

MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET DE LA RECHERCHE
SCIENTIFIQUE

ECOLE SUPERIEURE EN INFORMATIQUE 8 MAI 1945

EQUIPE :

ZINEB CHERKI

GHILES MEDDOUR

ANWAR SELMI

SIDI MOHAMED HICHAM ZEKRI

ENCADREE PAR:

MR SIDI MOHAMED BENSLIMANE

MME RIM SAKINA BENABI

2018 - 2019 :

RAPPORT DE REALISATION

D'un Outil d'Aide au Diagnostic du Cancer du Sein

Table des matières

| | |
|---|----|
| Remerciements. | 3 |
| Introduction : | 4 |
| L'informatique au service de la médecine : | 4 |
| Outil d'aide au diagnostic du cancer du sein : | 4 |
| Objectifs du projet. | 5 |
| 1. Analyse des besoins et recensement des données. | 6 |
| 1.1. Besoins et contraintes du projet. | 6 |
| 1.2. Diagrammes des cas d'utilisation. | 8 |
| 1.2.1. Diagramme des cas d'utilisation de la plateforme d'acquisition. | 8 |
| 1.2.2. Diagramme des cas d'utilisation de l'API-REST. | 11 |
| 1.3. Diagrammes de séquence. | 13 |
| 1.3.1. Diagramme de séquence de la plateforme d'acquisition. | 13 |
| 1.3.2. Diagramme de séquence de l'API-REST. | 14 |
| 1.4. Diagrammes d'activité. | 14 |
| 2. Conception de l'application. | 15 |
| 3. Prédiction et diagnostic du cancer du sein. | 16 |
| 3.1. Prédiction sur des <i>features</i> extraites de microphotographies d'échantillons par ponction à l'aiguille fine. | 16 |
| 3.1.1. La biopsie à l'aiguille fine. | 16 |
| 3.1.2. Le jeu de données | 17 |
| 3.1.3. Analyse des données. | 18 |
| 3.1.4. Construction du modèle. | 18 |
| 3.2. Prédiction sur des mammographies | 20 |
| 3.2.1. La mammographie. | 20 |
| 3.2.2. Le jeu de données | 21 |
| 3.2.3. L'apprentissage profond (Deep Learning)..... | 22 |
| 4. Architecture et Organisation. | 26 |
| 4.1. Architecture générale. | 26 |
| 4.1.1. Architecture N-tiers. | 26 |
| 4.1.2. Architecture API-REST. | 27 |
| 4.2. Diagramme de déploiement. | 28 |
| 5. Technologies utilisées. | 30 |
| 5.1. Django. | 30 |

| | | |
|------|---|------------------------------------|
| 5.2. | Flask Framework REST. | Erreur ! Signet non défini. |
| 5.3. | Bibliothèque PyTorch..... | 30 |
| 5.4. | Scikit-learn. | 31 |
| 5.5. | Angular7..... | 32 |
| 5.6. | Android Studio. | 32 |
| 5.7. | MySQL. | 33 |
| 6. | Ergonomie et Interface utilisateur..... | 34 |
| 6.1. | Plateforme d'acquisition..... | 34 |
| 6.2. | API Rest (utilisateur lambda). | 39 |
| 7. | Perspectives. | 42 |
| 8. | Conclusion..... | 42 |
| 9. | Références des figures..... | 43 |
| 10. | Bibliographie / webographie. | 45 |

Remerciements.

Dans le cadre du projet pluridisciplinaire que nous sommes amené.e.s à faire durant notre deuxième année du cycle supérieure au sein de l'Ecole Supérieure en Informatique 8 mai 1945, nous tenons à présenter nos remerciements, d'abord, à nos deux cher.e.s encadreurs : Monsieur Sidi Mohamed BENSLIMANE, et Madame Rim Sakina BENABI, pour le suivi, l'intérêt et l'assistance permanente qu'ils nous ont porté.e.s durant ces quelques mois de labeur.

Nous aimerions ensuite remercier les communautés internationales de développeurs et de passionnées, notamment : Django Community et Pytorch Community, la plateforme de MOOC Udacity, pour la formation PyTorch, ainsi que l'entreprise informatique JetBrains.

Merci.

Introduction :

Chez la femme, le cancer du sein reste le cancer le plus fréquent et le plus meurtrier. Chaque année, il y a près de 1,38 million de nouveaux cas et 458 000 décès dus au cancer du sein (IARC Globan, 2008). Ce type de cancer est très répandu aussi bien dans les pays développés que dans ceux en développement. En Algérie, 12 000 nouveaux cas sont enregistrés chaque année.¹

Dans le flou qui entoure encore les causes de ce cancer, le dépistage et le diagnostic précoce restent les principaux moyens de lutter contre ce mal. Un cancer du sein dépisté à un stade précoce peut être maîtrisé et soigné. À l'inverse, s'il est dépisté tardivement, il y a de faibles chances que le traitement curatif soit vraiment efficace.

La mammographie est le principal examen radiographique pour le dépistage et la détection d'éventuelles anomalies au niveau des tissus. Si l'intérêt de cet examen est indéniable, les interprétations et les suites à donner à ses résultats font encore débat. En effet, la question se pose sur le premier commentaire à faire, sur la nécessité de recourir à une biopsie ou encore sur l'intervalle approprié à fixer entre deux mammographies successives.

L'informatique au service de la médecine :

Même si le domaine de santé reste très vaste et très complexe, l'informatique médicale constitue aujourd'hui une science à part entière, puisque l'apport de l'informatique à la pratique médicale s'est graduellement solidifié avec l'introduction de l'outil informatique dans les établissements de santé, mais pas que. Cet apport peut être situé à deux niveaux : sur le plan de la pratique médicale, offrant au médecin de l'aide ou de l'assistance ; et sur le plan de la recherche en sciences médicales (laboratoires pharmaceutiques, centres de recherches en sciences médicales, etc.).

Parmi les cas d'utilisations les plus perceptibles de l'informatique médicale, nous pouvons citer : le dossier médical des patients ; l'optimisation économique et médicale des actes et des prescriptions (calcul de coût de l'ordonnance, proposition de génériques...) ; l'aide au diagnostic ; l'acquisition de données quantifiables ; messagerie électronique...etc.

Outil d'aide au diagnostic du cancer du sein :

« 98% de la santé, aujourd'hui, c'est du curatif. L'intelligence artificielle permettra de basculer sur une médecine plus préventive. » Laurent Schlosser, Microsoft.²

¹ Ahcene Tahraoui, « Lutte contre le Cancer du sein en Algérie et diagnostic précoce : Renforcer la formation pour le dépistage » in El Watan

² <https://experiences.microsoft.fr/business/intelligence-artificielle-ia-business/intelligence-artificielle-medecine/>

Les dernières prouesses de l'intelligence artificielle dans le domaine de l'imagerie en général et de l'imagerie médicale en particulier peuvent nous aider pour apporter des réponses à ces différentes questions et aider les spécialistes à poser leurs diagnostics.

L'apprentissage profond, notamment, permet de construire sur la base d'une grande base de données d'exemples des modèles prédictives très efficaces. Dans ce genre d'exercice trois problèmes sont posés : avoir une grande quantité de données utiles, trouver la meilleure architecture pour le modèle et disposer d'une capacité de calcul suffisante pour entraîner le modèle sur les exemples disponibles.

Objectifs du projet.

Nous avons essayé dans le cadre de notre projet de répondre à deux de ces problèmes. Nous avons d'un côté, construit différents modèles prédictives (mammographies et données numériques) et nous les avons entraînés sur des données disponibles en libre accès. D'un autre côté, nous avons développé une plateforme d'acquisition de données pour construire un jeu de données plus large.

Cette plateforme sert à héberger une grande quantité de mammographies tout en permettant à différents spécialistes de les labelliser. À la fin, pour chaque mammographie, toutes les labellisations associées sont croisées pour avoir une seule « labellisation collégiale ».

Cet outil d'aide au diagnostic du cancer du sein qui aura pour simple objectif :

- Offrir aux spécialistes un outil informatique d'aide au diagnostic du cancer du sein ;
- Encourager la prévention contre ce cancer, et dans les cas nécessaires, permettre un diagnostic positif fiable ;
- Offrir un diagnostic rapide ;
- Permettre l'enrichissement de cet outil pour une fiabilité en perpétuelle augmentation ;
- Réalisation d'une plateforme participative/collaborative pour la construction et l'ajout de nouveaux data-set.

1. Analyse des besoins et recensement des données.

1.1. Besoins et contraintes du projet.

1.2.1 *Besoins fonctionnels.*

Le médecin.

- S'authentifier (s'inscrire ou se connecter) à la plateforme d'acquisition.
- Visualiser les différentes images de mammographies ainsi que leurs labellisations.
- Ajouter des cas (en forme de mammographies) ainsi que leurs labellisations, pour agrandir le data-set.
- Labelliser des cas déjà existants.
- Contacter un l'administrateur.
- Contacter un médecin.
- Se connecter à l'interface principale d'aide au diagnostic.
- Entrer (uploader) des images de mammographies pour obtenir en sortie la classification faite par le modèle d'apprentissage (le résultat).

L'administrateur.

- S'authentifier (se connecter).
- Vérifier l'identité d'un médecin lors de son inscription (la valider) pour que celui-ci puisse apporter des modifications au data-set.
- Gestion des data-set, des labellisations et des droits d'accès des comptes.

Le visiteur.

- Le visiteur n'a accès qu'à la plateforme d'acquisition.
- Accéder à la plateforme d'acquisition pour visualiser les mammographies existantes ainsi que leurs labellisations.
- Exporter le data-set.
- Contacter l'administrateur.

1.2.2. *Besoins non fonctionnels.*

Il s'agit des besoins qui caractérisent le système. Ce sont des besoins en matière de performance, de type de matériel ou le type de conception.

L'application Web doit être réalisée avec les derniers standards web (HTML5, CSS3 et JavaScript) pour le *front-end*, le *framework* Python Django pour le *back-end*, et MySQL pour la gestion de la base de données. Elle doit aussi être compatible avec n'importe quel navigateur Web, accepter l'affichage mobile.

Hormis le data-set qui peut être téléchargeable, l'application doit être hautement sécurisée, car les informations ne devront pas être accessibles à tout le monde.

1.2.3. Contraintes.

Budget alloué au projet :

| dépenses | prix |
|---------------------------------------|--------------|
| <i>Serveur d'entraînement (Cloud)</i> | 14.99 €/mois |
| <i>Hébergement plateforme</i> | 7.99 €/mois |
| <i>Hébergement L'API-REST</i> | 3.99 €/mois |
| <i>Totale</i> | 26.99 €/mois |

1.2. Diagrammes des cas d'utilisation.

1.2.1. Diagramme des cas d'utilisation de la plateforme d'acquisition.



- Figure 01 : Diagramme des cas d'utilisation de la plateforme d'acquisition. –

Cas d'utilisation 01 : S'inscrire

Acteurs primaire : médecin
Acteurs secondaire : Administrateur

Entrées :

Nom, Prénom, spécialité, établissement, password , nom d'utilisateur

Sorties :

compte utilisateur

Conditions :

nom d'utilisateur unique, password différent du nom

Post conditions :

validation de l'administrateur

Description :

Le médecin doit s'inscrire avant de commencer l'utilisation de la plateforme, il doit fournir son nom, son prénom, son user-name, son mot de passe, sa spécialité, le nom de l'établissement dans lequel il exerce. Une vérification d'information est lancée avant la validation du compte. Les mots de passe sont hachés.

Cas d'utilisation 02 : Gérer les labellisations

Acteurs : médecin

Entrées :

background tissue, Class of abnormality present, Severity of abnormality, position d'un cercle, un id de mammographie, commentaire.

Sorties :

une labellisation ligne de data-set.

Conditions :

tous les champs sont obligatoires sauf la position du cercle et le commentaire

Description :

Un médecin peut ajouter une labialisation pour une mammographie précise, modifier une labellisation déjà existante, ou, éventuellement, la supprimer. Toutes ces opérations demandent une authentification à priori.

Cas d'utilisation 03 : Gérer les mammographies

Acteurs : médecin

Entrées :

Une image mammographie.

Sorties :

Une mammographie sans labellisation.

Conditions :

Image 1024*1024

Description :

Un médecin peut ajouter une mammographie, modifier une mammographie déjà existante, ou, éventuellement, la supprimer. Cette opération est étendue par le cas d'utilisation 02, Toutes ces opérations demandent une authentification à priori.

