PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
Ministry of Higher Education and Scientific Research

UNIVERSITY ECHAHID HAMMA LAKHDAR - EL OUED
FACULTY OF EXACT SCIENCES
Computer Science department

End of study memory
Presented for the Diploma of

ACADEMIC MASTER

Domain: **Mathematics and Computer Science**
Industry: **Computer Science**
Specialty: **Distributed Systems and Artificial Intelligence**

Theme

Artificial Intelligence for Disease Detection using Transformers

Presented by:
Safa Sahraoui
Karima Debbaha

Mr. Mohib Eddine KHEBBACHE MAA Supervisor Univ. El Oued

Miss. Nedjoua Houda KHOLLADI MAA Co-Supervisor Univ. El Oued

Academic year
2022/2023

إهداء

نهدي هذا العمل المتواضع للوالدين
لتشجيعهم لنا و لتضحياتهم كشهادة
على محبتنا لهم طوال سنوات السنة
الدراسية الأولى الى يومنا هذا ومع
خالص امتناننا نود أن نعبرلمشرفينا
الاستاذ خباش محب الدين بشكل
خاص وجميع اساتذة ودكاتره في قسم
الاعلام الالي وكلية العلوم الدقيقة
بجامعة الوادي عامة.

Résume

Le COVID-19 est une maladie respiratoire hautement contagieuse causée par le virus SARS-CoV-2. La détection précoce et précise des cas de COVID-19 joue un rôle crucial dans le contrôle de la propagation du virus et la fourniture de soins médicaux appropriés. Ces dernières années, les techniques d'apprentissage profond (Deep Learning ou DL) ont montré des résultats prometteurs dans les tâches d'analyse et de classification d'images médicales, ce qui aide à identifier et à diagnostiquer les maladies et les problèmes de santé avec une grande précision. Les réseaux de neurones convolutifs (Convolutional Neural Network ou CNN) et les transformers visuels (Visual Transformer ou ViT) sont deux architectures d'apprentissage profond populaires qui ont été appliquées avec succès à la classification des images médicales. L'objectif de ce travail est de faire une étude comparative entre l'exactutude des modèles VIT et des modèles CNN dans la classification COVID-19, comme un cas d'étude, afin de distinguer les cas COVID-19 et d'autres affections respiratoires. Cette étude est réalisée sur une base de données publique de covid-19 contanant des images CT-Scan.

Mots clés: Deep Learning, Convolutional Neural Network, Vision transformers, Medical Images, CT-Scan, COVID-19.

Abstract

COVID-19 is a highly contagious respiratory disease caused by the SARS-CoV-2 virus. Early and accurate detection of COVID-19 cases plays a crucial role in controlling the spread of the virus and providing appropriate medical care. In recent years, deep learning techniques have shown promising results in medical image analysis and classification tasks, helping to identify and diagnose diseases and health problems with great accuracy. Convolutional neural networks (CNNs) and visual transformers (ViTs) are two popular deep learning architectures that have been successfully applied to medical image classification. The aim of this work is to carry out a comparative study between the accuracy of ViT and CNN models in the classification of COVID-19, as a case study to distinguish between cases of COVID-19 and other respiratory diseases. This study is carried out on a public COVID-19 database containing CT-Scan images.

Keywords : Deep learning, Convolutional Neural Network, Vision transformers , Images Medical , CT-Scans , COVID-19

الملخص

COVID-19 هو مرض تنفسي شديد العدوى يسببه فيروس SARS-COV-2 يلعب الاكتشاف المبكر و الدقيق لحالات COVID-19 دورا مهما في السيطرة على انتشار الفيروس وتقديم الرعاية الطبية المناسبة .في السنوات الاخيرة ،اظهرت تقنيات التعلم العميق نتائج واعدة في تحليل الصور الطبية وتصنيف المهام التي تساعد في تحديد وتشخيص الامراض والمشاكل الصحية بدقة كبيرة. الشبكات العصبية التلافيفية (CNN) ومحولات الرؤية (ViTs) نموذجان من التعلم العميق التي تم تطبيقاها بنجاح على تصنيف الصور الطبية.
الهدف من هذا العمل اجراء دراسة شاملة ومقارنة بين دقة نماذج CNN وVIT في تصنيفCOVID-19 ، كدراسة حالة للتمييز بين حالات COVID-19 وامراض الجهاز التنفسي الاخرى .اجريت هذي الدراسة على قواعد البيانات عامة ل COVID-19 تحتوي على صور الاشعة المقطعية .**الكلمات الرئيسية:** التعلم العميق ، الشبكات العصبية التلافيفية، محولات الرؤية ،الصور الطبية، التصوير المقطعي ،كوفيد19

