Ecole Publique d'Ingénieurs en 3 ans

Rapport Projet 2A

APPRENTISSAGE SÉCURISÉ

le 30 mars 2024,

version 1.1

Étudiants :

Zeyd BOUMAHDI,
Noura OUTLIOUA,
Cécile LU,
Paul NGUYEN,
Anis AHMED ZAID

Tuteurs :

Christophe ROSENBERGER,
Tanguy GERNOT

www.ensicaen.fr

Table des matières

Table des figures

INTRODUCTION

1. Introduction

1.1. Contexte

L'utilisation de l'apprentissage machine sur des données sensibles et sécurisées, notamment celles qui sont chiffrées, présente un ensemble unique de défis.

Cela peut permettre à la gendarmerie de détecter des contenus illégaux chiffrés par les détenteurs ou encore de bénéficier de services sur un cloud sans compromettre ses données ou celles de ses clients.

En effet, certaines méthodes et techniques de chiffrement permettent de réaliser des calculs et donc de l'apprentissage statistique sur des données chiffrées, dont la confidentialité reste préservée.

Deux cas s'appliquent :

- Le modèle analyse de manière sécurisée les images médicales pour aider les médecins à diagnostiquer les maladies, tout en protégeant la confidentialité des patients. Le résultat est chiffré et seuls ceux ayant les clés peuvent déchiffrer le résultat.
- Le modèle analyse les images à caractère pedopornographique qui sont chiffrées, et donne le type d'image aux policiers.

Notre problématique principale est donc la suivante : est-il possible de réaliser des tâches de prédiction à partir de données protégées ? Quelles sont les conséquences sur les performances, à savoir la précision et le temps de calcul?

1.2. Objectifs

L'objectif de notre projet est de réaliser des tâches d'analyse et de prédictions sur des données protégées et chiffrées. Notre but est également de comparer les performances des différentes solutions trouvées ou mises en place sur une base de données commune et de les comparer avec un modèle qui travaillait sur des données non chiffrées.

Pour cela, nous devons mettre au point une base de données conséquente et qui permet d'illustrer les performances des différentes méthodes.

La partie la plus importante consiste à construire un modèle qui essaye de reconnaitre les images de notre base de données. Nous devons ensuite explorer les solutions plus abouties qui existent actuellement et les adapter à nos images, en essayant de maximiser les performances et de comparer les différentes approches selon un protocole précis.

2. Méthodologie

2.1. Organisation du projet

2.1.1. Tâches du projet

Les tâches du projet se répartissent en plusieurs catégories, comme suit :

- Revue de la littérature : Cette phase initiale consiste à analyser les recherches antérieures sur l'apprentissage sécurisé avec des données confidentielles. Elle implique également l'identification des méthodes principales utilisées et la réalisation d'un état de l'art sur le sujet.

- Mise en place de bases de données : Il s'agit de la description de deux ensembles de données distincts : une base de données d'entraînement et une autre pour la validation du modèle. L'accent est mis sur les méthodes utilisées pour garantir leur confidentialité.

- Modèle de prédiction : Notre projet principal vise à concevoir et adapter des modèles préexistants afin de développer une solution de prédiction qui soit à la fois performante et respectueuse des contraintes de confidentialité.

- Optimisation du modèle au niveau des performances et de la sécurité : Affiner les paramètres du modèle pour accroître la vitesse de traitement et la précision des prédictions.

- Démonstration Web : La visualisation de notre travail sera matérialisée par le biais d'une démonstration web qui montre l'application pratique des modèles développés et étudiés.

2.1.2. Planification

Notre projet démarre en novembre 2023 et se termine en avril 2024. On commence par la revue de la littérature et la méthodologie, avec des jalons à mi-tâche et des check-ins avec nos encadrants. Ensuite, on prépare les bases de données en janvier, avant de plonger dans le gros du travail en février et mars avec le développement des modèles de prédiction et la préparation de la démo web, qui durent chacun un mois. On évalue et compare les différents modèles.

