3ème année du cycle ingénieur CYTech, option Intelligence Artificielle
David Lasgleizes
Justine Ribas

JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation

AI based image processing - 2023/2024

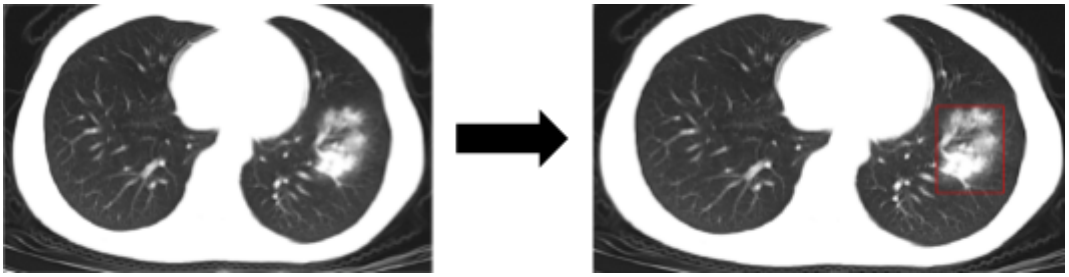
Sommaire :

Vous pouvez ajouter des titres (Format > Styles de paragraphe) qui apparaîtront dans votre table des matières.

Contexte

Dans le contexte d'une épidémie, la détection des personnes contaminées est d'importance capitale. Le premier moyen de détection utilisé pour le COVID-19 a été le test PCR. Ce test permet de détecter rapidement les cas plutôt avancés. Cependant, il se révèle inefficace pour des cas précoces. Pour déceler ces derniers, il faut faire un CT scan puis le faire lire par un professionnel de santé. Il est estimé que cette manœuvre prend 20 minutes par patient. Il a donc fallu trouver une méthode pour déterminer les traces de COVID-19 dans les scans.

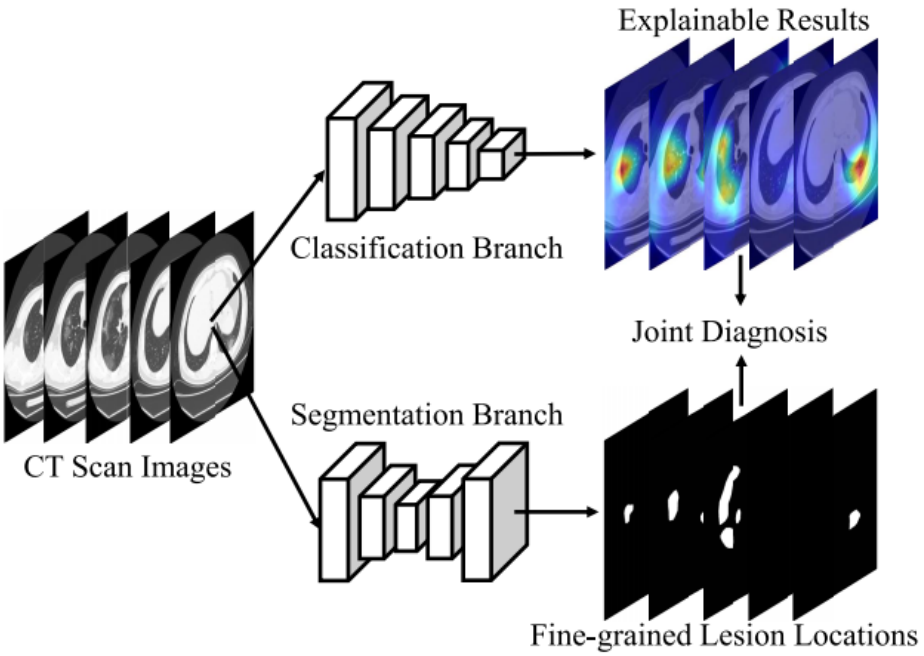
C'est donc dans cette situation que des modèles d'IA ont été implémentés. L'étude que nous allons détailler dans ce rapport se base sur un modèle JCS (Joint Classification and Segmentation). Nous allons détailler ce que sont la classification et la segmentation appliqués dans ce cas et comment les chercheurs ont joint ces deux méthodes pour faire un modèle performant pour la détection.