Contents

List of Figures

List of Tables

Introduction

General introduction

In recent years, the outbreak of the COVID-19 pandemic has posed a significant challenge to healthcare systems worldwide. One crucial aspect of managing the spread of the virus is early and accurate diagnosis. Computed Tomography (CT) scans have proven to be valuable diagnostic tools for identifying COVID-19-related lung abnormalities. With the advancements in artificial intelligence (AI), specifically Convolutional Neural Networks (CNN) and Vision Transformers (ViT), there has been considerable progress in automating the classification of CT scans for COVID-19.

CT scans provide detailed cross-sectional images of the internal structures of the body, allowing healthcare professionals to detect lung abnormalities associated with COVID-19, such as ground-glass opacities and consolidations. However, manually interpreting a large volume of CT scans can be time-consuming and subjective, leading to variations in diagnostic accuracy.

CNNs, a class of deep learning models, have been widely adopted for image classification tasks due to their ability to automatically learn relevant features from raw pixel data. CNNs can effectively capture spatial relationships in images, making them well-suited for the analysis of CT scans. By training CNN models on large datasets of annotated CT scans, it is possible to develop robust algorithms capable of accurately distinguishing COVID-19-related lung abnormalities from other lung conditions.

More recently, Vision Transformers (ViTs) have emerged as a promising alternative to CNNs for image classification tasks. Unlike CNNs, which rely on convolutional layers, ViTs utilize a self-attention mechanism to model global dependencies within an image. This attention mechanism allows ViTs to capture long-range interactions, enabling them to learn complex patterns and spatial relationships effectively.

When it comes to classifying CT scans for COVID-19, both CNNs and ViTs have demonstrated their effectiveness. However, few works have been done to compare the accuracy of CNNs and most recently variants of ViTs. To this end, the aim of our work is to carry out a comparative study between CNNs and various ViT variants, namely T2T-ViT (Tokens-To-Token Vision Transformer), VIT-Distillation and VIT-Simple.

This manuscript contain four chapters

- **Chapter 1:** provides a brief overview of the knowledge and information in Machine learning and deep learning algorithms and it techniques.
- **Chapter 2:** presents the types and characteristics of Image Medical and deep learning techniques applied in such domain.
- **Chapter 3:** provides experimental tests and results of classifying COVID-19 CT scan images using CNN and three models of Vision transformers.

Finally we conclude the manuscript by a general conclusion.

Chapter 1

Overview of deep learning models

1.1 Introduction

Within the recent years. Deep Learning (DL), a type of AI that is increasingly in use today, has the potential to transform the way that healthcare is provided in the future. Computer-aided detection has emerged as one of the most crucial methods for quickly diagnosing any diseases as a result of artificial intelligence algorithms. It involves the use of deep learning techniques in the field of image processing to improve the diagnosis . As a result, we cover the fundamentals of machine learning, Deep Learning, image classification techniques, medical imaging, and how to assess the system is effectiveness in this chapter.

1.2 Artificial Intelligence

Artificial intelligence pertains to the advancement of computer architectures or machines that can carry out tasks that typically need human expertise. It requires creating models and algorithms that enable robots to see.

Similar to human intellect, it can reason, acquire knowledge, and make judgments. The goal of artificial intelligence is to simulate human intellect in machines through the use of several subfields, including machine learning, natural language processing, computer vision, and advanced mechanics. Numerous industries, including education, transportation, healthcare, customer service, and many more, can benefit from the application of AI [1].

1.3 Machine learning

Machine learning, also known as ML, is an autonomous technology that enables machines to learn on their own and without direct user instruction. With minimal human intervention, machine learning (ML) uses data and experience to distinguish between examples and create predictions. It is a focus area within the larger field of research on artificial intelligence (simulated reasoning) [2].