2.2. Revue de l'État de l'art

Dans le cadre de ce projet, une compréhension approfondie des développements actuels et des recherches antérieures est cruciale pour établir une base solide sur laquelle nous pouvons construire. La section suivante résume l'état de l'art, soulignant les avancées significatives dans le domaine, les solutions existantes, ainsi que les défis et les lacunes qui persistent. Cette revue systématique vise non seulement à encadrer notre travail dans le spectre des connaissances actuelles mais également à identifier les opportunités d'innovation et d'amélioration.

Article	Année	Principes	Base de données	Précision
CryptoNets: Apply-ing Neural Networks to Encrypted Data with High Throughput and Accuracy	2016	Leveled Homomorphic Encryption and Neural Networks	MNIST database	99% accuracy and around 59000 predictions per hour
Privacy Preserving Training and Evaluation with Homomorphic Encryption	2021	Implementation and test machine learning algorithms, Fully Connected Neural Network(FCNN), and Convolutional Neural Network(CNN).	MNIST database	About 98% accuracy with EncFCNN on 50 images and 100% with EncCNN on 100 images
Privacy-Preserving Classification on Deep Neural Network	2017	Application of secure computation in the context of machine learning	MNIST database	99.59% accuracy

DÉVELOPPEMENT

3. Base de données et apprentissage

3.1. Conception et Développement de la Base de Données

L'élaboration de la base de données (BDD) a été obtenue via la plateforme Kaggle, qui offre un large éventail de données variées pour les travaux de recherche et de développement en intelligence artificielle.

La structure de la BDD est conçue pour répondre spécifiquement aux besoins de notre projet de classification d'images. Elle comprend deux classes principales : chats et chiens.

Le dossier d'entraînement contient 25000 images, tandis que le dossier de validation est composé de 2000 images. Cette répartition vise à fournir une quantité suffisante de données pour entraîner efficacement le modèle tout en conservant un ensemble distinct pour évaluer ses performances.

La phase de traitement des images a impliqué une étape de prétraitement pour préparer les données à l'apprentissage automatique :

Les données ont été ajustées en niveau de gris et redimensionnées à une taille plus petite. Les images ont ensuite été annotées avec leur classe respective et les caractéristiques pertinentes ont été extraites.