Explication de la méthode utilisée

L’architecture globale du modèle :

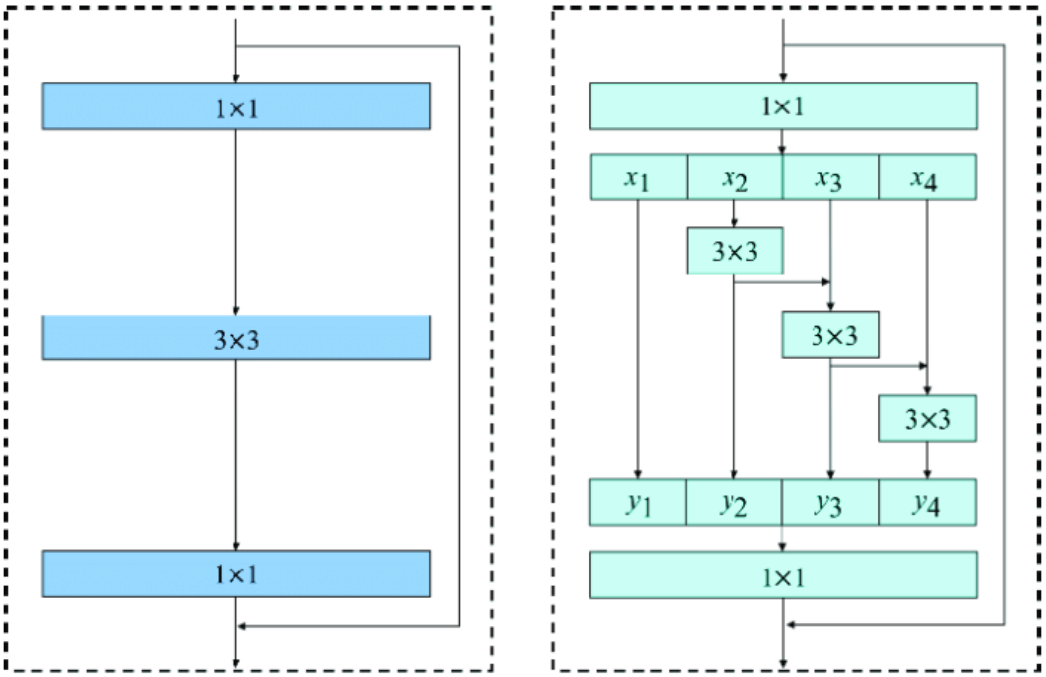
Le modèle est composé de deux branches. Une branche qui concerne une tâche de classification permettant de déterminer si un patient est atteint de COVID-19 ou non. La deuxième branche correspond à la tâche de segmentation qui cartographie sur l’image les lésions qui témoignent d’un cas atteint et ainsi d’effectuer un diagnostic explicable. Ainsi, ces deux branches, Classification et Segmentation, sont conçues pour travailler ensemble et fournir des résultats diagnostiques informatifs pour les patients atteints de la COVID-19, en combinant la capacité de prédire l'infection par la COVID-19 avec la capacité de cartographier les opacifications pulmonaires.



La partie de Classification :

La tâche de classification consiste à prédire si le patient est atteint de la COVID-19 ou non, basée sur les images CT.

La branche de classification exploite un modèle basé sur le réseau **Res2Net**, une amélioration du réseau ResNet. ResNet (Residual Network) et Res2Net (Residual Networks of Residual Networks) sont des architectures de réseaux de neurones profonds utilisés principalement dans des tâches de vision par ordinateur, telles que la classification d'images. Ils sont conçus pour résoudre le problème du "problème de disparition des gradients", qui se produit lors de l'entraînement de réseaux de neurones profonds.



ResNet est composé de nombreuses couches empilées, formant un réseau très profond. L'architecture principale de ResNet est basée sur des unités résiduelles, également appelées blocs résiduels. Chaque bloc résiduel comporte deux "voies" (ou chemins) : un chemin direct qui passe les informations à travers le bloc sans aucune transformation significative, et un chemin résiduel où les transformations sont appliquées aux données.
Les couches dans ResNet sont conçues de manière à permettre au réseau d'apprendre des représentations résiduelles, c'est-à-dire les différences entre les valeurs prédites et les valeurs réelles. Les couches résiduelles permettent d'éviter le problème de la disparition des gradients, car même si une couche ne parvient pas à apprendre une transformation utile, elle peut toujours transmettre l'information non transformée au bloc suivant. Cela facilite la formation de réseaux très profonds.

