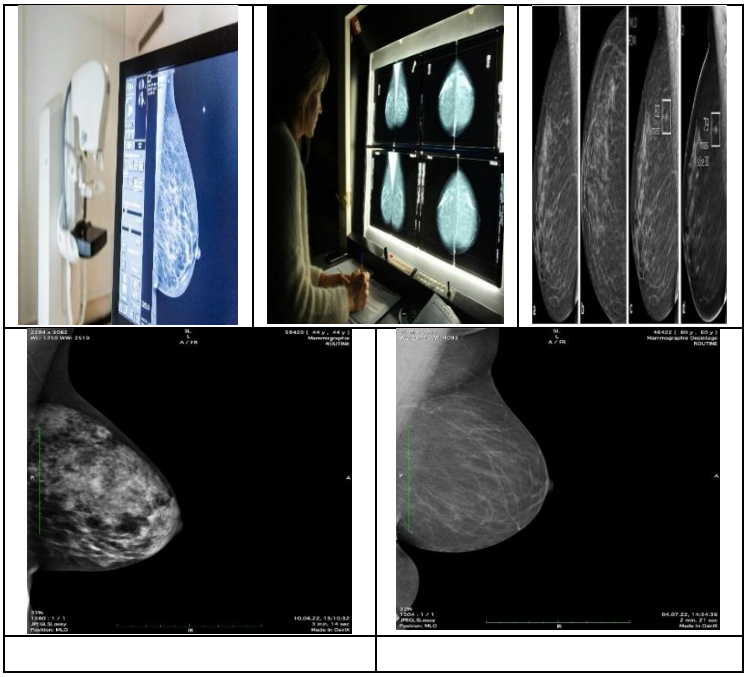

Figure 7: AI algorithms help radiologists detect abnormalities

b) Digital breast tomosynthesis (DBT)

3D mammography reduces the effect of breast tissue overlap. It increases cancer detection, reduces false positives and is particularly useful for dense breasts.

Figure 8: Digital mammography for Helianthus DBT breast tomosynthesis

c) Polygenic Risk Scores (PRS)

PRS combine hundreds of genetic variants to assess individual risk. They enable patients to be stratified into risk groups, paving the way for personalised screening. Limitation: variable performance depending on the population.

Figure 9: Viola DBT full-field digital mammography (2)

d) Liquid biopsies

Analysis of circulating biomarkers (circulating tumour DNA, circulating tumour cells). Potential: to detect cancer before it is visible on imaging and to monitor recurrence.

Figure 10: Liquid biopsies

e) Radiomics and multiparametric imaging

Multiparametric MRI (anatomical, diffusion, perfusion) combined with radiomic signatures improves the distinction between benign and malignant lesions.

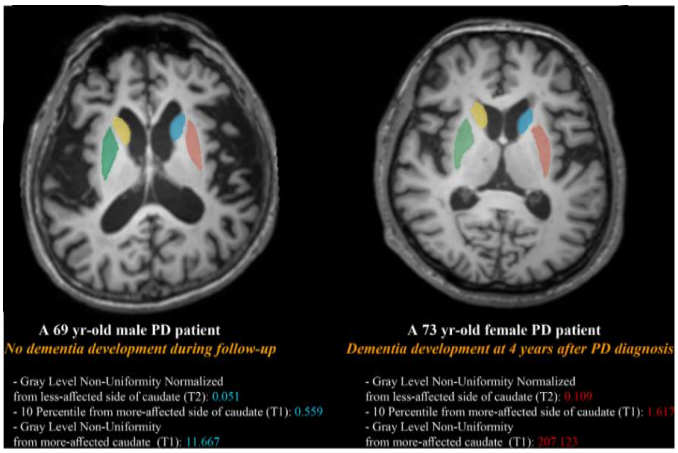

Figure 11: Predict dementia conversion

f) Personalised risk-based screening

This approach combines age, breast density, family history and PRS to tailor the frequency and modality of screening. It improves the effectiveness of screening while limiting over-diagnosis.

Figure 12: Breast cancer: towards more personalised screening

7. Analysis

7.1 Data exploration

Exploratory analysis of the relationships between different variables to better understand the correlations between them.

7.2 Preparing data for exploration

Before exploring the data, it must be cleaned and prepared. This includes:

- **Removing duplicates**: Eliminating redundant or incorrect entries. We do not have any duplicates in this dataset because the BCSC did some preliminary work on data preparation.

- **Treatment of missing values**: Imputation of missing values using statistical methods or deletion of incomplete observations. Note that, as in the previous point, we do not have any missing values thanks to the preliminary work carried out by BCSC.

7.2.1 Descriptive statistical analysis

This step consists of analysing the main characteristics of the data using basic statistical methods:

- **Univariate statistics**: Calculation of measures of central tendency (mean, median) and dispersion (standard deviation, variance) for each variable. All continuous variables are grouped together, so we are dealing with categorical variables. It is therefore not necessary to measure these trends.
- **Visualisation of distributions**: This involves visualising the distribution of the variables.

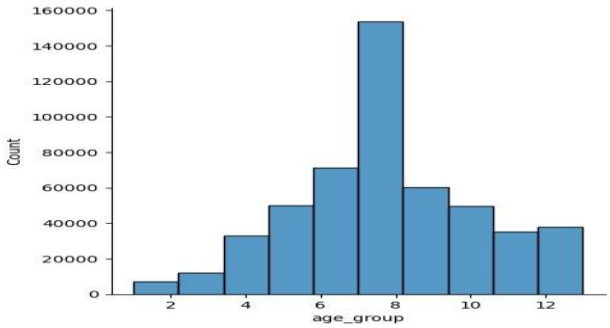

Figure 13: Distribution of the age variable

Note that the distribution across age groups shows us that the majority of women analysed are in groups 6, 7, 8 and 9. However, group 7 has the highest number of entries.