Cas d'utilisation 04 : Gérer le data set

Acteurs : administrateur

Entrées :

Un nouveau data-set.

Sorties :

Data-set plus adapté.

Conditions :

L'image claire labellisation conforme.

Post conditions :

Exposer aux médecins pour les labelliser

Description :

Un administrateur peut : charger un nouveau data-set afin qu'il soit labellisé par d'autres médecins, archiver le data set existant et/ou l'exporter. Ces opérations demandent une authentification à priori.

Cas d'utilisation 05 : Gérer les comptes utilisateurs

Acteurs : administrateur

Entrées :

Modification sur un compte (Créer, modifier ou supprimer).

Sorties :

Un compte utilisateur.

Conditions :

Champs obligatoires.

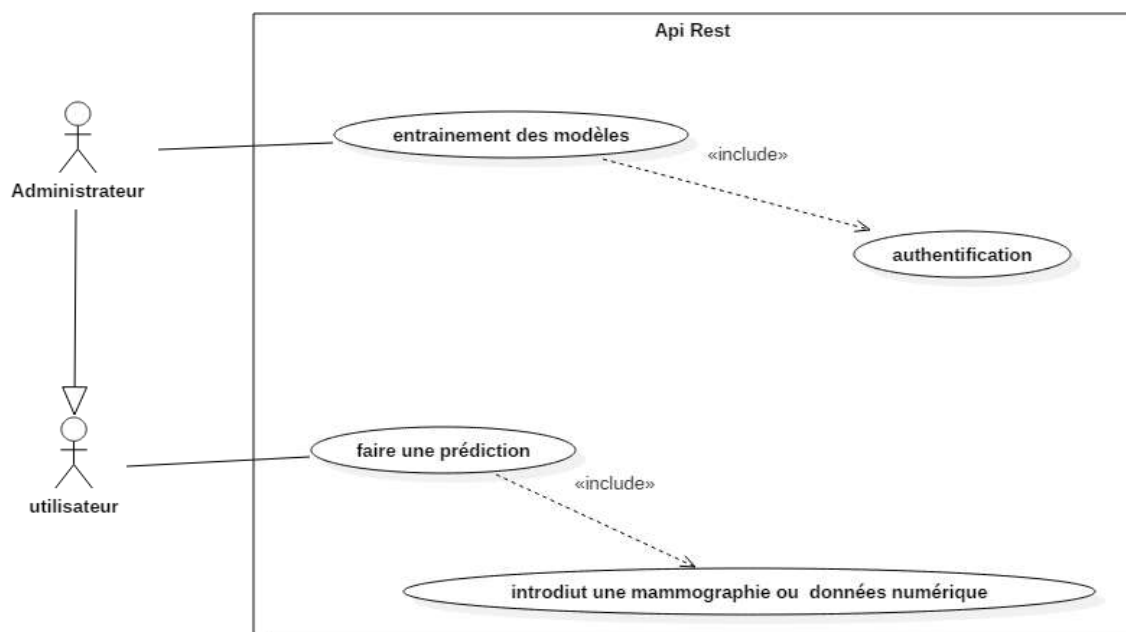
Post-conditions :

Validation et confirmation.

Description :

Un administrateur peut : ajouter de nouveaux utilisateurs, supprimer ou modifier des comptes déjà existants, ainsi qu'attribuer ou enlever des droit d'accès aux utilisateurs. Toutes ces opérations demandent une authentification à priori.

1.2.2. Diagramme des cas d'utilisation de l'API-REST.



- Figure 02 : Diagramme des cas d'utilisation de l'API-REST. –

Cas d'utilisation 01 : Entrainement du modèle

Acteurs : administrateur

Entrées :

Un data-set.

Sorties :

Modèle entraîné.

Conditions :

Data-set conforme.

Post-conditions :

Une précision meilleure.

Description :

Un administrateur peut entraîner le modèle et l'injecter au niveau de l'Api pour améliorer la prédiction. Ces opérations demandent une authentification à priori.

Cas d'utilisation 02 : faire une prédiction

Acteurs : utilisateur

Entrées :

Une image ou des données numériques.

Sorties :

La prédiction.

Préconditions :

La taille de l'image 1024*1024, les champs des données numérique obligatoires.

Post-conditions :

Validation d'un médecin spécialiste

Description :

N'importe quel utilisateur peut faire une prédiction à travers deux méthodes :

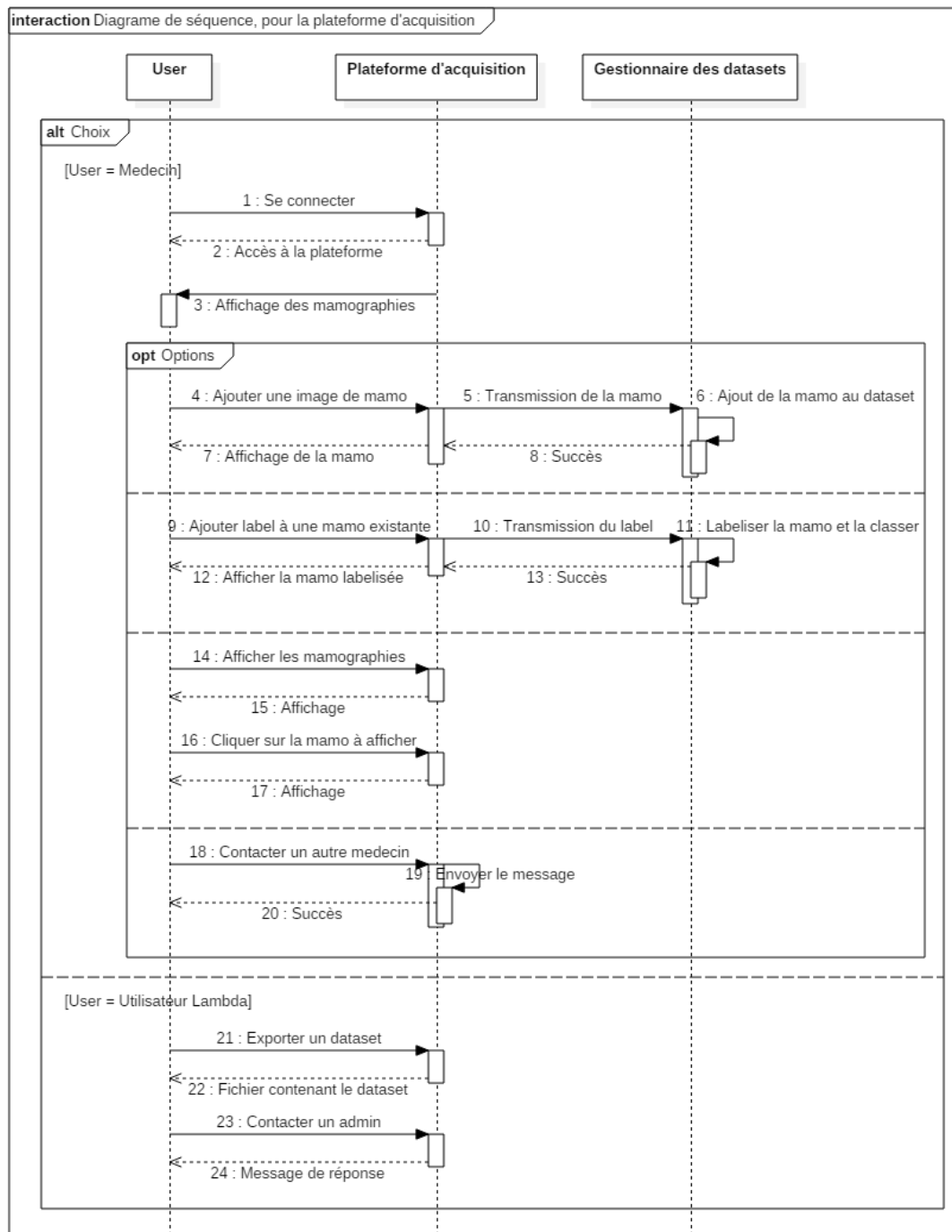
1/- par des données numériques ;

2/- par une image de mammographie.

Cette opération ne requiert pas d'authentification.

1.3. Diagrammes de séquence.

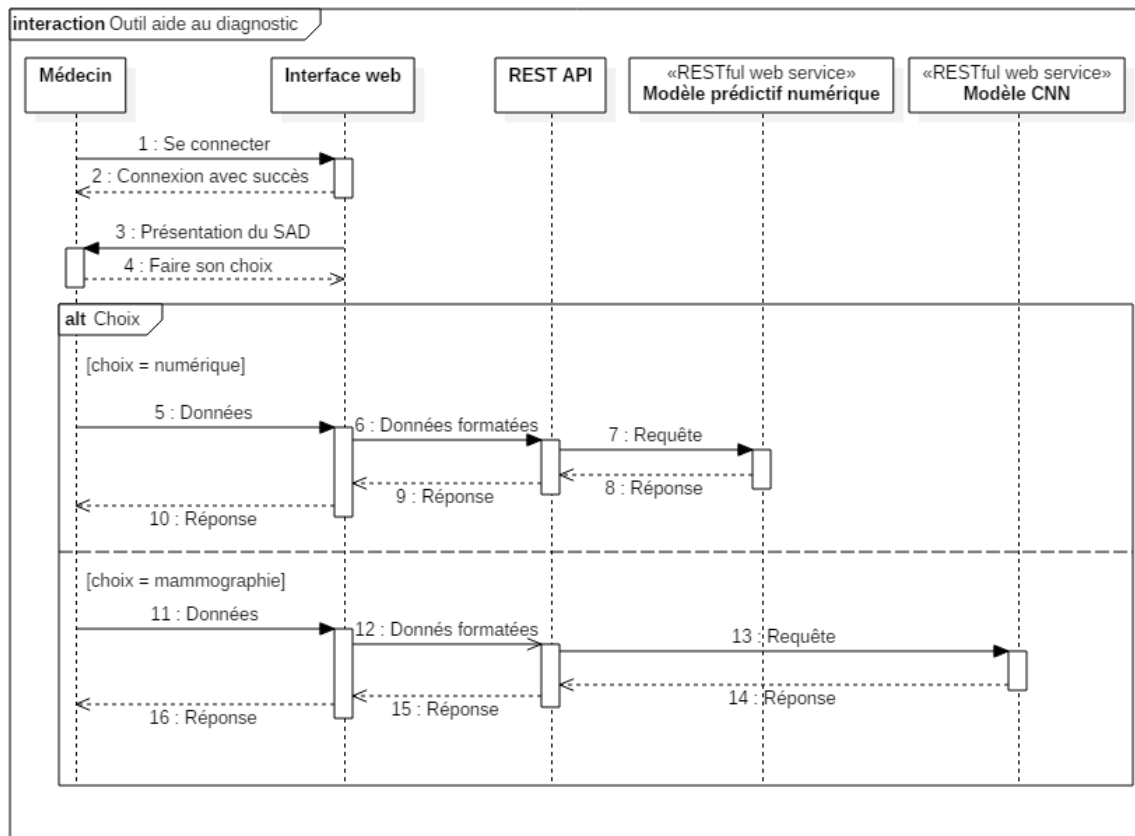
1.3.1. Diagramme de séquence de la plateforme d'acquisition.



- Figure 03 : Diagramme de séquence de la plateforme d'acquisition. -

Après s'être authentifié, un médecin peut ajouter des mammographies. Il peut également les labelliser, ou labelliser ceux d'autres médecins, comme montré dans la figure 03.

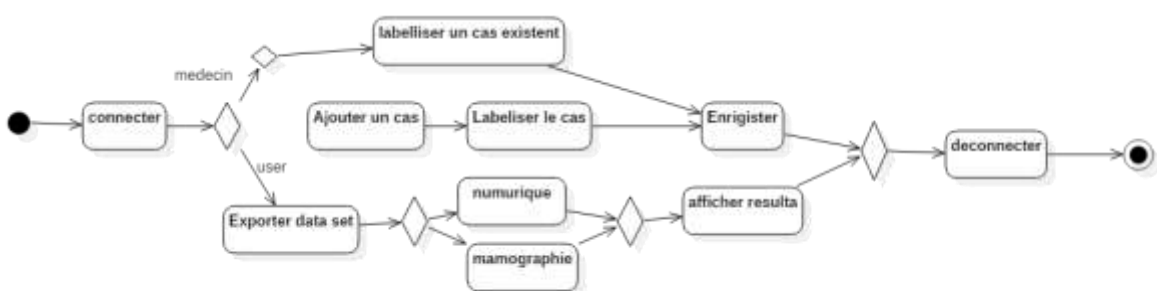
1.3.2. Diagramme de séquence de l'API-REST.