Res2Net est une évolution de ResNet. Dans Res2Net, les unités résiduelles traditionnelles sont remplacées par des unités résiduelles modifiées qui ont une topologie de connexion plus complexe. Contrairement à ResNet, où chaque unité résiduelle a deux chemins (direct et résiduel), dans Res2Net, chaque unité résiduelle a plusieurs sous-chemins résiduels. Les couches dans Res2Net ont pour rôle de permettre aux informations de se propager plus librement à travers le réseau, en utilisant des connexions résiduelles multiples. Cela améliore la capacité du réseau à apprendre des représentations hiérarchiques et à capturer des détails à différentes échelles spatiales. En d'autres termes, Res2Net est conçu pour mieux modéliser les dépendances entre les pixels dans une image, en utilisant des connexions résiduelles multiples pour gérer différentes échelles d'information.

En résumé, Res2Net utilise des connexions résiduelles multiples pour mieux capturer les dépendances spatiales à différentes échelles, ce qui peut être particulièrement utile dans les tâches de vision par ordinateur nécessitant une analyse fine des images.

La dernière couche convolutive du réseau de classification est suivie par une couche de **Global Average Pooling** (GAP) et une couche entièrement connectée. La couche GAP permet de réduire la taille des caractéristiques spatiales. La couche entièrement connectée est dotée de deux canaux distincts, l'un pour la probabilité d'infection par la COVID-19 et l'autre pour la probabilité d'absence d'infection. Si la probabilité du canal d'infection est supérieure à celle du canal d'absence d'infection, le patient est diagnostiqué comme étant atteint de la COVID-19.

Pour comprendre les prédictions, la technique d'**activation mapping** a été utilisée. Cette technique fait référence à la visualisation ou à la représentation graphique des activations de neurones à l'intérieur d'une couche particulière d'un réseau neuronal pendant le processus d'entraînement ou d'inférence. L'activation mapping est obtenue en calculant le poids pour chaque canal d'activation en utilisant la méthode de gradient.

La partie de segmentation :

La tâche de segmentation vise à identifier et à cartographier les zones d'opacification dans les images CT des patients atteints de la COVID-19.

Le modèle comprend un **encodeur-décodeur U-Net-like** pour la segmentation des opacifications. L'encodeur est basé sur l'architecture **VGG-16** avec des ajustements pour produire des caractéristiques de différentes échelles.

L'architecture VGG-16 est un modèle de réseau neuronal convolutif (ConvNet) qui a été développé par le Visual Geometry Group (VGG) à l'Université d'Oxford. Il fait partie de la série des réseaux VGG, qui sont connus pour leur profondeur et leur performance dans la reconnaissance d'images. VGG-16 est ainsi nommé en référence à ses 16 couches de réseau, qui comprennent 13 couches de convolution et 3 couches entièrement connectées.

L'architecture VGG-16 commence par une couche d'entrée qui prend des images de taille fixe, généralement 224x224 pixels en couleur (RVB). Le cœur de VGG-16 est composé de 13 couches de convolution, où chaque couche est suivie d'une couche de regroupement (max pooling). Les couches de convolution sont relativement petites (filtres 3x3), mais elles sont empilées les unes sur les autres, ce qui permet au réseau d'apprendre des caractéristiques complexes à partir des données. Après les couches de convolution et de regroupement, VGG-16 comprend 3 couches entièrement connectées, également appelées couches "fully connected". Ces couches sont utilisées pour effectuer la classification finale des caractéristiques extraites par le réseau.

VGG-16 utilise la fonction d'activation ReLU (Rectified Linear Unit) après chaque couche de convolution et de pooling, sauf pour les couches de sortie. Cela permet au réseau d'apprendre des caractéristiques non linéaires à partir des données.

Le décodeur utilise une stratégie d'**Attentive Feature Fusion** pour agréger les caractéristiques de différentes échelles.Une stratégie d'Attentive Feature Fusion, également appelée fusion de caractéristiques attentive, fait référence à une technique dans le domaine de l'apprentissage automatique et de la vision par ordinateur qui vise à combiner de manière adaptative les caractéristiques ou les informations provenant de différentes sources ou couches du réseau neuronal. Cette technique repose sur des mécanismes d'attention, qui permettent au modèle de donner plus d'importance à certaines caractéristiques ou informations en fonction du contexte de la tâche.