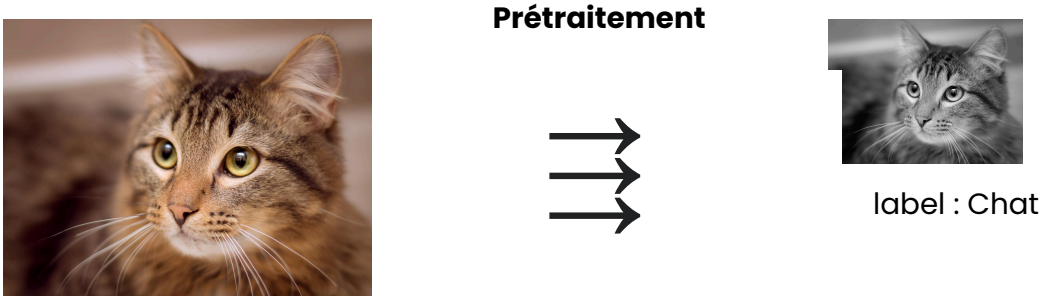


Figure 1 - Schématisation de l'étape de prétraitement

3.2. Apprentissage

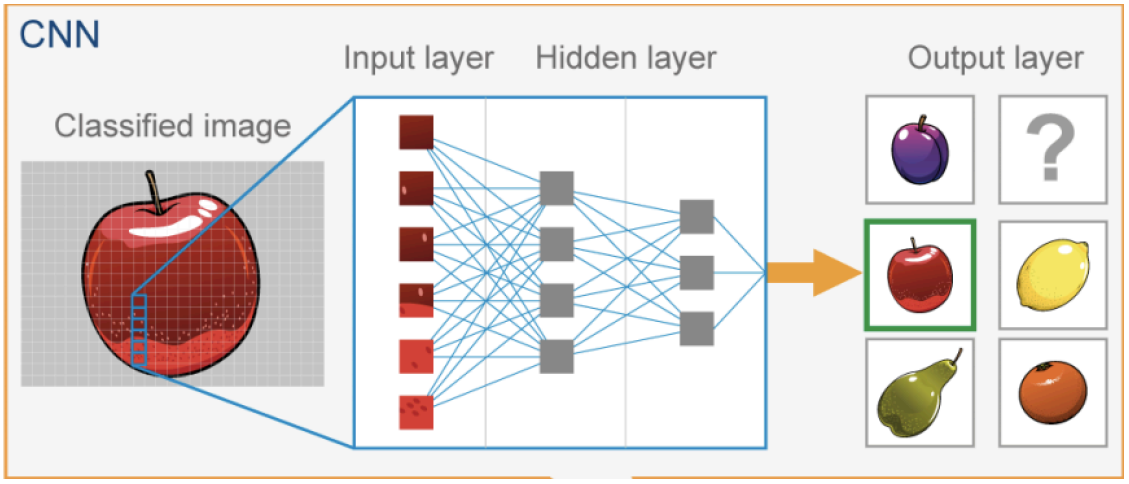


Figure 2 - Classification d'image par réseau de neurones convolutifs (CNN)

Le processus d'apprentissage a été subdivisé en trois étapes principales : Entraînement, Validation, et Test.

- Entraînement : Les images ont subi diverses transformations (rotation, zoom, etc.) pour augmenter la généralité du modèle. Elles ont été étiquetées comme appartenant à l'une des deux classes : chat ou chien.
- Validation : Le modèle a été évalué sur un nouvel ensemble de données non transformées pour tester sa capacité à généraliser à partir de nouvelles données.
- Test : Une évaluation finale a été réalisée pour déterminer la précision du modèle dans des conditions similaires à celles qu'il rencontrerait en production.

Le modèle est composé de couches de convolution. L'entraînement consiste à faire apprendre à ces couches des filtres qui capturent des caractéristiques spécifiques des données d'entrée pour améliorer la reconnaissance de patterns dans des images.

4. Types de chiffrement

4.1. Homomorphic Encryption (HE)

Le chiffrement homomorphique (Homomorphic Encryption - HE) est une technique de chiffrement permettant de réaliser des opérations sur des données chiffrées sans avoir besoin de les déchiffrer préalablement. Cela garantit la confidentialité des informations tout en permettant le traitement

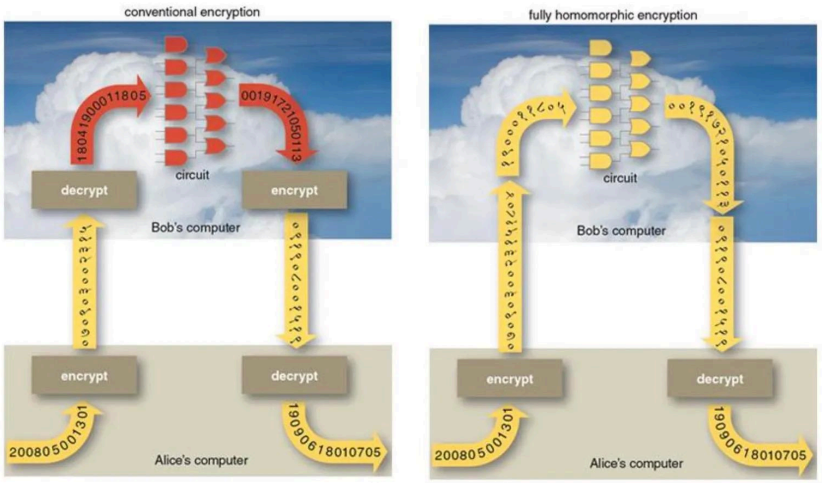
sécurisé des données, notamment dans des domaines sensibles tels que le calcul en nuage.

En particulier, le schéma de chiffrement homomorphique CKKS (pour "Cheon-Kim-Kim-Song") est une implémentation spécifique de HE, principalement utilisée dans le contexte du calcul sécurisé en nuage. CKKS est conçu pour prendre en charge des opérations sur des données chiffrées tout en préservant l'évolutivité et l'efficacité. Il permet notamment d'évaluer des fonctions polynomiales sur des données chiffrées, ce qui le rend adapté à une large gamme d'applications, y compris le traitement de données sensibles dans des environnements distribués. Son architecture est conçue pour maintenir un bon compromis entre performance et sécurité, tout en offrant une certaine flexibilité pour les opérations supportées.