Because ML algorithms create models from training data, sometimes referred to as sample data, they are able to anticipate or make judgments without the need for explicit programming. These algorithms are frequently utilized in challenging circumstances where conventional algorithms find it difficult to fulfill the necessary tasks [2].

1.3.1 Machine learning methods

AI models fall into three essential classes.

Supervised machine learning

The basis of supervised machine learning reside on using labeled datasets to train algorithms for accurate data classification or outcome prediction. As input data are added to the model, the weights are tuned until they fit the data correctly. To ensure that the model does not overfit or underfit, this adjustment is applied during the cross-validation procedure. Supervised learning enables businesses to address a range of real-world issues on a large scale, such as putting spam into a different folder from the inbox. Among the supervised learning methods are neural networks, naive Bayes, linear regression, logistic regression, random forests, and support vector machines (SVM)[3].

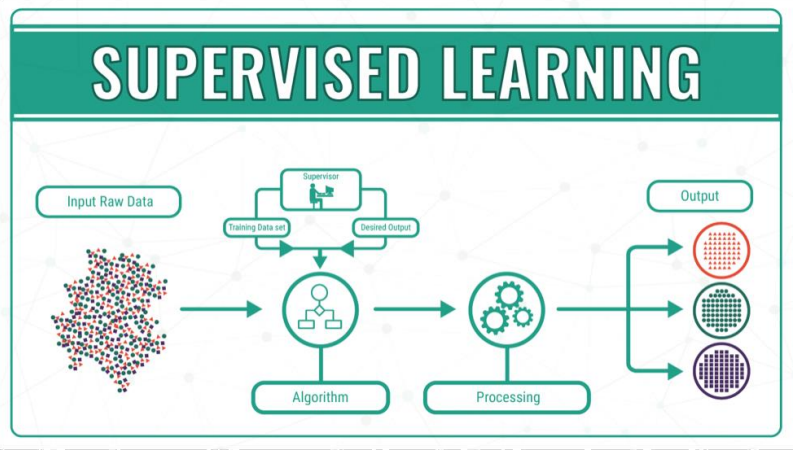

Figure 1.1: Supervised Learning [4]

Unsupervised machine learning

Unsupervised learning, also known as unsupervised machine learning, is used to evaluate and classify unlabeled information. These algorithms independently find hidden patterns and data clusters without assistance from a human. Because it can identify similarities and contrasts in information, this strategy is useful for exploratory data analysis, cross-selling tactics, consumer segmentation, and jobs like picture and pattern recognition. Unaided learning is also used to reduce dimensionality, which reduces the number of highlights in a model. Common techniques used for this include head component analysis (PCA) and single value decay (SVD). Additionally, neural networks, k-means clustering, and probabilistic clustering are employed in unsupervised learning[3].

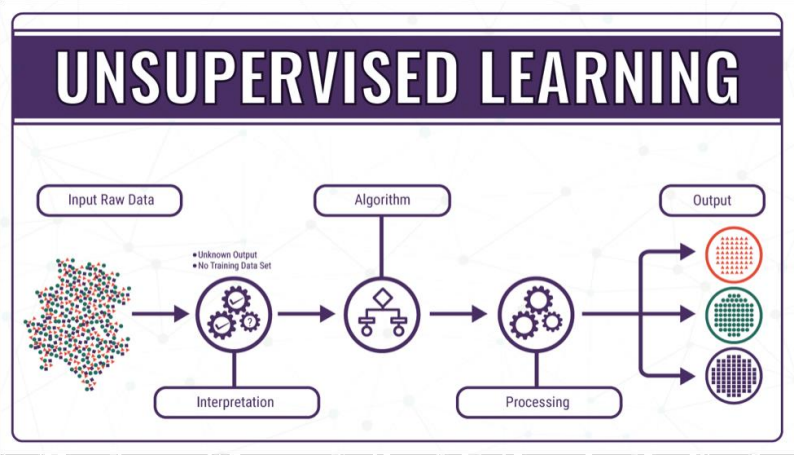

Figure 1.2: Un Supervised Learning [4]

Semi-Supervised Learning

A strategy that creates a balance between supervised and unsupervised learning approaches is semi-supervised learning. During the training process, it directs the classification and feature extraction from a larger amount of unlabeled data using a smaller set of labeled data. Semi-supervised learning is an excellent solution for the problem of sparse labeled data that may hinder supervised learning algorithms. It also proves helpful when classifying a sufficient volume of data becomes expensive or resource-intensive[3].