Le réseau neuronal effectue des opérations d'extraction de caractéristiques à partir de différentes sources, telles que différentes couches du réseau ou différents canaux de données. Une couche d'attention est ensuite utilisée pour calculer l'importance ou la pondération de chaque caractéristique extraite en fonction du contexte de la tâche. Cela peut être fait à l'aide de mécanismes d'attention tels que les réseaux de neurones récurrents (RNN), les réseaux de neurones récurrents à attention (LSTM ou GRU avec attention), les mécanismes d'attention transformer (comme dans les réseaux Transformer), ou d'autres mécanismes d'attention sur mesure. Les caractéristiques extraites sont ensuite fusionnées en utilisant les poids d'attention calculés. Les caractéristiques qui ont reçu une plus grande attention auront un impact plus important sur la fusion globale. Les caractéristiques fusionnées sont ensuite utilisées pour générer la sortie du modèle, que ce soit pour une tâche de classification, de détection d'objets, de segmentation d'images, ou toute autre tâche.

L'avantage de cette approche est qu'elle permet au modèle d'apprendre de manière adaptative quelles caractéristiques sont pertinentes pour la tâche en cours. Cela peut améliorer la capacité du modèle à gérer des données complexes, à prendre en compte des informations importantes et à ignorer les informations inutiles.

Ainsi le modèle de segmentation produit plusieurs sorties latérales avec différentes résolutions.

La Fonction de perte (Loss Function) :

Une stratégie de supervision profonde a été appliquée pour toutes les sorties latérales de la segmentation. Pour chaque sortie latérale, la loss est calculée en combinant la **perte de cross-entropie binaire** (BCE) et la **perte de Dice**.

La **perte de cross-entropie binaire**, également appelée perte de logistique ou perte log loss, est une fonction de coût couramment utilisée dans les problèmes de classification binaire en apprentissage automatique. Elle mesure la disparité entre les prédictions d'un modèle et les vraies étiquettes (0 ou 1) pour chaque exemple d'apprentissage. La formule générale de la perte de cross-entropie binaire est la suivante :

$$L(y,\hat{y}) = -(y.\log(\hat{y}) + (1-y.\log(1-\hat{y}))$$

où :

- $L(y,\hat{y})$ est la perte de cross-entropie binaire pour un exemple donné.
- y est la vraie étiquette binaire (0 ou 1) pour l'exemple.
- $\hat{y}$ est la prédiction du modèle, qui est une valeur continue entre 0 et 1, représentant la probabilité que l'exemple appartienne à la classe 1.

La perte de cross-entropie binaire mesure à quel point les prédictions du modèle correspondent aux étiquettes réelles. Plus précisément si y=1, la perte dépend du logarithme de la prédiction du modèle Plus la prédiction est proche de 1 (c'est-à-dire, le modèle est confiant que l'exemple appartient à la classe 1), moins la perte est élevée. En revanche, si la prédiction est proche de 0, la perte est élevée. Si y=0, la perte dépend du logarithme de $1-\hat{y}$. Dans ce cas, plus la prédiction est proche de 0 (c'est-à-dire, le modèle est

confiant que l'exemple n'appartient pas à la classe 1), moins la perte est élevée. Si la prédiction est proche de 1, la perte est élevée.

L'objectif de l'apprentissage dans un problème de classification binaire est de minimiser la perte de cross-entropie binaire. Cela revient à ajuster les paramètres du modèle de manière à ce qu'il produise des prédictions qui se rapprochent le plus possible des étiquettes réelles.

La **perte de Dice**, ou coefficient de Dice, est une métrique et une fonction de perte couramment utilisée dans les tâches de segmentation d'images, en particulier dans le domaine de la vision par ordinateur et de l'apprentissage profond. Elle est utilisée pour évaluer la similarité entre les masques de segmentation produits par un modèle et les masques de segmentation de référence (les masques étiquetés par des humains). La perte de Dice est souvent utilisée comme fonction de coût pour l'entraînement de modèles de segmentation. Le coefficient de Dice est défini comme suit :

$$Dice = \frac{2\ x\ Aire\ de\ recouvrement\ de\ la\ prédiction\ et\ de\ la\ vérité}{Aire\ de\ la\ prédiction + Aire\ de\ la\ vérité}$$

où :

- L'aire de recouvrement de la prédiction et de la vérité est la quantité d'intersection entre la prédiction du modèle et la vérité (c'est-à-dire, la zone correctement segmentée par le modèle).
- L'aire de la prédiction est la surface totale couverte par la prédiction du modèle.
- L'aire de la vérité est la surface totale couverte par le masque de référence.