- **Frequency assessment**: Verification of the distribution of categorical variables.

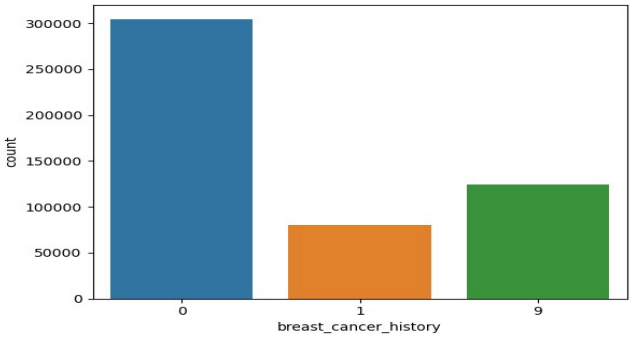

Figure 14: Distribution of the dependent variable in the dataset

7.2.2 Correlation analysis

Correlation analysis between variables allows important relationships to be identified:

Figure 15: Correlation matrix between different variables

7.2.3 Detection of anomalies

Data mining also helps to identify outliers or anomalies that may distort the model:

- **Identification of outliers**: Use techniques such as box plots or interquartile distance to detect abnormal data points.

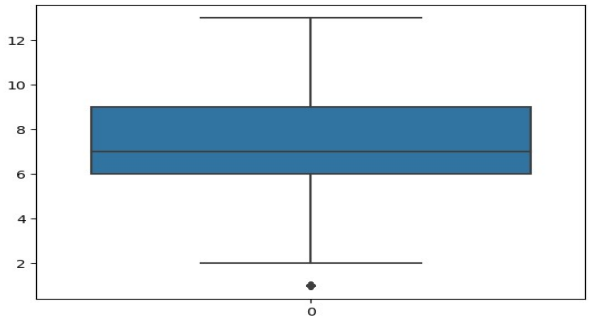

Figure 16: Interquartile distance of the age variable

- **Decision on how to handle outliers**: Either remove them or transform them to make them more consistent with the rest of the data.

We do not have outliers in this dataset thanks to the work carried out upstream by the BCSC.

Transformation of variables: Normalisation or standardisation of continuous variables and transformation of categorical variables into numerical values if necessary.

a. Separation of data into features X and target label y:

```
X = df.drop('breast_cancer_history',axis=1)
y = df['breast_cancer_history']
```

b. Perform a train-test split on the data, with a test size of 10%:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=101)
```

7.2.4 Modelling

The modelling phase involves designing, training and optimising the predictive model.

- **Choice of algorithms**: Justification of the machine learning algorithms selected.

For this project, we will use two of the most effective machine learning algorithms for classification problems: SVM (Support Vector Machine) and KNN (K-Nearest Neighbours).
After evaluation, only one will be selected for our model.

- **Training, optimisation and validation**: Description of the model training process using the collected data and cross-validation to measure performance.

1. SVM (Support Vector Machine):

   Model creation:

```
from sklearn.svm import SVC
svc = SVC(class_weight='balanced')
```

   Use of GridSearchCV to perform a grid search to find the best C and gamma parameters:

```
from sklearn.model_selection import GridSearchCV

param_grid = {'C':[0.001,0.01,0.1,0.5,1],'gamma':['scale','auto']}
grid = GridSearchCV(svc,param_grid)
```

   Model training:

```
grid.fit(X_train,y_train)
```

GridSearchCV

GridSearchCV(estimator=SVC(class_weight='balanced'),
param_grid={'C': [0.001, 0.01, 0.1, 0.5, 1],
'gamma': ['scale', 'auto']})

estimator: SVC

SVC(class_weight='balanced')

SVC

Figure 17: SVM model training

   Best parameters:

```
grid.best_params_
{'C': 1, 'gamma': 'auto'}
```

   Evaluation:

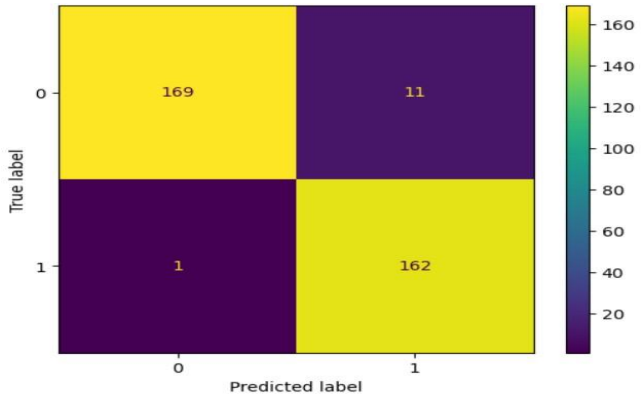

Figure 18: SVM confusion matrix

	precision	recall	f1-score	support
0	0.99	0.94	0.97	180
1	0.94	0.99	0.96	163
accuracy			0.97	343
macro avg	0.97	0.97	0.96	343
weighted avg	0.97	0.97	0.97	343

Figure 19: SVM classification report

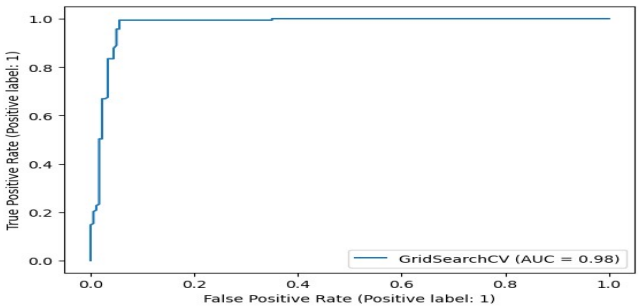

Figure 20: SVM ROC curve

2. KNN (K-Nearest Neighbours):

Model creation:

```
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier()
operations = [('knn',knn)]
from sklearn.pipeline import Pipeline
pipe = Pipeline(operations)
```

Use of GridSearchCV to perform a grid search to find the best parameters:

```
from sklearn.model_selection import GridSearchCV
k_values = list(range(1,30))
param_grid = {'knn__n_neighbors': k_values}
full_cv_classifier = GridSearchCV(pipe,param_grid,cv=5,scoring='accuracy')
```

Model training:

```
full_cv_classifier.fit(X_train,y_train)
```

```
GridSearchCV(cv=5, estimator=Pipeline(steps=[('knn', KNeighborsClassifier())]),
             param_grid={'knn__n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,
                                              12, 13, 14, 15, 16, 17, 18, 19,
                                              20, 21, 22, 23, 24, 25, 26, 27,
                                              28, 29]},
             scoring='accuracy')
estimator: Pipeline
Pipeline(steps=[('knn', KNeighborsClassifier())])
KNeighborsClassifier
```

Figure 21: Training the KNN model

Best parameters:

```
full_cv_classifier.best_estimator_.get_params()
```

```
{'memory': None,
 'steps': [('knn', KNeighborsClassifier(n_neighbors=9))],
 'verbose': False,
 'knn': KNeighborsClassifier(n_neighbors=9),
 'knn__algorithm': 'auto',
 'knn__leaf_size': 30,
 'knn__metric': 'minkowski',
 'knn__metric_params': None,
 'knn__n_jobs': None,
 'knn__n_neighbors': 9,
 'knn__p': 2,
 'knn__weights': 'uniform'}
```

Figure 22: Best KNN parameters

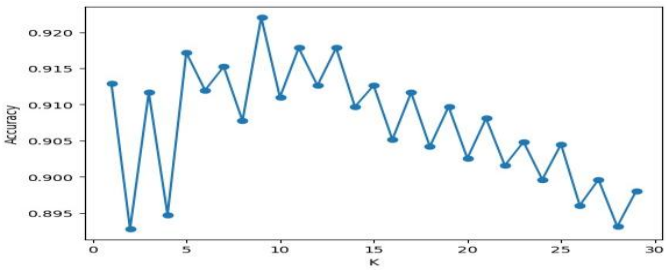

Figure 23: Model performance evaluation curve with KNN based on the K parameter

Evaluation:

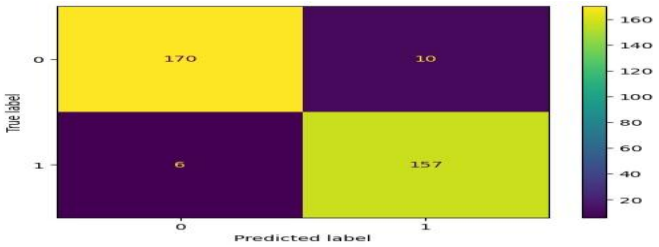

Figure 24: KNN confusion matrix

	precision	recall	f1-score	support
0	0.97	0.94	0.96	180
1	0.94	0.96	0.95	163
accuracy			0.95	343
macro avg	0.95	0.95	0.95	343
weighted avg	0.95	0.95	0.95	343

Figure 25: KNN classification report

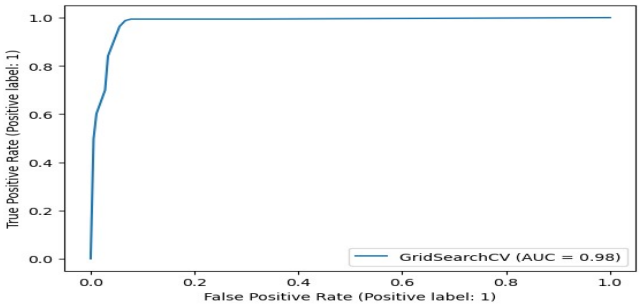

Figure 26: KNN ROC curve

The evaluations of the two models give us:

The proportion of correct predictions out of all predictions (**accuracy**), representing overall precision, is 0.97 for SVM and 0.95 for KNN. The proportion of correct positive predictions out of all positive predictions made (**precision**) for macro avg and weighted avg is 0.97 for SVM and 0.95 for KNN.

The proportion of true positives captured among the truly positive examples (**recall**), representing the sensitivity of the model, is 0.97 for SVM and 0.95 for KNN. The harmonic mean of accuracy and recall (**F1-score**), an overall measure of performance for unbalanced classes, is 0.97 for SVM and 0.95 for KNN.

7.2.5 Deployment

Once the model has been validated, it is deployed for use.

- We begin by saving our model using joblib so that we can load it later into the API.

```
from joblib import dump, load
dump(final_model, 'svn_cancer_model.joblib')
```

- Create a Flask API:

The API will serve as an intermediary between the model and the human-machine interfaces that doctors will use.

