République algérienne démocratique et populaire
الجمهورية الجزائرية الديمقراطية الشعبية
Ministre de l'enseignement supérieur et de la recherche scientifique
وزارة التعليم العالي و البحث العلمي

Université 20 août 1955-Skikda
Faculté des sciences
Département d'informatique

Mémoire de fin d'études

En vue de l'obtention du diplôme de Master en Informatique

Spécialité : Systèmes Informatique

Thème

Vision-Transformers pour la classification des images médicales

Présenté Par : Fekrache Ismahane

Jury :

- Dr. benjedou Zineb	*Président*
- Dr. ali guechi Farida	*Examinateur*
- Dr. hazmoune Samira	*Encadreur*

Année universitaire 2023/2024

Remerciements

“

Je tiens tout d'abord à remercier Dieu le tout puissant et miséricordieux, qui ma a donné la force et la patience d'accomplir ce Modeste travail.

Je tiens à remercier chaleureusement ma directrice de mémoire ***Dr. Hazmoune Samira*** *qui m'a honoré en acceptant mon encadrement et pour ces efforts, sa soutien et sa guidance tout au long de ce projet de recherche.*

Je tiens aussi à remercier les membres du jury pour avoir accepté d'examiner et d'évaluer ce travail.

Enfin, je souhaite remercier ma famille pour leur soutien constant et leur encouragement durant toute cette année d'études.

”

Dédicace

“

À toute ma famille

Aucune expression ne saurait exprimer mon respect et ma considération pour votre soutien et encouragements. Je vous dédie ce travail en reconnaissance de l'amour que vous m'offrez quotidiennement et votre bonté exceptionnelle. Que Dieu le Tout Puissant vous garde et vous procure santé et bonheur.

”

- Ismahane

Résumé

L'intelligence Artificielle (IA) se développe de plus en plus dans divers domaines, notamment celui de l'imagerie médicale, où elle devient un outil significatif dans la routine des radiologues. L'augmentation des données médicales a bouleversé la charge de travail des médecins, réduisant le temps passé avec les patients et augmentant le risque d'erreurs d'interprétation. L'IA, et en particulier les Vision Transformers (ViT), offrent une assistance au diagnostic en soulignant les anomalies pathologiques. Le Vision transformer est considéré comme une technologie de classification d'images de pointe.

Ce projet présente un modèle de Transformer utilisant ViT pour la classification d'images médicales. Des expériences ont été réalisées sur deux jeux de données distincts, les tumeurs cérébrales et les radiographies pulmonaires, pour affiner la précision du modèle en sélectionnant les hyperparamètres les plus adaptés au ViT. Les taux de précision de classification atteints pour ces bases de données étaient respectivement de 98,40 % et 90,68%. Des analyses comparatives avec différentes architectures de réseaux neuronaux convolutifs (CNN) et des études antérieures sur les mêmes jeux de données ont mis en évidence les performances supérieures du ViT par rapport aux méthodes classiques d'apprentissage automatique et profond.

Mots clés : Vision Transformer (ViT), Imageries médicales, Classification des images midicales, Brain-tumor-dataset, Chest X-Ray Images (Pneumonia).

Abstract

Artificial intelligence (AI) is increasingly being used in a variety of fields, including medical imaging, where it is becoming a significant tool in radiologists' routines. The increase in medical data has had a major impact on doctors' workloads, reducing the time spent with patients and increasing the risk of interpretation errors. AI, and Vision Transformers in particular, offer diagnostic assistance by highlighting pathological anomalies. The Vision Transform is considered to be a cutting-edge image classification technology.

This project presents a transformer model using the Vision transformer (ViT) for medical image classification. Experiments were performed on two separate datasets, brain tumours and chest X-rays, to affiner the accuracy of the model by selecting the hyperparameters best suited to ViT. The classification accuracy rates achieved for these databases were 98.40% and 90.68% respectively. Comparative analyses with different convolutional neural network (CNN) architectures and previous studies on the same datasets highlighted the superior performance of ViT compared with conventional machine and deep learning methods.

Keywords : Vision Transformer (ViT), Medical imaging, Classification of midical images, Brain-tumor-dataset, Chest X-Ray Images (Pneumonia).

ملخص

يتزايد استخدام الذكاء الاصطناعي (AI) في مجموعة متنوعة من المجالات، بما في ذلك التصوير الطبي، حيث أصبح أداة مهمة في روتين أطباء الأشعة. وقد كان للزيادة في البيانات الطبية تأثير كبير على أعباء عمل الأطباء، مما يقلل من الوقت الذي يقضونه مع المرضى ويزيد من مخاطر أخطاء التفسير. يقدم الذكاء الاصطناعي، ومحوّلات الرؤية على وجه الخصوص، المساعدة في التشخيص من خلال تسليط الضوء على الحالات المرضية الشاذة. تعتبر محولات الرؤية تقنية متطورة لتصنيف الصور.

يقدم هذا المشروع نموذج محول باستخدام محول الرؤية (ViT) لتصنيف الصور الطبية. أُجريت التجارب على مجموعتي بيانات منفصلتين، أورام الدماغ والأشعة السينية للصدر، وذلك لتحديد دقة النموذج من خلال اختيار المقاييس الأنسب لـ ViT. بلغت معدلات دقة التصنيف المحققة لقاعدتي البيانات هاتين 98,40% و 90,68% على التوالي. أبرزت التحليلات المقارنة مع بنيات مختلفة للشبكة العصبية التلافيفية (CNN) والدراسات السابقة على مجموعات البيانات نفسها الأداء المتفوق لـ ViT مقارنةً بأساليب التعلم الآلي التقليدي والتعلم العميق.

كلمات مفتاحية : محـول الرؤيـة (ViT) ، الصـور الطبيـة ، تصنيـف الصـور الطبيـة ، مجموعة بيانات أورام الدماغ، صور الأشعة السينية للصدر (التهاب رئوي).

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

IA	*Intelligence Artificielle*
GPS	*Global Positioning System*
KNN	*K-Nearest Neighbors*
SVM	*Machine à Vecteur de Support*
ANN	*Artificielle Neural Network*
MHA	*Multi-Head Attention*
MLP	*Multi-Layer Perceptron*
RNN	*Recurrent Neural Networks*
LSTM	*Long Short-Term Memory*
GRU	*Gated Recurrent Unit*
CNN	*Convolutional Neural Network*
FC	*Fully Connected layers*
NLP	*Natural Language Processing*
GPT	*Generative Pre-trained Transforme*
BERT	*Bidirectional Encoder Representations from Transformers*
ReLU	*Rectified Linear Unit*
ViT	*Vision Transformer*
DieT	*Data-efficient image Transformers*

IRM *Imagerie par Résonance Magnétique*

CPU *Central Processing Unit*

RAM *Random Access Memory*

GPU *Graphics Processing Unit*

Introduction Générale

Contexte

Certaines maladies requièrent un diagnostic précis, réalisable uniquement par un radiologue, pour évaluer la cause et déterminer le traitement approprié . L'imagerie médicale offre aux médecins des images détaillées des structures internes du corps, facilitant le diagnostic et le traitement de diverses affections. Cependant, l'augmentation du volume et de la complexité des données d'imagerie médicale entraîne des défis dans l'analyse et l'interprétation précise et rapide de ces informations étendues.

Problématique et objectif

Dans le secteur médical, l'information et en particulier l'imagerie médicale, est d'une importance capitale. Le traitement approprié de ces images est vital pour garantir un diagnostic précis dans des délais raisonnables. Effectivement, l'augmentation des données médicales a provoqué une perturbation dans la charge de travail des radiologues et des médecins, restreignant ainsi le temps consacré au patient et augmentant le taux d'erreur d'interprétation.

L'objectif principal de ce travail est de développer un modèle capable à améliorer les performances de classification des images médicales en introduisant les technique de l'intelligence artificielle notamment les Transformers.

Motivation et contribution

Les Transformers ont démontré des performances remarquables dans diverses tâches de traitement du langage naturel , et leur utilisation pour les tâches de vision par ordinateur est une évolution relativement nouvelle. Le Vision Transformer (ViT) est un modèle fondé sur les transformers, conçu spécifiquement pour la classification d'images. ViT traite les images comme des séquences de patchs. Cela permet au modèle de s'ajuster aisément à des tailles d'image variées. Le ViT fait aussi usage de mécanismes d'attention multi-têtes pour se focaliser sur les patchs pertinents de l'image et pour apprendre des représentations d'image performantes. Des performances exceptionnelles ont été obtenues par ViT dans divers standards de classification d'images, ce qui met en évidence l'efficacité des modèles basés sur les transformers pour la vision par ordinateur. C'est pour cette raison que

nous l'avons choisi pour la classification des images médicales dans notre système. Le modèle développé à donné des résultats prometteurs dans deux domaines d'applications de l'imagerie médicale, à savoir, la détection des tumeurs cérébrales à partir d'image IRM, et le diagnostic des affections thoraciques à partir des radiographies.

Organisation du mémoire

Ce mémoire est constitué d'une introduction générale, de trois chapitres et d'une conclusion générale :

- **Chapitre 1 :: Techniques d'apprentissage automatique**

  Ce chapitre donne un aperçu général des méthodes de l'apprentissage automatique ainsi que celles de l'apprentissage profond. Nous y présentons en particulier, le ViT qui est la technique utilisée dans cette étude.

- **Chapitre 2 :: État de l'art sur la classification des images médicales**

  Ce chapitre présente des généralités sur l'imagerie médicale. En effet, nous donnons la définition, les différents types, les objectifs, ainsi que les applications de l'imagerie médicale. Nous y dressons, enfin l'état de l'art du domaine de classification des images médicales.

- **Chapitre 3 :: Un système-basé Transformers pour la classification des images médicales**

  Ce chapitre regroupe la description de notre système, ainsi que les différentes expérimentations réalisées et l'analyse des résultats obtenus. Nous présentons, dans un premier temps, l'architecture de notre système de classification des images médicales. Puis, nous réalisant une série d'expérimentations afin de régler les hyperparamètres de notre modèle et d'évaluer ses performances. En fin, nous comparons les performances du modèle conçu avec celles de quelques architectures CNN et de certains travaux de la littérature.

Chapitre 1

Techniques d'apprentissage automatique

1.1 Introduction

Aujourd'hui, l'intelligence Artificielle (IA) s'intègre de plus en plus dans le quotidien des individus : Que ce soit dans la navigation par GPS, la gestion des émails, ou encore via les assistants vocaux sur les smart phones. Le secteur de la santé n'est pas en reste et doit s'adapter aux progrès technologiques. En Imagerie médicale, l'IA est déjà utilisée pour analyser les images médicales et apporter des informations essentielles aux professionnels de santé.

Dans ce chapitre, nous présentons d'abord, les différentes techniques de l'intelligence artificielle fréquemment utilisées. Ces techniques sont divisées en trois catégories principales : l'apprentissage automatique, l'apprentissage profond et l'apprentissage par transfert, dans ce chapitre aussi nous présentons une étude détaillée sur les transformers, en mettant l'accent sur les vision transformers (ViT) que nous avons sélectionnés pour la classification des images médicales dans notre système.

1.2 L'intelligence artificielle (IA)

L'IA ou l'intelligence artificielle est un domaine de la technologie qui vise à rendre les machines intelligentes. Plus précisément, il s'agit d'un ensemble de techniques informatiques qui permettent aux ordinateurs de développer des algorithmes et des systèmes informatiques qui peuvent effectuer des tâches qui nécessitent habituellement l'intelligence humaine, telles que la reconnaissance de la parole, la compréhension du langage naturel, la prise de décision et la résolution de problèmes complexes. L'IA est fondée sur l'idée que les machines peuvent être programmées pour simuler l'intelligence humaine, et elle se base sur des disciplines telles que la théorie de l'information, La logique mathématique, la statistique et les réseaux de neurones pour développer ses algorithmes. L'IA est actuellement utilisée dans de nombreuses applications pratiques, telles que la reconnaissance vocale et visuelle, les systèmes de recommandation, les chatbots, les robots industriels, les systèmes de contrôle autonome, etc (LeCun 2016).

1.3 Apprentissage automatique (Machine Learning)

Le domaine de l'apprentissage automatique est une discipline qui permet aux ordinateurs d'acquérir la capacité d'apprendre sans être explicitement programmés (Samuel 1959). Il fait partie de l'intelligence artificielle. Ces algorithmes élaborent des modèles en utilisant des données, aussi connues sous le nom de données d'entraînement, pour effectuer des prédictions ou des prises de décisions sans nécessiter une programmation explicite à cet occasion. Les algorithmes d'apprentissage sont couramment employés dans différents secteurs tels que la médecine et la vision par ordinateur (Hu et al. 2020).

1.3.1 Les différents types d'apprentissage

Dans le domaine de l'apprentissage automatique, on distingue principalement l'apprentissage supervisé, l'apprentissage non supervisé, l'apprentissage par renforcement et l'apprentissage par transfert.

1.3.1.1 L'apprentissage supervisé

L'apprentissage supervisé est la tâche de machine learning consistant à apprendre une fonction qui associe une entrée vers une sortie basée sur des paires d'entrée-sortie, par exemple. Il déduit une fonction à partir de données de d'entraînement étiquetées, constituées d'un ensemble d'exemples de d'entraînement. Les algorithmes d'apprentissage automatique supervisés sont ceux qui ont besoin d'aide externe. Le groupe de données d'entrée est divisé en un ensemble de données d'entraînement et de test. L'ensemble des données d'entraînement a une variable de sortie qui doit être prédite ou classée. Tous les algorithmes apprennent un certain type de modèles à partir de l'ensemble de données d'entraînement et les appliquent à l'ensemble de données de test pour la prédiction ou la classification. En d'autres termes, les algorithmes sont entraînés sur un ensemble de données étiquetées, où chaque exemple est associé à une étiquette ou à une sortie attendue. L'objectif est de développer un modèle capable de prédire la sortie correcte pour de nouvelles données non étiquetées. Les problèmes de classification et de régression sont deux domaines principaux dans lesquels l'apprentissage automatique supervisé est bénéfique. La classification implique de sélectionner une valeur d'entrée et de la transformer en une valeur discrète. En général, lorsqu'il s'agit de problèmes de classification, la sortie est une classe ou une catégorie (Mahesh 2020).

1.3.1.2 L'apprentissage non supervisé

L'apprentissage non supervisé est une méthode d'apprentissage automatique où un algorithme est entraîné à trouver des motifs ou des structures dans un ensemble de données sans l'aide d'étiquettes ou de réponses prédéfinies. Les algorithmes d'apprentissage non supervisé sont souvent utilisés pour explorer et analyser des données, détecter des anomalies, regrouper des données similaires et réduire la dimensionnalité des données (Arulkumaran et al. 2017).

1.3.1.3 Lapprentissage par renforcement

Les algorithmes apprennent à prendre des décisions séquentielles en interagissant avec un environnement et en recevant des récompenses ou des pénalités en fonction de leurs actions. L'objectif est de maximiser la récompense cumulative au fil du temps. L'apprentissage par renforcement est l'un des trois paradigmes de base de l'apprenant automatique, aux côtés de la formation supervisée et de celle sans supervision (Arulkumaran et al. 2017).

1.3.1.4 L'apprentissage par transfert

L'apprentissage par transfert est un concept essentiel en apprentissage automatique. Il consiste à transférer des connaissances acquises lors de la résolution d'une tâche source vers une tâche cible similaire. En d'autres termes, au lieu d'entraîner un modèle à partir de zéro pour chaque nouvelle tâche, on utilise un modèle **pré-entraîné** (par exemple, un réseau de neurones convolutif entraîné sur ImageNet) comme point de départ. Ensuite, on adapte ce modèle aux spécificités de la tâche cible en ajustant ses poids par **fine-tuning** ou en utilisant ses couches intermédiaires comme caractéristiques pour un nouveau modèle (Ying et al. 2018).

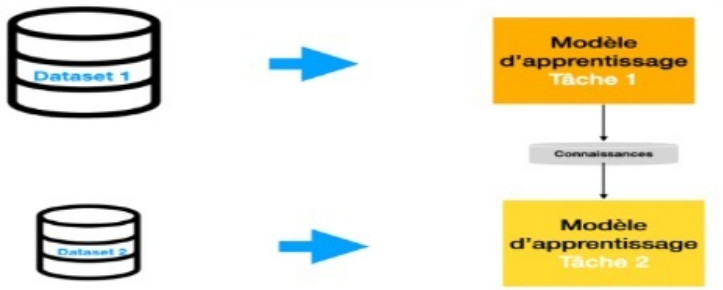

Fig. 1.1 : L'apprentissage par transfert (Chiter et al. 2022).

En résumé, l'apprentissage par transfert permet d'économiser du temps et des ressources en capitalisant sur les connaissances déjà acquises par un modèle pré-entraîné. Cela facilite également l'adaptation à de nouvelles tâches sans recommencer l'apprentissage à partir de zéro.

1.3.2 Méthodes d'apprentissage automatique supervisé

Nous nous concentrerons uniquement sur les méthodes d'apprentissage supervisé les plus fréquemment utilisées, étant donné que notre recherche porte sur un problème de classification.

1.3.2.1 K-nearest neighbors (KNN)

L'algorithme des K plus proches voisins ou K-nearest neighbors (kNN) est un algorithme de Machine Learning qui appartient à la classe des algorithmes d'apprentissage supervisé simple et facile à mettre en œuvre qui peut être utilisé pour résoudre les problèmes de classification et de régression, Les algorithmes KNN utilisent des données et classifient les nouveaux points de données en fonction de mesures de similarité (fonction de distance). Le classement se fait à la majorité de ses voisins. Les données sont affectées à la classe qui a les voisins les plus proche La méthode KNN est donc une méthode à base de voisinage, non-paramétrique Ceci signifiant que l'algorithme permet de faire une classification sans faire d'hypothèse qui relie a variable dépendante aux variables indépendantes (Mathieu-Dupas 2010).

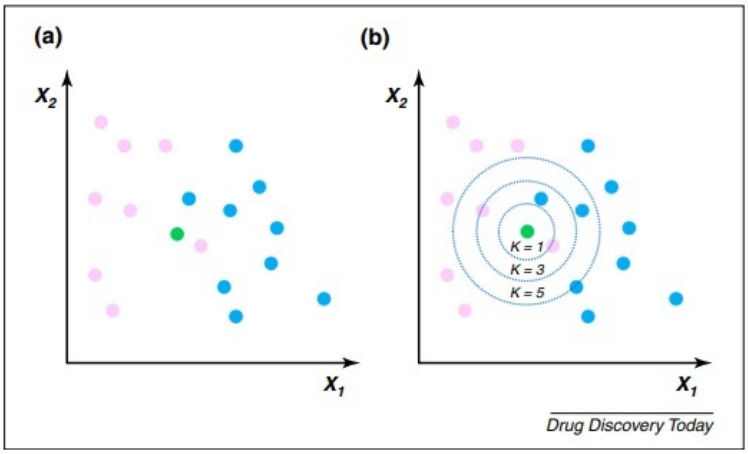

FIG. 1.2 : K plus proches voisins (LAVECCHIA 2015).