- Figure 04 : Diagramme de séquence de l'API REST. -

Un utilisateur peut interroger l'API à travers une interface qui offre la possibilité d'introduire des données numériques ou des photos de mammographies. Les données sont envoyées vers l'API en format JSON, comme montré dans la figure 04.

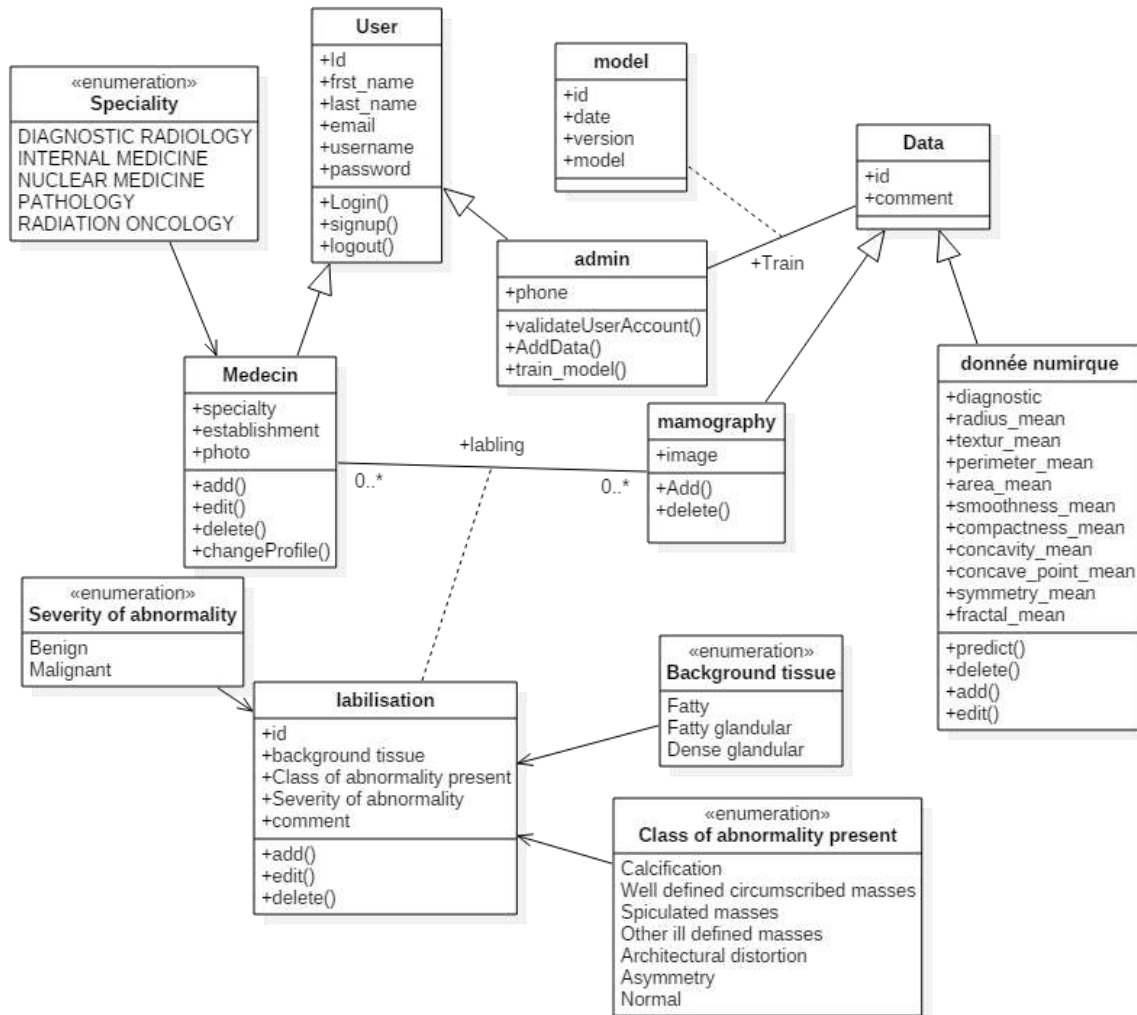
1.4. Diagrammes d'activité.



- Figure 05 : Diagramme d'activité de la plateforme d'acquisition. -

La figure 05 traduit le déroulement des tâches/activités initiées par l'utilisateur.

2. Conception de l'application.



- Figure 06 : Diagramme de classe de l'application. -

3. Prédiction et diagnostic du cancer du sein.

3.1. Prédiction sur des *features* extraites de microphotographies d'échantillons par ponction à l'aiguille fine.

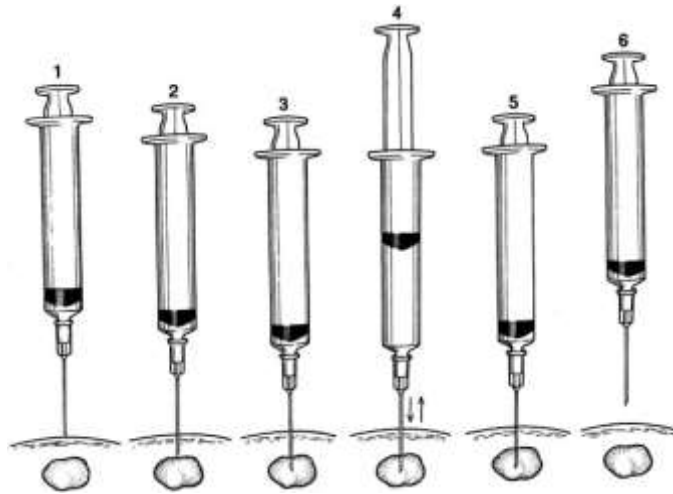
3.1.1. La biopsie à l'aiguille fine.

« La biopsie à l'aiguille fine (BAF) ou cytoponction (figure 07), est une méthode de diagnostic consistant à prélever des cellules et du tissu dans des nodules ou du liquide dans des kystes et des ganglions lymphatiques à l'aide d'une aiguille très fine (figure 08) montée sur une seringue pour effectuer ensuite un examen cytologique du prélèvement au microscope, comme montré dans la figure 09. C'est une procédure chirurgicale de diagnostic sûre et non invasive, qui évite souvent les biopsies chirurgicales majeures. Contrairement à la biopsie par forage, la biopsie à l'aiguille fine est généralement indolore et ne provoque que peu de complications (saignement, infection). Elle est le plus souvent employée pour la détection des cancers du sein, de la thyroïde ou des ganglions lymphatiques dans le cou, l'aîne ou l'aisselle. [...] La précision de cette procédure dans le diagnostic de nombreux cancers est supérieure à 90%. »³

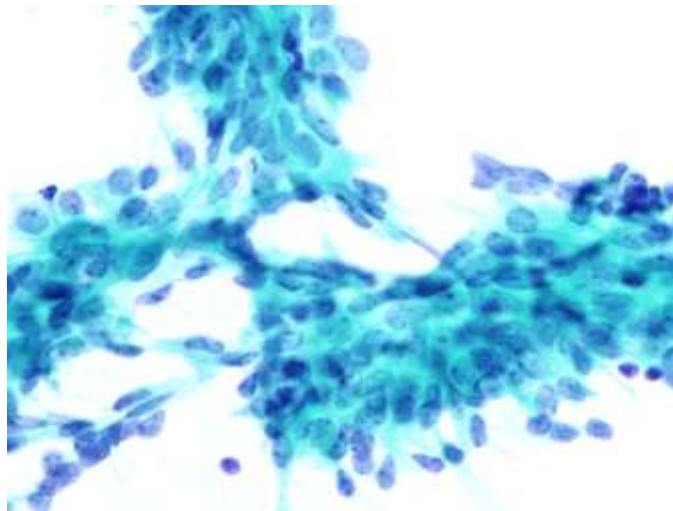


- Figure 07 : Médecin faisant un prélèvement à l'aiguille fine sur une patiente -

³ https://en.wikipedia.org/wiki/Fine-needle_aspiration



- Figure 08 : Biopsie à l'aiguille fine –



- Figure 09 : Microphotographie d'un échantillon par ponction à l'aiguille fine –

3.1.2. Le jeu de données

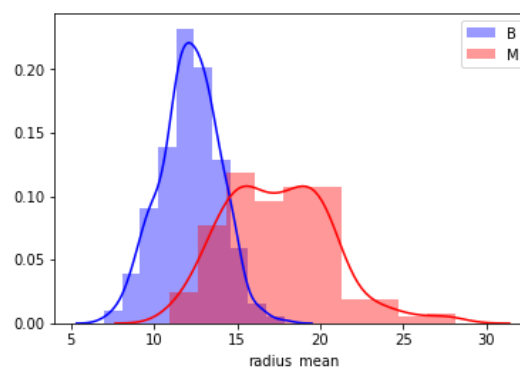
Nous avons travaillé dans cette partie sur un data-set très célèbre : le *Breast Cancer Wisconsin (Diagnostic) Data Set*. Comme décrit par les auteurs, les features sont extraites à partir d'images numérisées de biopsie à l'aiguille fine (FNA) sur une masse mammaire. Ces features décrivent les caractéristiques du noyau cellulaire visible sur l'image.

Le jeu de données est composé de 569 lignes et de 32 attributs. Le label est binaire : bénin ou main. Nous nous sommes intéressées dans notre étude uniquement à 10 attributs qui correspondent à la moyenne de la caractéristique mesurée. Ces attributs sont listés ci-dessous.

- Radius (mean of distances from center to points on the perimeter);

- Texture (standard deviation of gray-scale values);
- Perimeter ;
- Area ;
- Smoothness (local variation in radius lengths);
- Compactness ($\text{perimeter}^2 / \text{area} - 1.0$) ;
- Concavity (severity of concave portions of the contour);
- Concave points (number of concave portions of the contour);
- Symmetry ;
- Fractal dimension ("coastline approximation" - 1).

3.1.3. Analyse des données.



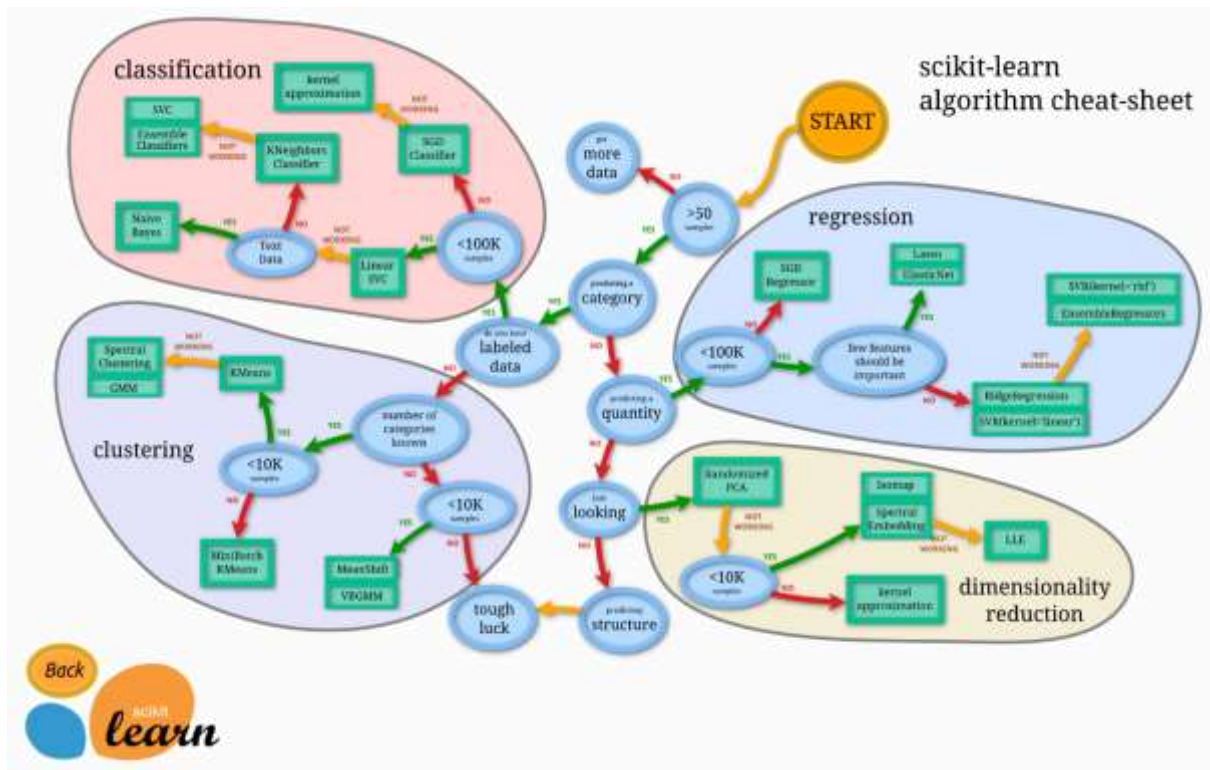
- Figure 10 : la taille du rayon pour chaque cas -

Une brève analyse du jeu de données nous a permis d'appréhender les données, de voir quelques liens entre les attributs et de poser un certain nombre d'hypothèses mais il est impossible d'aller plus loin dans l'interprétation sans une connaissance profonde du domaine ou l'intervention d'un spécialiste. La figure 10 par exemple indique qu'un rayon plus petit serait annonciateur d'un cas malin mais ne nous pouvons en tirer aucune conclusion formelle.