Le coefficient de Dice est une valeur qui varie de 0 (absence de recouvrement entre la prédiction et la vérité) à 1 (recouvrement parfait entre la prédiction et la vérité). Plus le coefficient de Dice est proche de 1, meilleure est la correspondance entre la prédiction et la vérité. La perte de Dice, utilisée comme fonction de coût, est la complémentaire du coefficient de Dice. Elle est définie comme : DiceLoss=1−Dice

L'objectif lors de l'entraînement d'un modèle de segmentation est de minimiser la perte de Dice, ce qui revient à maximiser la similarité entre les masques de segmentation prédits et les masques de référence. Cela conduit à des prédictions de segmentation plus précises.

Ainsi, la BCE mesure la confiance de la prédiction, tandis que la perte de Dice mesure la précision de la segmentation.

Les performances du modèle :

Les performances du modèle de classification sont les suivantes. On remarque que le modèle atteint une sensibilité de 95.0% et une spécificité de 93.0% lorsque le seuil est réglé à 25.

TABLE IV

SENSITIVITY AND SPECIFICITY OF OUR CLASSIFICATION MODEL UNDER DIFFERENT THRESHOLDS. WE SET THE THRESHOLD AS 25 (THE GRAY ROW) IN THE FINAL SETTING

No.	Threshold	Sensitivity	Specificity
1	15	96.0%	91.5%
2	20	95.0%	92.0%
3	*25*	*95.0%*	*93.0%*
4	30	94.5%	93.5%

Les performances du modèle de segmentation sont les suivantes :

TABLE V

ABLATION STUDY FOR THE PROPOSED EFM AND AFF IN THE SEGMENTATION MODEL. THE BASELINE IS THE VGG16-BASED SEGMENTATION MODEL WITHOUT EFM&AFF (NO. 1). WE ADD EFM AND AFF SEPARATELY AND SHOW THE EFFECTIVENESS OF THEM (NO. 2 AND NO. 3). THE NO. 4 RESULT IS THE COMPLETE VERSION OF THE SEGMENTATION MODEL

No.	EFM	AFF	Dice	IoU	E_ϕ
1			71.0%	57.7%	88.0%
2	✔		74.3%	61.4%	88.9%
3		✔	75.9%	63.4%	90.9%
4	✔	✔	**77.5%**	**65.4%**	**92.0%**

Le modèle proposé obtient des améliorations significatives par rapport aux autres modèles de pointe sur trois métriques principales : le Dice score, l'Intersection over Union (IoU) et l'Enhanced Alignment Measure (Eφ).

TABLE VII

QUANTITATIVE RESULTS ON OUR SEGMENTATION TEST SET

Methods	Publication	Dice	IoU	E_ϕ
U-Net [58]	MICCAI'15	65.1%	54.1%	79.7%
DSS [73]	TPAMI'19	65.7%	51.7%	79.9%
EGNet [75]	ICCV'19	69.3%	55.4%	83.6%
PoolNet [74]	CVPR'19	69.7%	55.9%	83.9%
JCS (Ours)	Submit'20	**78.5%**	**66.4%**	**92.7%**

Les améliorations par rapport au deuxième meilleur modèle sont les suivantes :

- Dice score : Amélioration de 8.8%
- IoU : Amélioration de 10.5%
- Eφ : Amélioration de 8.8%

Les résultats montrent que le modèle de segmentation proposé surpasse de manière significative les autres méthodes de pointe en termes de précision de segmentation. Des analyses statistiques montrent que le modèle de segmentation est stable et fournit des résultats fiables sur la plupart des images CT de la base de test.