```
# Importer les bibliothèques nécessaires
from flask import Flask, request, jsonify
import joblib
import numpy as np

# Créer une application Flask
app = Flask(__name__)

# Charger le modèle SVM sauvegardé
model = joblib.load('svm_model.pkl')
# Route pour prédire la classe d'un exemple
@app.route('/predict', methods=['POST'])
def predict():
    # Récupérer les données JSON envoyées avec la requête
    data = request.json
    # Extraire les caractéristiques (features) pour la prédiction
    features = np.array(data['features']).reshape(1, -1)

    # Faire la prédiction
    prediction = model.predict(features)
    probability = model.predict_proba(features).max()

    # Retourner les résultats sous forme de JSON
    return jsonify({
        'prediction': int(prediction[0]),
        'probability': float(probability)
    })
# Démarrer l'application Flask
if __name__ == '__main__':
    app.run(debug=True)
```

Figure 27: Backend (API)

7.2.6 Customer acceptance

With an overall performance of 0.97, the model meets the main objective defined in the planning phase. The next step is to ensure that stakeholders (doctors, healthcare professionals, patients) are involved in the process of validating and adopting the model. To this end, a follow-up form will be provided to gather feedback from stakeholders. The template is shown below:

Follow-up Form for Doctor Feedback
Name:
_____
Speciality:
_____
Date template was used: _____
Section 1: Use of the template

1. Frequency of use:
How often did you use the predictive model during consultations?
- [ ] Daily
- [ ] Weekly
- [ ] Less than once a week
- [ ] First use

2. Clinical relevance:
Was the model relevant in assessing patients at risk?
- [ ] Very relevant
- [ ] Relevant
- [ ] Not very relevant
- [ ] Not relevant at all

3. Ease of use:
How would you rate the ease of use of the template?
- [ ] Very simple
- [ ] Simple
- [ ] Moderately simple
- [ ] Complex

4. Interpretation of results:
Did you find the results provided by the model clear and easy to interpret?
- [ ] Very clear
- [ ] Clear
- [ ] Moderately clear
- [ ] Difficult to interpret

5. Relationship to clinical examinations:
Were the model results consistent with the results of other clinical examinations (mammography, ultrasound)?
- [ ] Always consistent
- [ ] Often consistent
- [ ] Sometimes consistent
- [ ] Rarely consistent

Section 2: Experience with patients

6. Acceptance by patients:
How did patients react to the results provided by the model?
- [ ] Very well accepted
- [ ] Accepted
- [ ] With scepticism
- [ ] Rejected

7. Impact on decision-making:
Will the model influence your clinical decisions regarding patients?
- [ ] Yes, significantly
- [ ] Yes, moderately
- [ ] No, little influence
- [ ] Not at all

Section 3: Suggestions for improvement

8. Suggestions for improvement of the user interface:
_____

9. Suggested improvements for interpreting results:
_____

10. Other comments:
_____

Signature:_____Date: _____

8. Sample

Breast cancer is the most common malignant tumour in women worldwide and one of the leading causes of cancer-related mortality. More than 1.2 million cases are diagnosed each year, affecting 10 to 12% of the female population and accounting for nearly 500,000 deaths per year worldwide. This geographical variability in incidence and mortality rates reflects the diagnostic and therapeutic challenges of breast cancer in resource-limited countries. **(Ingala et al., 2024)**

a) Incidence and Mortality

In Africa, breast cancer caused approximately 74,072 deaths [95% CI, 67,345-81,472 with 168,690 (152,332–186,804 cases) estimated to occur in 2018. The age-standardised incidence rate (ASIR) was 37.9/100,000 in Africa, while the age-standardised mortality rate was 17.2/100,000 in 2018.

Nigeria had the highest absolute burden with 26,310 [23,610 to 29,919] cases and 11,564 [10,302 to 12,981] deaths, followed by Egypt with 23,081 [21,734 to 24,512] new cases and (9,254 [8,707-9,836) deaths in 2018 **(Fig. 1a and b).** Geographically, high age incidence rates are also reflected in the high number of deaths across all ages, but age-standardised rates were quite asymmetrical between incidence and mortality in Africa. **(Journal of Public Health, 2018)**

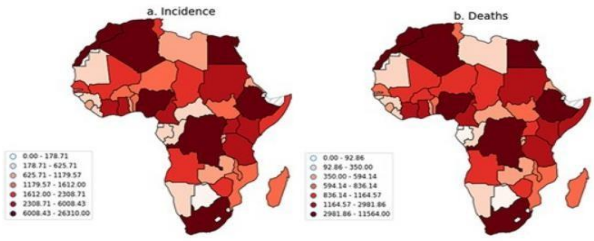

Figure 28: Geographical distribution of the burden of breast cancer in Africa: a) Incidence b) Deaths. Data source: GLOBOCAN 2018 (IARC).