1.3.2.2 Machines à vecteurs de support

Les machines à vecteurs de support (SVM) peuvent être définies comme étant une famille d'algorithmes d'apprentissage permettant de résoudre efficacement des problèmes de classification, ou de régression, en exploitant les résultats de la théorie de l'apprentissage statistique développée en grande partie par Vladimir Vapnik. Le principe sous-jacent aux SVM consiste à utiliser une transformation non linéaire pour redécrire les données d'apprentissage dans un espace de plus grande dimension, dans lequel elles pourront être traitées efficacement de manière linéaire. Dans le cas d'un problème de classification binaire, l'objectif est alors de déterminer dans le nouvel espace, que 1' on nomme espace de représentation, un hyperplan qui permet de séparer les données d'apprentissage de manière optimale (MILGRAM 2007).

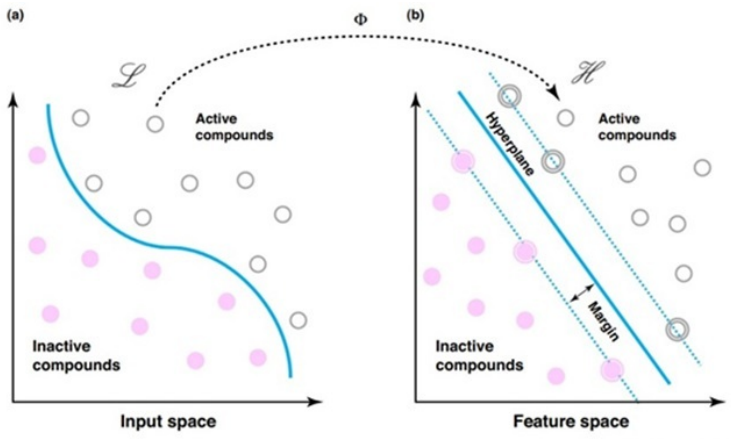

FIG. 1.3 : Machine à vecteurs de support (LAVECCHIA 2015).

1.3.2.3 Arbres de Décision

Les arbres de décision sont composés d'une structure hiérarchique en forme d'arbre. Cette structure est construite grâce à des méthodes d'apprentissage par induction à partir d'exemples. L'arbre ainsi obtenu représente une fonction qui fait la classification d'exemples, en s'appuyant sur les connaissances induites à partir d'une base d'apprentissage. En raison de cela, ils sont aussi appelés arbres d'induction (Induction Décision Trees). Une définition un peu plus formelle des arbres de décision est la suivante : un arbre de décision est un graphe orienté, sans cycles, dont les noeuds portent une question, les arcs des réponses, et les feuilles des conclusions, ou des classes terminales. Traditionnellement, un arbre de décision se construit à partir d'un ensemble d'apprentissage par raffinements de sa structure. Un ensemble de questions sur les attributs est construit afin de partitionner l'ensemble d'apprentissage en sous-ensembles qui deviennent de plus en plus petits jusqu'à ne contenir à la fin que des observations relatives à une seule classe. Les résultats des tests forment les branches de l'arbre et chaque sous-ensemble en forme les feuilles. Le classement d'un nouvel exemple se fait en parcourant un chemin qui part de la racine pour aboutir à une feuille, l'exemple appartient à la classe qui correspond aux exemples de la feuille (Osório 1998).

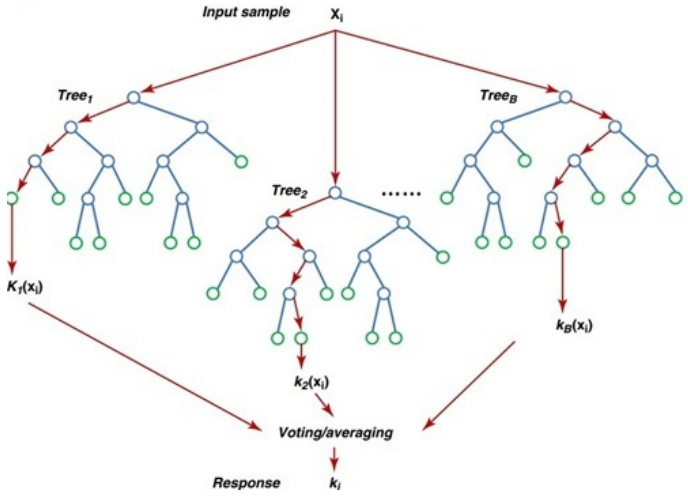

Fig. 1.4 : Exemple d'illustration d'un arbre de décision (Lavecchia 2015).

1.3.2.4 Naïve Bayes

Le classifieur Bayésien naïf est une méthode d'apprentissage supervisé qui repose sur une hypothèse simplificatrice forte : les variables sont indépendantes conditionnellement à la classe à prédire. Cette hypothèse naïve ne permet pas de modéliser les interactions entre différentes variables. Cependant, sur de nombreux problèmes réels, cette limitation n'a que peu d'impact. Un classifieur naïf de Bayes est un classifieur probabiliste basé sur l'application du théorème de Bayes avec l'hypothèse naïve, c'est-à-dire que les variables explicatives (Xi) sont supposées indépendantes conditionnellement à la variable cible (C). Malgré cette hypothèse forte, ce classifieur s'est avéré très efficace sur de nombreuses applications réelles et est souvent utilisé sur les flux de données pour la classification supervisée (Salperwyck et al. 2014).

1.3.2.5 Réseaux de neurones artificiels (ANN)

Un réseau de neurones artificiels (ANN) est constitué d'unités disposées en couches successives, chacune connectée aux couches adjacentes. Les ANN sont inspirés des systèmes biologiques, comme le cerveau, et de leur manière de traiter l'information. Ils sont composés d'un grand nombre d'éléments de traitement interconnectés qui travaillent ensemble pour résoudre des problèmes spécifiques. Les ANN apprennent aussi par l'exemple et l'expérience et sont particulièrement efficaces pour modéliser des relations non linéaires dans des données volumineuses ou quand la relation entre les variables d'entrée est complexe. La structure d'un réseau neuronal inclut une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Le nombre de couches cachées et de neurones dans chaque couche varie selon la complexité du problème à résoudre (EL MASSARI 2023).

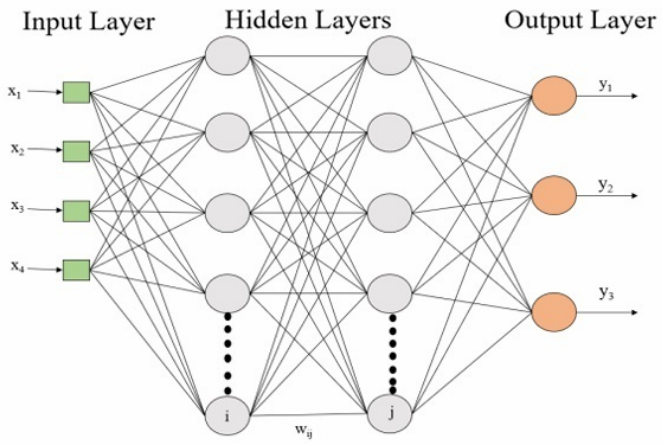

Fig. 1.5 : Structure générale du réseau de neurones artificiels (Moghaddamnia et al. 2009).

Le perceptron multi-couches (MLP)

Un modèle couramment utilisé est le perceptron multicouche (MLP). Le MLP est constitué d'une couche d'entrée, d'une ou de plusieurs couches cachées et d'une couche de sortie. Chaque couche contient plusieurs neurones, et chaque neurone d'une couche est relié aux neurones de la couche suivante par des poids variés, ce qui détermine l'influence d'une unité sur une autre. La tâche d'un MLP est de produire une sortie désirée à partir d'entrées de certains types. À cette fin, l'apprentissage du perceptron est effectué. Les entrées et les sorties souhaitées sont introduites dans le modèle, et l'erreur de sortie est calculée. Les paramètres du système, c'est-à-dire les poids des interconnexions, sont ajustés durant la phase d'apprentissage de l'ANN pour minimiser l'écart entre la sortie réelle et la sortie souhaitée. Le MLP est particulièrement adapté à la résolution de problèmes non linéairement séparables, à la classification et à l'approximation de fonctions continues (Taud et al. 2018).

1.4 Apprentissage profond (Deep Learning)

Deep Learning est une branche de Machine Learning qui permet de générer des prédictions à partir d'une grande quantité de données . Deep Learning utilise des Réseaux de Neurones Artificielle (ANN) inspirés par les connaissances en biologie humaine du cerveau, ce qui donne naissance à des algorithmes extrêmement efficaces pour résoudre des problèmes de classification. En termes généraux, deep learning est une composante de l'intelligence artificielle qui utilise des réseaux neuronaux hiérarchiques complexes et de nombreux données précises afin de rendre les machines capables d'apprendre automatiquement des choses (comme l'apprend l'être humain). Deep Learning est le point de départ de l'intelligence artificielle , et il s'agit d'une des technologies les plus enthousiasmantes de la dernière décennie. Actuellement, ce genre de méthodes d'apprentissage est utilisé dans divers domaines tels que classification des images. Le Deep Learning est un domaine à croissance rapide, et de nouvelles architectures variantes aux algorithmes d'apparaissage. Dans cette section, nous presentons un bref aperçu des structures communes que l'on retrouve dans de nombreux réseaux profonds. (Tyagi et al. 2020).

1.4.1 Les réseaux de neurones récurrents (RNN)

Les réseaux neuronaux récurrents (RNN) ont obtenu des résultats impressionnants dans des problèmes de génération de séquences tels que le NLP. Dans le système RNN, il y a des retours de neurones à leurs entrées ou aux précédentes classes. Le but de ce feedback est d'intégrer le concept de mémoire dans le réseau neuronal. En d'autres termes, le processus de formation ne repose pas seulement sur les données actuelles qui sont fournies à la couche d'entrée, mais aussi sur les valeurs précédentes des entrées. Dans le domaine de la RNN, on calcule un état passé sur l'unité RNN, appelé état invisible (h), en fonction de son type, mais en général, il dépend des données historiques et est utilisé pour obtenir le prochain état. Par exemple, la Figure 1.6 illustre les différentes états pour l'unité de base RNN, également connue sous le nom de vanilla RNN unit. Vanilla RNN présente de nombreux inconvénients tels que les gradients de disparition, ce qui a conduit les chercheurs à développer des types de RNN plus sophistiqués pour faire face à ces inconvénients. Parmi ces catégories, on peut citer Long short-term memory (LSTM) et Gated Recurrent Unit(GRU) . Par la suite, nous présenterons une brève présentation de la structure LSTM et GRU (Farrag et al. 2021).

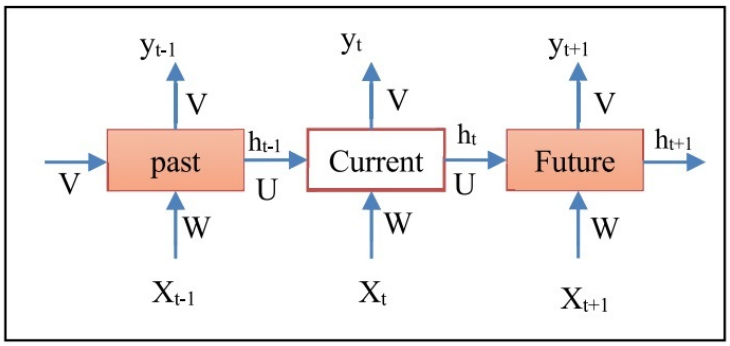

FIG. 1.6 : Les différentes états pour l'unité de base RNN (FARRAG et al. 2021).

1.4.1.1 Le Long Short Term Memory (LSTM)

Les réseaux LSTM ont été créés afin de résoudre la problématique de la perte de mémoire à long terme des réseaux de neurones récurrents (RNN). Leur conception vise à représenter des liens à longue distance dans une séquence en utilisant une mémoire interne appelée "cell state" présente dans chaque neurone. Lorsqu'un LSTM arrive à l'unité courante, la sortie de l'unité neuronale précédente traverse trois portes : la porte d'oubli, la porte d'entrée et la porte de sortie. L'information est gérée par ces portes en utilisant des fonctions d'activation comme des sigmoïdes et des tangentes hyperboliques. Un schéma de cellule LSTM est présenté dans la figure 1.7. La porte i, appelée porte d'entrée, reçoit des informations sur les valeurs qui seront actualisées dans le vecteur d'état. La porte F, également connue sous le nom de porte d'oubli, reçoit les données de l'état précédent qui peuvent être refusées. Lorsque ces deux couches sont émises, le réseau génère un vecteur de nouvelles valeurs d'état candidates c. Finalement, la porte O, appelée porte d'échappement, Explore les informations qui seront transmises à la sortie H (CHENG et al. 2016).

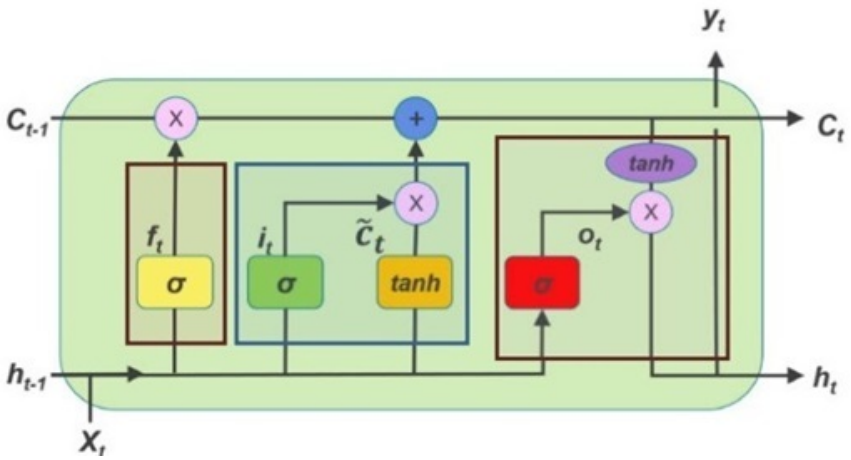

FIG. 1.7 : Cellule Long Short-Term Memory (LSTM) (CHENG et al. 2016).

1.4.1.2 Gated Recurrent Unit(GRU)

Le GRU, est une alternative aux unités LSTM pour réduire la complexité. Le GRU a moins de paramètres ajustables car il ne possède pas de couches de sortie comme le LSTM. Bien que les LSTM et les GRU soient très similaires, ils présentent des différences notables Le GRU dispose de deux portes (réinitialisation et mise à jour), alors que le LSTM en a trois (entrée, sortie et oubli). La porte de réinitialisation dans le GRU gère la combinaison des nouvelles entrées avec la mémoire antérieure, et la porte de mise à jour détermine la quantité de l'état antérieur à conserver. La porte de mise à jour dans le GRU remplit les fonctions des portes d'entrée et d'oubli dans le LSTM. À la différence du LSTM, le GRU n'intègre pas de mémoire cellulaire c_t dans chaque unité. Il est donc clair que le GRU, tout en étant très similaire au LSTM, présente moins de complexité et moins de paramètres, ce qui en fait une architecture intéressante pour comparer ses performances avec celles du LSTM dans des scénarios de séquençage (Nosouhian et al. 2021).

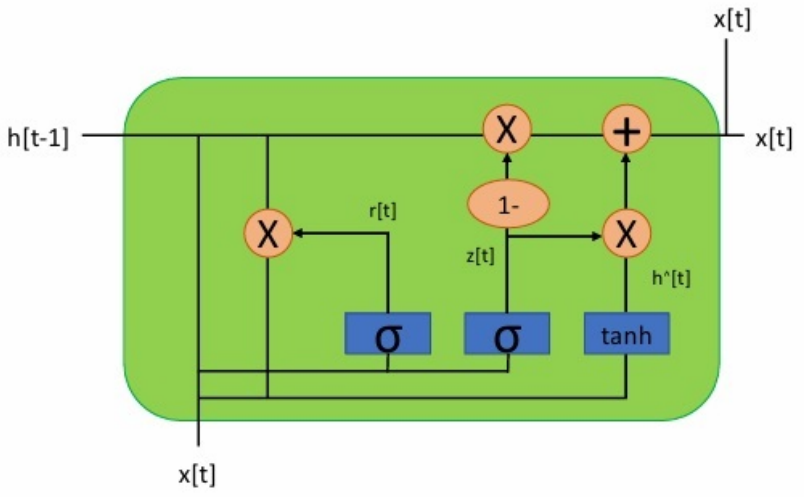

Fig. 1.8 : Cellule Gated Recurrent Unit(GRU) (Nosouhian et al. 2021).

1.4.2 Les réseaux de neurones convolutifs (CNN)

Les réseaux de neurones convolutionnels ou (CNN pour Convolutional Neural Network en anglais), sont des réseaux de neurones multicouches spécialisés dans la classification des images et la reconnaissance de formes. Ils comportent deux parties distinctes, première partie c'est la partie convolutionnelle, son architecture repose sur des couches de convolution alternant avec des couches d'agrégation (pooling). L'image en entrée est fournie sous forme de matrice de pixels, elle passe par une succession de filtres au niveau de chaque couche de convolution, créant des nouvelles images appelées cartes de convolutions. La couche d'agrégation (pooling) réduit la taille de l'image par une fonction de max pooling. Au final, les cartes de convolutions sont mises à plat et concaténées en un vecteur de caractéristiques, appelé code CNN. Ce code CNN en sortie de la partie convolutionnelle est ensuite branché en entrée d'une deuxième partie, constituée de couches entièrement connectées (perceptron convolutionnels multicouche). Le rôle de cette partie

est de combiner les caractéristiques du code CNN pour classer l'image. La sortie est une dernière couche comportant un neurone par catégorie. Les valeurs numériques obtenues sont généralement normalisées entre 0 et 1, de somme 1, pour produire une distribution de probabilité sur les catégories (Hakim et al. 2018).

Le CNN se compose de trois couches neuronales principales, chacune jouant un rôle différent : Couche de convolution (CONV), Couche de Pooling (POOL) et Couche Fully-Connected (FC) comme illustré dans la figure 1.9.

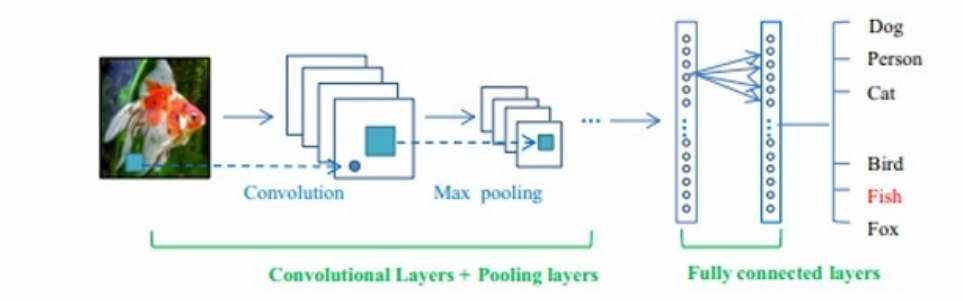

Fig. 1.9 : Schéma de l'architecture globale CNN (Guo et al. 2017).