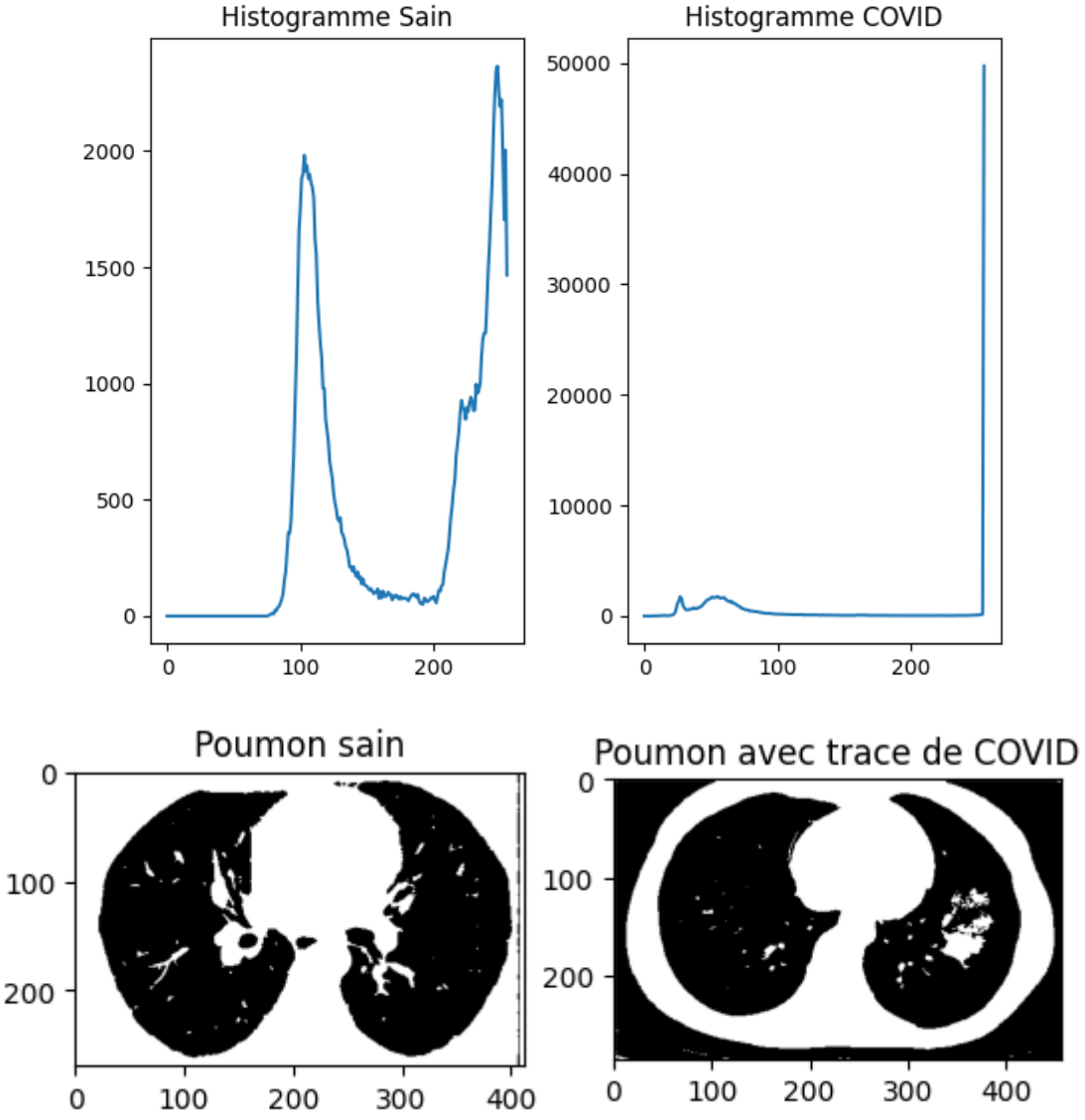
En ce qui concerne la rapidité de diagnostic, le modèle de classification du système JCS prend environ 1.0 seconde pour déterminer si le patient est infecté. Si le patient est infecté, le modèle de segmentation prend environ 21.0 secondes pour effectuer une segmentation fine des lésions. Le système JCS prend donc environ 22.0 secondes pour chaque cas infecté ou 1.0 seconde pour chaque cas non infecté. Ces temps de traitement sont considérablement plus rapides que les méthodes de diagnostic conventionnelles, telles que le test RT-PCR et le diagnostic CT par un radiologue.

Notre application

Le seuillage par histogramme

Le seuillage par histogramme est une technique de traitement d'image utilisée pour segmenter une image en deux ou plusieurs régions en fonction des niveaux de gris (ou des valeurs de pixels) de l'image. Le but du seuillage par histogramme est de simplifier l'image en regroupant les pixels en deux catégories : pixels d'intérêt et pixels de fond, en fonction de la valeur de gris de chaque pixel par rapport à un seuil prédéfini.

Le processus de seuillage par histogramme repose sur l'histogramme de l'image, qui est une représentation graphique de la distribution des valeurs de gris dans l'image. L'histogramme affiche le nombre de pixels ayant une certaine valeur de gris. En fonction de la forme de cet histogramme, vous pouvez choisir un seuil pour distinguer les pixels d'intérêt des pixels de fond.