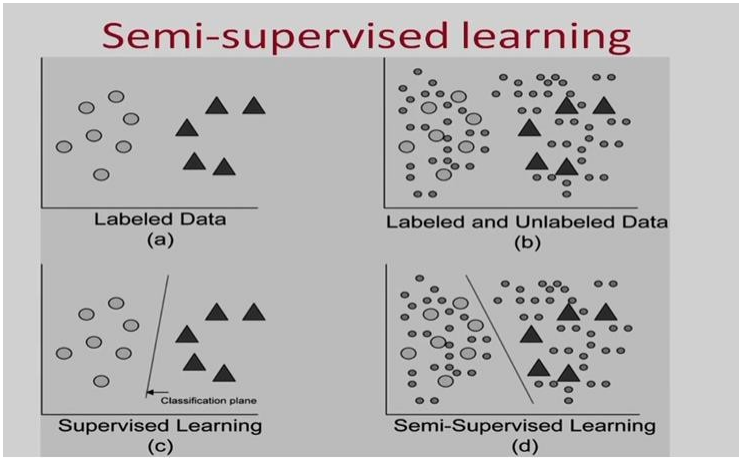

Figure 1.3: Semi-Supervised Learning [5]

1.4 Computer Vision

Computer vision is an area of artificial intelligence that focuses on giving computers the ability to understand and interpret visual data. A study and application field is computer

vision. It also entails the creation of algorithms and techniques that let computers mine digital films and images for information that is relevant to them. By combining machine learning, pattern recognition, and image processing, computer vision systems can analyze and interpret visual data. This makes it possible for them to finish tasks like classifying images, segmenting images, identifying faces, and understanding scenarios. Computer vision has many practical applications in a wide range of fields, including autonomous cars, surveillance systems, medical imaging, augmented reality, robotics, and quality control in manufacturing[6].

1.4.1 Computer Vision Applications

Computer vision is also in demand across several industries due to the growing need for artificial intelligence and AI breakthroughs. On the retail, automotive, security, healthcare, and agricultural sectors, it has a big impact. The applications of computer vision listed below are some of the more popular ones [7]:

- Detection of defects through Computer Vision.
- Optical Character Recognition (OCR) using Computer Vision.
- Monitoring and analysis of crops.
- Utilizing Computer Vision for analyzing X-rays, MRI, and CT scans.
- Monitoring road conditions.
- Building 3D models with the help of Computer Vision.
- Cancer detection through Computer Vision.
- Detection of plant diseases using Computer Vision.

1.5 Natural language processing (NLP)

The study of computing methods for understanding, analyzing, and producing human language is the focus of the subject of research known as natural language processing, or NLP. Making it feasible for robots to accurately comprehend and process natural language in a range of formats, including spoken words, written text, and structured data, is its main objective. NLP incorporates ideas from linguistics, computer science, and artificial intelligence to help humans and machines communicate and extract meaning.

NLP encompasses a wide range of practical applications and tasks, some of which include language translation, sentiment analysis, information retrieval, question answering, text summarization, and speech recognition. Using algorithms and models, it examines linguistic patterns, semantic links, syntactic structures, and contextual cues in textual data. By implementing these techniques to enhance language comprehension and

human-machine communication, NLP aims to close the gap between natural language and computational systems[8].

1.6 Deep learning

Deep learning, a branch of machine learning, represents complicated data abstractions using multi-layered neural networks with complex structures and non-linear transformations. Larger datasets and enhanced computer power have led to the adoption and widespread use of neural networks with more complicated topologies in a range of industries[9].

1.6.1 Neural Networks

The brain network is designed to mimic how the human brain functions. The deep neural network architecture consists of three layers: the info layer, the result layer, and the secret layer, which is sandwiched between the info and result levels. It is founded on the principles of deep learning, a branch of machine learning that employs AI-based techniques for data classification and organizing.

Layers of hubs are included in counterfeit brain organizations (ANNs), including an info layer, at least one hidden layer, and a result layer. Each node, often referred to as an artificial neuron, is connected to other nodes and has a weight and threshold. When the output value of a node crosses the threshold, the node becomes active and transmits data to the network's next tier. If the output is less than the threshold, no data is sent to the following layer[10].