Breast cancer incidence and age-specific deaths increased up to the 45-49 age group and declined thereafter (Fig. 28). This pattern was consistent across almost all African countries, with incidence being highest in the middle age groups (30-49).

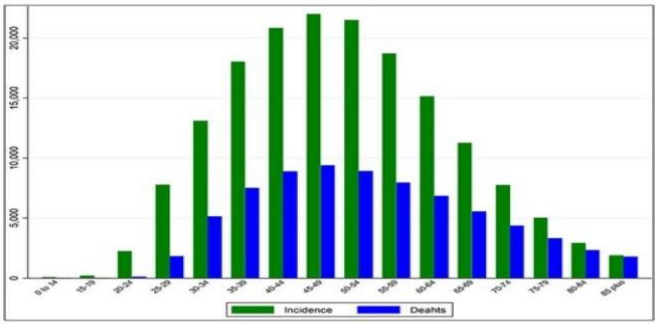

Figure 29: Age-specific distribution of breast cancer incidence and deaths in 2018. Incidence: New cases by age; Deaths: Number of deaths by age. Data source: GLOBOCAN 2018 (IARC).

In Africa, the probability of developing breast cancer was 1 in 25 women (probability: 3.94%) during their lifetime (before the age of 75), while 1 in 57 women (probability: 1.77%) were expected to die from breast cancer during their lifetime (Fig. 3). The risk of developing breast cancer was 10 times higher in Africa; it was highest in Mauritius (7.27%, probability: 1 in 14) and lowest in Gambia (0.71%, probability: 1 in 141). The probability of dying from breast cancer varied sevenfold in Africa, from 0.42% (approximately 1 in 240) in The Gambia to Somalia (2.97%, probability: 1 in 34). **(Journal of Public Health, 2018)**

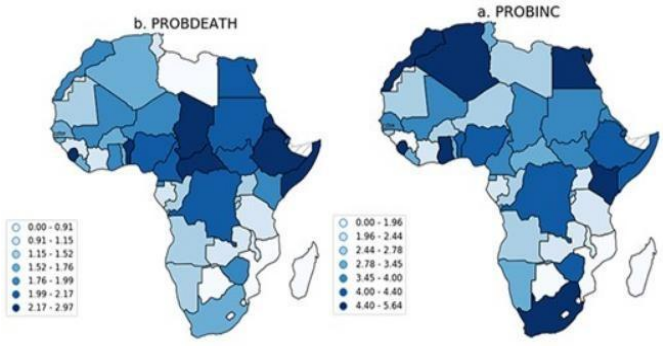

Figure 30: Geographical distribution of the risk of developing and dying from breast cancer in Africa a) PROBINC (b) PROBDEATH PROBINC: Probability of developing breast cancer during one's lifetime (before the age of 75); PROBDEATH: Probability of dying from breast cancer

In the Democratic Republic of Congo (DRC), breast cancer is a major public health problem due to its morbidity and mortality rates. In a study on the epidemiological, clinical and molecular profile of breast cancer in hospitals in the city of Kinshasa, Kingu M *et al.* found a frequency of 24% and that 94% of women were diagnosed at an advanced stage. **(Kingu M et al., n.d.)**

Table 1: The impact of breast cancer in the DRC

Country	Incidence	ASIR	Deaths	ASMR	MIR
Democratic Republic of Congo	6,149 [4,896–7,723]	24.6	3,263 [2,631-4,048]	13.6	0.40

Table 2: Risk of incidence and death from breast cancer in 2018 in the DRC

Country	Incidence	Deaths
Democratic Republic of Congo	2.61 (1 in 38)	1.46 (1 in 69)

b) Risk factors and prediction

There are many different risk factors for breast cancer. These include biological, environmental and behavioural factors. The main ones are as follows:

1. Age and gender

Breast cancer occurs mainly in women, although 0.5 to 1 per cent of cases occur in men. Age is an important factor, with an increased risk after the age of 40. Most cases are diagnosed in women over the age of 50. **(World Health Organisation, n.d.)**

2. Family history and genetics

Family history, particularly when a first-degree relative (mother, sister, daughter) has had breast cancer, increases the risk. Genetic mutations in the **BRCA1** and **BRCA2** genes are strongly associated with an increased risk of developing breast cancer. Other less common genetic mutations, such as PALB2, TP53 and CHEK2, also increase the risk. **(Breast Cancer Hub, n.d.)**

3. Hormonal and reproductive factors

Prolonged exposure to hormones, particularly oestrogen, is a significant risk factor. This includes factors such as early age at first menstruation (before age 12), late menopause (after age 55), and use of hormone replacement therapy after menopause. Not having children or having your first child after the age of 30 may

also increase your risk. **(Global Cancer Observatory, n.d.)**

4. Lifestyle

Lifestyle factors such as alcohol consumption, obesity, and lack of physical activity increase the risk of developing breast cancer. Exposure to X-rays or radiation (e.g., after treatment for another cancer) is also a known risk factor. **(Cancer Info Hub, n.d.)**

8.1 Empirical examination

Research related to breast cancer detection has increased over the last decade. Much of this work has focused on detecting the presence of cancerous tissue in the breast and classifying tumours. The approaches used come from several fields: probability and statistics, connectionism, and other tools from artificial intelligence and cognitive science.

8.1.1 Probabilistic and statistical approaches

Statistical and probabilistic methods are frequently mentioned in the literature, often proposing improved versions of classical approaches , such as Bayesian networks and the k-nearest neighbours rule.

In **(Subhash et al., 2003)**, the authors propose an approach based on a generalisation of the k-nearest neighbour rule for classification in the context of breast cancer screening. The first database was partitioned and classification was performed on each of its partitions. Table I illustrates an example of the classification results for the tenth partition of this data set.