3.1.4. Construction du modèle.

Nous avons fait le choix de travailler avec la bibliothèque Scikit-learn pour construire nos modèles et pour les entraîner sur le jeu de données.

Un aperçu rapide de la figure ci-dessous, la figure 11 (fournie dans la documentation scikit-learn) permet de sélectionner les modèles les plus adaptés à nos données et à nos besoins.



- Figure 11 : carte des algorithmes Scikit-learn -

Nous avons testé cinq modèles basés sur cinq algorithmes différents : LinearSVC, KNeighborsClassifier, SVC, RandomForestClassifier, ExtraTreesClassifier. Nous avons partagé le jeu de données en trois (entraînement, validation et test) et nous avons, pour chaque modèle, effectué un entraînement, une validation et enfin un test.

| Modèle | Précision |
|-----------------------------|---------------|
| Extra Trees | 97.94% |
| Random Forest | 94.74% |
| KNeighborsClassifier | 89.47% |
| NN | 87.51% |
| SVC | 76.32% |
| LinearSVC | 73.68% |

Nous avons gardé les trois meilleurs modèles. Dans la fonction de prédiction, une prédiction (bénigne ou maligne) est faite par chacun des trois modèles, puis la prédiction la plus redondante est faite.

3.2. Prédiction sur des mammographies

3.2.1. La mammographie.

La mammographie est un examen radiographique du sein. Elle a pour but de détecter d'éventuelles anomalies au niveau des tissus, notamment au niveau de la glande mammaire. Elle est prescrite, entre autres, dans le cadre du dépistage organisé du cancer du sein aux femmes à partir de 50 ans.⁴

La procédure de la mammographie est illustrée dans la figure 12.



- Figure 12 : Chacun des deux seins est placé dans le mammographe et est progressivement comprimé pour effectuer l'examen⁵ -

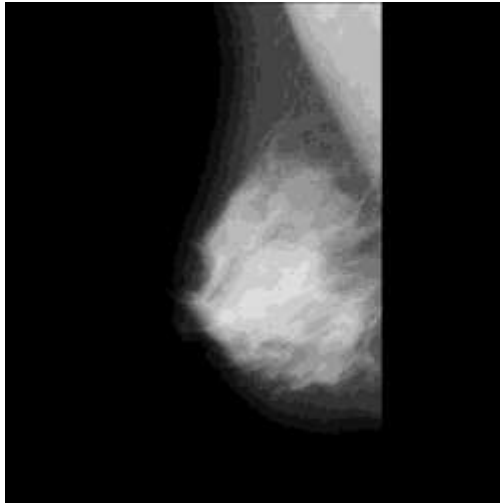
La mammographie étudie la glande mammaire et permet de dépister des anomalies des tissus, notamment un cancer du sein même à un stade précoce. Elle recherche des anomalies telles que des opacités, des micro-calcifications.⁶

Le résultat de la mammographie est illustré dans la figure 13.

⁴ <http://www.doctissimo.fr/html/sante/imagerie/mammographie.htm>

⁵ doctissimo

⁶ <http://www.doctissimo.fr/html/sante/imagerie/mammographie.htm>



- Figure 13 : Exemple d'une mammographie produite par un mammographe -

3.2.2. Le jeu de données

Nous avons travaillé sur un data-set (The mini-MIAS database of mammograms⁷) ouvert qui contient 322 mammographies labellisées. Les labellisations sont contenues dans le fichier "info.txt". Chaque mammographie est labellisée sur plusieurs colonnes. Nous nous sommes intéressés ici uniquement à la troisième et à la quatrième colonne.

La troisième colonne représente la classe de l'anomalie. Il y a sept classes en tout :

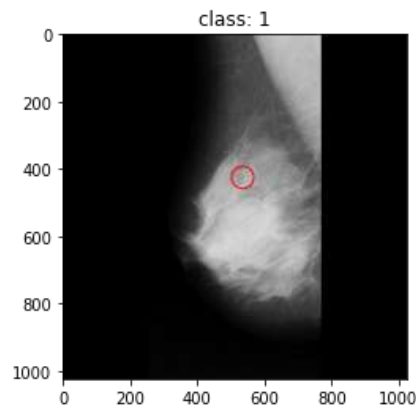
- **CALC** Calcification
- **CIRC** Well-defined/circumscribed masses
- **SPIC** Spiculated masses
- **MISC** Other, ill-defined masses
- **ARCH** Architectural distortion
- **ASYM** Asymmetry
- **NORM** Normal

La quatrième colonne quant à elle représente la sévérité de l'anomalie.

- **B** Benign
- **M** Malignant

⁷ <http://peipa.essex.ac.uk/info/mias.html>

En cas d'absence d'anomalie (NORM dans la troisième colonne), la question de la sévérité ne se pose naturellement pas. La quatrième colonne est donc vide dans ce cas.



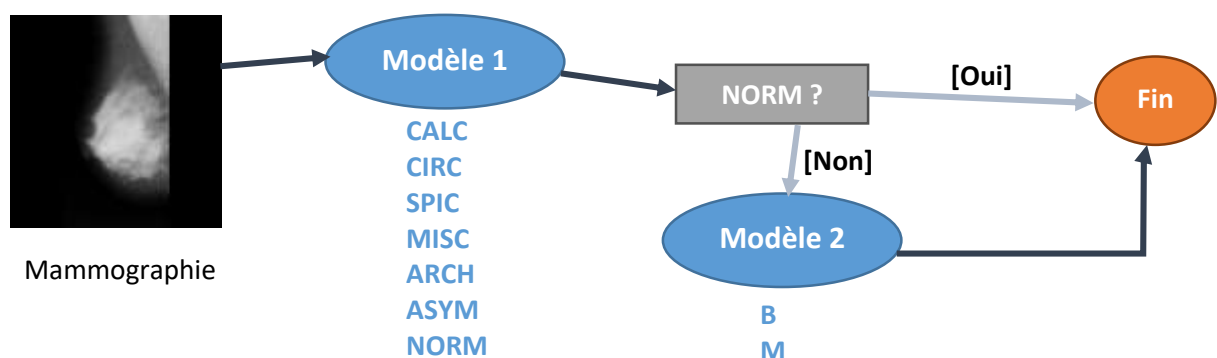
- Figure 14 : Mammographie extraite du jeu de données avec ses labellisations -

Les colonnes 5, 6 et 7 indiquent respectivement les deux coordonnées et le rayon d'un cercle qui entoure la zone suspecte. La figure ci-dessus, par exemple, montre un cercle en rouge autour de la zone, cette information est extraite des colonnes citées.

3.2.3. L'apprentissage profond (Deep Learning)

Le deep learning a montré sa grande efficacité dans de nombreux domaines, la vision par ordinateur notamment. Il est même le premier outil à faire mieux que l'homme dans certains problèmes de classification d'images. L'analyse automatique des images médicales tire aussi profit de cet outil et donne des résultats jamais égalés auparavant.

Nous utiliserons le deep learning pour construire des modèles capables de détecter la classe et la sévérité d'une anomalie mammaire à partir d'une mammographie. Nous ferons un diagnostic à deux niveaux : d'abord la classe parmi sept classes prédéfinies, ensuite la sévérité (bénin ou malin), comme montré dans la figure 15.



- Figure 15 : Schéma illustrant le déroulement de la prédiction au niveau des modèles. -

3.2.3.1. Architecture

Le choix d'une architecture est un problème difficile ; il n'existe aucune méthode pour arriver au meilleur modèle pour un problème donné. Il faut généralement faire un état de l'art, se documenter et s'inspirer de travaux similaires ou proches pour aiguïser son intuition. Ensuite, il faut partir sur une première architecture, faire des tests et effectuer des changements au fur et à mesure. Pour affiner le modèle, il existe de nombreuses techniques qui ont fait leur preuve. Il faut utiliser ces techniques au bon endroit et à la juste mesure.

| Entrées | Couches |
|---|--------------------------------------|
| (28, 28, 32) | FC 7 Softmax |
| | FC 1024 Relu |
| | FC 1024 Relu |
| (56, 56, 16) | Max pooling 2 x 2 |
| | Conv 3 x 3, 32 channels to 32 Relu |
| | Conv 3 x 3, 16 channels to 32 Relu |
| (112, 112, 8) | Max pooling 2 x 2 |
| | Conv 3 x 3, 16 channels to 16 Relu |
| | Conv 3 x 3, 8 channels to 16 Relu |
| (244, 244, 3) | Max pooling 2 x 2 |
| | Conv 3 x 3, 8 channels to 8 Relu |
| | Conv 3 x 3, 3 channels to 8 Relu |
| | Input 244 x 244 x 3 |
| FC : fully connected, Conv : convolution, Relu : Rectified Linear Units | |

- Figure 16 : Architecture du réseau de neurones -

Après plusieurs tests, nous sommes arrivés à l'architecture ci-dessus. Une architecture plus profonde amènerait le modèle à mieux connaître les données mais pourrait causer d'autres problèmes (comme l'évanescence du gradient). Ce modèle donne une précision de **54%** après un entraînement de 300 époques.

Il faut noter que les mammographies ont été redimensionnées de 1024*1024 à 244*244. Ceci fait perdre un certain nombre de détails mais réduit considérablement le temps de calcul.

La même architecture a été reprise pour le deuxième modèle (bénin / malin) en changeant la dernière couche (un FC avec deux sorties). La précision est de **91%**.

3.2.3.2. Optimiseur

Un optimiseur met à jour les poids pour minimiser la fonction objective (loss function). La fonction objective indique à l'optimiseur la direction à prendre pour minimiser l'erreur et atteindre le minimum global.

Les derniers optimiseurs proposés sont adaptifs, c'est-à-dire que le taux d'apprentissage est ajusté dynamiquement.

| | Adaptif | Erreur |
|----------|---------|-------------|
| Adam | Oui | $2e^{-2}$ |
| Adadelta | Oui | $2.5e^{-2}$ |
| RMS | Oui | $6e^{-2}$ |
| MOMENTUM | Non | 10.256 |
| ADAGRAD | Non | 15.584 |
| SGD | Non | 32.856 |

- Figure 17 : L'erreur de chaque optimiseur après 300 époques -

3.2.3.3. L'apport du Dropout

Le Dropout a d'abord été proposé comme une technique pour éviter le sur-apprentissage (overfitting) mais les vertus de ce dernier se sont avérées plus grandes. En effet, le Dropout est l'une des techniques de régularisation les plus puissantes.

| Taux du Dropout | Architecture 1 | |
|-----------------|-------------------------|-------------------|
| | Précision en validation | Précision en test |
| 0.0 | 41% | 39% |
| 0.1 | 40% | 40% |
| 0.2 | 38% | 42% |
| 0.3 | 38% | 49% |

| | | |
|-----|-----|-----|
| 0.4 | 38% | 49% |
| 0.5 | 36% | 52% |
| 0.6 | 35% | 53% |
| 0.7 | 35% | 47% |

- *Figure 18 : La précision de deux architectures avec différents taux de Dropout après 100 époques -*

Le Dropout améliore presque toujours l'entraînement mais quand le taux dépasse un certain seuil, son apport devient négatif.