1.4.2.1 Couche de convolution (CONV)

La convolution est un outil mathématique utilisé pour simplifier des équations plus complexes et pour faire du traitement de l'image et du signal numérique, car elle permet de faire l'extraction des caractéristiques à partir des images d'entrées, afin d'appliquer le bon filtre. Le filtre (aussi connu sous le nom du noyau de convolution) consiste en des poids appliquer à une image. La sortie de la couche de convolution est l'image entrée mais avec certaines modifications en constituant ainsi une carte des caractéristiques. La couche de convolution fonctionne de manière très différente des autres couches du réseau neuronal. Cette couche n'utilise pas de poids de connexion et de somme pondérée. Au lieu de cela, elle contient des filtres qui convertissent les images. Nous appellerons ces filtres des filtres de convolution (Ounissi et al. 2020).

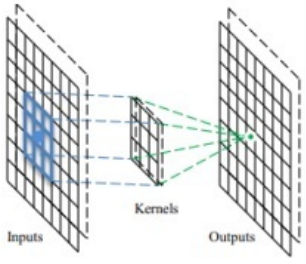

Fig. 1.10 : L'opération de convolution (Guo et al. 2017).

1.4.2.2 Couche de Pooling (POOL)

Une couche de pooling suit une couche de convolution permettre de réduire la dimension de chaque carte de caractéristiques, tout en conservant les informations pertinentes pour le système. Différentes méthodes de sous-échantillonnage (pooling) existent mais les plus usitées sont les fonctions mathématiques de type maximum et moyenne . Cette étape est importante, puisqu'elle permet de garantir une réduction du nombre de paramètres, qui peut augmenter très rapidement avec les couches de convolution (Macary 2022).

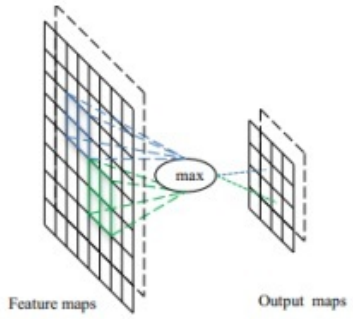

Fig. 1.11 : L'opération de Max pooling (Guo et al. 2017).

1.4.2.3 Couche Fully-Connected (FC)

Après plusieurs couches de convolution et de max-pooling, le raisonnement de haut niveau dans le réseau neuronal se fait via des couches entièrement connectées. Les neurones dans une couche entièrement connectée ont des connexions vers toutes les sorties de la couche précédente (comme on le voit régulièrement dans les réseaux réguliers de neurones). Le vecteur appelé code CNN obtenu en sortie de la partie convolutive est ensuite branché en entrée d'une deuxième partie de classification, son objectif est d'attribuer à chaque échantillon de données une étiquette décrivant sa classe d'appartenance. (Malki et al. 2019).

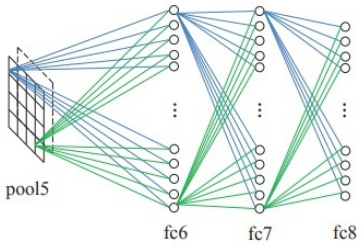

Fig. 1.12 : Couche Fully-Connected (Guo et al. 2017).

1.4.3 Les Transformers

Les Transformers sont issus des travaux de (Vaswani et al. 2017) introduite dans le document "Attention is all you need". Il s'agit d'un modèle qui utilise une architecture encodeur décodeur mais qui met en place des mécanismes d'attention différents. Elle correspond à un empilement de plusieurs encodeurs et décodeurs. Le nombre d'encodeurs et de décodeurs est identique au départ. Maintenant ce n'est plus forcément le cas. Dans le cas des Transformers, un encodeur est composé d'un bloc d'auto-attention (self-attention), suivie d'une couche linéaire. Le décodeur est lui aussi complété par une couche d'attention. De plus, les Transformers ont introduit les mécanismes d'attention à plusieurs têtes (multi-head attention). Ils sont très performants mais également très coûteux à l'apprentissage. En effet, ils sont composés de beaucoup de paramètres du fait de l'architecture, ce qui se traduit par un long temps de convergence, et un besoin de très grandes quantités de données (Vaswani et al. 2017).

Mécanisme d'attention

L'attention est une méthode qui reproduit l'attention cognitive. L'objectif de la technique de l'attention est de se focaliser sur les éléments clés d'une entrée et de négliger les informations qui ne sont pas aussi pertinentes. Une fonction d'attention permet d'associer un vecteur de requête et un ensemble de paires de vecteurs clé-valeur à un vecteur de sortie. L'attention mise en échelle par produit scalaire (Scaled Dot-Product Attention), introduite par (Vaswani et al. 2017), peut être formellement exprimée comme suit :

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1.1}$$

Où Q, K et V représentent respectivement les ensembles de requêtes, de clés et de valeurs, regroupés sous forme de matrices, et d_k représente la dimension des requêtes et des clés, tandis que d_v est la dimension des valeurs. Le produit est mis à l'échelle par $\frac{1}{\sqrt{dk}}$ pour compenser le fait que les produits scalaires peuvent devenir très grands lorsque d_k est élevé, ce qui saturerait la fonction softmax et rendrait ses gradients très faibles (Vaswani et al. 2017).

Multi-Head Attention

L'attention multihead est une sorte de mécanisme d'attention utilisé dans certains modèles de réseaux de neurones. L'utilisation de plusieurs têtes ou processus d'attention permet au modèle de se concentrer sur plusieurs aspects de ses informations à la fois. Ceci est bénéfique pour des tâches telles que le traitement du langage naturel où le modèle doit comprendre les liens entre différents mots dans une phrase. Un modèle d'attention multi-head transforme l'entrée en plusieurs espaces de représentation distincts avant d'appliquer un mécanisme d'attention séparé à chaque espace de représentation. Avant de passer les vecteurs Q, K et V à la fonction Attention, ceux-ci sont projetés linéairement en utilisant des matrices compréhensibles, transformant les entrées en vecteurs de dimensions d_k, d_k et d_v respectivement. De plus, au lieu de réaliser cette opération une seule fois, elle est effectuée h fois, chaque projection étant simultanément transmise à la fonction Attention, ce qui constitue le MultiHead

Attention. Après ce traitement parallèle, les matrices obtenues sont concaténées puis à nouveau projetées linéairement à l'aide d'une matrice compréhensible. Ce traitement parallèle permet au modèle de gérer efficacement les informations issues de différentes représentations des entrées de manière formelle.

$$\begin{aligned} MultiHead(Q,K,V) &= Concat(head1,...,headh)W^O, \\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{1.2}$$

$W_i^Q \in \mathbb{R}^{d_{model}d_k}$, $W_i^K \in \mathbb{R}^{d_{model}d_k}$, $W_i^V \in \mathbb{R}^{d_{model}d_v}$ et $W^O \in \mathbb{R}^{hd_vd_{model}}$

Enfin, nous désignons par self-attention le cas particulier où $K = V$ et, par analogie, nous appelons Multi-Head Self-Attention (MSA) une Multi-Head Attention lorsque $K = V$ (Vaswani et al. 2017).

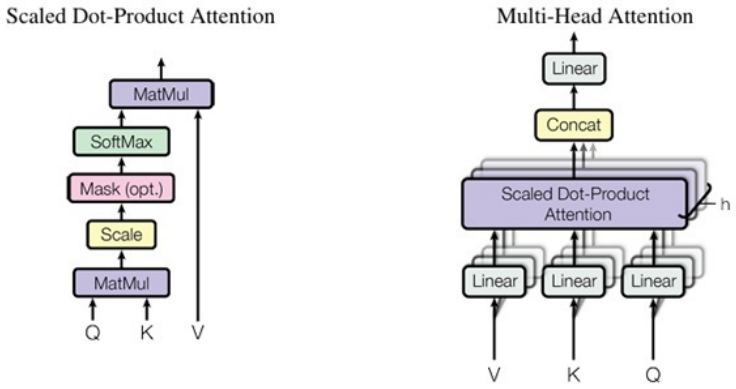

Fig. 1.13 : Schéma représentant les différentes étapes de scaled dotproduct attention et de multihead attention (Vaswani et al. 2017).

Encodage positionelle

Le Transformer opère sur une séquence de données simultanément en ignorant son ordre. Pour éviter la perte d'informations séquentielles, un prétraitement qui ajoute des vecteurs de position supplémentaires à l'entrée est utilisé, d'où le terme "encodage positionnel". Il existe plusieurs options pour réaliser l'encodage positionnel. Par exemple, les fonctions sinus et cosinus de diverses fréquences sont un choix courant comme montrée dans l'équation 1.3 où *pos* indique la position et i est l'indice de patch courant :

$$\begin{aligned} PE_{(pos,2i)} &= \sin(pos/10000^{2i/d_{model}}) \\ PE_{(pos,2i+1)} &= \cos(pos/10000^{2i/d_{model}}) \end{aligned} \tag{1.3}$$

Le bloc MLP et le bloc transformer global

Une fois que le self-attention pour les entrées a été calculé, le résultat est transmis à un bloc Multi-Layer Perceptron (MLP) qui est entièrement connecté à une couche cachée. En termes formels, en considérant une entrée x, des poids et biais apprenables W1, b1, W2, b2, ainsi qu'une fonction d'activation, on peut définir le bloc MLP comme suit :

$$MLP(x) = ((xW1 + b1)W2 + b2). \tag{1.4}$$

emploient l'unité linéaire rectifiée (ReLU) en tant que fonction d'activation. Ils mettent également en œuvre la normalisation de couche après les connexions résiduelles MSA et MLP (Ba et al. 2016). En résumé, étant donné une entrée x_l vers le l^{ime} bloc transformer, un bloc de MSA, un bloc MLP et une normalisation de couche bloc LN, le bloc transformateur global peut être exprimé comme suit :

$$\begin{aligned} x'_l &= LN(x_l + MSA(x_l)) \\ x_l + 1 &= LN(x_{\ l} + MLP(x_{\ l})). \end{aligned} \tag{1.5}$$

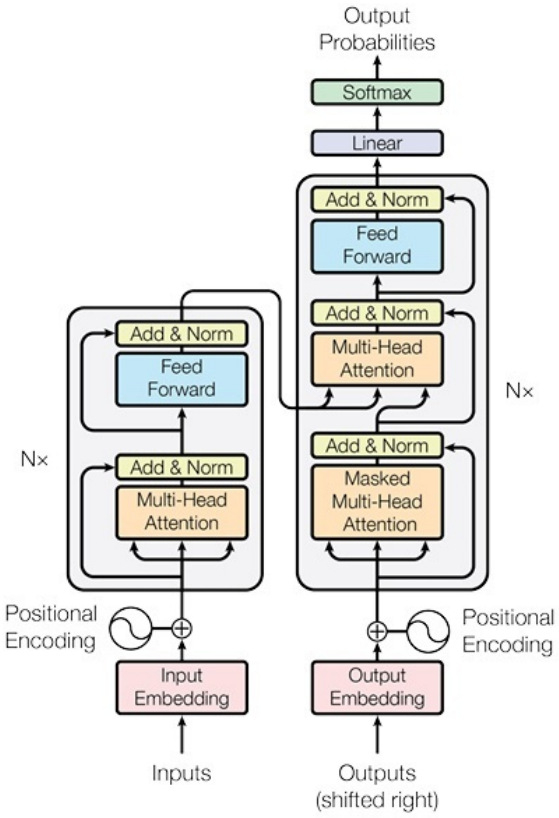

Fig. 1.14 : La structure globale de l'architecture Transformer d'origine (Vaswani et al. 2017).

Il existe deux modles de transformers couramment utilisés. transformers pour NLP et transformers pour la vision par ordinateur (CV) Voici une brève explication de chacun :

1.4.3.1 Transformers pour NLP

Les transformers en NLP ont conduit à l'émergence de nombreux modèles fondés sur cette architecture. Parmi les modèles présentés dans cette section, il y a le GPT (Générative Pré-trained Transformer), axé sur la génération de texte, et le BERT (Bidirectional Encoder Representations from Transformers), qui se focalise sur les représentations bidirectionnelles d'encodeurs.

- **Générative Pré-training (GPT)**

  Le GPT, développé par (Radford et al. 2018), représente une technologie de pointe dans le domaine du traitement du langage naturel (NLP). Utilisé pour diverses tâches de NLP telles que la génération de texte, la traduction automatique et le résumé de texte, il se compose d'une série de décodeurs de type transformer. Autrement dit, il est construit à partir de blocs décodeurs transformer. Dans GPT, le module d'attention multi-têtes avec attention croisée est omis du décodeur transformer, car il n'y a pas d'encodeur. Ainsi, les blocs décodeurs de GPT incluent uniquement le codage de position, le module d'encodage positionnel, l'attention auto-dirigée multi-têtes masquée et le réseau de neurones feedforward, complétés par des fonctions d'addition, de normalisation par couche et d'activation. GPT est une méthode non supervisée, ce qui lui permet d'être entraînée sur d'immenses volumes de données disponibles sur Internet (Ghojogh et al. 2020).

- **Bidirectional Encoder Representations from Transformers (BERT)**

  Les représentations bidirectionnelles d'encodeurs à partir de transformateurs (BERT), par (Devlin et al. 2018), ont utilisé des encodeurs dans un transformateur comme sous-structure pour pré-entraîner des modèles pour des tâches de traitement du langage naturel (NLP) telles que l'analyse de sentiments, les systèmes de questions-réponses et le résumé de texte. BERT fonctionne en deux étapes : l'entraînement préliminaire pour comprendre le langage et le réglage fin pour des tâches spécifiques. Il apprend la langue en utilisant les mécanismes de modélisation de langage masqué (MLM) et de prédiction de la prochaine phrase (NSP). Avec le MLM, BERT apprend les contextes bidirectionnels en masquant certains mots dans les phrases et en les reconstruisant à partir du contexte environnant. Il peut aussi prendre deux phrases en entrée et déterminer si la seconde suit logiquement la première, ce qui lui permet de réaliser la NSP et de maintenir des relations à longue distance dans le texte. BERT a été entraîné sur 16 Go de textes issus de BooksCorpus et de Wikipedia en anglais. Après l'entraînement préliminaire, le modèle est finement ajusté sur une tâche NLP spécifique en utilisant un apprentissage supervisé sur un ensemble de données et en remplaçant la couche de sortie entièrement connectée de BERT par un nouvel ensemble de couches. BERT s'entraîne plus rapidement car seuls les paramètres de la nouvelle couche de sortie sont appris à partir de zéro. Il existe deux versions de BERT : BERT-base et BERT-large (Acheampong et al. 2021).

1.4.3.2 Transformers pour la vision

Les transformers ont connu un véritable succès. dans le domaine du NLP, ce qui a conduit à leur application dans les tâches de vision par ordinateur. Les transformers de classification d'images les plus récents sont Image GPT (iGPT) et Vision Transformers (ViT).

- **Image GPT**

  La génération d'images à l'aide de GPT (Chen et al. 2020), utilise GPT pour entraîner le modèle sur ImageNet, mais elle présente des limites en raison de sa puissance de calcul élevée et de sa qualité d'image inférieure (Usman et al. 2022).

- **Vision Transformers (ViT)**

  Le Vision Transformer est une architecture de réseau de neurones conçue pour la classification d'images. Elle transforme une image en une séquence de vecteurs, ce qui permet de saisir les relations à longue terme entre les différentes parties de l'image. Cette méthode innovante a démontré des performances exceptionnelles dans le domaine de la vision par ordinateur.

1.5 Vision Transformers (ViT)

Inspiré par son succès dans le domaine du traitement du langage naturel, plusieurs tentatives ont été faites pour adapter l'architecture de Transformer aux traitements d'images, La première tentative réussie a été faite par l'équipe Google brain. Ils ont présenté Vision Transformers (ViT) (Dosovitskiy et al. 2020), Semblable à la séquence d'incorporation de mots utilisée lors de l'application de Transformers au texte, ViT représente une image d'entrée comme une séquence de patchs et tente de prédire ses étiquettes de classe par le biais d'une architecture d'encodeur uniquement. Sur la figure 1.15 un aperçu du modèle vision transformer.

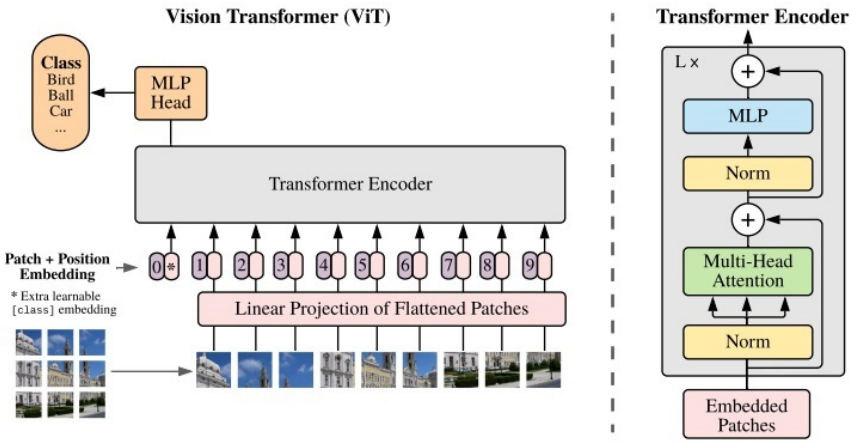

Fig. 1.15 : Aperçu du modele Vision Transformer(ViT) (Dosovitskiy et al. 2020).

1.5.1 L'architecture de (ViT)

La structure de l'architecture Transformer peut être aisément adaptée aux tâches de vision par ordinateur, comme le montre (Dosovitskiy et al. 2020). Cette architecture adaptée est nommée Vision Transformer (ViT). Elle permet notamment de découper les images en patchs non superposés, traités ensuite comme des tokens. L'avantage majeur de l'attention dans les tâches de vision est sa capacité à évaluer l'importance relative de différentes parties d'une image, favorisant ainsi une approche globale plutôt que locale, contrairement aux convolutions. Il n'est pas obligatoire d'utiliser une structure encodeur-décodeur : (Dosovitskiy et al. 2020). emploient une structure similaire à celle de l'encodeur, soit l'intégration des entrées suivie d'une succession de blocs MSA (Multi-Head Self-Attention) et MLP (Multi-Layer Perceptron). Enfin, un "MLP Head" est ajouté au terme du dernier bloc.

- La couche Multi-Head Self-Attention (MSA) concatène linéairement toutes les sorties d'attention pour correspondre aux dimensions appropriées. Les Multi-Head Self-Attention facilitent l'apprentissage des dépendances locales et globales au sein d'une image.

- Couche de Multi-Layer Perceptron (MLP) : Cette couche contient deux couches avec l'unité linéaire d'erreur gaussienne (GELU).

- MLP Head : Un MLP qui prend en entrée un jeton spécial appelé "jeton [classe]" ([class] Token), résultant du dernier bloc d'attention, et le mappe à la classe prédite pour l'entrée.