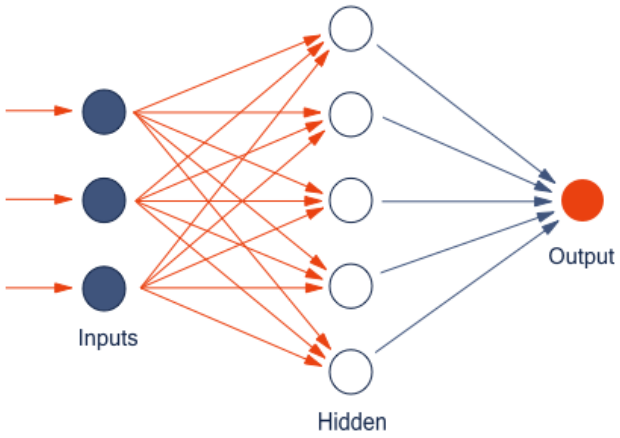

Figure 1.4: Neural Networks global architecture[11]

1.6.2 Convolutional Neural Network

Contrary to ordinary neural networks that rely on matrix multiplications, a convolutional Neural Network (CNN/ConvNet) is a type of deep artificial neural network (DANN/DNN)

that uses a unique technique called as convolution. So, the convolution is a mathematical technique that combines two functions to create a third function.

The primary objective of a ConvNet is to preserve essential features necessary for precise predictions while transforming input data , such as images, into a form that is easier to process and find hidden patterns in the representative feature space of raw data with two or more dimensions [12]. The Figure(1.5) presents the typical architecture of CNN.

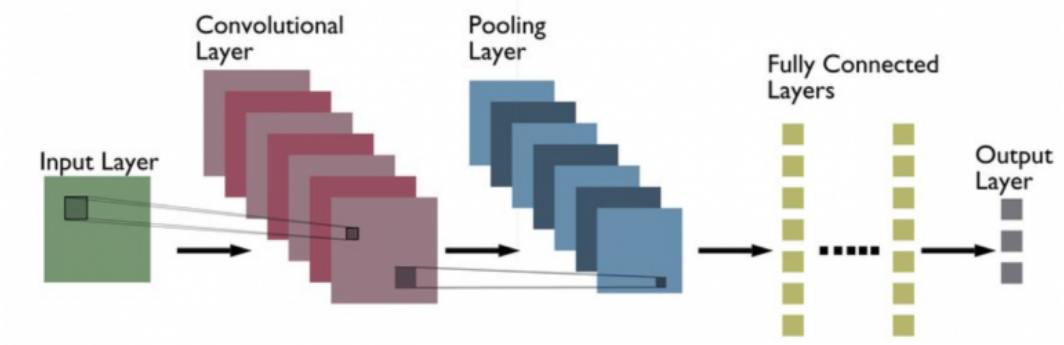

Figure 1.5: Typical-CNN-architecture[13]

As described in Fig(1.5)typical Convolutional Neural Network (CNN) architecture consists of multiple layers designed to process and extract features from input data. CNNs are particularly well-suited for tasks like image classification, object detection, and image segmentation. Is detailed bellow [14]:

Input Layer

input layer receives the raw input data, which is usually an image represented as a grid of pixel values. The dimensions of the input layer depend on the size and shape of the input images.

Convolution Layer

A convolutional layer multiplies the input using a set of weights, just like a regular neural network does. For CNNs, however, the input is two-dimensional. This multiplication is based on a filter or kernel, a two-dimensional array of weights, and an array of input data.

The filter or kernel, which is always smaller than the input, traverses the picture matrix. It multiplies its values by the values of the original pixels to produce a single integer. The channel flows in advances, often to one side and downward, with varying numbers of advances. The output matrix is projected to be smaller than the input matrix.