Le seuillage par K means

Le seuillage par K-means est une technique de seuillage d'image qui repose sur l'algorithme de regroupement de données K-means. L'algorithme K-means est généralement utilisé pour effectuer une segmentation non supervisée des données en clusters ou groupes en fonction de leurs caractéristiques. Lorsqu'il est appliqué à des images, il peut être utilisé pour segmenter l'image en plusieurs régions en se basant sur les niveaux de gris (ou les valeurs des pixels) de l'image.

Image originale poummon sain

Image originale poumon COVID

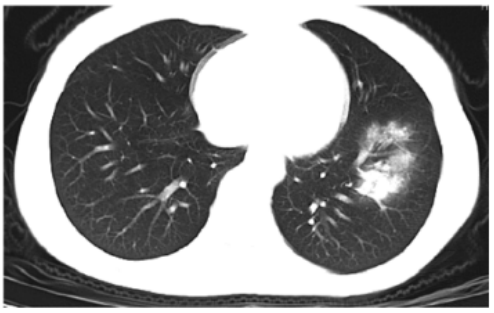


Image seuillée, K=4

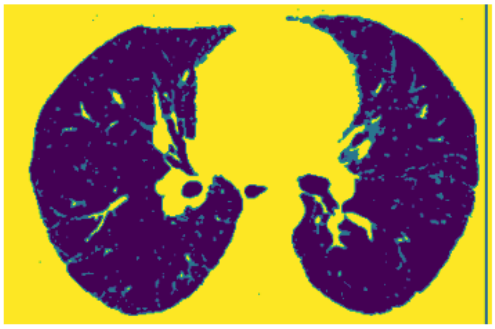
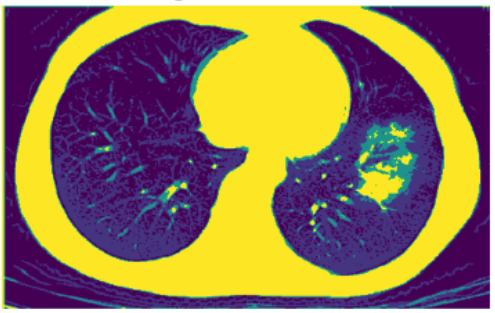
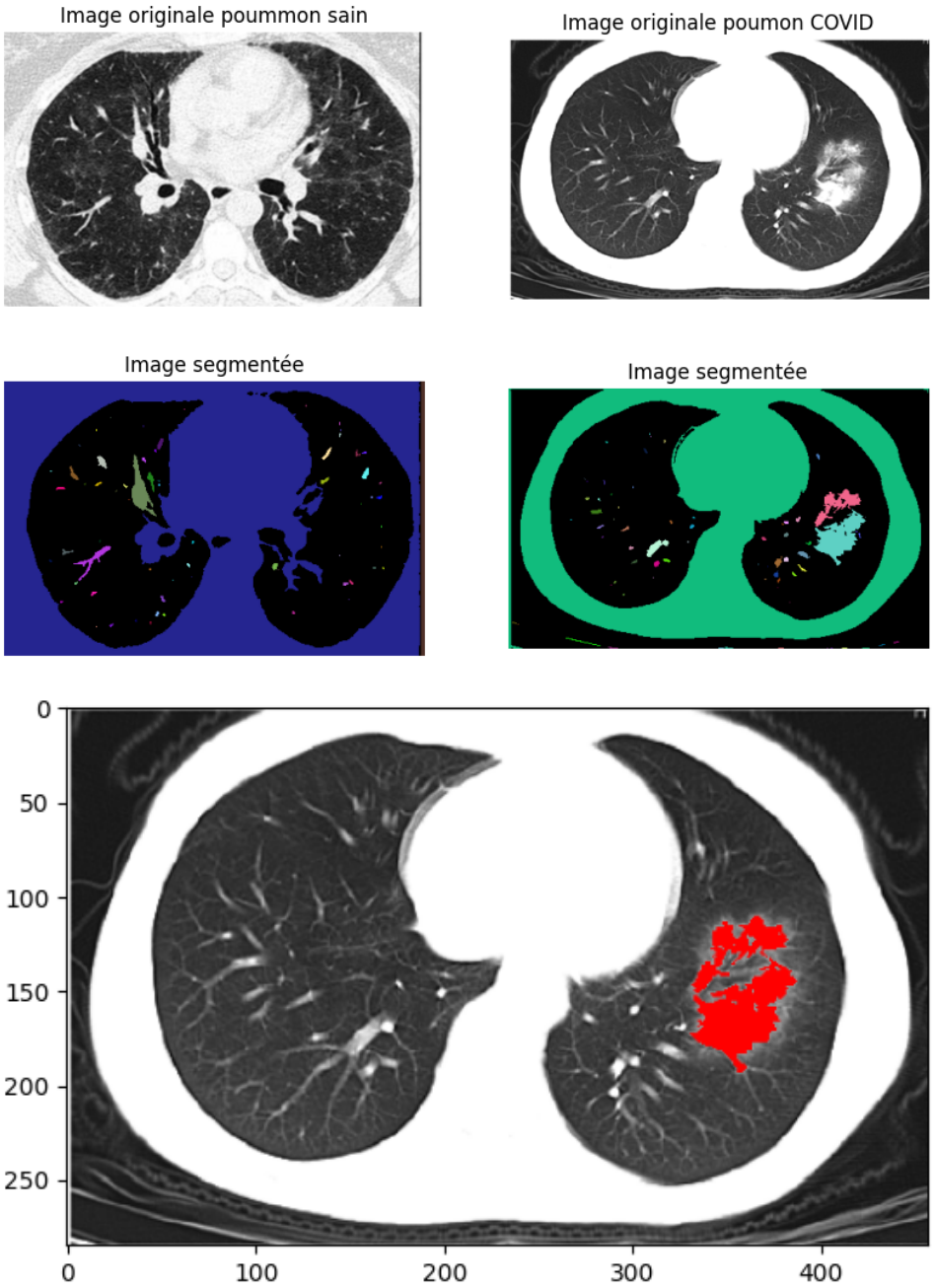