3.2.3.4. L'apprentissage par transfert

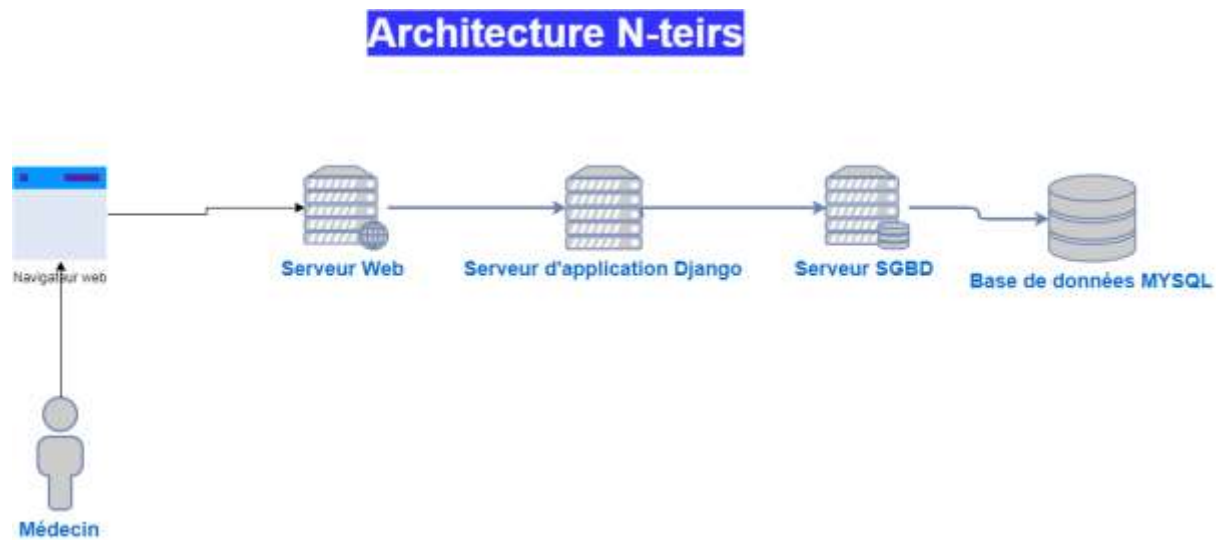
Il s'agit d'utiliser un modèle pré-entraîné en modifiant son classifieur et en l'entraînant sur nos données pour quelques époques supplémentaires. Cette technique donne d'excellents résultats et fait l'économie de beaucoup de temps de calcul. Puisqu'il s'agit d'entraîner le modèle une seule fois sur une grande quantité de données puis de le réutiliser pour résoudre des problèmes complètement différents en l'entraînant pour seulement quelques époques supplémentaires.

Pour notre part, nous avons utilisé un ResNet18 pré-entraîné sur ImageNet. Ce modèle donne pour nos deux modèles une précision de respectivement **63%** et **98%**.

4. Architecture et Organisation.

4.1. Architecture générale.

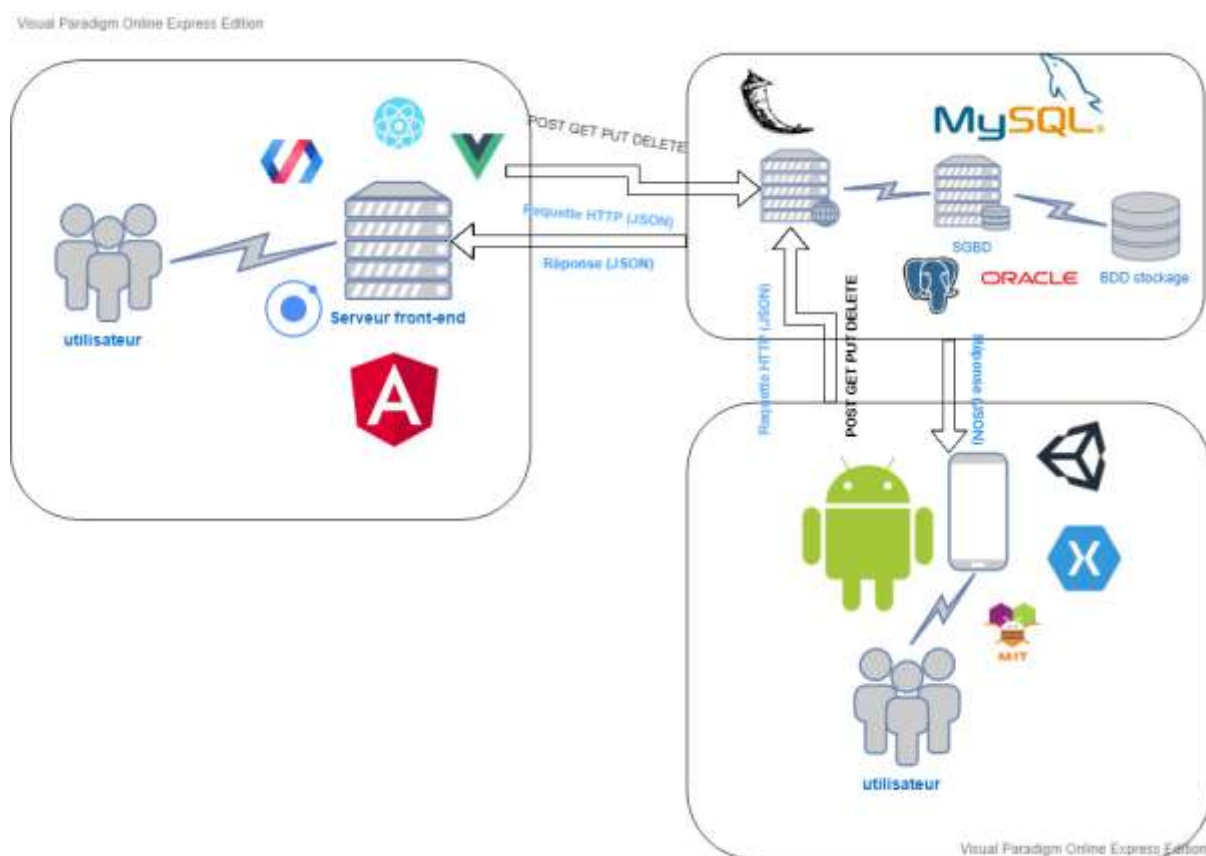
4.1.1. Architecture N-tiers.



- Figure 19 : L'architecture N-tiers de l'application. –

Le médecin est connecté à un serveur web, qui, à son tour, est connecté à un serveur d'application qui contient la plateforme pour ajouter de nouvelles données à la base de données, comme illustré dans la figure 19.

4.1.2. Architecture API-REST.



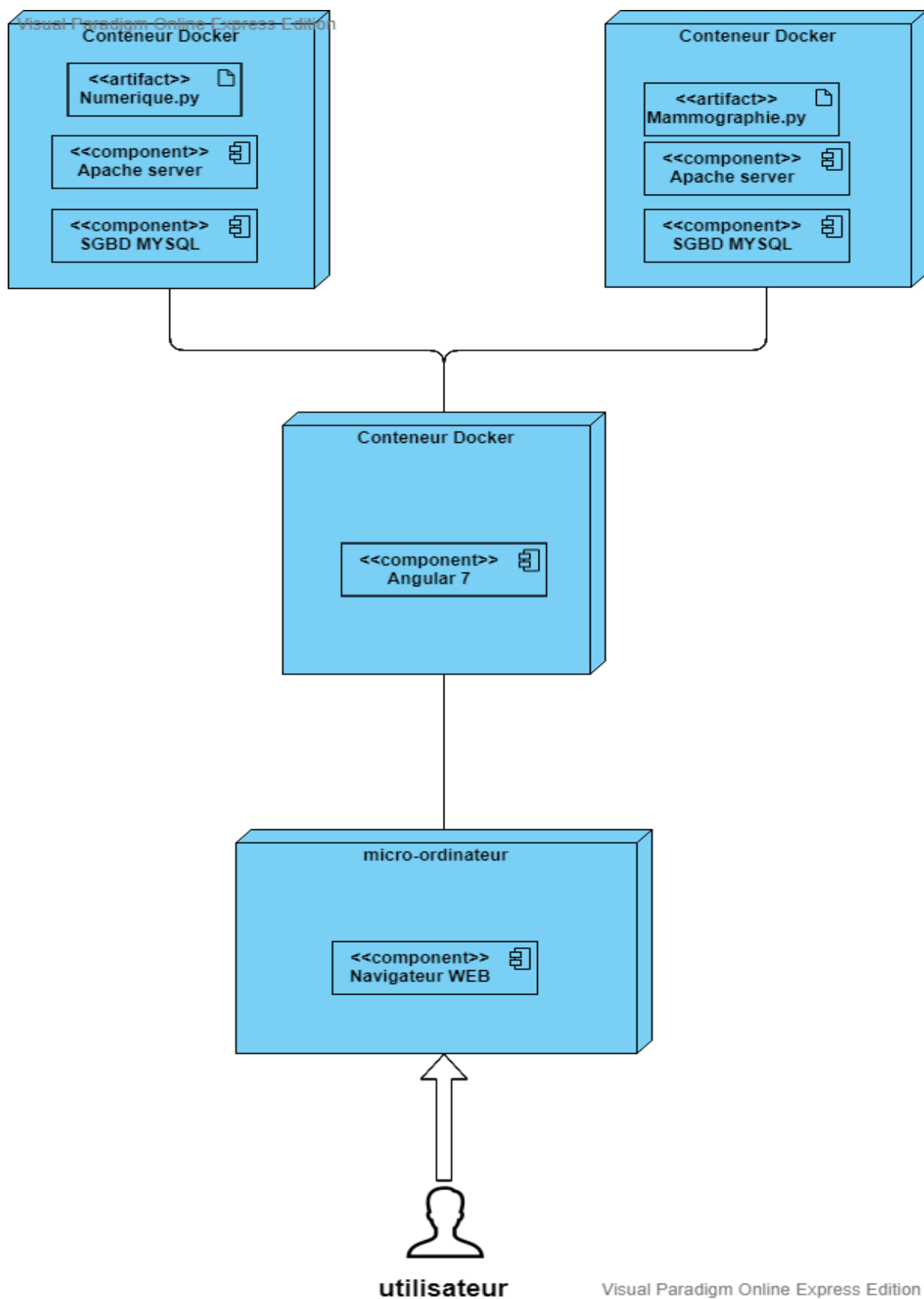
- Figure 20 : Architecture et fonctionnement de l'API-REST. –

Un utilisateur peut se connecter à travers une application web ou une application mobile à un serveur Django Rest via des requête http (POST, GET, PUT, DELETE), qui se chargera d'effectuer un traitement sur les données introduites par l'utilisateur, pour ensuite renvoyer une réponse (JSON) à l'utilisateur contenant le résultat, ainsi qu'enregistrer la nouvelle donnée dans la base de données. Cette architecture est illustrée dans la figure 20.

4.1.2.1 Division Micro Services :

On diviser notre solution (la partie visualisation) en deux micros services le premier pour la prédiction à base des mammographies et l'autre a base des données numérique c'est deux sont exposer via une API REST.

4.2. Diagramme de déploiement.



- Figure 21 : Diagramme de déploiement / Docker. –

Afin de faciliter la tâche de déploiement, pour les micros services nous avons utilisé des conteneurs docker qui comportent le langage python, le Framework Web Django, et les installations des librairies nécessaires au développement (Pillow, Django Rest Framework... etc.), et pour l'interface nous avons utilisé un conteneur qui comporte un serveur Angular 7.

5. Technologies utilisées.

5.1. Django.



Django est un cadre de développement web open source en Python. Il a pour but de rendre le développement web 2.0 simple et rapide. Pour cette raison, le projet a pour slogan « Le framework web pour les perfectionnistes sous pression ». Développé en 2003 pour le journal local de Lawrence (Kansas), Django a été publié

sous licence BSD à partir de juillet 2005.

Depuis juin 2008, la Django Software Foundation s'occupe du développement et de la promotion du cadre. En plus de cette promotion régulière, des conférences entre développeurs et utilisateurs de Django sont organisées deux fois par an depuis 2008. Nommées DjangoCon, une se déroule en Europe et l'autre aux États-Unis.⁸

Django a été utilisé pour développer la partie Web de notre application.