Table 3: Recognition results for the 10th partition of WDBC (Subhash et al., 2003)

k	k-RNN rule				k-NN rule			
	Confusion matrix		Prob. of false positive false negative	Avg. error rate	Confusion matrix		Prob. of false positive false negative	Avg. error rate
1	335	9	0.026	0.041	337	7	0.020	0.033
	11	128	0.079		9	130	0.065	
2	334	10	0.029	0.039	338	6	0.017	0.048
	9	130	0.065		17	122	0.122	
3	334	10	0.029	0.037	335	9	0.026	0.037
	8	131	0.056		9	130	0.065	
4	334	10	0.029	0.037	338	6	0.017	0.041
	8	131	0.056		14	125	0.101	
5	334	10	0.029	0.037	337	7	0.020	0.035
	8	131	0.056		10	129	0.072	
6	334	10	0.029	0.039	338	6	0.017	0.041
	9	130	0.065		14	125	0.101	

The results obtained through the experiments were highlighted in comparison with those obtained using the conventional k-nearest neighbours rule. The best recognition rate obtained was 98.1% for the WDBC database and 97% with WBC. In **(Fei et al., 2003)**, the authors propose a wide margin separation method for classification in the context of breast cancer screening, namely a support vector machine (SVM). The maximum margin hyperplane between classes is obtained by minimising the loss function:

$$L(\alpha) = -\sum_i \alpha_i + \frac{1}{2}\sum_{ij} \alpha_i \alpha_j y_i y_j K(x_i, x_j);$$

Where:
$\{(x_i, y_j)\}_{i=1}^{N}$ represents the pairs of two-class input-outputs such that *Yi* = ± l, *K(Xi, Xj)* is a kernel function applied to the inputs, and $\{\alpha_i\}_{i=1}^{N}$ represents multiplicative update values.

Table 4: SVM classification error rates (breast cancer database) (Fei et al., 2003)

Kernel	Polynomial		Radial		
Data	$k=4$	$k=6$	$\sigma=0.3$	$\sigma=1.0$	$\sigma=3.0$
Sonar	9.6%	9.6%	7.6%	6.7%	10.6%
Breast cancer	5.1%	3.6%	4.4%	4.4%	4.4%

The classification results are shown in Table 2. The experiments were conducted using a polynomial kernel function and a radial basis function. The authors assigned degrees k=4 and k=6 to the polynomial function, and variances σ of 1.0 and 3.0 to the radial basis function. The coefficients $\alpha_i$were initialised to a value of **1** uniformly in each experiment. The classification error rates obtained for the "breast cancer" database vary between 3.6% and **5.1**%.

Table 5: Classification recall rates with the "breast cancer" database (Huang 2004).

Kernel	BMPM					MPM			
	α		Accuracy			α	Accuracy		
	α_x	α_y	TSA_x	TSA_y	TSA	α	TSA_x	TSA_y	TSA
Linear(%)	90.0±0.3↑	50.0±0.0	99.9±0.1↑	92.0±0.2	94.9±0.2	84.2±0.3	96.9±0.4	97.1±0.5	96.9±0.3
Gaussian(%)	97.6±0.3↑	50.0±0.0	100.0±0.0↑	88.9±0.2	92.8±0.2	90.1±0.3	96.6±0.2	97.1±0.3	96.8±0.2

Classification was performed on an unidentified database using the classic *MPM* approach and the approach proposed by the authors **(Huang 2004)**, BMPM, in order to highlight its performance. The simulations were performed with two types of functions, one linear and the other non-linear, i.e. a Gaussian function. More than 94% classification accuracy was achieved with the proposed method. It is important to note that the classifier proposed in **(Huang 2004)** is biased and binary. We are particularly interested in the performance relative to the breast cancer database (WBCD).

Table 6: Recognition rates of the three conventional algorithms and the MBBC approach proposed in (Madden 2002)

	Naïve	TAN	K2	MBBC
Chess	87.63± 1.61	91.68± 1.09	94.03± 0.87	**97.03± 0.54**
WBCD	**97.81± 0.51**	**97.47± 0.68**	**97.17± 1.05**	**97.30± 1.01**
LED-24	**73.28± 0.70**	**73.18± 0.63**	**73.14± 0.73**	**73.14± 0.73**
DNA	94.80± 0.44	94.75± 0.42	**96.22± 0.64**	**95.99± 0.42**
Lymph.	83.60± 9.32	**85.47± 9.49**	81.47± 10.4	83.47± 9.45
Nursery	90.48± 0.41	**94.16± 0.33**	92.63± 0.67	**94.16± 0.33**
SPECT	71.70± 6.56	**81.25± 4.78**	**80.19± 4.66**	**80.75± 4.97**
TTT	70.69± 1.94	75.08± 1.86	74.04± 3.51	**77.37± 4.37**

Based on the results obtained by the author using the proposed *MBBC (Markov Blanket Bayesian Classifier)* method, it can be seen that, with the breast cancer database, the *Markov-Blanket* Bayesian classifier performed comparably to the naive Bayes, naive Bayes with tree (TAN), and conventional Bayesian network (K2) algorithms **(Madden 2002)**.

9. Result