4.2. Fully Homomorphic Encryption (FHE)

Le Fully Homomorphic Encryption (FHE) va au-delà du chiffrement homomorphique (Homomorphic Encryption - HE) en permettant l'évaluation de fonctions arbitraires sur des données chiffrées. Contrairement à certaines implémentations spécifiques d'HE comme le schéma CKKS, FHE n'est pas limité en termes d'opérateurs ou de types de fonctions pouvant être évalués. Cela signifie qu'avec FHE, pratiquement toutes les opérations peuvent être effectuées sur des données chiffrées, offrant ainsi une flexibilité maximale pour le traitement sécurisé des données dans divers domaines d'application.

Cependant réaliser des calculs sans limiter le nombre d'additions et de multiplications peut avoir de sérieuses répercussions sur les performances et nécessiter une puissance de calcul supplémentaire.



Source : Brian Hayes, *American Scientist* (www.americanscientist.org), septembre 2012

Figure 3 - Comparaison du chiffrement conventionnel et du chiffrement homomorphe

5. Développement de modèles d'apprentissage chiffrés

L'utilisation de modèle d'apprentissage classique de détection d'image ne peut pas être applicable dans le cas de données chiffrées. L'aléatoire appliquée sur les données, rend la détection, de certaines formes récurrentes d'un objet, plus détectable par un modèle d'apprentissage. L'objectif donc de cette partie est de mettre en place des modèles capables de réaliser des prédictions sur des données chiffrées.

5.1. Modèle EncCNN

L'EncCNN, pour "Encrypted Convolutional Neural Network", est un modèle de réseau de neurones convolutifs spécialement développé pour être utilisé sur des données chiffrées, en utilisant le framework TenSEAL pour le chiffrement homomorphe CCKS. Ce modèle offre la possibilité de prédire les données sans les déchiffrer, garantissant ainsi la confidentialité des informations traitées.

La structure du modèle vise à maximiser l'efficacité des opérations sur des données chiffrées. La couche convolutive effectue la première étape en appliquant des filtres sur les images afin d'en extraire les caractéristiques essentielles, tandis que les couches complètement connectées qui suivent permettent d'interpréter ces caractéristiques pour aboutir à une prédiction finale.

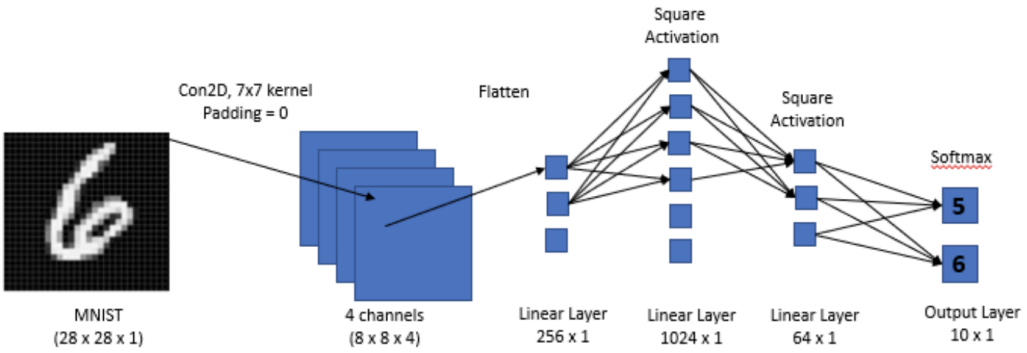


Figure 4 - Modèle de réseau de neurones sur données non chiffrées (CNN)

Le modèle EncCNN repose sur la conversion d'un modèle en clair (CNN), développé pour être utilisé avec des données non chiffrées, en une version qui fonctionne avec des données chiffrées. En d'autres termes, les poids et les entrées du modèle initial sont mesurés afin de pouvoir prédire des données chiffrées sans les déchiffrer, garantissant ainsi la confidentialité des informations traitées tout au long du processus d'inférence.

Ainsi, l'EncCNN est un modèle répondant à notre objectif de réaliser un apprentissage et des prédictions sur des données chiffrées.