Pooling Layer

The goal of layer pooling is to reduce the spatial dimensionality of the representation, which will reduce the number of boundaries and computational complexity in the orga-

nization. Controlling overfitting is aided by pooling layers. There are two major groups of pooling layers. Maximum and average pooling:

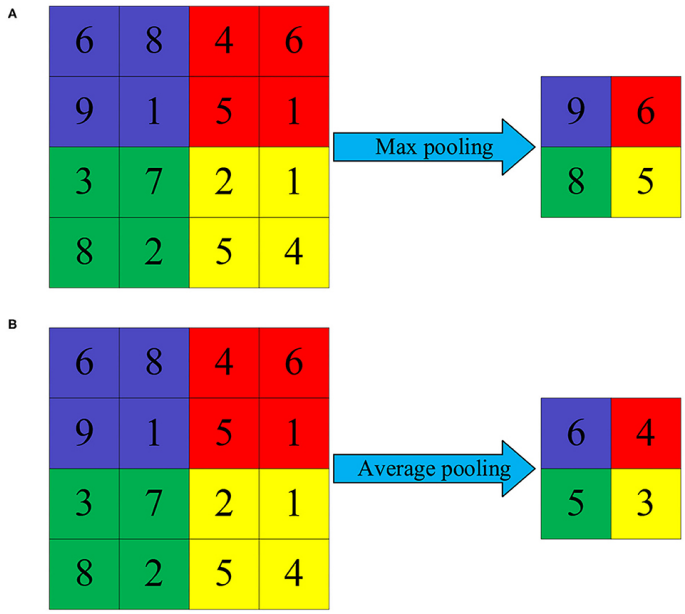

Figure 1.6: Max Pooling And Average Pooling Calculate the maximum and average values for each patch of the feature map[15].

Fully-Connected Layer

Neurons in a completely connected layer create connections with all activations in the layer preceding it, just like in traditional neural networks. The source of the input for the fully connected layer is the output of the preceding layer, whether it be a pooling layer or a convolutional layer. This input is flattened into a one-dimensional vector and fed into the fully linked layer. When a three-dimensional layer is input, this layer converts it into a format that satisfies the specifications of a fully connected layer.

1.7 Transformers

Transformers are a type of profound learning model used in tasks involving understanding natural language.They have a complex neural network architecture and were trained using a lot of annotated text data. The self-attention mechanism, which enables the model to concentrate on different elements of the input sequence when producing predictions, is one of the most crucial components of transformer models. Transformers have generally demonstrated cutting-edge performance on a wide range of NLP tasks and have grown to be one of the most well-known categories of deep learning models in this discipline.

1.7.1 Attention mechanism

People commonly use their "attention mech-anism" to explicitly sift through irrelevant information and concentrate on the crucial components of practical knowledge while they are examining data. In response to this finding, deep learning researchers created attention mechanisms that allow homogeneous material to be sorted through with a focus on the most important features or components[16].

1.7.2 Attention mechanism in computer vision

Similar approaches have been established in the field of CVs. For example, Squeeze-and-Excitation, a unique attention mechanism, was devised to perform feature re-calibration. This mechanism prioritizes informative characteristics pertinent to a certain visual task while downplaying the significance of the remaining features[16].

Self-attention.

The attention mechanism was re-defined as a function that operated on queries, keys, and values derived from the module's input vectors, in contrast to Bahdanau attention. The result is described as a weighted sum of values, where each value's weight is established by allocating attention between queries and keys.

The self-attention procedure is typically carried out in a matrix to expedite parallel computations. To give a quick explanation of the notion, we start by breaking down self-attention into its component parts.

For each inpu $x_i \in Ri = 1, ..., n$ the corresponding query $q_i \in R_q^d$ key $K_i \in R_k^d$, and value $v_i \in R_v^d$ vectors are generated through the parameters $W^q, W^k, andW^v$,respectively d_q, d_k, d_v are the sizes of q_i, k_i, v_i and also the number of features that are learned from x_i .

$$q_i = x_i \times W^q, W^q \in R^{c \times d_q}, \tag{1.1}$$

$$k_i = x_i \times W^k, W^k \in R^{c \times d_k}, \tag{1.2}$$

$$q_i = x_i \times W^q, W^q \in R^{c \times d_q}, \tag{1.3}$$

$$v_i = x_i \times W^v, W^v \in R^{c \times d_v}, \tag{1.4}$$

$$d_q = d_k, \tag{1.5}$$

The output is also a probability calculated as the weighted sum of the calculated weighting values:

$$\alpha_{ij} = Softmax(\frac{\alpha_{ij}^{'}}{\sqrt{d_k}}) \tag{1.6}$$

$$\alpha_{ij}^{'} = q_i \times K_j^T, \tag{1.7}$$

Where $\alpha_{ij}^{'}$ measures the contribution of the j^{th} element of the input to the i^{th} element of the output.Through this operation,$\alpha_{ij}^{'}$ can be regarded as the attention assigned to the element v_i. Thereby the final output attentions can be computed as a weighted sum of all values as follows:

$$Z_i = \sum_j \alpha_{ij} \times v_j \tag{1.8}$$

The element-wise self-attention can be feasibly extended to matrices. In most cases, the query q_i, key k_i and value v_i for each input x_i are generated using parallel matrix computation.x_i,q_i,k_i , v_i can be stacked together to matrices,respectively. Let$X \in R^{s\times c}$ denote the input matrix,Q denote the query matrix K denote the key matrix, and V denote the value matrix, where s is the number of the samples and each matrix is consisted of the elements,i,e.... $X = [x_1; x_2; .....x_s]^T$.Similarly, we compute the attention matrix A and output matrix Z as follows:

$$A = Softmax(\frac{Q \times K^T}{\sqrt{d_k}}) \in R^{s\times s}, \tag{1.9}$$

$$Z = A \times V \in R^{s\times d_v}. \tag{1.10}$$

Multi-head self-attention

Hierarchical features might be more effectively captured by self-attentions to the same input. These self-attention layers function similarly to several convolution layer kernels. The module outputs the final result by concatenating the calculated attentions, given h self-attentions (heads):

$$Z_i = Attention(Q \times W_i^Q, K \times W_i^K, V \times W_i^V), \tag{1.11}$$

$$MultiHead(Q, K, V) = Contact(Z_1.....Z_h)W^O \tag{1.12}$$

where W_i^Q, W_i^J, W_i^V,denote linear projection matrices that map matrices Q, K, V into different subspaces, respectively.W^O is an output projection matrix that concatenates self-attention outputs of all attention heads

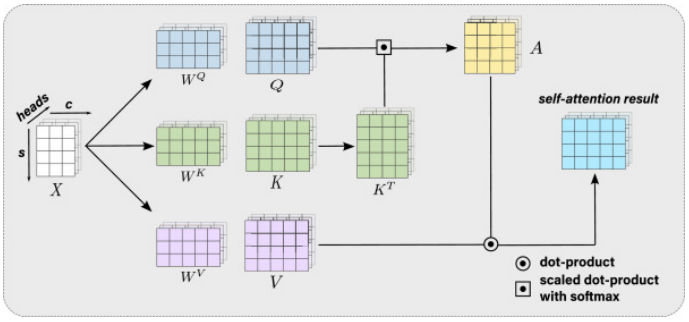

Figure 1.7: brief illustration of a self-attention mechanism[16]

Here is an overview of how the self-attention mechanism works as descibed in fig(1.7):

The self-attention mechanism takes as input a sequence of vectors, often representing words or tokens in a sentence. Each vector is called an "embedding," and these embeddings are typically obtained through a combination of word embeddings .
For each word (or token) in the input sequence, the self-attention mechanism computes three vectors: Query (Q), Key (K), and Value (V). These vectors are derived from linear projections of the original embeddings. The weight matrices used for these projections are learned during training.

The core of the self-attention mechanism is to compute a set of similarity scores between the query vector of a word and the key vectors of all other words in the sequence. This is done for all words in the sequence, producing a set of attention scores. The similarity between a query and a key is often computed as the dot product between the two vectors, but it can also be scaled to prevent gradients from becoming too large.

Attention Weights: The attention scores are then transformed into attention weights using a softmax function. This ensures that the attention weights sum up to 1 for each query and control how much each word contributes to the output for the given query.

Finally, the attention weights are used to calculate a weighted sum of the value vectors. This step combines information from all words in the sequence, with the attention weights determining the importance of each word's contribution. The result is the output of the self-attention mechanism for the current word.The outputs from multiple attention heads are typically concatenated and passed through another linear transformation to produce the final output of the self-attention layer. This output is then passed to subsequent layers in the neural network.

The self-attention mechanism enables the model to weigh the importance of different words or tokens in a sequence based on their relevance to the current word, which is particularly effective in capturing long-range dependencies and relationships in data. It has been a crucial innovation in the success of Transformer-based models in various NLP tasks like machine translation, text generation, and sentiment analysis.