Tokenisation d'entrée et encodage de position : Considérer chaque pixel comme un jeton et calculer l'attention entre eux serait irréalisable, car l'opération d'attention requiert une complexité de $O(n^2)$ tant en mémoire qu'en temps d'exécution. C'est pourquoi (Dosovitskiy et al. 2020) segmentent l'image en patchs d'entrée non chevauchants. Ces patchs sont convertis en jetons, réduisant ainsi le nombre total d'entrées nécessaires à l'opération d'attention. Ils sont ensuite intégrés en les projetant linéairement en vecteurs de dimensionnalité R^d_{modle}. En pratique, cette projection est réalisée efficacement via une couche de convolution avec un pas égal à la taille du patch. Une autre distinction de l'architecture de (Dosovitskiy et al. 2020) par rapport au Transformer original (Vaswani et al. 2017) est que les codages de position sont apprenables plutôt que fixes et prédéterminés.

Class token : Après la génération des jetons d'entrée, un vecteur est inséré au début de la séquence de jetons et traité avec ces derniers. Ce vecteur, nommé jeton [classe], est par la suite extrait des résultats du dernier bloc d'attention et transmis à la tête de classification qui détermine la classe prédite pour l'entrée. L'état initial de ce vecteur c'est-à-dire avant traitement est un paramètre apprenable du modèle. Le jeton de classe est conçu pour focaliser l'attention sur certaines parties.

1.5.2 Cas d'utilisation et applications réels du VITs

Les vision transformes ont de nombreuses applications dans les tâches de reconnaissance d'images populaires telles que la détection d'objets, la segmentation, la classification d'images et la reconnaissance d'actions. De plus, les ViT sont appliqués dans la modélisation générative et les tâches multi-modèles, y compris l'ancrage visuel, la réponse visuelle aux questions et le raisonnement visuel. La prévision vidéo et la reconnaissance d'activité sont toutes des parties du traitement vidéo qui nécessitent ViT. De plus, l'amélioration de l'image, la colorisation et la super-résolution de l'image utilisent également des modèles ViT. Enfin, les ViT ont de nombreuses applications dans l'analyse 3D, telles que la segmentation et la classification des nuages de points (Web 01).

1.6 Approches de l'apprentissage par transfert

L'apprentissage par transfert peut être appliqué aux problèmes de classification. Il existe deux méthodes couramment utilisées :

- **Modèle pré-entraîné**

  Un modèle pré-entraîné est un modèle d'apprentissage, souvent un réseau neuronal profond important, qui a été entraîné pour une tâche générique sur un vaste ensemble de données. Le modèle pré-entraîné est ensuite sauvegardé et utilisé ultérieurement pour résoudre un problème similaire grâce à l'application de l'apprentissage par transfert. Essentiellement, , il utilise les connaissances accumulées pendant son entraînement initial pour s'adapter à des tâches spécifiques sans commencer à apprendre de zéro. Parmi les exemples de modèles pré-entraînés de grande taille, on peut mentionner MobileNet (pour la vision par ordinateur), BERT, ainsi que la série des modèles linguistiques GPT-x. Il y a aussi des modèles avec l'architecture CNN tels que Resnet, VGG, AlexNet, etc., et des modèles d'architecture ViT comme ViT base, Swin T et Diet-ViT de notre choix.

- **Fine-tuning**

  Le fine-tuning est une technique courante d'apprentissage par transfert. Elle consiste à prendre un modèle pré-entraîné et à le continuer à s'entraîner sur des données spécifiques à la tâche cible. Par exemple, un modèle de traitement du langage naturel (NLP) pré-entraîné sur un vaste corpus textuel peut être fine-tuné avec des données spécifiques à un secteur (telles que des critiques de produits) afin d'en optimiser les performances pour cette tâche spécifique.

1.7 Conclusion

Dans ce chapitre, nous avons d'abord examiné les différentes méthodes de l'intelligence artificielle, telles que l'apprentissage automatique, l'apprentissage profond, l'apprentissage par transfert, ainsi que les Transformers et leurs divers types. Ces techniques d'apprentissage sont utilisées dans de nombreux domaines, notamment l'imagerie médicale, qui sera le sujet du chapitre suivant.

Le prochain chapitre exposera une étude théorique sur l'imagerie médicale, suivie d'une présentation de l'état de l'art concernant la classification des images médicales.

Chapitre 2

État de l'art sur la classification des images médicales

2.1 Introduction

Dans le domaine médicale, l'information joue un rôle très important en particulier l'image médicale. Cette dernière nécessite un traitement spécifique et particulière afin d'avoir le bon diagnostic et dans un temps acceptable. Malgré les avancées technologiques dans le domaine de la médecine, notamment en imagerie médicale, l'analyse de ses images continue de faire l'objet de recherches et d'intérêt constants.

Ce chapitre portera une étude théorique sur l'imagerie médicale et présentera un état de l'art sur la classification des images médicales.

2.2 Définition d'imagerie médicale

L'imagerie médicale est définie comme étant l'ensemble des moyens d'acquisition et de restitution d'images du corps humain en se basant sur divers phénomènes physiques tels que la résonance magnétique nucléaire, la réflexion d'ondes ultrasons, l'absorption des rayons X ou la radioactivité . Grâce au progrès de l'informatique, ces techniques ont révolutionné la médecine en permettant de visualiser l'anatomie, le métabolisme et la physiologie du corps humain. L'objectif de l'imagerie médicale est non seulement d'offrir un meilleur diagnostic des maladies, suivre leur évolution, découvrir le fonctionnement de certains organes mais d'offrir aussi de nouveaux espoirs de traitement pour plusieurs pathologies. Développées comme moyens de diagnostique, les techniques de l'imagerie médicale sont aussi largement utilisées dans la recherche biomédicale afin de mieux comprendre le fonctionnement de l'organisme. Elles sont également appliquées dans différents domaines tels que la sécurité et l'archéologie (Dali 2022).

2.3 Objectif de l'imagerie médicale

L'imagerie médicale a pour objectif de produire une représentation visuelle compréhensible d'une information à caractère médical pour (Web 02) :

- **Aide au diagnostic** :

  L'imagerie médicale peut être utilisée en première intention, c'est le cas dans le dépistage systématique des cancers du sein par mammographie (radiographie), ou pour confirmer ou infirmer un diagnostic supposé. L' imagerie par résonance magnétique (IRM) peut apporter des arguments en faveur du diagnostic de sclérose en plaques ou de maladie d'Alzheimer tandis que le scanner mettra en évidence un rétrécissement des artères coronaires en cas de douleurs thoraciques ou d' infarctus du myocarde.

- **Évaluer la sévérité d'une maladie** :

  Par l'imagerie, le diagnostic est affiné. Grâce à la scintigraphie on peut, par exemple, repérer des métastases et donc mesurer le niveau de dissémination d'un cancer dans l'organisme. En cardiologie, la scintigraphie dite de perfusion évalue le débit sanguin

au niveau du cœur au repos, ou lors d'un effort, afin de statuer sur le niveau de dysfonctionnement de certaines artères

- **Aide à l'intervention** :

  Des ponctions effectuées chez des patients le sont parfois sous échographie afin de bien visualiser la zone à prélever, notamment lorsqu'elle n'est pas palpable. Des injections d'anti-inflammatoires ou des drainages peuvent également être pratiqués avec l'aide de l' imagerie.

- **aide à la prise en charge et au suivi thérapeutique** :

  La comparaison de clichés pris à des temps différents offre au corps médical un moyen de suivre l'évolution d'une maladie ou encore d'une fracture osseuse. Très utilisée en cancérologie, la scintigraphie permet de vérifier l'efficacité d'un traitement en visualisant le niveau d'activité des cellules tumorales ou de détecter des métastases et poser ainsi l' indication d'une chimiothérapie. Dans 30% à 40% cas le support de l' imagerie a permis de modifier l'attitude thérapeutique, au bénéfice des patients.

- **Améliorer les connaissances** :

  L'imagerie a également contribué à faire avancer à grands pas la connaissance de l'activité cérébrale chez l'homme. Ainsi, grâce à l' IRM fonctionnelle on en sait davantage sur les mécanismes de l'adduction ou de maladies mentales telles que l'autisme.

2.4 Les différents types d'imagerie médicale

Il existe différents types d' imagerie médicale qui reposent sur l'utilisation des rayons X, des ultrasons, du champ magnétique ou de la radioactivité naturelle ou artificielle.

2.4.1 La radiographie

La radiographie a été découverte il y a plus d'un siècle, elle est basée sur l'utilisation des rayons X. Ces rayons sont capables de traverser les tissus de façon plus ou moins importante selon leur densité. Ainsi, devant le corps à radiographier est située une source de rayons X et à l'arrière du corps est situé un détecteur. Le corps est traversé par les photons émis qui sont plus ou moins absorbés par les tissus croisés sur leur chemin. Cela assure la différenciation entre les os et les muscles sur le cliché final (Dali 2022).

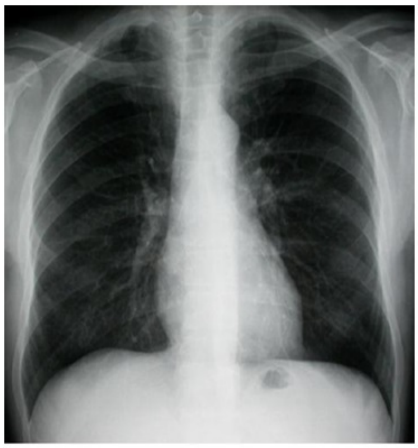

Fig. 2.1 : Image Radiographie (Chiter et al. 2022).

2.4.2 La mammographie

Une mammographie, également appelée mastographie, est une étude radiologique des seins. Il offre la possibilité de prendre des images de l'intérieur du sein en utilisant des rayons X, ce qui permet de repérer certaines anomalies. L'examen mammographique est effectué dans deux cas : pour un dépistage ou un diagnostic précoce du cancer mammaire. Quoi qu'il en soit, deux photographies (photos) par sein sont prises, une de face et une en oblique, permettant ainsi de comparer les deux côtés de chaque sein. La plupart du temps, elle constitue le premier examen d'imagerie (Chiter et al. 2022).

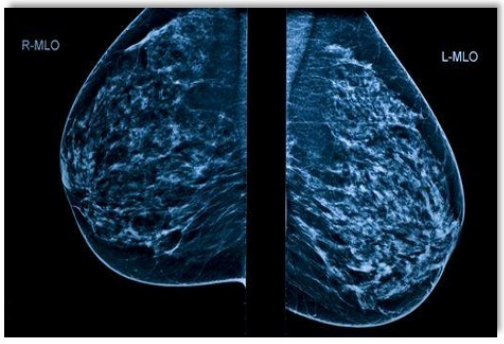

Fig. 2.2 : Image Mammographie (Chiter et al. 2022).

2.4.3 La Tomodensitométrie (scanner)

Le scanner (appelé aussi tomodensitométrie) est également basé sur l'utilisation d'une source de rayons X et d'un détecteur de part et d'autre du corps en question. Il permet une rotation simultanée de la source émettrice de rayons X et du détecteur autour du corps pour l'obtention d'images 3D. La zone à explorer est balayée et des images en coupes fines, ou "tranches" de l'organisme sont reconstituées, ce qui permet la détermination précise de la localisation et l'étendue d'une lésion (Dali 2022).

Fig. 2.3 : Imagerie par Scanner (BARKA et al. 2021).

2.4.4 L'échographie ultrasonore

L'échographie ultrasonore est une technique d'imagerie médicale basée sur l'exposition du corps à des ondes ultrasonores et sur la réception de leur écho. Un faisceau d'ultrasons est envoyé par une sonde dans la région du corps à explorer. Ces ondes sonores sont réfléchies d'une manière plus ou moins importante. Ces ondes traversent les tissus et y font écho d'une façon différente selon leur densité : plus un tissu est dense, plus l'écho est puissant. On arrive à visualiser les organes observés suite au traitement de ces échos. Ainsi, lors d'une échographie réalisée dans le cas d'une grossesse par exemple, on arrive à différencier les organes du fœtus, de son squelette, du liquide amniotique, etc (Dali 2022).

Fig. 2.4 : Image d'échographie (BARKA et al. 2021).

2.4.5 Imagerie par Résonance Magnétique (IRM)

L'imagerie par Résonance Magnétique (IRM) a la possibilité de visualiser des détails qui sont invisibles sur les radiographies standards, le scanner ou l'échographie . Il s'agit d'une technique d'imagerie qui se base sur les propriétés magnétiques des molécules d'eau qui remplit le corps humain à plus de 80%. Les atomes d'hydrogène contenus dans les molécules d'eau possèdent un "moment magnétique", appelé également spin, qui fonctionne comme un aimant. Une bobine permet à la machine IRM de créer un champ magnétique puissant (B0). Le patient est situé au centre de ce champ magnétique. La totalité des molécules d'eau qui composent le corps vont s'orienter suivant B0. Une antenne placée sur la région du corps étudiée va assurer l'émission et la réception de certaines fréquences. Lors de l'émission, les molécules vont basculer par la fréquence induite dans un plan perpendiculaire à B0. Suite à l'arrêt de l'émission de l'antenne, les molécules retrouvent leur position initiale et se mettent à émettre à leur tour une fréquence captée par l'antenne. Des logiciels permettent ensuite d'analyser et de traiter cette fréquence comme un signal électrique. Le signal est différent selon la contenance d'eau dans les tissus (Dali 2022).

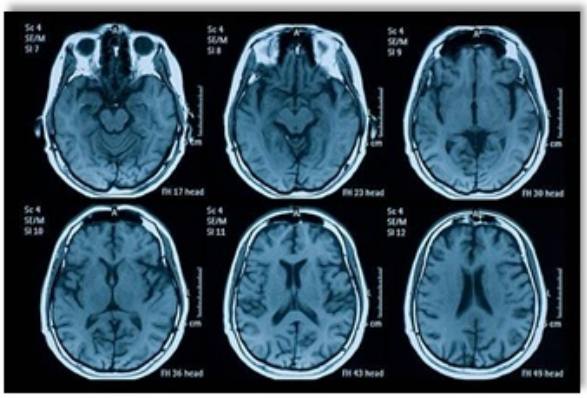

Fig. 2.5 : Image IRM du cerveau (BARKA et al. 2021).

2.4.6 La scintigraphie mono photonique

La scintigraphie mono photonique est une technique qui fait intervenir la médecine nucléaire. Cette technique d'imagerie repose sur l'utilisation d'un traceur radioactif administré au patient et une caméra sensible aux rayons gamma. Le traceur, marqué par un atome radioactif émettant des photons dans toutes les directions, va être capté par l'organe à analyser. Le corps du patient va être traversé par les photons émis, jusqu'à la gamma caméra. Cette caméra est équipée d'un collimateur ne laissant passer que les rayons parallèles aux zones aménagées à cet effet. L'objectif est de délimiter les points d'émission des photons. Autour du patient tourne la gamma caméra pour obtenir des images 3D de l'organe étudié suite à une reconstitution informatique (Dali 2022)

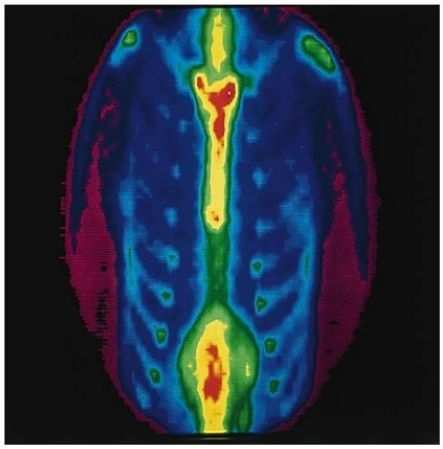

FIG. 2.6 : Image Scintigraphique normal du foie (BARKA et al. 2021).

2.4.7 L'endoscopie

Une endoscopie, aussi appelée fibroscopie, est une opération médicale qui permet à un médecin d'examiner l'intérieur d'un organe ou d'une cavité corporelle en introduisant un endoscope. Un endoscope est un tube étroit et souple rempli de fibres optiques, relié à l'extrémité par une lumière et une petite caméra (BARKA et al. 2021).

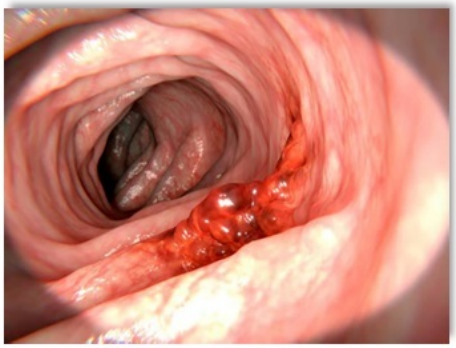

FIG. 2.7 : Image d'endoscopie de colon (CHITER et al. 2022).

2.5 Applications de l'imagerie médicale

- La première est la procédure guidée par imagerie :il s'agit d'une représentation dont le but est de diriger les processus thérapeutiques et minimiser leurs dommages. Dans ce cas, il y a par exemple la biopsie pulmonaire guidée par TDM qui est une méthode efficace, complémentaire de la bronchoscopie souple, pour le diagnostic du cancer broncho-pulmonaire. L'imagerie avec la méthode susmentionnée vise à diagnostiquer une tumeur, à surveiller les changements de traitements au fur et à mesure qu'ils surviennent et à surveiller l'échec ou le succès du traitement (AHMAD et al. 2022).
- L'autre domaine d'application très connu est l'imagerie fonctionnelle où selon la nature des recherches, on distingue celles qui fournissent des propriétés structurelles de

la zone étudiée (IRM, Rayons X, ...), de celles qui restituent des aspects fonctionnels (TEP, IRM Fonctionnelle...). L'un des cas possibles grâce à l'imagerie médicale est la surveillance de l'activité cérébrale ainsi que le moment de sa réaction aux stimuli moteurs, émotionnels et mentaux. En effet, l'activité de différents centres cérébraux est caractérisée par une consommation accrue d'oxygène et, à l'aide de l'IRM, les changements causés par l'augmentation de la consommation peuvent être observés, avec le moment de leur apparition (Ahmad et al. 2022).

L'imagerie médicale a donc beaucoup évolué en très peu de temps. Souvent essentielle lors de l'élaboration d'un diagnostic, elle se démarque par sa poly-valence et la vaste étendue de ses champs d'application.

2.6 Etat de l'art sur la classification des images médicales

Cette partie vise à donner une vue d'ensemble des travaux de recherche réalisés dans le domaine de la classification des images médicales.

2.6.1 Bases de données

Une base de données, aussi connue sous le nom de jeu de données ou dataset, est un ensemble de données qui sert à entraîner et à tester un modèle d'apprentissage automatique. Il s'agit d'un ensemble organisé ou non organisé d'échantillons de données qui servent à résoudre un problème particulier d'apprentissage automatique. Dans le jeu de données, on retrouve habituellement un ensemble de caractéristiques ou de variables qui servent d'entrées pour l'algorithme d'apprentissage automatique et qui permettent de générer une sortie ou une prédiction.