5.2. Modèle EncFCNN

Le EncFCNN, pour "Encrypted Fully Connected Neural Network", est une architecture de réseau de neurones entièrement connecté, conçu pour fonctionner avec des données chiffrées grâce au chiffrement homomorphe CKKS, en s'appuyant sur le framework TenSEAL. Ce modèle illustre la capacité de mener à bien des prédictions sur des données chiffrées, assurant ainsi la confidentialité et la sécurité des informations manipulées.

Le modèle EncFCNN réalise les opérations de réseau entièrement connecté (ou dense) sur des données chiffrées, en évitant de les déchiffrer. Ce modèle, au même titre que le EncCNN peut être employé pour des tâches telles que la classification sur des données sensibles sans compromettre la confidentialité des données.

Le modèle FCNN se compose de plusieurs couches linéaires (ou entièrement connectées) qui transforment l'entrée en sortie à travers une série d'opérations matricielles et d'activations non linéaires. Chaque couche linéaire applique une transformation linéaire suivie d'une activation non linéaire carrée, adaptée au contexte du chiffrement homomorphe, permettant ainsi de maintenir les opérations compatibles avec le type de chiffrement utilisé.

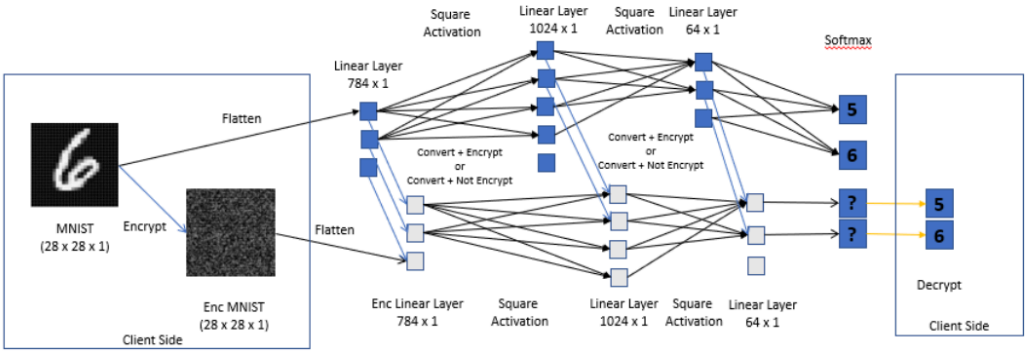


Figure 5 - Modèle de réseau de neurones CNN sur données chiffrées

Le modèle EncFCNN repose ainsi également sur la conversion d'un modèle en clair (FCNN) en un modèle chiffré qui réalise des prédictions sur des données chiffrées.

5.3. Modèle IBM-FHE

5.4. Présentation du toolkit

La librairie (ou kit de développement logiciel) 'HElayers' développée par IBM permet de réaliser des tâches de prédiction et supporte les régressions linéaires, régressions logistiques et les réseaux de neurones. Cette troisième option nous intéresse particulièrement car c'est avec celle-ci que nous avons les meilleurs résultats pour un modèle qui reconnaît des images claires non chiffrées de chiens et de chats. Cette librairie se base sur un système cryptographique à chiffrement homomorphe, cela permet de conserver les propriétés des opérations dans le domaine du chiffre (on peut additionner, multiplier effectuer des rotations sur des vecteurs chiffres et le résultat peut être déchiffré et sera correct avec une perte de précision due au chiffrement).