The implementation of the model consists of moving from the development phase to the production phase, thus ensuring that the model can be used in real clinical environments to anticipate the risk of breast cancer in patients. The result brings us to the objective of presenting the implementation process of the breast cancer predictive model developed in this study, as well as the results obtained through user interfaces and concrete use cases.

The predictive model, designed using machine learning algorithms and based on breast cancer risk factors, aims to provide doctors with a decision-making tool to identify high-risk patients and improve early cancer detection. This chapter describes the different stages of model deployment, the tools used to make it accessible to end users, and the integration of the model into a computer system.

We will then detail the user interfaces developed to enable smooth interaction with the model, illustrating the practical use of the system through case studies. Finally, the model's performance in production will be analysed, and feedback from physician users will be discussed to assess the system's acceptance and identify areas for improvement.

9.1 Model production process

9.1.1 Production environment

The production environment is a key aspect in ensuring that the breast cancer predictive model is accessible, effective, and usable in a real clinical setting. This section describes the technical components, infrastructure, and tools used to deploy and run the model in production.

9.1.2 Server and cloud infrastructure

The model has been deployed in a cloud environment to ensure continuous accessibility and efficient processing of breast cancer risk factors. The infrastructure is based on servers configured to run the model responsively, enabling it to respond to a high number of simultaneous requests while ensuring fast execution.

9.1.3 Model integration

The predictive model is integrated into this environment via RESTful APIs that allow user interfaces (frontend) to interact with the model in real time. Each user request, such as the submission of patient data to assess cancer risk, is processed by the API, which returns the prediction results in a matter of seconds.

9.1.4 Security and authentication

Given the sensitive nature of the data, advanced security measures have been put in place, including the use of HTTPS protocols for secure data transfer. Access to the system's web s restricted to authorised users, primarily healthcare professionals.

9.1.5 Model deployment

The deployment of the breast cancer prediction model represents a crucial step in the production process. This section presents the main actions taken to integrate the model into the production environment, ensuring its availability, robustness and effectiveness for clinical use.

9.1.6 Preparing the model for deployment

Prior to deployment, the model was optimised and extensively tested to ensure maximum accuracy and reliable performance in a production environment. The tests included:

- **Validation of the model** on test datasets not used during training, in order to evaluate actual performance in terms of accuracy, recall and F1-score.

- **Optimisation of hyperparameters** to improve model performance, including techniques such as grid search and Bayesian optimisation.
- **Robustness and stability testing** to ensure that the model responds correctly to the various patient data configurations that may be submitted by clinicians.

9.1.7 Cloud infrastructure for deployment

The model has been deployed on a cloud infrastructure, ensuring flexible, scalable and secure deployment. The cloud offers the advantage of high availability and allows for dynamic scaling according to needs:

- **Scalability**: In the event of an increase in the number of users or workload, additional instances of the model can be activated automatically, ensuring constant responsiveness.
- **Security**: The cloud environment has been configured to meet medical data security requirements, including the protection of patient information through encryption protocols and the isolation of computing environments.

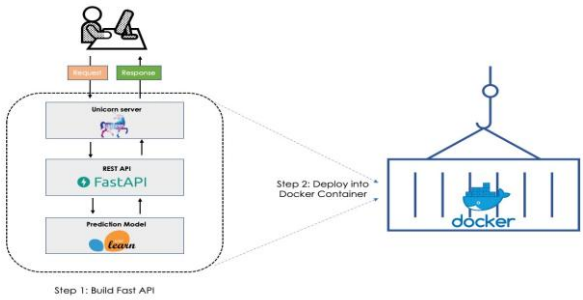

Figure 31: Model deployment diagram

9.1.8 User interfaces

Presentation of the graphical interfaces developed to enable doctors to interact with the predictive model.

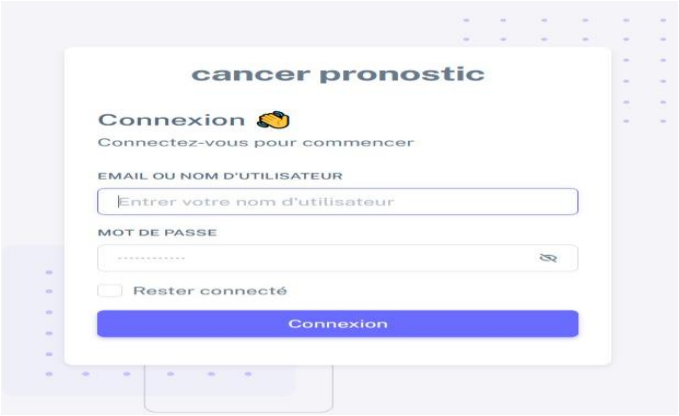

Figure 32: Login page

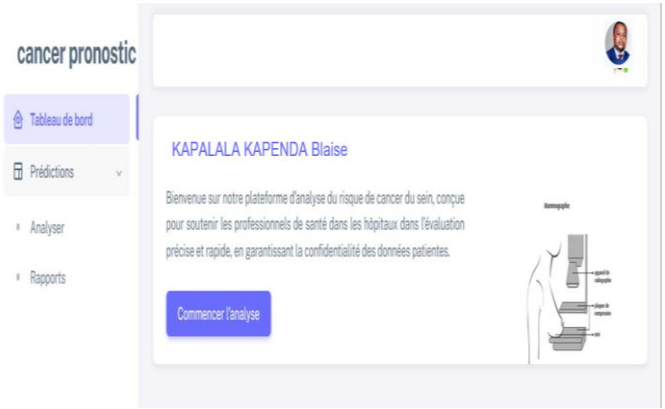

Figure 33: Home page

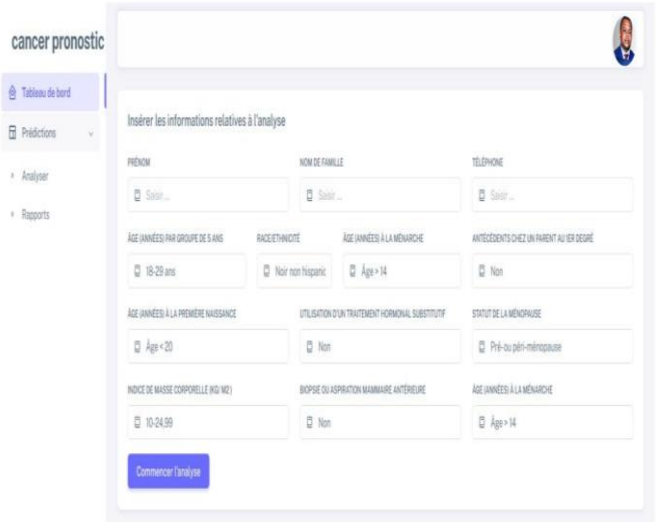

Figure 34: Analysis form

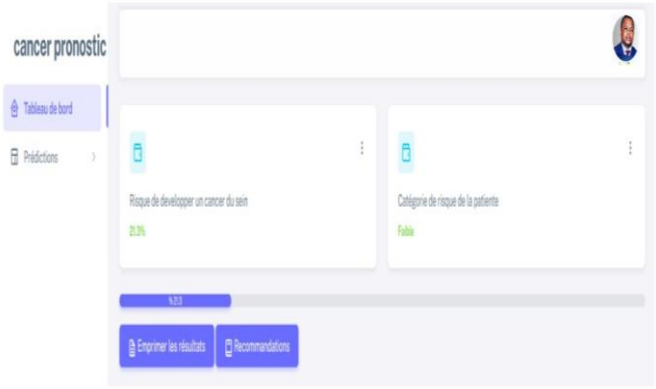