2.6.1.1 Chest X-Ray Images (Pneumonia)

L'ensemble de données est organisé en 2 dossiers (train, test) et contient des sous-dossiers pour chaque catégorie d'image (Pneumonia/Normal). Il y a 5 856 images X-Ray (JPEG) et 2 catégories (Pneumonia/Normal). Le nombre des images Pneumonie est 4273 images et le nombre des images Normal est 1583 images. Les radiographies thoraciques (avant-arrière) ont été sélectionnées à partir de cohortes rétrospectives de patients pédiatriques âgés d'un à cinq ans du Centre médical des femmes et des enfants de Guangzhou. Toutes les radiographies thoraciques ont été effectuées dans le cadre des soins cliniques de routine des patients.Pour l'analyse des images de rayons X thoraciques, tous les radiographies thoraciaires ont été examinées pour le contrôle de la qualité en supprimant toutes les scans de mauvaise qualité ou non lisibles. Les diagnostics pour les images ont ensuite été évalués par deux médecins experts avant d'être clarifiés pour la formation du système d'IA. Afin de tenir compte des erreurs de notation, un troisième expert a également vérifié l'ensemble d'évaluation(Web 03).

2.6.1.2 Brian Tumor Dataset

L'ensemble est une collection complète de données d'imagerie médicale conçues aux fins de la détection et de la classification des tumeurs du cerveau. Il englobe 4600 images par résonance magnétique (IRM) du cerveau humain, les catégorisant en deux classes principales : Brain Tumor (tumeurs cérébrales présentes) et Healthy (aucun tumeur cérébrale). Cet ensemble de données est destiné à être utilisé dans l'apprentissage automatique, l'imagerie médicale et la recherche en soins de santé pour aider à élaborer des modèles de diagnostic et de classification avancés. Composition des ensembles de données : Classe Brain Tumor (présentation de tumeurs cérébrales) : Cette catégorie comprend des scans IRM de patients. Ces scannings montrent la présence de tumeurs à différents stades et endroits dans le cerveau 2513 images. Classe Healthy (Pas de tumeur cérébrale) : Cette classe contient des scans IRM d'individus sans tumeurs du cerveau détectables. Ces scannings servent de données de référence pour les structures et conditions cérébrales saines 2087 images.(Web 04)

2.6.1.3 Breast Ultrasound Images Dataset

Les données recueillies à la base comprennent des images à ultrasons du sein chez des femmes âgées de 25 à 75 ans. Ces données ont été recueillies en 2018. Le nombre de patients est de 600 femmes. L'ensemble de données se compose de 780 images avec une taille moyenne d'image de 500*500 pixels. Les images sont en format PNG. Les images de vérité sont présentées avec des images originales. Les images sont classées en trois classes, qui sont normales, bénignes et malignes (Web 05).

2.6.1.4 Padchest

L'ensemble de données Padchest (a large chest x-ray image dataset) se compose de 160 868 images de radiographie thoracique étiquetées de 69 882 patients, obtenues dans une seule institution entre 2009 et 2017. Les patients ont eu une moyenne de 1,62 études de radiographie thoracique effectuées à différents moments. Chaque étude contient une ou plusieurs images correspondant à différentes vues de position, principalement P-A et latérale, et est associée à un seul rapport radiographique décrivant les résultats de toutes les vues en position dans une section de texte commune. La moyenne des images de rayons X thoraciques était de 2,37 par patient. L'ensemble de données généré fournit deux types de champs pour chaque image à rayons chest-x : les champs avec le suffixe DICOM 6 contiennent les valeurs du champ d'origine dans la norme DICom, et les autres champs 5 enrichissent la base de données PadChest avec des informations supplémentaires traitées (Bustos et al. 2020).

2.6.1.5 Digital Database for Screening Mammography

Le DDSM est une base de données mammographique complète qui a été créée dans le cadre d'un projet de collaboration entre de nombreuses universités de médecine et d'ingénierie médicale de premier plan . Le DDSM se compose d'un nombre total de 2620 cas avec

des classes normales, cancéreuses et bénignes. Les mammogrammes ont été scannés par les dispositifs médicaux (DBA, LUMYSIS et HOWTEK) avec des résolutions différentes (Altan 2020).

2.6.1.6 Ayyaz et al. 2021

Ayyaz et al. (2021) collecte l'ensemble de données de vidéos endoscopiques d'un spécialiste médicale a Sargodha. Au départ, l'ensemble de données comprenait 118 vidéos de 80 patients. ils ont sélectionné 50 vidéos en fonction de la qualité et les avons converties en cadres en utilisant MATLAB . Après cela, ils ont étiqueté l'ensemble de données avec l'aide d'un spécialiste médicale et assures que le groupe de données final était dans une forme normalisée. L'ensemble de données finales était composé de cinq classes. Les noms des classes étaient gastrite, ulcère, saignement, esophagite, et saine. Il y avait respectivement 527, 519, 514, 519, et 511 images de chaque classe. En outre, ils divisent les données en ensembles d'entraînement et d'essai en utilisant une technique de validation croisée à dix pattes (Ayyaz et al. 2021).

2.6.2 Quelques travaux récents

Nous présentons dans cette section quelques travaux récents dans le domaine de la classification des images médicales.

2.6.2.1 Seetha et al. 2018

Dans cette étude, Seetha and Raja proposent une détection automatique des tumeurs cérébrales en utilisant la classification par réseaux de neurones convolutifs (CNN). Une architecture plus profonde a été conçue avec l'utilisation de petits noyaux, et un faible poids a été attribué aux neurones. Les résultats expérimentaux révèlent que leurs modèle CNN atteint un taux de précision de 97,5 %, tout en présentant une complexité réduite par rapport aux autres méthodes avancées.

2.6.2.2 Swati et al. 2019

Dans ce travail Swati et al ont utilisent un modèle de CNN profond pré-entraîné et ont proposent une stratégie de réglage de blocs basée sur l'apprentissage des transferts. La méthode proposée est évaluée sur un ensemble de données de référence de résonance magnétique améliorée par contraste pondérée T1. la méthode est plus générique car elle n'utilise aucune caractéristique artisanale, nécessite un pré-traitement minimal et peut atteindre une précision moyenne de 94,82 sous cinq fois la validation croisée. ils comparent les résultats non seulement avec l'apprentissage automatique traditionnel, mais aussi avec les méthodes d'apprenant en profondeur en utilisant CNNs. Les résultats expérimentaux montrent que la méthode proposée dépasse la classification de pointe sur le groupe de données CE-MRI.

2.6.2.3 Çınarer et al. 2019

Dans cette étude, les performances des méthodes de classification des tumeurs pour la classification des caractéristiques de l'image cérébrale IRM comme n/a, multifocale, multicentrique et gliomatose ont été analysées. Dans le processus de classification, les propriétés statistiques des images d'entrée ont été analysées et les données ont été systématiquement divisées en différentes catégories. Ces données ont été testées avec des algorithmes d'apprentissage automatique KNN (k plus proche voisin), RF (forêt aléatoire), SVM (machines à vecteurs de support) et LDA (analyse discriminante linéaire). L'algorithme SVM (machines à vecteurs de support) avec un taux de précision de 90 % s'est avéré meilleur que les autres algorithmes.

2.6.2.4 Ramdlon et al. 2019

Le niveau de précision dans le diagnostic du type de tumeur par les résultats de l'IRM est nécessaire pour établir un traitement médical approprié. Les résultats de l'IRM peuvent être examinés par calcul à l'aide de la méthode du K-Nearest Neighbor, une application scientifique fondamentale et une technique de classification du traitement d'images. Le système de classification des tumeurs est conçu pour détecter les tumeurs et les œdèmes dans les séquences d'images T1 et T2, ainsi que pour marquer et classer le type de tumeur. L'interprétation des données d'un tel système provient uniquement de la section axiale des résultats d'IRM, qui est classée en trois classes : astrocytome, glioblastome et oligodendrogliome. Pour détecter la zone tumorale, une technique de traitement d'image de base est utilisée, comprenant l'amélioration de l'image, la binarisation de l'image, l'image morphologique et le bassin versant. La classification des tumeurs est appliquée après le processus de segmentation de la fonction d'extraction de forme. Les résultats de la classification des tumeurs obtenus étaient de 89,5 %, ce qui permet de fournir des informations plus claires et plus spécifiques sur la détection des tumeurs.

2.6.2.5 Saraiva et al. 2019

Cet article décrit une comparaison de deux réseaux neuronaux, le perceptron multilayer et le réseau neuronal, pour la détection et la classification de la pneumonie. La base de données utilisée était l'ensemble de données Chest-X-Ray fourni par (Kermany et al., 2018) avec un total de 5840 images, avec deux classes, normale et avec pneumonie. Pour valider les modèles utilisés, une validation croisée de k-fold a été utilisée. Les modèles de classification étaient efficaces, avec une précision moyenne de 92,16% avec le Perceptron Multilayer et de 94,40% avec la Convolution Neural Network.

2.6.2.6 Amin et al. 2020

Dans ce document, une méthode automatisée est proposée pour distinguer facilement entre la résonance magnétique cancéreuse et non-cancéreuse (IRM) du cerveau. Différentes techniques ont été appliquées pour la segmentation des lésions candidates. Ensuite, un ensemble de caractéristiques est choisi pour chaque lésion du candidat en utilisant la forme,

la texture et l'intensité. À ce stade, le classificateur Support Vector Machine (SVM) est appliqué avec différentes validations croisées sur les caractéristiques définies pour comparer la précision du cadre proposé. La méthode proposée est validée sur trois ensembles de données de référence tels que Harvard, RIDER et Local. La méthode a obtenu une précision moyenne de 97,1 %, une zone sous la courbe de 0,98, une sensibilité de 91,9 % et une spécificité de 98 %. Il peut être utilisé pour identifier la tumeur plus précisément en moins de temps de traitement par rapport aux méthodes existantes.

2.6.2.7 Jain et al. 2020

Ce travail présente des modèles de Réseau Neural Convolutionnel pour détecter la pneumonie à l'aide d'images à rayons X. Plusieurs réseaux neuronaux convolutifs ont été formés pour classer les images de rayons X en deux classes, à savoir la pneumonie et la non-pneumonie, en modifiant divers paramètres, hyperparametres et nombre de couches convolutives. Six modèles ont été formés. Les modèles premier et deuxième sont composés de deux et trois couches convolutives, respectivement. Les autres modèles sont les VGG16, VGG19, ResNet50 et Inception-v3. Les premiers et les deuxièmes modèles ont rapporté une précision de validation de 85,26% et 92,31% respectivement. La précision de VGG16, VGG19, ResNet50 et Inception-v3 est de 87,28%, 88,46%, 77,56% et 70,99% respectivement.

2.6.2.8 Jiang et al. 2021

Cette étude présente un modèle combiné ViT-CNN pour classifier les images de cellules cancéreuses et les images de cellules saines, contribuant ainsi au diagnostic de la leucémie aiguë lymphob lastique. Le modèle ViT-CNN associe le modèle de transformateur visuel et le réseau de neurones convolutif (CNN) pour extraire les caractéristiques des images cellulaires de deux façons distinctes, améliorant les résultats de classification. L'ensemble de données utilisé est déséquilibré et bruyant, pour lequel ils proposent une méthode d'amélioration des données appelée échantillonnage aléatoire à amélioration différentielle DERS, créant un nouvel ensemble de données équilibré et employant la fonction de perte symétrique d'entropie croisée pour minimiser l'impact du bruit. La précision de classification du modèle ViT-CNN sur le jeu de données de test ISBI 2019 a atteint 99,03, démontrant par comparaison expérimentale que le modèle ViT-CNN surpasse les autres modèles existants.

2.6.2.9 Wang et al. 2023

Wang et al ont introduisent un modèle nommé PneuNet, basé sur le Vision Transformer (ViT), pour un diagnostic précis via des images radiographiques pulmonaires, en utilisant une attention basée sur les canaux. Contrairement aux approches traditionnelles, leurs attention multi-têtes se concentre sur les patchs de canaux plutôt que sur les patchs de caractéristiques. Ce document détaille des techniques spécifiquement adaptées à l'usage médical des réseaux neuronaux profonds et du ViT. leurs tests approfondis révèlent que PneuNet peut atteindre une précision de 94,96% dans la classification tripartite.

2.6.2.10 Singh et al. 2024

Ce document examine en détail une méthode de pointe pour la détection de la pneumonie mise en œuvre sur l'architecture Vision Transformer (ViT) sur un ensemble de données publiques de rayons X thoraciques disponibles sur Kaggle. Afin d'acquérir des relations spatiales et du contexte mondial à partir d'images à rayons X thoraciques, le cadre proposé déploie le modèle ViT, qui intègre des mécanismes d'auto-attention et une architecture de transformateur. Selon leurs expériences avec le cadre proposé basé sur Vision Transformer, il atteint une précision supérieure de 97,61%, une sensibilité de 95%, et une spécificité de 98% dans la détection de la pneumonie à partir de rayons X thoraciques. Le modèle ViT est préférable pour capturer le contexte global, comprendre les relations spatiales et traiter des images qui ont des résolutions différentes. Le cadre établit son efficacité en tant que solution robuste de détection de la pneumonie en dépassant les architectures basées sur les réseaux neuronaux convolutifs (CNN).

2.6.2.11 Pacal 2024

cette étude, présente une approche avancée de l'apprentissage profond basée sur le transformateur Swin. La méthode proposée introduit un nouveau module Hybrid Shifted Windows Multi-Head Self-Attention (HSW-MSA) ainsi qu'un modèle réévolué. les auteurs ont évalué le modèle Proposed-Swin sur un ensemble de données MRI du cerveau publiquement disponible. Les performances du modèle sont améliorées grâce à l'application de techniques d'apprentissage de transfert et d'augmentation des données pour une formation efficace et robuste. Le modèle proposé-Swin atteint une précision remarquable de 99,92%, dépassant les précédents modèles de recherche et d'apprentissage profond.

2.6.2.12 Huang et al. 2024

Dans le diagnostic des maladies thyroïdiennes, le diagnostic de nodules thyroïdes bénins et malins basé sur la classification TI-RADS est fortement influencé par le jugement subjectif des ultrasonographes, et en même temps, il apporte également une charge de travail extrêmement lourde aux ultrasonographies. Pour y remédier, les auteurs ont proposé Swin-Residual Transformer (SRT) dans ce document, qui intègre les blocs résiduels et les pertes triplées dans Swin Transformer (SwinT). Il améliore la sensibilité aux caractéristiques globales et localisées des nodules thyroïdiens et distingue mieux les petites différences de caractère. Dans leurs expériences exploratoires, le modèle SRT atteint une précision de 0,8832 avec une AUC de 0,8660, dépassant les modèles de réseaux neuronaux convolutionnaires (CNN) et de transformateurs. En outre, les expériences d'ablation ont démontré l'amélioration des performances dans la tâche de classification des nodules thyroïdiens après l'introduction de blocs résiduels et de triple perte. Ces résultats valident le potentiel du modèle SRT proposé pour améliorer le diagnostic des ultrasons des nodules thyroïdiens. Il fournit également une garantie viable pour éviter l'échantillonnage excessivement perforé des nodules thyroïdiens dans le diagnostic clinique futur.

2.7 Conclusion

Dans ce chapitre, nous nous sommes intéressés aux images médicales en particulier. Nous avons présenté l'imagerie médicale, ses objectifs, leurs types, tels que le scanner, la radiographie, l'IRM, etc., et ses applications. Enfin, nous avons présenté un état de l'art sur la classification des images médicales, en mettant en lumière des travaux récents. Nous avons observé qu'ils s'appuient sur le deep learning traditionnel tel que les CNN, et que ces dernières années, il y a une tendance à introduire les transformers, compte tenu de leur succès en NLP et de leurs résultats performants.

Dans le chapitre suivant, nous appliquerons les Vision Transformers pour concevoir un système de classification des images médicales.

Chapitre 3

Un système basé-transformers pour la classification des images médicales

3.1 Introduction

Certaines maladies nécessitent l'utilisation de l'imagerie médicale et l'expertise essentielle pour l'interprétation des images, un processus susceptible d'entraîner des inexactitudes diagnostiques. Par conséquent, il est impératif d'améliorer la précision et la rapidité du diagnostic par imagerie médicale pour obtenir de meilleurs résultats cliniques. L'objectif principal de notre étude est de construire un modèle adapté à la classification des images médicales avec une marge d'erreur minimale. Par conséquent, le présent chapitre dévoile un système de classification des images médicales basé sur une architecture Vision Transformers (ViT).

Ce chapitre est divisé en deux parties principales. Premièrement, nous donnons un aperçu général sur notre système suivi d'une exploration des différentes étapes de conception de notre modèle ViT. Par la suite, nous présentons les expérimentations menées et les résultats qui en ont résulté.

3.2 Architecture globale du système

Dans cette partie, nous exposons notre système de classification des images médicales. Avant de commencer, nous abordons les spécificités du système en présentant d'abord son architecture globale. Le système de classification des images médicales comporte des étapes fondamentales au sein de sa structure, comme le montre le schéma illustré à la Figure 3.1.

Au début de notre approche, nous utilisons une base d'apprentissage qui renferme les images indispensables pour l'entraînement du modèle ViT. Par la suite, nous effectuons le prétraitement des données afin d'améliorer la qualité des images. Ensuite, le modèle est entraîné sur une partie des données d'apprentissage traitées. Ce modèle produit sera ensuite employé lors de la phase de test ou de classification afin de classer les nouvelles images en deux catégories : celles qui sont normales et celles qui présentent une anomalie.

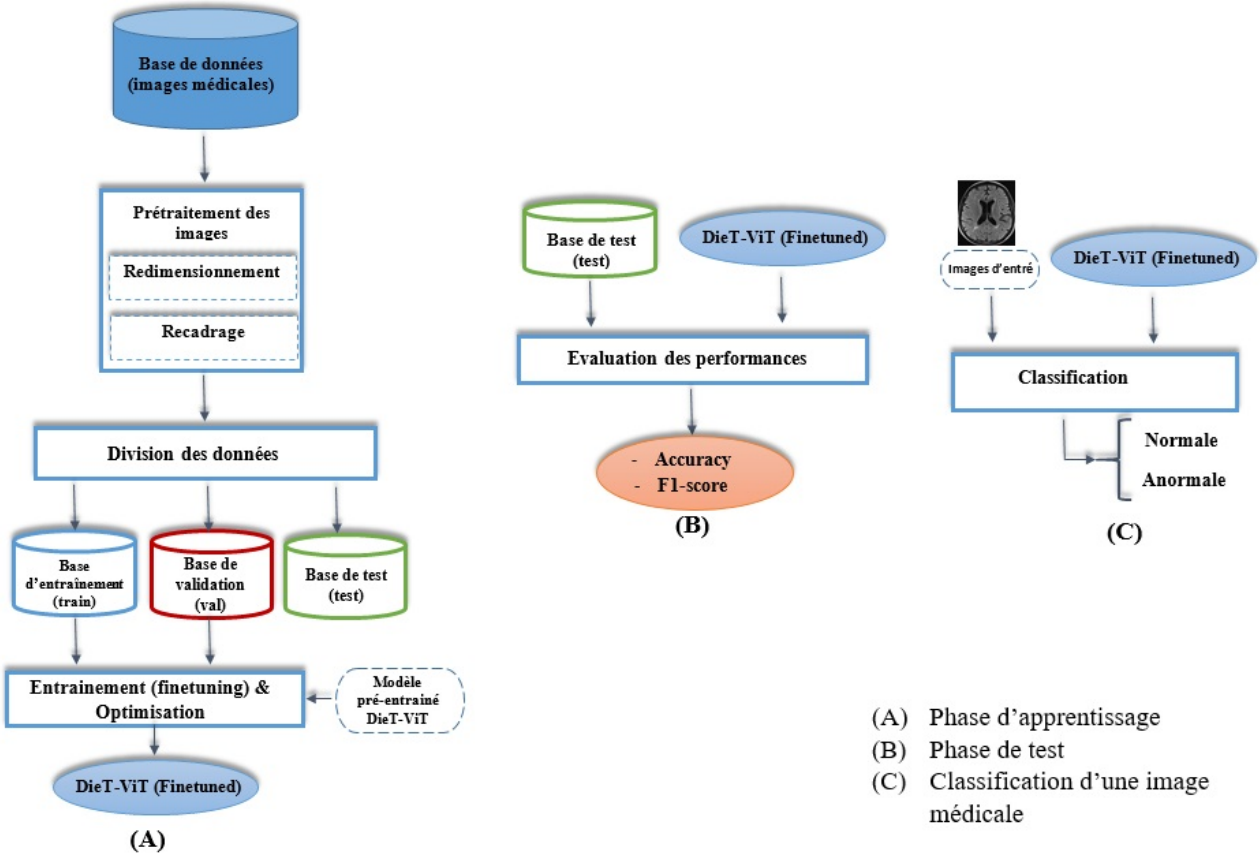

FIG. 3.1 : Architecture de notre système

3.3 Présentation détaillée du système

Pour construire notre modèle, il est essentiel de disposer de base de données qui décrivent les entités faisant l'objet de l'étude. nous avons utilisé deux bases de données qui seront présentée en détail dans la sous-section 3.4.1.