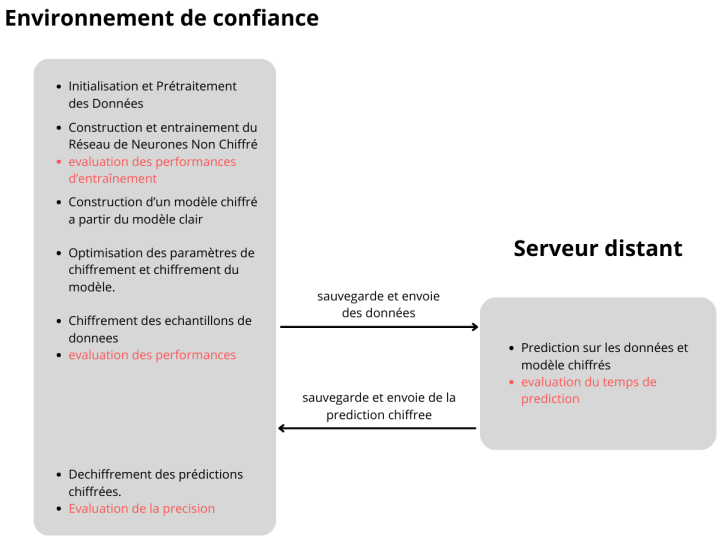


Figure 6 - Schéma d'utilisation de la librairie HElayers

La librairie 'HElayers' permet comme décrit par le schéma de chiffrer un modèle pré-entraîné (on chiffre les poids de notre réseau de neurone), de chiffrer les samples de données et de réaliser des prédictions, on peut ensuite déchiffrer les prédictions. Cela permet notamment de profiter des avantages d'un cloud ou d'un serveur distant en lequel on a pas totalement confiance et d'utiliser sa puissance de calcul pour réaliser des tâches de prédictions. Cela intéresse énormément d'entreprises ou institutions qui tiennent à maintenir un niveau élevé de confidentialité sur leurs données.

HElayers utilise soit HEaaN (Homomorphic Encryption for Arithmetic of Approximate Numbers) soit sur SEAL (Microsoft) pour réaliser les chiffrements et calculs homomorphes.

5.4.1. Résultats obtenus

Nous avons donc construit un réseau de neurone compatible avec HElayer pour ensuite chiffrer ses poids et tenter de faire des prédictions sur des images chiffrées. Les modèles compatibles avec le chiffrement homomorphique du toolkit helayers sont limités car certaines couches ne sont pas disponibles, de même que les réseaux pré-entraînés, le dropout, etc.

Dans un programme séparé, on chiffre le modèle chargé, on chiffre les données de test et on fait faire les projections chiffrées sur les données protégées. La figure 7 présente les matrices de confusion pour des prédictions sur des batch size de taille 4 et 8, on peut voir qu'après déchiffrement, les prédictions s'avèrent être majoritairement correctes.

$$\begin{pmatrix} 3 & 1 \\ 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 7 & 1 \\ 0 & 0 \end{pmatrix}$$

Figure 7 - Matrices de confusion des prédictions sur des données chiffrées

6. Développement des modèles d'apprentissage chiffrés

6.1. Comparaison des modèles

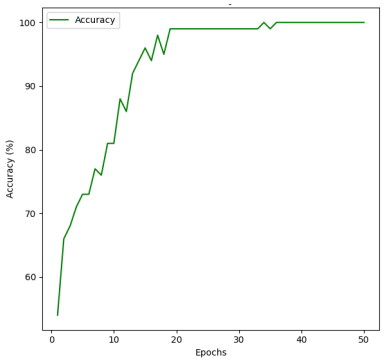


Figure 8 - Accuracy over epoch of FCNN model

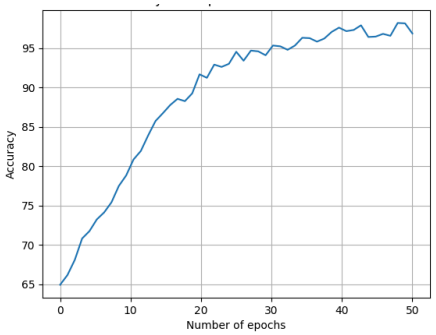


Figure 9 - Accuracy over epoch of CNN model

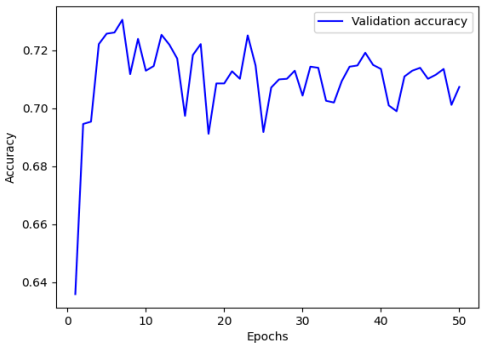


Figure 10 - Accuracy over epoch of HElayer model

Model	CNN	EncCNN	FCNN	EncFCNN	HElayer
Accuracy	97%	97%	100%	50%	~70%
Prediction Time	0.43 s	2.4 s	0.41 s	19 s	0,48 s

Figure 11 - Tableau de comparaison de différents modèles

D'après la figure 11, on peut voir que le modèle EncCNN est celui qui a de meilleures performances. En effet, même si le modèle FCNN a une accuracy très élevée, le modèle EncFCNN reste très peu performant. De ce fait, le modèle EncCNN est le modèle répondant le plus à la problématique.