5.1. FLASK Framework.



Flask est un framework open-source de développement web en Python. Son but principal est d'être léger, afin de garder la souplesse de la programmation Python, associé à un système de templates. Il est distribué sous licence BSD.



Flask a été créé initialement par Armin Ronacher comme étant un poisson d'avril.

Le souhait de Ronacher était de réaliser un framework web contenu dans un seul fichier Python mais pouvant maintenir des applications très demandées.⁹

⁸ [https://fr.wikipedia.org/wiki/Django_\(framework\)](https://fr.wikipedia.org/wiki/Django_(framework))

⁹ [https://fr.wikipedia.org/wiki/Flask_\(framework\)](https://fr.wikipedia.org/wiki/Flask_(framework))

5.2. Bibliothèque PyTorch.



Deep Learning with PyTorch

PyTorch est une bibliothèque logicielle Python open source d'apprentissage machine qui s'appuie sur Torch développée par Facebook.

PyTorch permet d'effectuer les calculs tensoriels nécessaires notamment pour l'apprentissage profond (deep learning). Ces calculs sont optimisés et effectués soit par le processeur (CPU) soit, lorsque c'est possible, par un processeur graphique (GPU) supportant CUDA. Il est issu des équipes de recherche de Facebook, et avant cela de Ronan Collobert dans l'équipe de Samy Bengio à l'IDIAP.

PyTorch est dérivé d'un logiciel antérieur, Torch, qui s'utilisait avec le langage Lua. PyTorch est indépendant de Lua et se programme en Python.

PyTorch permet de:

- manipuler des tenseurs (tableaux multidimensionnels), de les échanger facilement avec Numpy et d'effectuer des calculs efficaces sur CPU ou GPU (par exemple, des produits de matrices ou des convolutions);
- calculer des gradients pour appliquer facilement des algorithmes d'optimisation par descente de gradient. PyTorch utilise la bibliothèque autograd.¹⁰

PyTorch, en association avec Python, a été utilisée pour concevoir les modèles d'entraînement et d'apprentissage (la partie machine learning), c'est-à-dire le cœur de l'application.

5.3. Scikit-learn.



Machine Learning with Scikit-Learn

Scikit-learn est une bibliothèque libre Python destinée à l'apprentissage automatique.

Elle est développée par de nombreux contributeurs notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche

comme Inria³ et Télécom ParisTech.

¹⁰ <https://fr.wikipedia.org/wiki/PyTorch>

Elle comprend notamment des fonctions pour estimer des forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. Elle est conçue pour s'harmoniser avec d'autres bibliothèques libres Python, notamment NumPy et SciPy.¹¹

5.4. Angular7.



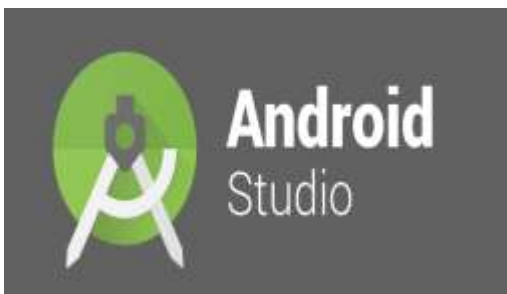
Angular7 est un framework Javascript côté client qui permet de réaliser des applications de type "Single Page Application". Il est basé sur le concept de l'architecture MVC (Model View Controller) qui permet de séparer les données, les vues et les différentes actions que l'on peut effectuer.

Depuis 2009 jusqu'à aujourd'hui, Google a sorti trois différentes versions d'Angular: AngularJS (2009), Angular 2 (2016), Angular 4 (2017) et Angular 7 (2018). A partir de la version 2, le framework a été complètement réécrit. AngularJS n'est donc pas compatible avec les versions ultérieures. En revanche, il est tout à fait possible de migrer entre les versions 2, 4, et 7.

Le code source d'Angular est écrit en TypeScript. Le TypeScript est une couche supérieure au Javascript développé par Microsoft qui se compile en JavaScript simple. Etant un langage typé, il permet de créer des classes, des variables, des signatures de fonction et l'utilisation de modules. Il est important de noter que l'utilisation du TypeScript est facultative, on peut tout à fait utiliser du JavaScript dans un fichier TypeScript.

Angular7 a été utilisé pour développer la partie front-end de l'API de l'application.

5.5. Android Studio.



Android Studio est un environnement de développement pour développer des applications mobiles Android. Il est basé sur IntelliJ IDEA et utilise le moteur de production Gradle. Il peut être téléchargé sous les systèmes d'exploitation Windows, macOS et Linux.

Android Studio a été utilisé pour développer l'application mobile, qui permet la visualisation des résultats fournis par les modèles d'apprentissage.¹²

¹¹ <https://fr.wikipedia.org/wiki/Scikit-learn>

¹² https://fr.wikipedia.org/wiki/Android_Studio



5.6. MySQL.

MySQL est un système de gestion de bases de données relationnelles (SGBDR) libre et open source. Il est distribué sous une double licence GPL et propriétaire. Il fait partie des logiciels de gestion de base de données les plus utilisés au monde, autant par le grand public (applications web principalement) que par des professionnels, en concurrence avec Oracle, PostgreSQL et Microsoft SQL Server.

SQL fait référence au Structured Query Language, le langage de requête utilisé. Il est davantage orienté vers le service de données déjà en place que vers celui de mises à jour fréquentes et fortement sécurisées. Il est multi-thread et multi-utilisateur.

Deux moteurs principaux sont présents dans MySQL, MyISAM et InnoDB. MyISAM, contrairement à InnoDB, ne supporte ni transactions ni intégrité automatique des tables, il n'est pas destiné aux applications dont la cohérence des données est critique ; cependant, ses performances le font adopter pour des applications ayant besoin d'une base de données simple et peu onéreuse à mettre en œuvre.¹³

¹³ <https://fr.wikipedia.org/wiki/MySQL>*

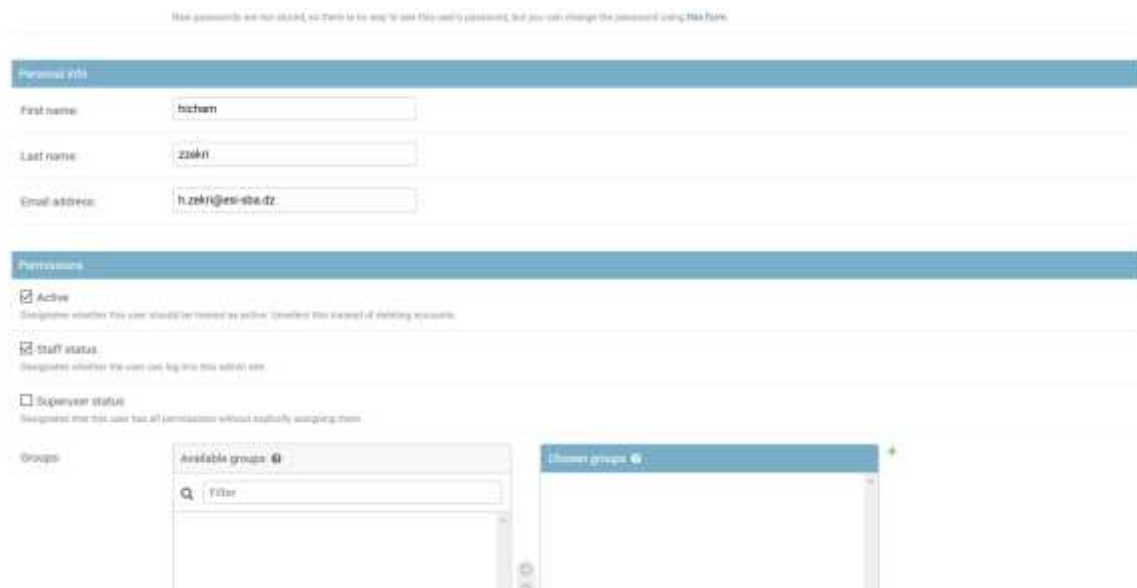
6. Ergonomie et Interface utilisateur.

6.1. Plateforme d'acquisition.

6.1.1. Administrateur.



- Figure 22 : Gestion de la Base de données –



- Figure 23 : Gestion des utilisateurs. –

| | version | date | duration | accuracy | Type | file |
|--------------------------|---------|--------------------------|-----------------|----------|-------------|---------------------------|
| <input type="checkbox"/> | 1 | July 18, 2019, 5:51 p.m. | 0:00:12 | 60.0% | Numérique | mod/model_knn.csv |
| <input type="checkbox"/> | 2 | July 20, 2019, 4:52 p.m. | 0:16:40 | 55.0% | Mammography | mod/data-equilibrated.csv |
| <input type="checkbox"/> | 3 | July 20, 2019, 4:52 p.m. | 1 day, 19:12:35 | 25.0% | Numérique | mod/data-no-norm.csv |

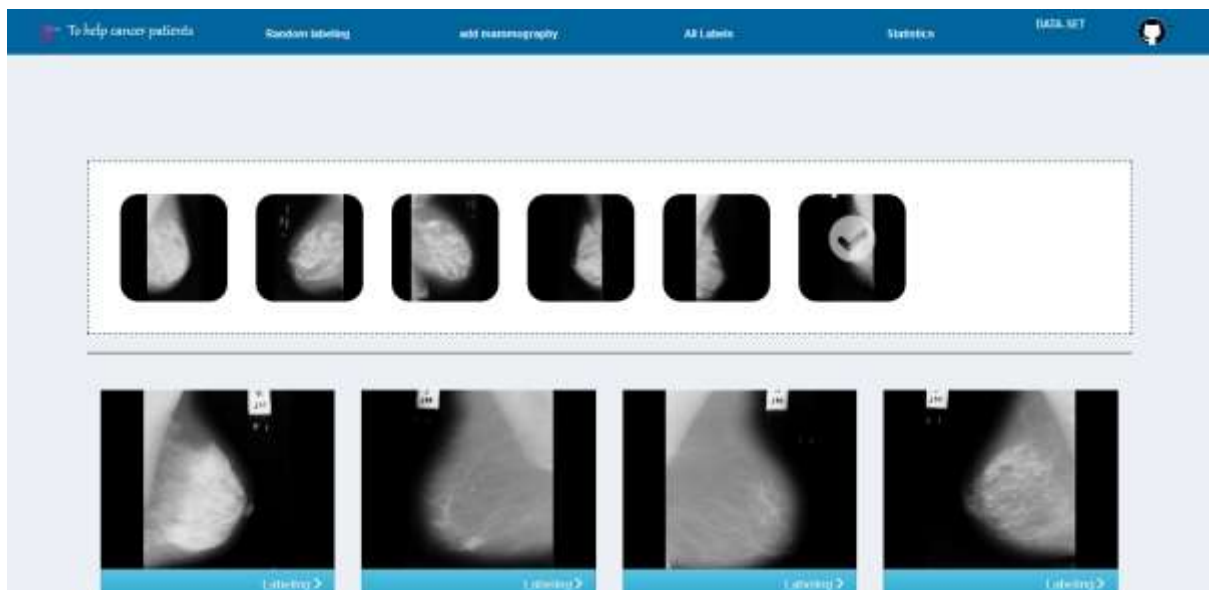
- Figure 24 : Affichage des versions des entraînements. –

6.1.2. Utilisateur / médecin.

The screenshot shows a 'Create account' form with the following fields and elements:

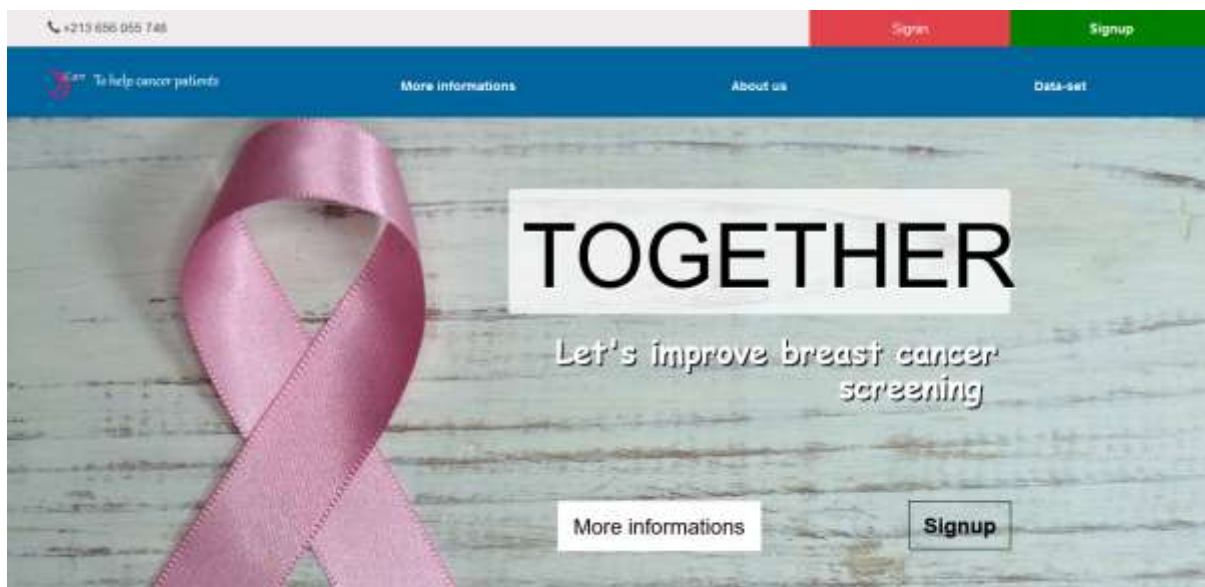
- Phone number: +213 656 055 746
- Buttons: Signin, Signup
- Form title: Create account
- Fields:
 - Votre Nom
 - Votre Prénom
 - Choisir un nom d'utilisateur
 - Required: 100 characters or more. Letters, digits and @/./+/_ only
 - Votre Email
 - Votre mot de passe
- Password requirements:
 - Your password can't be too similar to your other personal information.
 - Your password must contain at least 8 characters.
 - Your password can't be a commonly used password.
 - Your password can't be entirely numeric.
- Confirmation: Confirmez votre mot de passe
- Background image: A woman's face with the text 'cancer screening' and a 'Sign up' button.

- Figure 25 : Authentification. –



- Figure 26 : Ajouter une mammographie. –

- Figure 27 : Modifier le profil. –



- Figure 28 : Interface d'accueil. –

To help cancer patients
Radicon labeling
add mammography
24 Labels
Statistics
RADIA-BCI

Background tissue
Fatty

Abnormality present
Calcification

Severity of abnormality
Benign

Positions
207,115,76,576766876306

Commentaire

Save

Label
Fatty
Calcification
Benign
Commentaire

Next image >
Cancel ✕

Label
Fatty
Asymmetry
Benign
Commentaire

Next image >
Cancel ✕

Label
Fatty
Well-defined / circumscribed masses
R
Commentaire

Next image >
Cancel ✕

Label
Fatty
Normal
None
Commentaire

Next image >
Cancel ✕

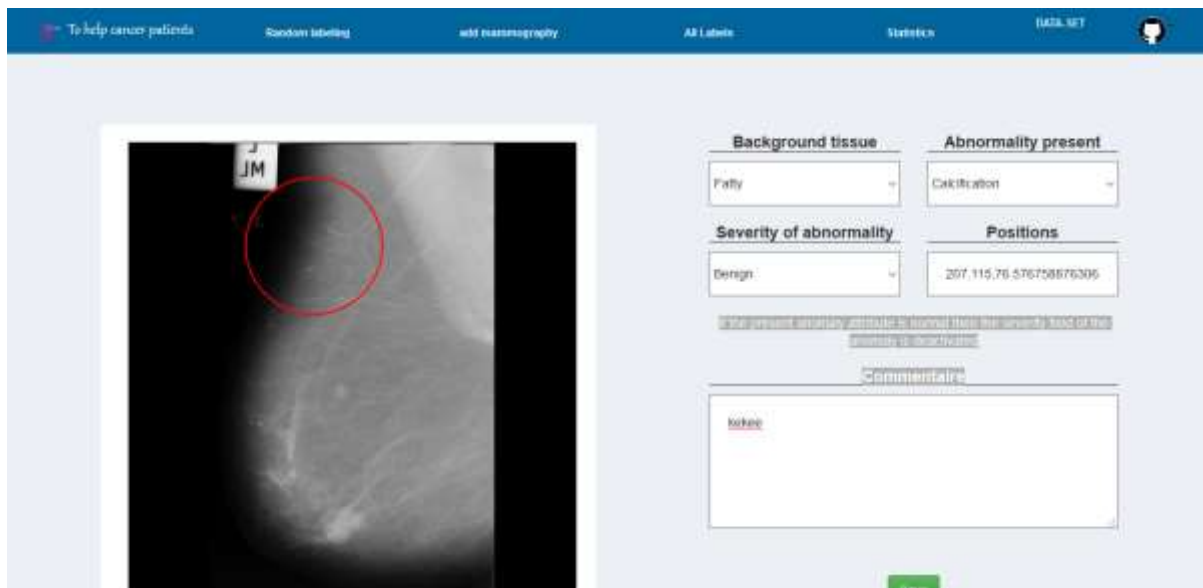
Label
Fatty
Spiculated masses
Benign
Commentaire

Next image >
Cancel ✕

Label
Fatty
Calcification
Malignant
Commentaire

Next image >
Cancel ✕

- Figure 29 : Labelliser une mammographie ajoutée par le médecin. –

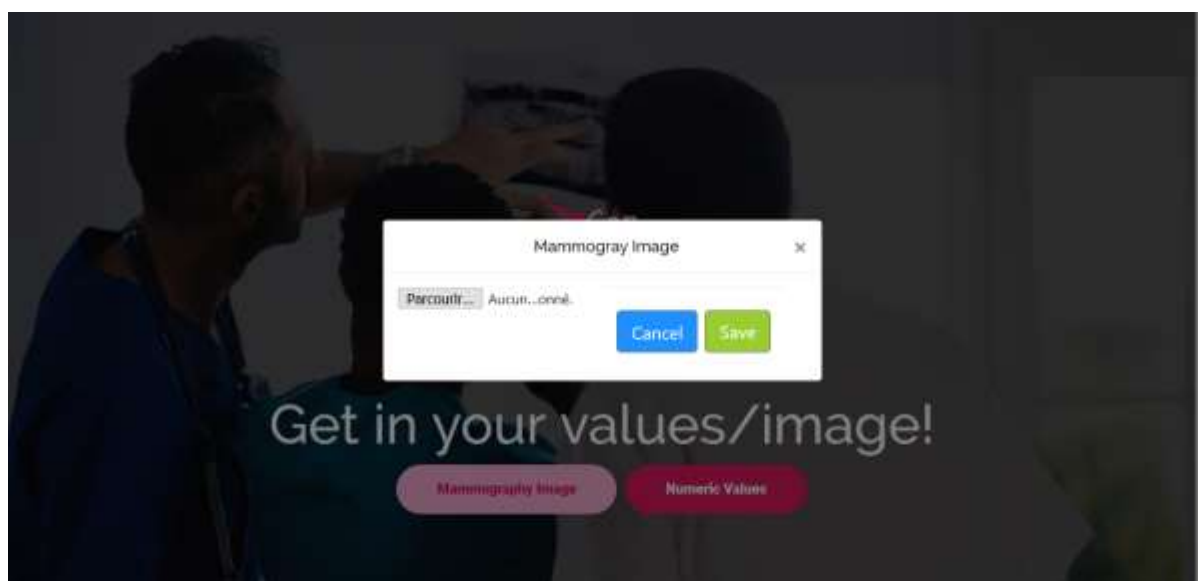


- Figure 30 : Afficher une mammographie non-labélisée aléatoirement, et la labelliser. –

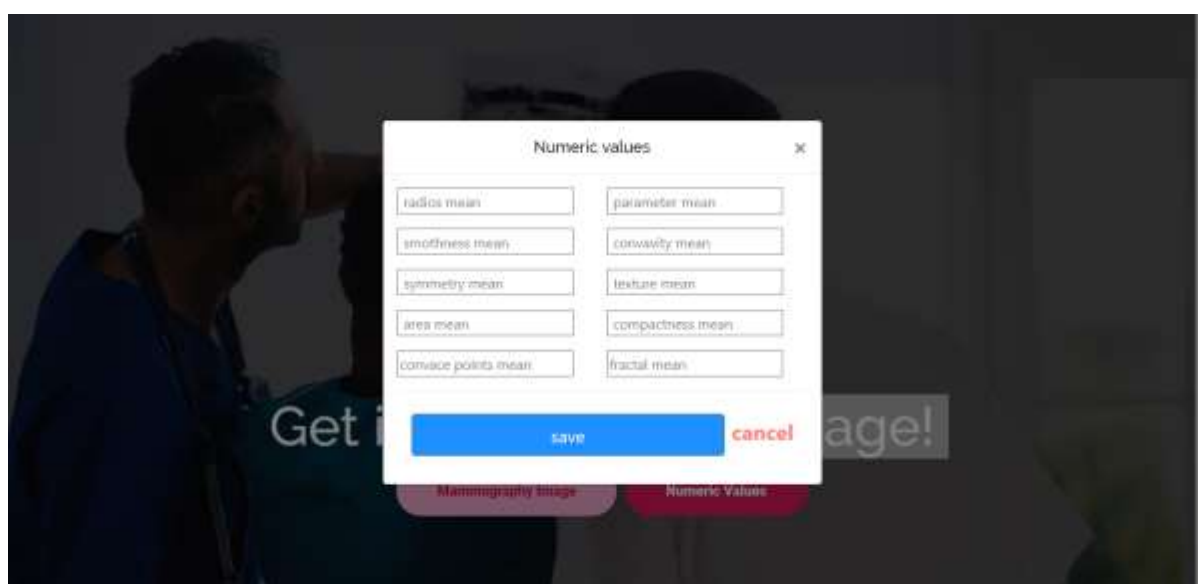


- Figure 31 : Afficher les statistiques. –

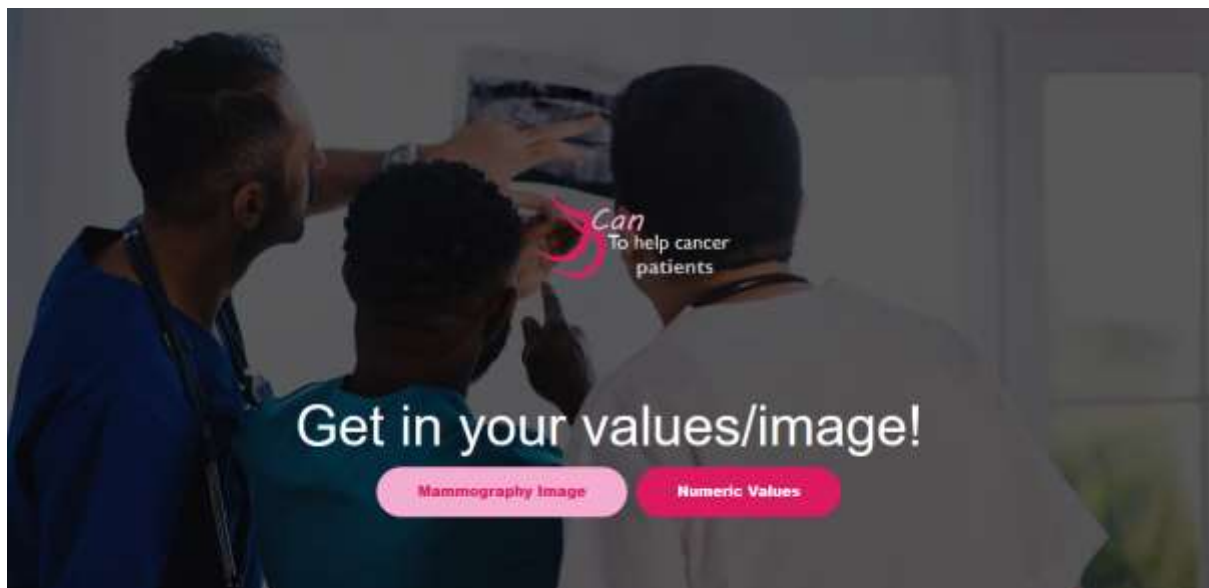
6.2. API Rest (utilisateur lambda).