La conception de notre système se fait en deux phase principales : la phase d'apprentissage où notre modèle ViT est entraîné sur nos données suivant une approche par Transfer Learning, et la phase de test, où le modèle entraîné est évalué en utilisant différents métriques, à savoir, l'accuracy et le F1-score.

3.3.1 Phase d'apprentissage

L'apprentissage vise à créer un modèle ViT qui sera ensuite employé pour la classification d'images médicales. Il suit les étapes suivantes :

- **Prétraitement des données :** La qualité des données est essentielle dans le domaine de l'apprentissage automatique, car elle a un impact direct sur la réussite du projet. Ainsi, le traitement préliminaire des données est généralement perçu comme la première étape à réaliser dans un projet. Au sein de notre système, nous effectuons des tâches de prétraitement sur les images pour les préparer à l'entraînement du modèle. Parmi ces tâches, on peut effectuer des rotations aléatoires horizontales et verticales, ajuster la taille de l'image, effectuer un recadrage centré, convertir en tenseur et normaliser les valeurs des pixels. L'objectif de ces actions est d'accroître la variété des données, de préparer les images pour l'entraînement du modèle et de normaliser les valeurs des pixels afin d'améliorer la convergence de l'apprentissage.

- **Division des données :** L'ensemble de données a été subdivisé en trois sous-ensembles : l'ensemble d'entraînement, l'ensemble de validation et l'ensemble de test (train/val/test).

  - L'ensemble d'entraînement (train) : C'est une partie de l'ensemble de données utilisée lors de la phase d'apprentissage du modèle. Il sert à ajuster les paramètres du modèle pour qu'il puisse efficacement classer des nouvelles données jamais vues.
  - L'ensemble de validation (val) : une partie de l'ensemble de données utilisée lors de la phase de validation. Il sert à mesurer les performances du modèle à diverses étapes de l'entraînement et à ajuster les hyperparamètres du modèle, comme le pas d'apprentissage, afin d'optimiser ses performances en évitant les problèmes d'overfitting et d'underfitting.
  - L'ensemble de test (test) : une partie de l'ensemble de données utilisée pour tester et évaluer le modèle final. Il est utilisé pour mesurer les performances du modèle sur des données indépendantes et inconnues, afin de déterminer son efficacité dans des situations réelles

- **Création des data loaders :** A cette étape, nous mettons en place les dataloaders pour charger les données par batchs durant l'entraînement et l'évaluation du modèle. Cela simplifie le traitement par batchs des images et augmente l'efficacité du modèle.

- **Entraînement :** Le but principal d'un système de classification d'images médicales basé sur le modèle ViT est de créer un système capable de classer diverses caractéristiques et motifs présents dans les images médicales en se basant sur une approche d'apprentissage par transfert (Transfer Learning). Suivant cette approche, le modèle pré-entraîné DieT (une version optimisé de ViT) est adapté à nos données particulières en évitant à construire un modèle à zéro (Fine-tuning). Les principales étapes de l'entraînement de notre modèle ViT (DeiT) sont les suivantes :

  - **Initialisation du modèle ViT :** En premier lieu, nous initialisons le modèle DeiT Tiny ViT qui a été pré-entraîné sur l'ensemble de données ImageNet 2012, qui est une base de données d'images contenant plus d'un million d'images réparties en 1 000 classes différentes. Elle a été largement utilisée pour l'entraînement et l'évaluation des modèles de vision par ordinateur. Ensuite, nous effectuons la réentraînement des dernières couches sur un jeu de données plus

restreint et précis(dans notre cas brain-tumor-detection et chest-xray-images), ce qui constitue le processus de fine-tuning. Ce réentraînement en fine-tuning permet au modèle de s'ajuster aux spécificités de la tâche tout en conservant les représentations globales des images acquises lors du pré-entraînement sur un jeu de données plus vaste.

- **Extraction des caractéristiques :** Dans l'architecture ViT que nous avons développée, le modèle DieT-ViT a été identifié comme un composant essentiel responsable de l'extraction des caractéristiques de l'image. La structure globale de ViT est composée de deux composants principaux : l'encodeur patch et l'encodeur transformer.
  - **L'encodeur de patches :** L'encodeur de patch comprend deux couches consécutives la couche patch embedding et la couche de projection linéaire, qui prétraitent collectivement l'image d'entrée. Dans la couche patch embedding les images sont divisées en 256 patchs mesurant 16 x 16 chacun comme ulistrer dans la figure 3.2 ; ces patchs ne sont pas superposés et sont ensuite utilisés comme jetons(Tokens).

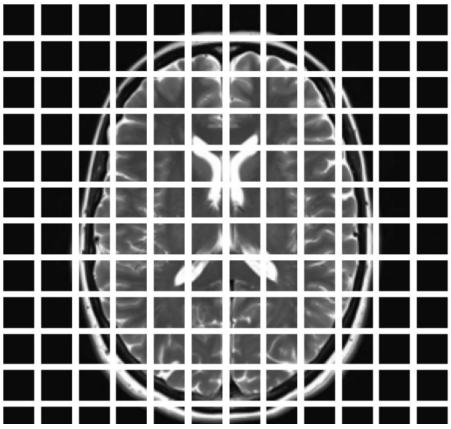

FIG. 3.2 : Image divisées en 256 patches de taille 16x16.

Au niveau de la couche de projection linéaire, les patchs sont soumis à une projection linéaire dans un espace vectoriel de dimension inférieure afin de générer une série de vecteurs, chacun désignant un patch. Étant donné que DieT-ViT traite les images 2D en entrée, l'aplatissement de l'image en une séquence 1D devient impératif, suivi de la projection linéaire sur chaque séquence de patch avant son intégration dans une architecture de transformer. Cette méthodologie permet de réduire les dimensions et les variables d'entrée, contribuant ainsi au traitement efficace des patchs individuels par les couches encodeurs du transformer. À la suite de cette projection, les vecteurs représentant les patchs sont traités par diverses couches de transformation, ce qui permet d'acquérir des interconnexions complexes entre les patchs, tout en comprenant les caractéristiques générales et les corrélations spatiales entre eux. La représentation qui en résulte facilite l'exécution de la classification des images de manière fluide.

* **L'encodeur transformer :** L'encodeur du transformer acquiert les vecteurs des patchs et les transmet à travers une série de couches. Les couches d'encodeur de transformer représentent une séquence de nombreuses couches de transformers chargées d'acquérir des connaissances sur les connexions entre les patchs d'image. Chaque couche Transformer est composée de deux sous-couches : une couche Multi-Head Attention et une couche Feedforward. La couche self-attention multi-head permet au modèle de comprendre les connexions entre les patchs à différents niveaux, tandis que la couche feedforward fournit des fonctions non linéaires à la représentation de chaque patch.

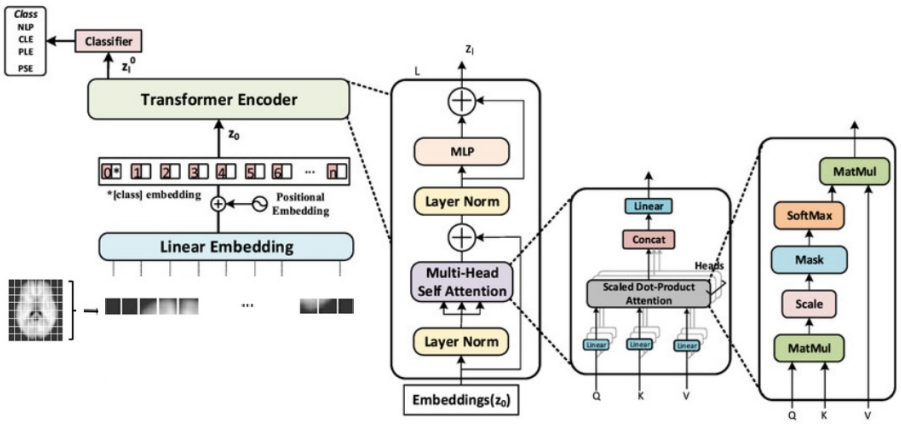

Fig. 3.3 : L'Architecture principale du modèle ViT et les spécifications du bloc encodeur transformer

Plus précisément, Pour optimiser les paramètres du modèle et accélérer l'entraînement, la sortie de l'encodeur Transformer est d'abord traitée par une couche de normalisation par batch. Puis, une couche de classification linéaire transforme les données en un vecteur unidimensionnel, suivie d'une couche Dropout qui une technique de régularisation qui aide à prévenir le sur-apprentissage en limitant la dépendance du modèle à certaines caractéristiques de l'ensemble de données. Dans le bloc MLP la partie classification intègre une couche Dense d'un grand nombre de neurones est suivie d'un Dropout. La sortie finale, générée par une couche Dense dotée de deux neurones, permet la classification binaire des images pour déterminer si une image contient une anomalie ou si elle est normale. La sortie de cette couche est une probabilité reflétant la confiance du modèle dans sa prédiction : une valeur proche de 1 indique la haute certitude du modèle quant à une image anormale, tandis qu'une valeur proche de 0 indique un état normal.

– **Ajustement des paramètres du modèle :** Notre modèle est entraîné en utilisant une boucle d'entraînement qui se déroule sur les données d'entraînement pendant plusieurs époques. Nous procédons aux étapes suivantes pour chaque batch de données :

* Les images sont passées dans le modèle pour obtenir des prédictions

* la perte est calculée par une comparaison entre ces prédictions et les étiquettes réelles.
* Ensuite, une rétropropagation est effectuée pour déterminer les gradients et modifier les poids du modèle. Le pas d'apprentissage est définie comme un hyperparamètre d'optimisation clé, essentiel pour régir la vitesse et la direction des mises à jour des poids du modèle.
* Les poids sont ajustés à l'aide de l'algorithme d'optimisation choisi (AdamW).

Après chaque époque d'entraînement, une évaluation des performances du modèle sur l'ensemble de données de validation est réalisée. La précision est ensuite calculée pour évaluer les capacités prédictives du modèle et identifier les problèmes potentiels d'overfitting. Afin d'améliorer les performances du modèle, les hyperparamètres sont ajustés en fonction des résultats du processus de validation.

3.3.2 Phase de test

Après avoir terminé l'entraînement, nous évaluons le modèle final sur l'ensemble de test afin d'obtenir une estimation objective de ses performances sur des données inconnues. Toutes les images de l'ensemble de test ont été intégrées au modèle afin d'obtenir des prédictions de classification. Les prévisions se présentent sous la forme de probabilités pour chaque catégorie. Les métriques d'évaluation sont calculées et les résultats obtenus sont analysés afin d'évaluer la performance et la capacité de généralisation du modèle.

3.3.3 Classification

Après avoir réentraîné notre modèle, nous l'utilisons pour faire des prédictions sur les étiquettes de classe pour les nouvelles images. Le modèle calcule les probabilités pour chaque classe, tout en attribuant une valeur spécifique à chaque classe. Ensuite, en se basant sur ces probabilités calculées, il tire une conclusion décisive pour identifier quelle étiquette attribuer à l'entrée donnée en fonction de la probabilité la plus élevée.

3.4 Résultats expérimentaux et discussion

Pour l'apprentissage et l'évaluation de modèles DieT-ViT, nous avons utilisé deux bases de données qui sont d'abord présentées. Ensuite, des expérimentations sont menées pour déterminer la meilleure combinaison d'hyperparamètres. Enfin, les meilleurs résultats expérimentaux obtenus sont comparés à ceux de différentes architectures CNN, ainsi qu'à ceux de quelques travaux de la littérature de la classification d'images médicales.

3.4.1 Bases de données

Dans cette étude et lors de notre entraînement et évaluation de notre modèle, deux bases de données ont été utilisées :

3.4.1.1 Brain-tumor-dataset

La base de donnée est une collection complète de données d'imagerie médicale conçues aux fins de la détection et de la classification des tumeurs du cerveau. Elle englobe 4600 images par résonance magnétique (IRM) du cerveau humain, elle contient deux classes de données Classe Brain Tumor (présentation de tumeurs cérébrales). Cette catégorie comprend des scans MRI de patients. Ces scans montrent la présence de tumeurs à différents stades et endroits dans le cerveau (2513 images). Classe Healthy (pas de tumeur cérébrale) Cette classe contient des scans MRI d'individus sans tumeurs du cerveau détectables. Ces scannings servent de données de référence pour les structures et conditions cérébrales saines (2087 images).

3.4.1.2 Chest X-Ray Images (Pneumonia)

L'ensemble de données est organisé en 2 dossiers (train, test) et contient des sous-dossiers pour chaque catégorie d'image (Pneumonia/Normal). Il y a 5 856 images X-Ray (JPEG) et 2 catégories (Pneumonia/Normal). Les radiographies thoraciques (avant-arrière) ont été sélectionnées à partir de cohortes rétrospectives de patients pédiatriques âgés d'un à cinq ans du centre médical des femmes et des enfants .

Pour plus de détails sur les deux bases de données, consultez le chapitre précédent.

3.4.2 Métriques d'évaluation utilisées

Dans nos expérimentations, deux métriques importantes sont utilisées, dont l'accuracy et le F1 score.

L'accuracy

L'accuracy est une mesure couramment utilisée pour évaluer la performance d'un modèle de machine learning. Elle indique la proportion de prédictions correctes par rapport au nombre total d'exemples.la formule de

$$Exactitude = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.1}$$

Cependant, il est important de noter que l'accuracy peut être trompeuse dans certains cas, notamment lorsque les classes sont déséquilibrées ou lorsque certaines erreurs sont plus graves que d'autres. Dans ces situations, d'autres métriques telles que, le rappel et le F1-score sont souvent utilisées pour obtenir une vue plus complète de la performance du modèle.

Le F1-score

Le F1-score est une métrique couramment utilisée pour évaluer les performances d'un modèle de classification, en particulier dans les problèmes de classification binaire.

$$F1 - score = \frac{2 \cdot Prcision \cdot Recall}{Prcision + Recall} \tag{3.2}$$

tel que la Précision et le Recall (ou Rappel) sont deux métriques complémentaires qui peuvent être décrit, respectivement, par la fraction des exemples assignés à la classe positive qui appartiennent à la classe positive, et la façon dont la classe positive a été prédite. Elles sont calculées selon les deux éauqtions suivantes :

$$Prcision = \frac{VP}{VP + FP} \tag{3.3}$$

$$Recall = \frac{VP}{VP + FN} \tag{3.4}$$

Une valeur de 1 sur le F1-score indique une classification parfaite, tandis qu'une valeur de 0 indique une classification mauvaise. Par conséquent, un score F1 plus élevé signifie que le modèle de classification fonctionne mieux.

Micro-F1 : Agrège les classifications par échantillon et calcule le score F1 global.

Macro-F1 : Moyenne arithmétique des scores F1 par classe.

Weighted-F1: Moyenne pondérée en fonction de la proportion de chaque classe.

La fonction de perte (Loss)

La fonction de perte est un concept essentiel en apprentissage automatique. Elle permet de quantifier l'écart entre les prédictions d'un modèle et les observations réelles dans le jeu de données utilisé pour l'entraînement. Elle est un outil essentiel pour guider l'apprentissage du modèle vers des prédictions plus précises.

3.4.3 Ajustement des hyperparamètres

Cette sous-section présente les résultats expérimentaux obtenus avec le modèle DieT-ViT pour la classification d'images médicales. Des expérimentations ont été effectuées pour évaluer la performance de l'architecture et identifier les éléments et hyperparamètres contribuant aux meilleurs résultats. Les hyperparamètres analysés incluent la taille du batch, le taux de dropout et le pas d'apprentissage.

3.4.3.1 Taille de batch

La taille de batch est un hyperparamètre essentiel dans l'apprentissage des modèles de machine learning. Elle détermine le nombre d'échantillons (images, dans ce cas) qui sont regroupés et traités simultanément lors de l'entraînement et de l'évaluation du modèle. Le tableau 3.1 présente les résultats issus d'expérimentations menées pour analyser l'impact de la taille de batch sur l'accuracy en utilisant différentes tailles de batch .

Tab. 3.1 : Résultats expérimentaux pour différentes tailles de batch

	Brian-tumor-dataset			Chest X-Ray		
Taille batch	Train acc	Val acc	Test acc	Train acc	Val acc	Test acc
64	96.39%	98.50%	97.50%	89.94%	90.89%	**90.74%**
32	97.28%	97.25%	98.95%	89.89%	90.77%	88.88%
16	97.62%	98.70%	**98.98%**	89.80%	89.97%	89.93%
8	97.45%	97.68%	98.40%	89.90%	90.89%	88.41%

Le tableau 3.1 présente les résultats expérimentaux pour différentes tailles de batch sur les deux ensembles de données : "Brain-tumor-dataset" et "Chest X-Ray". Chaque ensemble de données a trois colonnes principales : "Train Acc" (accuracy d'entraînement), "Val Acc" (accuracy de validation) et "Test acc" (accuracy de test).

Pour le "Brain-tumor-dataset", les tailles de batch expérimentées sont 64, 32, 16 et 8. Par exemple, avec un batch de 64, l'accuray d'entraînement est de 96,39 %, celle de validation est de 98,50 % et celle de test est de 97,50 %. De même, pour le "Chest X-Ray", les mêmes tailles de batch sont utilisées. Par exemple, avec un batch de 64, l'accuracy d'entraînement est de 89,94 %, celle de validation est de 90,89 % et celle de test est de 90,74 %.