6.2. Optimisation du modèle EncCNN

Le modèle EncCNN est un modèle dont certains paramètres peuvent être modifiés afin de rendre le modèle plus performant. En effet comme dit dans l'explication du modèle EncCNN, la première couche de celui-ci est une couche de convolution. De ce fait, elle dépend de certains paramètres comme le nombre de filtres et la taille de ceux-ci. Ces paramètres contrôlent la capacité du modèle à extraire et à apprendre des caractéristiques pertinentes des données d'entrée telles que les formes ou encore des couleurs spécifiques.

Ensuite, les images sont prétraitées, car l'apprentissage est réalisé sur des images en niveaux de gris et de 28 pixels. Ces deux transformations sont des obstacles pour obtenir de meilleures performances. Effectivement, des images colorées permettraient au modèle de repérer les couleurs liées à une certaine catégorie (chien ou chat), tandis que des images de taille plus importantes favorisent un apprentissage de meilleure qualité car les photos seront de meilleure qualité.

De ce fait, ce modèle peut être grandement amélioré et optimisé ce qui peut permettre une meilleure précision des prédictions et un temps plus faible pour une prédiction.

7. Démonstration Web

Afin de mettre en évidence le travail effectué dans ce projet, une démonstration en ligne a été mise en place. Cette démonstration pédagogique illustre l'intégration et l'application pratique de l'apprentissage sécurisé dans des scénarios réels grâce à l'utilisation de Streamlit, un outil puissant pour créer rapidement des applications web dédiées à la science des données.

Dans cette démonstration web, les utilisateurs ont la possibilité de télécharger des images de chats ou de chiens et de choisir diverses méthodes de chiffrement pour les traiter. L'interface offre également la possibilité de visualiser les probabilités que l'image soit attribuée à l'une ou l'autre des catégories (chat ou chien), ce qui permet de rendre la décision du modèle plus claire. En effet, l'utilisation d'une jauge graphique permet de visualiser de manière intuitive la confiance du modèle dans ses prédictions, ce qui renforce l'interactivité et la confiance de l'utilisateur envers l'application. Les utilisateurs peuvent aussi évaluer l'effet du chiffrement sur la complexité computationnelle et la latence du modèle en comparant le temps d'exécution et les performances de chaque méthode de chiffrement.

BILAN

8. Difficultés rencontrées

Dans le développement de notre projet d'apprentissage sécurisé, nous avons dû faire face à plusieurs difficultés..

Pour ce qui est de la complexité du chiffrement homomorphe, nous avons dû faire face à des défis de taille, notamment en termes d'optimisation des performances computationnelles.

En outre, la taille considérable de notre base de données a allongé le temps nécessaire pour entraîner les modèles, principalement en raison du besoin de puissance de calcul plus élevé pour gérer de grands ensembles de données.

Malgré ces défis, notre engagement envers le projet nous a permis de développer des solutions et d'avancer vers la réalisation de nos objectifs.

9. Conclusion

En conclusion, notre projet d'apprentissage sécurisé a été une expérience enrichissante. Bien que nous ayons rencontré des défis, tels que la complexité du chiffrement homomorphe et la gestion de grandes bases de données, nous avons pu développer un modèle chiffré pouvant donner des prédictions sur des données chiffrées.

De plus, la création d'une démonstration web a permis de concrétiser nos travaux et de pouvoir visualiser la prédiction d’une image chiffrée . Nous sommes reconnaissants envers nos tuteurs pour leurs aides et leur disponibilité pour répondre à nos questions. En résumé, notre projet offre une perspective intéressante sur les possibilités de l'apprentissage sécurisé, et nous sommes impatients de voir comment ces résultats pourront être utilisés dans l'avenir.