Image seuillée, K=4



La segmentation

La segmentation d'image est un processus de traitement d'image qui vise à diviser une image en régions ou segments distincts, chaque région étant une partie cohérente de l'image qui partage certaines caractéristiques communes, telles que la couleur, la luminosité, la texture ou la forme. L'objectif de la segmentation d'image est de simplifier et de comprendre le contenu de l'image en regroupant les pixels ou les régions similaires, ce qui facilite l'analyse ultérieure de l'image.



Grâce au résultat de la segmentation, nous pouvons facilement identifier la zone de lésion qui témoigne du Covid 19 chez le patient.

Conclusion et perspectives

Le système JCS de détection du COVID-19 a fait preuve d'une grande précision dans ses performances. Selon les informations fournies, le système a atteint une sensibilité de 95,0 % et une spécificité de 93,0 % sur l'ensemble de tests de classification de l'ensemble de données COVID-CS. Cela indique que le système est capable d'identifier avec précision les cas positifs à la COVID-19 et de les distinguer des cas non infectés.

En outre, le modèle de segmentation du système JCS a obtenu un score Dice de 78,5 % sur l'ensemble de test de segmentation, surpassant les méthodes de segmentation de pointe précédentes. Ces résultats démontrent que le système JCS est efficace pour diagnostiquer avec précision les cas de COVID-19 et identifier les zones de lésions à grain fin dans les CT scan.

Nous avons réussi à reproduire les grandes lignes du programme des chercheurs. On remarque que la méthode des Kmeans permet une identification primaire des traces tandis que la segmentation permet réellement de percevoir les nuances entre la forme des poumons ou bien traces de COVID.

Références

Article de recherche :
https://ieeexplore.ieee.org/abstract/document/9357961

ResNet vs Res2Net :
https://www.researchgate.net/figure/Comparison-between-the-ResNet-block-and-Res2Net-block_fig2_341175426

ResNet :
https://fr.wikipedia.org/wiki/R%C3%A9seau_neuronal_r%C3%A9siduel#:~:text=Un%20r%C3%A9seau%20neuronal%20r%C3%A9siduel%20

Res2Net :
https://arxiv.org/abs/1904.01169
https://paperswithcode.com/method/res2net

Global Average Pooling :
https://blog.paperspace.com/global-pooling-in-convolutional-neural-networks/

VGG-16 :
https://datacorner.fr/vgg-transfer-learning/#:~:text=VGG16%20est%20un%20mod%C3%A8le%20de,Large%2DScale%20Image%20Recognition%C2%BB.

Attentive Feature Fusion
https://www.isca-speech.org/archive/pdfs/interspeech_2022/liu22f_interspeech.pdf

La perte de cross-entropie binaire et perte de dice :
https://www.quantmetry.com/blog/choix-fonction-de-perte-en-computer-vision/

Informations complémentaires sur les poumons :
https://www.info-radiologie.ch/poumon.php