Les conclusions suivantes sont tirées de cette expérimentation :

1. **Impact de la taille du batch (batch size) sur l'accuracy :** Plus la taille du batch est grande, plus l'estimation du gradient est stable, ce qui peut conduire à une convergence plus rapide lors de l'entraînement. Cependant, des batchs trop grands peuvent entraîner une utilisation excessive de la mémoire et des temps de calcul plus longs. Il est important de trouver un équilibre entre la stabilité du gradient et l'efficacité computationnelle.

2. **Sur-apprentissage (overfitting) et sous-apprentissage (underfitting) :** Lorsque la taille du batch est petite (par exemple, 8), le modèle peut sur-apprendre les données d'entraînement. Une taille de batch plus grande (par exemple, 64) peut aider à réduire le sur-apprentissage en moyennant les gradients sur un plus grand nombre d'exemples. Cependant, si la taille du batch est trop grande, le modèle risque de sous-apprendre (sous-ajustement) et de ne pas généraliser correctement.

3. **Choix de la taille du batch :** Il n'y a pas de taille de batch universelle qui fonctionne bien pour tous les problèmes. Il est recommandé d'expérimenter différentes tailles de batchs pour trouver celle qui donne les meilleures performances sur notre ensemble de données.

Ces résultats sont intéressants pour comprendre comment la taille du batch affecte les performances du modèle en termes d'accuracy lors des différentes phases (entraînement/-validation/test) sur ces deux ensembles de données liés à l'imagerie médicale.

3.4.3.2 Taux du Dropout

Le concept de Dropout repose sur la désactivation aléatoire de neurones spécifiques dans le réseau neuronal durant l'apprentissage. Cette technique force le modèle à apprendre de différentes combinaisons de neurones, réduisant ainsi la dépendance envers certains neurones spécifiques. Les Dropouts servent donc de méthode de régularisation efficace, diminuant le risque de sur-apprentissage et renforçant la généralisation du modèle. L'objectif de cette recherche est de mettre au point un modèle ViT efficace, performant à la fois sur les données d'entraînement utilisées pendant la phase d'apprentissage et sur les nouvelles données utilisées à des fins prédictives. Notre objectif est que le modèle puisse appliquer les connaissances acquises à des situations nouvelles et exceptionnelles. L'objectif de cette expérimentation est de déterminer le taux de Dropout optimal afin de réduire le sur-apprentissage et d'améliorer la généralisation du modèle ViT.

Tab. 3.2 : Résultats expérimentaux pour déterminer le taux de Dropout

	Brian-tumor-dataset			Chest X-Ray		
Dropout	Train acc	Val acc	Test acc	Train acc	Val acc	Test acc
0.1	96.29%	96.81%	97.11%	88.18%	88.20%	87.98%
0.2	97.51%	97.83%	98.35%	89.59%	88.27%	89.30%
0.3	98.50%	98.84%	**98.40%**	90.43%	89.98%	**90.68%**
0.4	97.16%	97.54%	97.81%	89.89%	89.18%	90.08%

Le tableau 3.2 présente les résultats expérimentaux des deux ensembles de données médicales, pour différentes valeurs de taux de dropout (0,1, 0,2, 0,3 et 0,4). Voici ce que nous pouvons en déduire :

1. **Brain-tumor-dataset :** L'accuracy en entraînement augmente à mesure que le taux de dropout diminue, ce qui est attendu. Cela signifie que le modèle apprend mieux sur les données d'entraînement lorsque le taux de dropout est plus bas. L'accuracy en validation atteint un maximum à un taux de dropout de 0,3, puis diminue légèrement. Cela suggère que le modèle généralise bien avec un taux de dropout de 0,3. L'accuracy en test est également la plus élevée à un taux de dropout de 0,3, ce qui confirme que ce taux est optimal pour ce modèle et cet ensemble de données.

2. **Chest X-Ray :** Les tendances sont similaires à celles de Brain-tumor-dataset, mais avec des valeurs d'accuracy légèrement différentes. L'accuracy en test est plus élevée que l'accuracy en validation, ce qui peut indiquer que le modèle a une bonne capacité de généralisation.

En résumé, il est important de trouver un équilibre entre le surapprentissage et le sous-apprentissage en ajustant le taux de dropout. Les résultats affichés dans le tableau 3.2 indiquent clairement que le Dropout fixé à 0.3 est optimal. Cette valeur implique l'élimination de seulement 30% des neurones à chaque époque. Il est aussi notable que l'accuracy sur les données d'entraînement atteint 98.50% pour le Brain-tumor-dataset et 90.43% pour le Chest X-Ray, tandis que sur les données de test, elle est de 98.40% et 90.68% respectivement. Cela démontre que notre modèle DieT-ViT a une excellente capacité de généralisation.

3.4.3.3 Le pas d'apprentissage

Le pas d'apprentissage désigne la mesure dans laquelle les poids du modèle sont mis à jour pendant l'entraînement. Il peut être ajusté de diverses façons pour améliorer les performances d'apprentissage. Les résultats obtenus en termes d'accuracy pour différents taux d'apprentissage sont présentés dans le tableau 3.3. Ce tableau présente les résultats expérimentaux pour différentes valeurs de pas d'apprentissage (0,1, 0,01, 0,001, 0,0001 et 0,00001).

Tab. 3.3 : Résultats expérimentaux pour déterminer le pas d'apprentissage

	Brian-tumor-dataset			Chest X-Ray		
Pas d'apprentissage	Train acc	Val acc	Test acc	Train acc	Val acc	Test acc
0.1	85.65%	90.28%	89.97%	88.34%	89.64%	88.94%
0.01	92.32%	96.52%	95.34%	88.46%	89.80%	89.78%
0.001	98.49%	98.84%	**98.40%**	90.39%	89.99%	**90.68%**
0.0001	92.84%	94.42%	95.49%	89.44%	89.50%	89.66%
0.00001	83.48%	86.23%	85.75%	88.18%	88.67%	88.58%

Voici ce que nous pouvons déduire de cette expérimentation :

- Pour l'ensemble des données Brain-tumor-dataset, l'accuracy d'entraînement (98.49%), de validation (98.84%) et de test (98.40%) est la plus élevée pour un pas d'apprentissage de 0.001.
- Pour l'ensemble de données Chest X-Ray, l'accuracy d'entraînement (90.39%), de validation (89.99%) et de test (90.68%) est la plus élevée pour un pas d'apprentissage de 0.001.
- Un pas d'apprentissage trop élevé peut entraîner une convergence rapide, mais il risque également de sauter par-dessus les minima locaux. D'autre part, un pas d'apprentissage trop bas peut ralentir la convergence. Dans notre tableau, nous avons testé différentes valeurs de pas d'apprentissage.

Sur la base de ces résultats, un pas d'apprentissage de 0.001 semble être un bon choix pour les deux ensembles de données. Cela permet d'obtenir de bonnes performances sans risquer de sauter par-dessus les minima locaux.

3.4.4 Evaluation des performances

Les tableaux 3.4, 3.5 illustrent les résultats de notre modèle en matière d'accuracy sur les bases de données Brain-tumor-dataset et Chest X-Ray respectivement. Alors que le tableau 3.6 montrent les résultats obtenus en termes de F1-score sur les deux bases de données.

TAB. 3.4 : Accuracy de notre modèle sur la base de données brian-tumor-dataset

Images	Train accuracy	Validation accuracy	Test accuracy
Taille de la base	3220	690	690
Accuracy	98.49%	98.84%	98.40%

TAB. 3.5 : Accuracy de notre modèle sur la base de données Chest X-Ray

Images	Train accuracy	Validation accuracy	Test accuracy
Taille de la base	4099	879	878
Accuracy	90.39%	89.99%	90.68%

TAB. 3.6 : Les performances du modèle DieT-ViT sur les données de test

	Brian-tumor-dataset			Chest X-Ray		
Modèle	Micro-F1	Macro-F1	Weighted-F1	Micro-F1	Macro-F1	Weighted-F1
DieT-ViT	0.98	0.98	0.98	O.90	0.50	0.86

Les conclusions des tableaux confirment pleinement les observations des expérimentations précédentes concernant la généralisation adéquate des données de test jamais vues.

3.4.5 Comparaison des résultats

Dans la suite, nous procéderons à la comparaison des résultats de notre système avec ceux obtenus par architectures CNN et ceux rapportés dans des travaux existants, en nous basant sur les deux bases de données sur lesquelles nous avons travaillées.

3.4.5.1 Comparaison avec des architectures CNN

Pour comparer, nous avons mis en place, en plus du modèle ViT, quelques architectures distinctes de CNN : Resnet50, InceptionV3, Seresnext101d_32x8d et Sent154 sur les deux bases de données. Les résultats obtenus en termes d'accuracy et de perte (loss) sont présentés dans les tableaux 3.7 et 3.8.

Brain-tumor-dataset

TAB. 3.7 : Comparaison de notre modèle DieT-ViT avec différentes architectures CNN sur Brain-tumor-dataset

Modèles	Test accuracy	Val accuracy	Test loss	val loss
Resnet50	90.34%	90.66%	25.55%	26.02%
Inception V3	75.86%	76.29%	34.70%	35.90%
Seresnext101d_32x8d	85%	86%	46.50%	47.43%
Sent154	88%	88%	47.26%	51.33%
DieT	**98.40%**	**98.84%**	**6.16%**	**5.19%**

Analysons les résultats du tableau 3.7, Voici ce que nous pouvons en déduire :

- Le modèle DieT-ViT se distingue nettement des autres modèles. Il affiche la meilleure accuracy de test (98,40 %) et la meilleure accuracy de validation (98,84 %).
- De plus, il présente des pertes de test et de validation beaucoup plus faibles (6,16 % et 5,19 % respectivement), ce qui indique une meilleure performance à la fois en termes d'accuracy et de métriques de perte.

Ces résultats démontrent l'efficacité du modèle DieT-ViT par rapport aux architectures traditionnelles de CNN lorsqu'il est appliqué à la détection de tumeurs cérébrales dans l'analyse d'imagerie médicale.

Chest X-Ray

TAB. 3.8 : Comparaison de notre modèle DieT-ViT avec différentes architectures CNN Chest X-Ray

Modèles	Test accuracy	Val accuracy	Test loss	val loss
Resnet50	86.92%	87.20%	44.85%	43.50%
Inception V3	62.99%	62.75%	68.58%	68.30%
Seresnext101d_32x8d	83%	84%	42.50%	42.03%
Sent154	90%	90%	40.70%	40.90%
DieT ViT	**90.38%**	**89.99%**	**29.38%**	**30.95%**

Analysons les résultats du tableau 3.8, nous pouvons déduire que le modèle DieT-ViT se distingue par l'accuracy la plus élevée aussi bien sur l'ensemble de test que sur celle de validation, ainsi que par ses pertes les plus faibles sur les deux ensembles parmi tous les modèles répertoriés. Cela indique qu'il pourrait être le modèle le plus efficace pour cette application.

En résumé, le modèle DieT-ViT se distingue par ses performances élevées sur les deux ensembles de données comparé au différentes architectures CNN testées.

3.4.5.2 Comparaison avec des travaux de la littérature

Pour évaluer les performances de notre système en comparaison avec d'autres études, nous avons mesuré son accuracy contre celle de travaux récents utilisant les mêmes bases de données. Les résultats de cette comparaison sont détaillés dans le tableau 3.9 pour Brain tumor dataset et dans le tableau 3.10 pour Chest X-Ray.

Afin de faciliter la compréhension, les résultats obtenus ont été illustrés visuellement dans les graphiques 3.4 et 3.5. Nous avons accompagné chaque tableau d'un graphique correspondant.

Brain tumor dataset

Tab. 3.9 : Comparaison des performances avec l'état de l'art brain tumor dataset

Systeme	Technique	Accuracy
Seetha et al. 2018	CNN	97.5%
Çinarer et al. 2019	SVM	90%
Ramdlon et al. 2019	KNN	89.5%
Amin et al. 2020	SVM	97.1%
Pacal 2024	**Swin T**	**99.92%**
Notre système.2024	DieT-ViT	98.40%

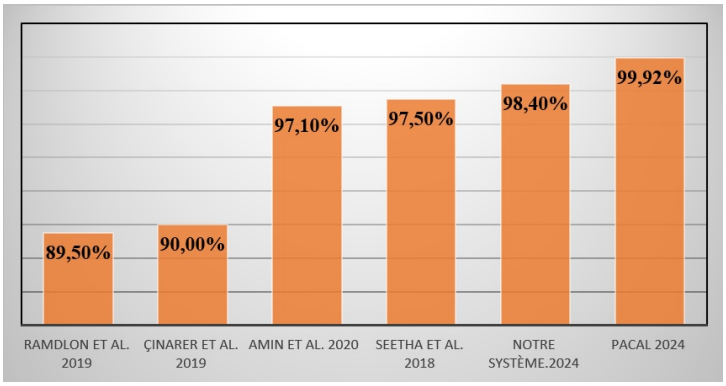

Fig. 3.4 : Comparaison des performances avec l'état de l'art brain tumor dataset

Selon le tableau 3.9 et le graphique 3.4, il est intéressant de noter que la base de données utilisée Brain tumor dataset a obtenu le meilleur résultat pour Swin Transformer, avec une accuracy de 99,92%. Il est aussi intéressant de constater que notre système offre des résultats compétitifs très prometteurs par rapport aux meilleurs résultats rapportés, avec une accuracy de 98,40%.

Chest X-Ray

Tab. 3.10 : Comparaison des performances avec l'état de l'art de la base de données chest X-Ray

Systeme	Technique	Accuracy
Jain et al. 2020	VGG16	87.28%
Jain et al. 2020	VGG19	88.46%
Jain et al. 2020	ResNet50	77.56%
Jain et al. 2020	Inception V3	70.99%
Wang et al. 2023	PneuNet-ViT	94.96 %
Singh et al. 2024	**ViT**	**97.61%**
Notre système.2024	DieT-ViT	90.68%

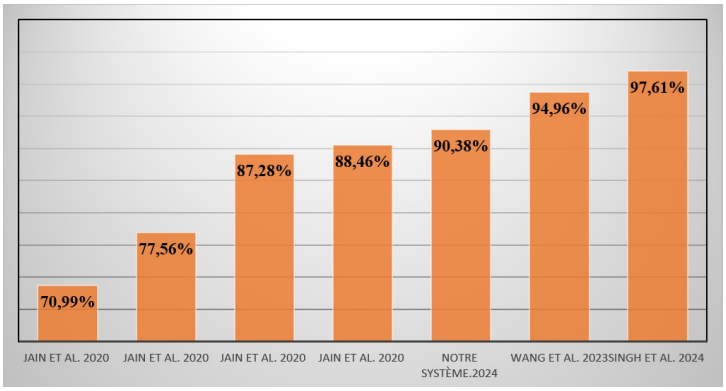

Fig. 3.5 : Comparaison des performances avec l'état de l'art Chest X-Ray

Selon le tableau 3.10 et le graphique 3.5, pour la deuxième base de données utilisée, Chest X-Ray, il est remarquable que le meilleur résultat est obtenu avec le ViT Transformer, affichant une accuracy de 97,61%. Il est également intéressant de noter que notre système produit des résultats très prometteurs en comparaison avec les meilleurs résultats déjà rapportés, atteignant une accuracy de 90,68%.

En conclusion et d'après les résultats obtenus, il apparaît clairement que les techniques basées sur les Transformers dépassent régulièrement les performances des CNN et des algorithmes d'apprentissage automatique traditionnels. Ces approches, souvent utilisées en conjonction avec des Vision Transformers, réussissent à monter en tête des classements pour la classification des images en général.

3.5 Conclusion

Au cours de ce chapitre, nous avons exposé les diverses étapes de la mise en place de notre modèle ViT qui permet de classer les images médicales. Les images IRM du cerveau humain et les images X-Ray (radiographies thoraciques) étaient utilisées comme exemples.

En outre, nous avons étudié l'impact de diverses configurations sur la précision du modèle. Afin d'accomplir cette tâche, nous avons effectué une série d'expérimentations qui nous ont permis de déterminer les valeurs optimales des hyperparamètres de notre modèle ViT. Elle a été évaluée sur les bases de données mentionnées et ses performances ont été comparées à d'autres travaux récents. Les expérimentations ont démontré que notre système basé sur des transformers est performant pour classer les images provenant de diverses bases de données et offre des perspectives prometteuses pour des applications cliniques.

Conclusion Générale

Conclusion

La classification des images médicales joue un rôle essentiel dans la détection précoce, la précision du diagnostic, la prise de décision en médecine, la réduction des temps de traitement et l'élargissement de l'accès aux soins. Son utilisation contribue à améliorer les résultats cliniques et à sauver des vies en offrant des informations précises et en aidant à l'administration de traitements adéquats.

Notre recherche se concentre sur la classification des images médicales. Nous visons spécifiquement à améliorer les performances des systèmes de classification pour diverses bases de données d'images. La complexité de cette tâche a incité de nombreux chercheurs à réaliser d'importantes études pour résoudre le problème de la classification. Néanmoins, les avancées scientifiques n'ont pas encore atteint le niveau de compétence humaine dans ce domaine.

Dans ce contexte, nous avons concentré nos efforts sur la classification d'images médicales en utilisant les Transformers de Vision (ViTs), une technique d'intelligence artificielle très récente qui a bouleversé le domaine de la vision par ordinateur.

Pour explorer les capacités du modèle ViT, une étude expérimentale a été réalisée afin d'identifier les hyperparamètres qui fournissent les meilleurs résultats en termes d'accuracy. Grâce à cette étude, nous avons établi un système hautement efficace, qui a atteint un taux de reconnaissance de 90,78% sur la base de données exemplaire Chest x-ray et 98.40% sur la base de données Brain tumor dataset.

Afin d'évaluer le système proposé basé sur ViT, nous l'avons comparé en termes d'accuracy avec différentes architectures de CNN, notamment resnest50d, Inception V3, Seresnext101d_32x8d et Sent154. Ensuite, nous avons mis en parallèle nos résultats avec ceux de travaux récents utilisant les même bases des données. Cette comparaison nous a permis de conclure que notre système figure avantageusement parmi les meilleurs, prouvant ainsi l'efficacité du ViT dans la classification d'images médicales.

Perspectives

Concernant nos perspectives pour les travaux à venir, plusieurs options sont envisageables pour améliorer et continuer nos recherches, parmi lesquelles les plus significatives sont :

- Nous prévoyons d'améliorer la généralisation du modèle de classification d'images médicales par l'emploi de techniques de data augmentation.
- À côté de ViT qui a déjà prouvé son efficacité, il serait intéressant d'examiner d'autres modèles basés sur des transformers. Les modèles proposés peuvent offrir de nouvelles perspectives en ce qui concerne la représentation et les capacités de classification, ce qui ouvre de nouvelles opportunités pour améliorer la précision du diagnostic.

Bibliographie

Acheampong, Francisca Adoma, Henry Nunoo-Mensah et Wenyu Chen (2021). « Transformer models for text-based emotion detection : a review of BERT-based approaches ». In : *Artificial Intelligence Review* 54.8, p. 5789-5829.

Ahmad, Haissam Haidar et al. (2022). « Intelligence artificielle et imagerie médicale ». In.

Altan, Gokhan (2020). « Deep learning-based mammogram classification for breast cancer ». In : *International Journal of Intelligent Systems and Applications in Engineering* 8.4, p. 171-176.