Figure 35: Results form

CONCLUSION

The development of a predictive model for breast cancer based on breast cancer risk factors represents a significant advance in improving the early detection of this disease. Throughout this thesis, we followed a conceptual methodology using the TDSP (Team Data Science Process) framework to structure all stages of the model's design, development, testing and deployment. This process ensured the scientific and technical rigour necessary to develop a reliable model.

The main results obtained during model training, particularly with the support vector machine (SVM) algorithm, showed an accuracy of 0.97, demonstrating the effectiveness of our approach in predicting breast cancer risk. In addition to its strong performance, the model's deployment in a cloud environment, combined with a RESTful API, allows for seamless integration into clinical environments, thereby promoting its adoption by healthcare professionals.

Our study also highlighted the importance of breast cancer risk factors in predicting complex diseases such as breast cancer. By leveraging information related to family history, hormonal factors and lifestyle, the model provides more personalised predictions, contributing to more informed decision-making by clinicians.

However, despite the model's positive performance, several limitations remain, including the availability and quality of data in certain clinical contexts, as well as challenges related to the management of sensitive information. Furthermore, the lack of large-scale clinical validation currently limits the model's use in real-world medical settings.

REFERENCES

1. **Alpaydin, E.** (2020). Introduction to Machine Learning. 4th edition , MIT Press. ISBN: 9780262043793.
2. **Breast Cancer Surveillance Consortium (BCSC)**. (2019). Risk Factor Dataset Download. Available at: https://bcsc-research.org/data/rfdataset.
3. **Docker Inc.** (2023). Docker Documentation. Available at: https://docs.docker.com.
4. **Furlan, A.** (2007). Metastases in breast cancer.
5. **Geron, A.** (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2nd edition, O'Reilly Media. ISBN: 978-1492032647.
6. **Goodfellow, I., Bengio, Y., & Courville, A.** (2016). Deep Learning. MIT Press. ISBN: 978-0262035613.
7. **Hunter, J. D.** (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90–95. DOI: 10.1109/MCSE.2007.55.
8. **Krizhevsky, A., Sutskever, I., & Hinton, G. E.** (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 25, 1097-1105.
9. **LeCun, Y., Bengio, Y., & Hinton, G.** (2015). Deep Learning. Nature, 521, 436-444. DOI:
10. 10.1038/nature14539.
11. **Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P.** (1998). Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, 86(11), 2278-2324.
12. **Louvain Medical**. (2020). Role of machine learning techniques in the fight against breast cancer.

13. **McKinney, W.** (2018). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. 2nd edition, O'Reilly Media. ISBN: 978-1491957660.
14. **Mitchell, T.** (2016). Machine Learning. Carnegie Mellon University.
15. **Muller, G., Taran, G., & Jameson, D.** (2020). Systematic Approaches in Machine Learning Projects: Enhancing Project Transparency and Outcomes. Machine Learning Journal, 22(1), 32-48. DOI: 10.1145/MLJ2020-01-007.
16. **Murphy, K. P.** (2012). Machine Learning: A Probabilistic Perspective. MIT Press. ISBN: 978-0262018029.
17. **Mitchell, T. M.** (1997). Machine Learning. McGraw-Hill.
18. **World Health Organisation**. (2024). Global Breast Cancer Initiative: Breast cancer awareness month.
19. **World Health Organisation**. (2024). Breast cancer.
20. **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., &**
21. **Duchesnay, E.** (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830. Available at : https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.
22. **Peterson, M., Smith, A., & Taylor, R.** (2019). Collaboration and Iteration in Data Science: The Role of Teamwork in Machine Learning Projects. Journal of Data Science and Technology, 14(2), 45–62. DOI: 10.1016/j.jdascit.2019.05.002.
23. **Raschka, S., & Mirjalili, V.** (2017). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow. 2nd edition, Packt Publishing. ISBN: 978-1787125933.
24. **Swinnen, G.** (2010). Learning to program with Python. Editions Eyrolles. ISBN: 9782212126955.
25. **Sutton, R., & Barto, A. G.** (2018). Reinforcement Learning: An Introduction. 2nd edition, MIT Press. ISBN: 978-0262039246.