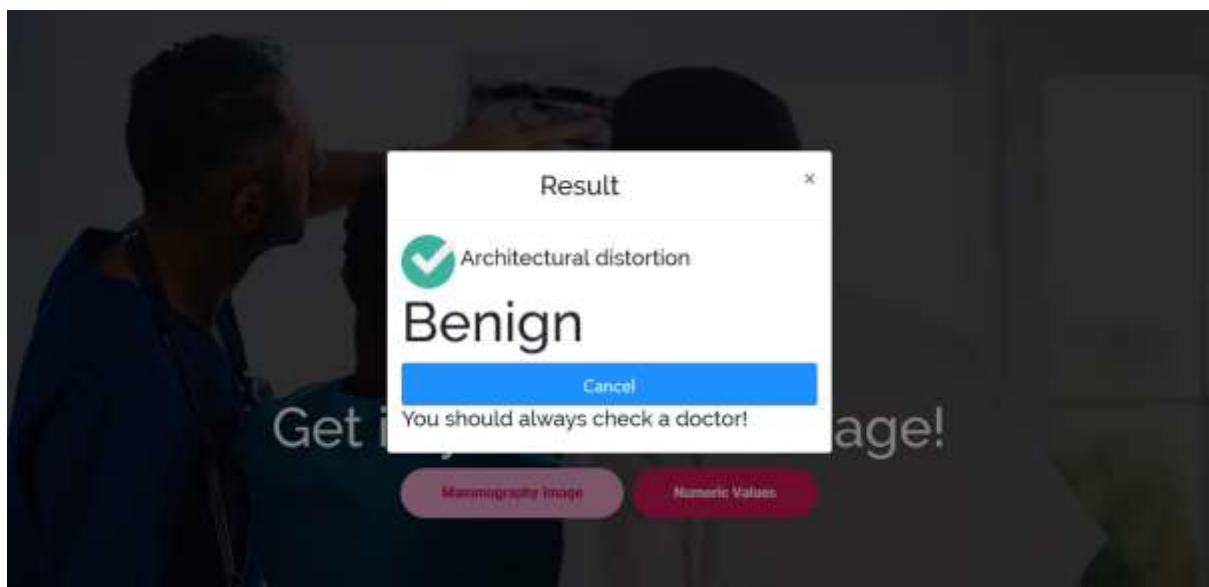
- Figure 32 : Uploader une image mammographique. -



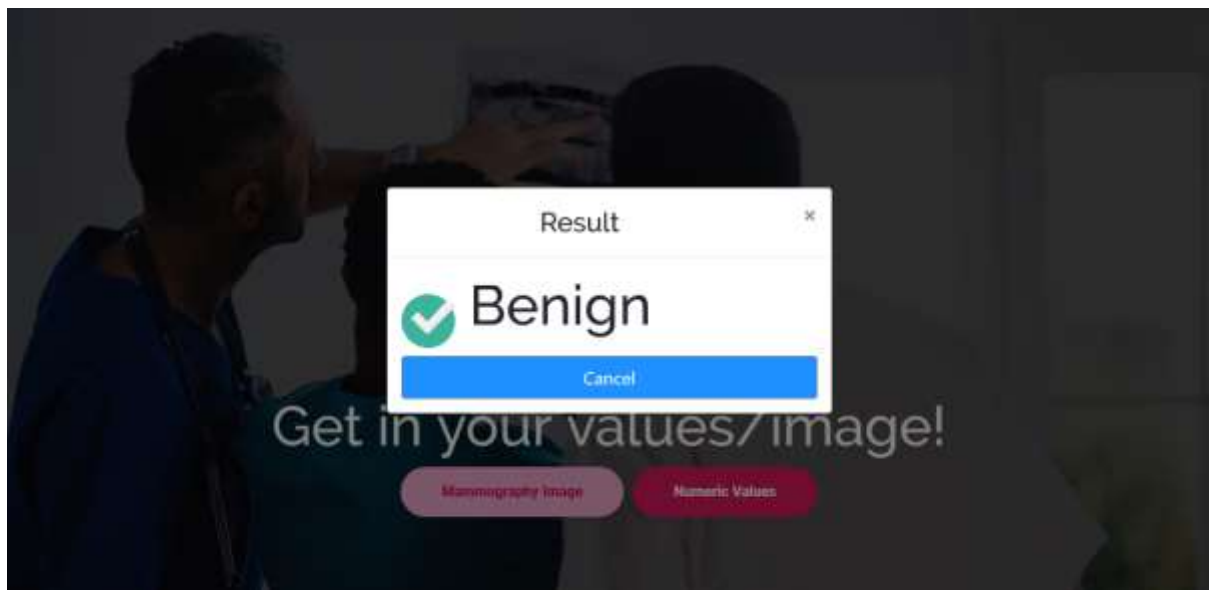
- Figure 33 : Introduire des données numérique.-



- Figure 34 : Interface utilisateur. –



- Figure 35 : Résultat de la prédiction dur la mammographie. –



- Figure 36 : Résultat de la prédiction sur les données numériques. -

7. Perspectives.

Nous aspirons que le développement et l'amélioration de cet outil ne s'en tient pas là, et qu'à l'avenir nous pourrions enrichir la plateforme avec de nouveaux sets de datas, améliorer nos modèles d'entraînement et de prédiction pour augmenter leur exactitude et leur précision (the Accuracy), ainsi que démocratiser un tel outil informatique pour participer, ne serait-ce que modestement, à la lutte contre le cancer du sein.

Nous aspirons également à introduire un prétraitement sur les images mammographiques, utiliser des mammographies colorées, ainsi que des images mammographiques à haute résolution. Ces aspirations n'en resteront évidemment pas unes, mais seront mises en pratiques au fur et à mesure que nous relevons de nouveaux défis.

8. Conclusion.

Ce projet s'est révélé très enrichissant dans la mesure où il a consisté en une approche concrète du métier d'ingénieur/développeur. En effet, la prise d'initiative, le respect des délais et le travail en équipe seront des aspects essentiels de notre futur métier.

De plus, il nous a permis d'appliquer nos connaissances en développement Web et en deep learning à un domaine pratique et très utile, à savoir le domaine médical, qui se révèle aujourd'hui d'intérêt général au vu de l'augmentation incessante de la mortalité du cancer du sein.

Les principaux problèmes que nous avons rencontrés concernaient la collecte de données, souvent indisponibles ou peu suffisantes. En effet, les data-sets qui nous servaient de base d'entraînement des modèles deep learning n'appartenaient souvent pas au domaine public ou n'étaient seulement pas disponibles sur le net. Ainsi, nous avons touché du doigt la difficulté d'extrapoler le jeu de données disponibles pour obtenir des valeurs cohérentes sur l'ensemble du domaine étudié, ce que pourtant nous serons vraisemblablement amenés à faire dans notre futur métier.

Les grèves qui ont eu lieu et le décalage de l'année scolaire qu'elles ont causé ont également été un challenge qui nous a permis de tester nos capacités à nous adapter aux imprévus et aux changements de dernières minutes. Revisiter les sprints et travailler sous pression a été une grande partie de notre quotidien ces derniers mois.

Toutefois, cela a constitué pour nous une expérience très enrichissante et un avant-gout assez représentatif de notre futur métier.

9. Références des figures.

| | |
|---|----|
| - FIGURE 01 : DIAGRAMME DES CAS D'UTILISATION DE LA PLATEFORME D'ACQUISITION. — | 8 |
| - FIGURE 02 : DIAGRAMME DES CAS D'UTILISATION DE L'API-REST. — | 11 |
| - FIGURE 03 : DIAGRAMME DE SEQUENCE DE LA PLATEFORME D'ACQUISITION. — | 13 |
| - FIGURE 04 : DIAGRAMME DE SEQUENCE DE L'API REST. — | 14 |
| - FIGURE 05 : DIAGRAMME D'ACTIVITE DE LA PLATEFORME D'ACQUISITION. — | 15 |
| - FIGURE 06 : DIAGRAMME DE CLASSE DE L'APPLICATION. — | 15 |
| - FIGURE 07 : MEDECIN FAISANT UN PRELEVEMENT A L'AIGUILLE FINE SUR UNE PATIENTE - | 16 |
| - FIGURE 08 : BIOPSIE A L'AIGUILLE FINE — | 17 |
| - FIGURE 09 : MICROPHOTOGRAPHIE D'UN ECHANTILLON PAR PONCTION A L'AIGUILLE FINE — | 17 |
| - FIGURE 10 : LA TAILLE DU RAYON POUR CHAQUE CAS - | 18 |
| - FIGURE 11 : CARTE DES ALGORITHMES SCIKIT-LEARN - | 19 |
| - FIGURE 12 : CHACUN DES DEUX SEINS EST PLACE DANS LE MAMMOGRAPHE ET EST PROGRESSIVEMENT COMPRI ME POUR EFFECTUER L'EXAMEN - | 20 |
| - FIGURE 13 : EXEMPLE D'UNE MAMMOGRAPHIE PRODUITE PAR UN MAMMOGRAPHE - | 21 |
| - FIGURE 14 : MAMMOGRAPHIE EXTRAITE DU JEU DE DONNEES AVEC SES LABELLISATIONS - | 22 |
| - FIGURE 15 : SCHEMA ILLUSTRANT LE DEROULEMENT DE LA PREDICTION AU NIVEAU DES MODELES. - | 22 |
| - FIGURE 16 : ARCHITECTURE DU RESEAU DE NEURONES - | 23 |
| - FIGURE 17 : L'ERREUR DE CHAQUE OPTIMISEUR APRES 300 EPOQUES - | 24 |
| - FIGURE 18 : LA PRECISION DE DEUX ARCHITECTURES AVEC DIFFERENTS TAUX DE DROPOUT APRES 100 EPOQUES - | 25 |
| - FIGURE 19 : L'ARCHITECTURE N-TIERS DE L'APPLICATION. — | 26 |
| - FIGURE 20 : ARCHITECTURE ET FONCTIONNEMENT DE L'API-REST. — | 27 |
| - FIGURE 21 : DIAGRAMME DE DEPLOIEMENT / DOCKER. — | 28 |
| - FIGURE 22 : GESTION DE LA BASE DE DONNEES — | 34 |
| - FIGURE 23 : GESTION DES UTILISATEURS. — | 34 |
| - FIGURE 24 : AFFICHAGE DES VERSIONS DES ENTRAINEMENTS. — | 34 |

| | |
|--|----|
| - FIGURE 25 : AUTHENTIFICATION. — | 35 |
| - FIGURE 26 : AJOUTER UNE MAMMOGRAPHIE. — | 35 |
| - FIGURE 27 : MODIFIER LE PROFIL. — | 36 |
| - FIGURE 28 : INTERFACE D’ACCUEIL. — | 36 |
| - FIGURE 29 : LABELLISER UNE MAMMOGRAPHIE AJOUTEE PAR LE MEDECIN. — | 37 |
| - FIGURE 30 : AFFICHER UNE MAMMOGRAPHIE NON-LABELISEE ALEATOIREMENT, ET LA LABELLISER. — | 38 |
| - FIGURE 31 : AFFICHER LES STATISTIQUES. — | 38 |
| - FIGURE 32 : UPLOADER UNE IMAGE MAMMOGRAPHIQUE. - | 39 |
| - FIGURE 33 : INTRODUIRE DES DONNEES NUMERIQUE.- | 39 |
| - FIGURE 34 : INTERFACE UTILISATEUR. — | 40 |
| - FIGURE 35 : RESULTAT DE LA PREDICTION DUR LA MAMMOGRAPHIE. — | 40 |
| - FIGURE 36 : RESULTAT DE LA PREDICTION SUR LES DONNEES NUMERIQUES. - | 41 |

10. Bibliographie / webographie.

Documentations officielles :

- Documentation officielle Django : <https://docs.djangoproject.com/en/2.2/>;
- Documentation officielle PyTorch : <https://pytorch.org/docs/stable/index.html>;
- Documentation officielle d'Angular7 : <https://angular.io/docs>;

Formations :

- Formation Deep Learning offerte par Udacity: Intro To Deep Learning With PyTorch: <https://www.udacity.com/course/deep-learning-pytorch--ud188>;

Sites:

- Stackoverflow.com
- GitHub.com

Articles:

- *Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov*; Dropout: A Simple Way to Prevent Neural Networks from Overfitting: 15 (Jun): 1929–1958, 2014: http://jmlr.org/papers/v15/srivastava14a.html?fbclid=IwAR2hYmi3LkU7F1RQ4eG7tGUxz5oAXbUz_SsjShr-gCoN17oBI53pAndvp40