Amin, Javeria et al. (2020). « A distinctive approach in brain tumor detection and classification using MRI ». In : *Pattern Recognition Letters* 139, p. 118-127.

Arulkumaran, Kai et al. (2017). « A brief survey of deep reinforcement learning ». In : *arXiv preprint arXiv :1708.05866*.

Ayyaz, M Shahbaz et al. (2021). « Hybrid deep learning model for endoscopic lesion detection and classification using endoscopy videos ». In : *Diagnostics* 12.1, p. 43.

Ba, Jimmy Lei, Jamie Ryan Kiros et Geoffrey E Hinton (2016). « Layer normalization ». In : *arXiv preprint arXiv :1607.06450*.

BARKA, IKRAM et SAIDA BERBIA (2021). « Analyse et classification d'image médicales par réseaux de neurone profond ». Thèse de doct. UNIVERSITE KASDI MERBAH OUARGLA.

Bustos, Aurelia et al. (2020). « Padchest : A large chest x-ray image dataset with multilabel annotated reports ». In : *Medical image analysis* 66, p. 101797.

Chen, Mark et al. (2020). « Generative pretraining from pixels ». In : *International conference on machine learning*. PMLR, p. 1691-1703.

Cheng, Jianpeng, Li Dong et Mirella Lapata (2016). « Long short-term memory-networks for machine reading ». In : *arXiv preprint arXiv :1601.06733*.

Chiter, Yasmine et Aicha Hafiane (2022). « Apprentissage profond pour l'analyse et la classification d'imageries médicales ». Thèse de doct. UNIVERSITY OF KASDI MERBAH OUARGLA.

Çinarer, Gökalp et Bülent Gürsel Emiroğlu (2019). « Classificatin of brain tumors by machine learning algorithms ». In : *2019 3rd international symposium on multidisciplinary studies and innovative technologies (ISMSIT)*. IEEE, p. 1-4.

Dali, Asma (2022). « Segmentation à base d'atlas d'images médicales utilisant l'apprentissage en ligne ». Thèse de doct. Ecole nationale supérieure Mines-Télécom Atlantique Bretagne Pays de la Loire.

Devlin, Jacob et al. (2018). « Bert : Pre-training of deep bidirectional transformers for language understanding ». In : *arXiv preprint arXiv :1810.04805*.

Dosovitskiy, Alexey et al. (2020). « An image is worth 16x16 words : Transformers for image recognition at scale ». In : *arXiv preprint arXiv :2010.11929*.

EL MASSARI, HAKIM (2023). « Proposition d'un modèle de prédiction basé sur Machine Learning et le web sémantique ». In.

Farrag, Tamer Ahmed et Ehab E Elattar (2021). « Optimized deep stacked long short-term memory network for long-term load forecasting ». In : *IEEE Access* 9, p. 68511-68522.

Ghojogh, Benyamin et Ali Ghodsi (2020). « Attention mechanism, transformers, BERT, and GPT : tutorial and survey ». In.

Hakim, Belhadjer et Sarouer Brahim (2018). « Classification des images avec les réseaux de neurones convolutionnels ». Thèse de doct. Université Mouloud Mammeri.

Hu, Junyan et al. (2020). « Voronoi-based multi-robot autonomous exploration in unknown environments via deep reinforcement learning ». In : *IEEE Transactions on Vehicular Technology* 69.12, p. 14413-14423.

Huang, Long et al. (2024). « SRT : Swin-residual transformer for benign and malignant nodules classification in thyroid ultrasound images ». In : *Medical Engineering & Physics* 124, p. 104101.

Jain, Rachna et al. (2020). « Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning ». In : *Measurement* 165, p. 108046.

Jiang, Zhencun et al. (2021). « Method for diagnosis of acute lymphoblastic leukemia based on ViT-CNN ensemble model ». In : *Computational Intelligence and Neuroscience* 2021.

Lavecchia, Antonio (2015). « Machine-learning approaches in drug discovery : methods and applications ». In : *Drug discovery today* 20.3, p. 318-331.

LeCun, Yann (2016). « Les enjeux de la recherche en intelligence artificielle ». In : *Accès https ://dataanalyticspost. com/wp-content/uploads/2017/04/ylecun_college_France. pdf*.

Macary, Manon (2022). « Analyse de données massives en temps réel pour l'extraction d'informations sémantiques et émotionnelles de la parole ». Thèse de doct. Le Mans Université.

Mahesh, Batta (2020). « Machine learning algorithms-a review ». In : *International Journal of Science and Research (IJSR).[Internet]* 9.1, p. 381-386.

Malki, Narimene et Tahar Guerram (2019). « Classification automatique des textes par Les réseaux de neurones à convolution ». In.

Mathieu-Dupas, Eve (2010). « Algorithme des k plus proches voisins pondérés et application en diagnostic ». In : *42èmes Journées de Statistique*.

Milgram, Jonathan (2007). « Contribution à l'intégration des machines à vecteurs de support au sein des systèmes de reconnaisance de formes : application à la lecture automatique de l'écriture manuscrite ». Thèse de doct. École de technologie supérieure.

Moghaddamnia, A et al. (2009). « Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques ». In : *Advances in Water Resources* 32.1, p. 88-97.

Nosouhian, Shiva, Fereshteh Nosouhian et Abbas Kazemi Khoshouei (2021). « A review of recurrent neural network architecture for sequence learning : Comparison between LSTM and GRU ». In.

Osório, Fernando Santos (1998). « INSS : un système hybride neuro-symbolique pour l'apprentissage automatique constructif ». Thèse de doct. Institut National Polytechnique de Grenoble-INPG.

Ounissi, Mohammed, Zahra Ilham Harnane et Kheireddine Lamamra (2020). « Modélisation et classification avec Deep Learning ». In.

Pacal, Ishak (2024). « A novel Swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in MRI images ». In : *International Journal of Machine Learning and Cybernetics*, p. 1-19.

Radford, Alec et al. (2018). « Improving language understanding by generative pre-training ». In.

Ramdlon, Rafi Haidar, Entin Martiana Kusumaningtyas et Tita Karlita (2019). « Brain tumor classification using MRI images with K-nearest neighbor method ». In : *2019 International Electronics Symposium (IES)*. IEEE, p. 660-667.

Salperwyck, Christophe, Vincent Lemaire et Carine Hue (2014). « Classifieur naïf de Bayes pondéré pour flux de données. » In : *EGC*, p. 275-286.

Samuel, Arthur L (1959). « Machine learning ». In : *The Technology Review* 62.1, p. 42-45.

Saraiva, Arata Andrade et al. (2019). « Models of learning to classify X-ray images for the detection of pneumonia using neural networks. » In : *Bioimaging*, p. 76-83.

Seetha, J et S Selvakumar Raja (2018). « Brain tumor classification using convolutional neural networks ». In : *Biomedical & Pharmacology Journal* 11.3, p. 1457.

Singh, Sukhendra et al. (2024). « Efficient pneumonia detection using Vision Transformers on chest X-rays ». In : *Scientific Reports* 14.1, p. 2487.

Swati, Zar Nawab Khan et al. (2019). « Brain tumor classification for MR images using transfer learning and fine-tuning ». In : *Computerized Medical Imaging and Graphics* 75, p. 34-46.

Taud, Hind et Jean-Franccois Mas (2018). « Multilayer perceptron (MLP) ». In : *Geomatic approaches for modeling land change scenarios*, p. 451-455.

Tyagi, Amit Kumar et G Rekha (2020). « Challenges of applying deep learning in real-world applications ». In : *Challenges and applications for implementing machine learning in computer vision*. IGI Global, p. 92-118.

Usman, Mohammad, Tehseen Zia et Ali Tariq (2022). « Analyzing transfer learning of vision transformers for interpreting chest radiography ». In : *Journal of digital imaging* 35.6, p. 1445-1462.

Vaswani, Ashish et al. (2017). « Attention is all you need ». In : *Advances in neural information processing systems* 30.

Wang, Tianmu et al. (2023). « PneuNet : deep learning for COVID-19 pneumonia diagnosis on chest X-ray image analysis using Vision Transformer ». In : *Medical & Biological Engineering & Computing* 61.6, p. 1395-1408.

Ying, Wei et al. (2018). « Transfer learning via learning to transfer ». In : *International conference on machine learning*. PMLR, p. 5085-5094.

Webographie

Web 01 : vision-transformer-vit. Retrie June 2 ,2024.website : https ://viso.ai/deep-learning/vision-transformer-vit/

Web O2 : DOSSIER-DOCUMENTAIRE-IMAGERIE-.pdf. Retrie June 2 ,2024.

website : https ://www.urml-normandie.org/wp-content/uploads/2018/04/ DOSSIER-DOCUMENTAIRE-IMAGERIE-.pdf

Web 03 : kaggle - Chest X-Ray Images (Pneumonia). Retrie June 2 ,2024. website : https ://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia.

Web 04 : kaggle - Brian Tumor Dataset. Retrie June 2 ,2024. website :https ://www.

kaggle.com/datasets/preetviradiya/brian-tumor-dataset

Web 05 : kaggle - Breast Ultrasound Images Dataset. Retrie June 2 ,2024. website : https ://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset.

Web 06 : Colaboratory. Retrie June 2 ,2024. website : https ://research.google.com/colaboratory/faq.html ?hl=fr

Web 07: Python.Retrie June 2,2024. https ://www.datarockstars.ai/glossary/python/

Web 08 : la bibliothèque Python la plus utilisée en Data Science. Retrie June 2 ,2024. website : https ://datascientest.com/numpy/

Web 09 : Pandas.Retrie June 2,2024. website : https ://datascientest.com/pandas python-data-science

Web 10 : A SCIENTIFIC COMPUTING FRAMEWORK FOR LUAJIT. Retrie June 2 ,2024. website : http ://torch.ch/

Web 11 : Miscellaneous operating system interfaces. Retrie June 2 ,2024 . website :https ://docs.python.org/fr/3/library/os.html

Annexe

Annexe A

Outils d'implémentation

A.1 Introduction

Au début de cette annexe, nous exposons les outils de développement, puis nous exposons les différentes étapes nécessaires pour mettre en place le système de classification des images médicales.

A.2 Plateforme et environnements de développement

Pour réaliser notre étude de manière concrète et évolutive, nous avons utilisé intensivement Google Colab, Ce service cloud de Google reproduit l'environnement Jupyter Notebook dans le cloud. Il met à disposition des processeurs graphiques (GPU) performants pour un grand nombre d'utilisateurs, ce qui est avantageux, en particulier pour ceux qui n'ont pas les moyens de développer des projets d'apprentissage automatique localement (Web 06).

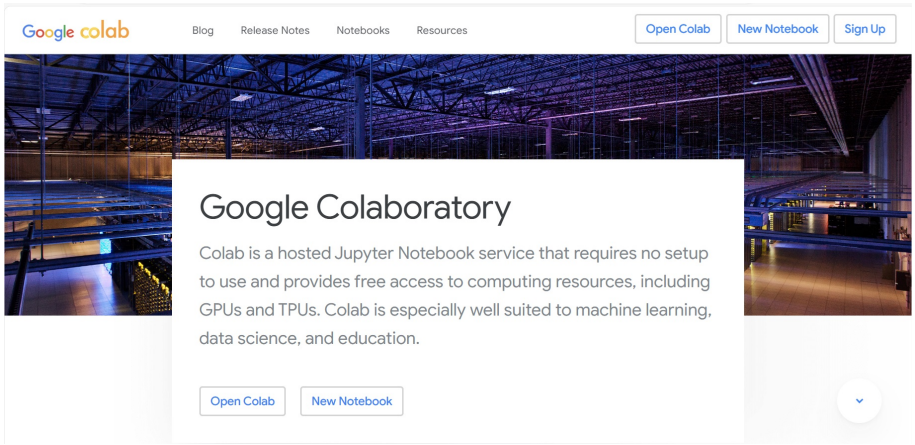

Fig. A.1 : Google Colab

Afin de mettre en œuvre ce projet, nous avons utilisé un ensemble de matériel dont les caractéristiques sont les suivantes :

Tab. A.1 : Caractéristiques de l'environnement

CPU Model	11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz 2.42 GHz
RAM	16 GB
Disque dur	500 GB
Système d'exploitation	Microsoft windows 11(64bits)

A.3 Le langage Python

Python est classé comme un langage de programmation open source appartenant au groupe des langages interprétés. Il permet aux développeurs de se concentrer sur les tâches qu'ils exécutent plutôt que sur la multitude de manières de les exécuter. Par rapport aux langages compilés, Python permet aux programmeurs de gagner beaucoup de temps. Dans le domaine de l'informatique, Python est largement utilisé, apprécié et très recherché. Python est reconnu comme un langage typé dynamiquement, permettant de modifier le type d'une variable. Les annotations sont utilisées pour délimiter les types de paramètres et les valeurs de sortie des fonctions. Python trouve des applications courantes dans le développement Web, l'analyse de données, l'intelligence artificielle, les réseaux de neurones, le calcul scientifique et divers autres domaines du calcul sophistiqué. Sans aucun doute, la programmation s'impose comme le principal domaine d'application de ce langage, adapté à des projets à la fois simples et complexes (Web 07).

Fig. A.2 : Logo Python

A.3.1 Les bibliothèques utilisées

Nous avons utilisé ces bibliothèques pour mettre en place et analyser les modèles de classification des images médicales basés sur les Transformers, ainsi que pour manipuler les données et évaluer les résultats.

A.3.1.1 NumPy

NumPy, abréviation de « Numerical Python », est une bibliothèque de programmation scientifique en Python, principalement destinée à la manipulation de données numériques. Elle propose des tableaux de données multidimensionnels, ainsi qu'un ensemble d'outils

intégrés pour simplifier la mise en œuvre en Python. NumPy allie principalement les caractéristiques du langage C et de Python, en utilisant des tableaux pour traiter les données numériques et réaliser des opérations multidimensionnelles et des manipulations de données (Web 08).

A.3.1.2 Pandas

La bibliothèque Python Pandas est employée pour l'analyse de données. Il a été développé en raison de la nécessité d'un outil puissant et souple pour l'analyse quantitative, ce qui l'a fait devenir l'une des bibliothèques Python les plus célèbres. Pandas jouit d'une communauté très active de contributeurs qui participent régulièrement à son amélioration et à son développement (Web 09).

A.3.1.3 TorCh

Torch est un framework informatique scientifique qui fournit un support complet aux algorithmes d'apprentissage automatique, en particulier en faveur des GPU. Il se caractérise par sa nature conviviale et sa grande efficacité. L'objectif principal de Torch est de fournir un haut niveau d'adaptabilité et de rapidité dans le développement d'algorithmes scientifiques, tout en garantissant un processus simple. Torch est soutenu par une vaste gamme de packages orientés vers la communauté, couvrant divers domaines tels que l'apprentissage automatique, la vision par ordinateur, le traitement du signal, le traitement parallèle, l'analyse d'images, le traitement vidéo, le traitement audio et la mise en réseau, entre autres, qui sont tous enracinés dans la communauté Lua. Au cœur de Torch se trouvent les bibliothèques de réseaux neuronaux et d'optimisation très appréciées, réputées pour leur facilité d'utilisation et la grande liberté qu'elles offrent dans la création de structures de réseaux neuronaux complexes. Ce cadre permet de construire divers graphes de réseaux neuronaux qui peuvent être parallélisés efficacement à la fois sur les processeurs CPU et les GPU (Web 10).

A.3.1.4 Os

Le module os est une bibliothèque Python qui offre une interface portable pour utiliser les fonctionnalités spécifiques au système d'exploitation. Cette bibliothèque est cruciale pour interagir avec le système d'exploitation, ainsi que pour gérer les fichiers, les répertoires et les variables d'environnement dans les programmes Python. (Web 11) .

A.4 Etapes d'implémentation

Le système de classification des images médicales peut être mis en place selon les étapes suivantes.

A.4.1 Importation des bibliothèques

Dans un premier temps, nous allons importer tous les modules nécessaires pour entraîner notre modèle. La figure A.3 présente un fragment de code qui permet d'intégrer les bibliothèques requises.

```
import numpy as np
import pandas as pd
import os
import torch
```

FIG. A.3 : Bibliothèques utilisées.

A.4.2 Prétraitement des données

Il est essentiel de traiter les données d'image et d'effectuer des opérations spécifiques pour prétraiter ces données avant de les intégrer dans le modèle Vision Transformer (ViT).

```
[ ]  def get_data_loaders(data_dir, batch_size , train = False):
        if train:
            transform = transforms.Compose([
                transforms.RandomHorizontalFlip(),
                transforms.RandomVerticalFlip(),
                transforms.RandomApply(transforms=[
                    transforms.GaussianBlur(kernel_size=(5, 9), sigma=(0.1, 5))
                ], p=0.2),
                transforms.Resize(256),
                transforms.CenterCrop(224),
                transforms.ToTensor(),
                transforms.Normalize((0.485, 0.456, 0.406), (0.229, 0.224, 0.225)),
            ])
```

FIG. A.4 : Les opérations pour prétraitement des données.

A.4.3 La division des données

la base de données a été subdivisée en trois sous-ensembles, l'un pour l'entraînement, l'autre pour la validation et l'autre pour le test (train/val/test). La répartition de ces données a été réalisée en utilisant le code suivant :

```
all_data = datasets.ImageFolder(data_dir, transform=transform)
train_data_len = int(len(all_data)*0.75)
valid_data_len = int((len(all_data) - train_data_len)/2)
test_data_len = int(len(all_data) - train_data_len - valid_data_len)
train_data, val_data, test_data = random_split(all_data, [train_data_len, valid_data_len, test_data_len])
train_loader = DataLoader(train_data, batch_size=batch_size, shuffle=True, num_workers=4)
return train_loader, train_data_len
```

FIG. A.5 : La division des données.

A.4.4 Création du modèle ViT

Afin de concevoir un modèle ViT pré-entraîné qui peut être utilisé pour des tâches de classification d'images, il est possible de prendre des images en entrée, extraire des caractéristiques visuelles et faire des prédictions sur les classes d'images présentes. La fonction utilisée est illustrée dans la figure A.6.

```
model = torch.hub.load('facebookresearch/deit:main', 'deit_tiny_patch16_224', pretrained=True)
```

Fig. A.6 : Chargement d'un modèle pré-entraîné ViT.

A.4.5 Apprentissage

La ligne de code qui suit effectue l'entraînement de notre modèle.

```
def train_model(model, criterion, optimizer, scheduler, num_epochs):
    since = time.time()
```

Fig. A.7 : Fonction d'entraînement.

A.4.6 Test

Dans nous bases de données,les données de test qui seront employées afin d'évaluer la qualité de notre modèle n'ont pas fait partie de la base d'apprentissage, ce qui signifie qu'elles sont nouvelles pour notre modèle.

```
test(model)
```

Fig. A.8 : Fonction d'accuracy.

A.5 Conclusion

Dans cette annexe, nous avons exposé les divers aspects de l'évolution de notre système. Nous avons fourni une description détaillée des outils et des étapes de mise en place de notre système, ainsi que quelques captures d'écran de notre code Python.