1.7.3 Architecture

Encoder

In a typical transformer architecture, the encoder is composed of n = 6 stacked blocks with the multi-head consideration layer and the feed-forward layer as its two types of layers. The normalization and residual connections layers are applied to these layers. Within each block, the multi-head attention is precisely computed, and the input and output are added to perform layer-wise normalization. The feed-forward layer follows this, and layer-wise standardization is again applied to the amount of information and outcome of the feed-forward layer[16].

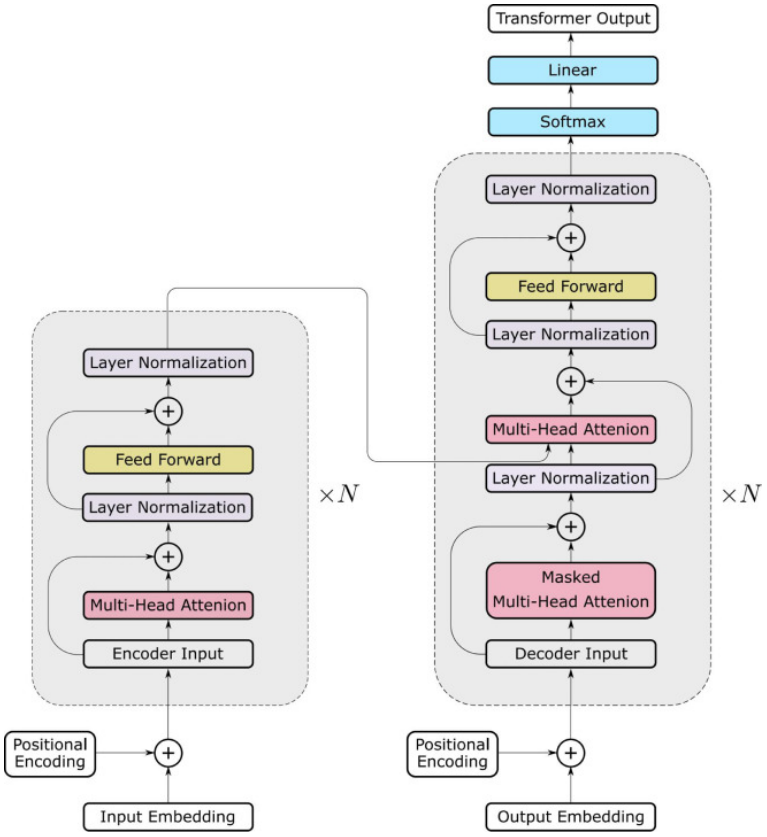

Figure 1.8: A brief illustration of a typical transformer architecture[16]

Decoder

The decoder has a structure considerably different from the encoder, consisting of $n = 6$ blocks. An additional layer of self-attention is layered on top of the encoded output. The initial self-attention layer employs masking to stop future locations from changing the current state because the prediction is based on a known state. The output of the decoder is followed by a linear layer and a Softmax layer to create the final output[16].

1.7.4 Vision transformers

The success of transformers in natural language processing (NLP) motivated researchers in the field of computer vision (CV) to look into how to adapt them for challenges involving vision. The DETR (detection transformer)[17], the Vision Transformer (ViT)[18], the data-efficient image transformer (DeiT)[19], are notable examples of transformer-based vision models that have advanced rapidly[16].

DETR

For the DETR was a ground-breaking use of transformers for object detection in computer vision. DETR's end-to-end detection approach uses a transformer encoder to capture the relationship between visual features recovered by a CNN backbone, in contrast to conventional object identification techniques that rely on manual feature engineering. A transformer decoder generates the object queries, and a feed-forward network determines the labels and bounding boxes of the objects that are found.

ViT

The essential design of the conventional transformer serves as the foundation for the ViT model for image classification. In ViT, the original image is divided into a number of patches, each of which is connected to a positional encoding technique that preserves spatial information by recording the positions of the patches. These patches and a class token are then fed into the transformer, which uses multi-head self-consideration (MHSA) to execute calculations and create embeddings for the patches. The multi-layer perceptron (MLP) is used to categorize the learned image representation. The image representation obtained via ViT is represented by the class token in the subsequent state.

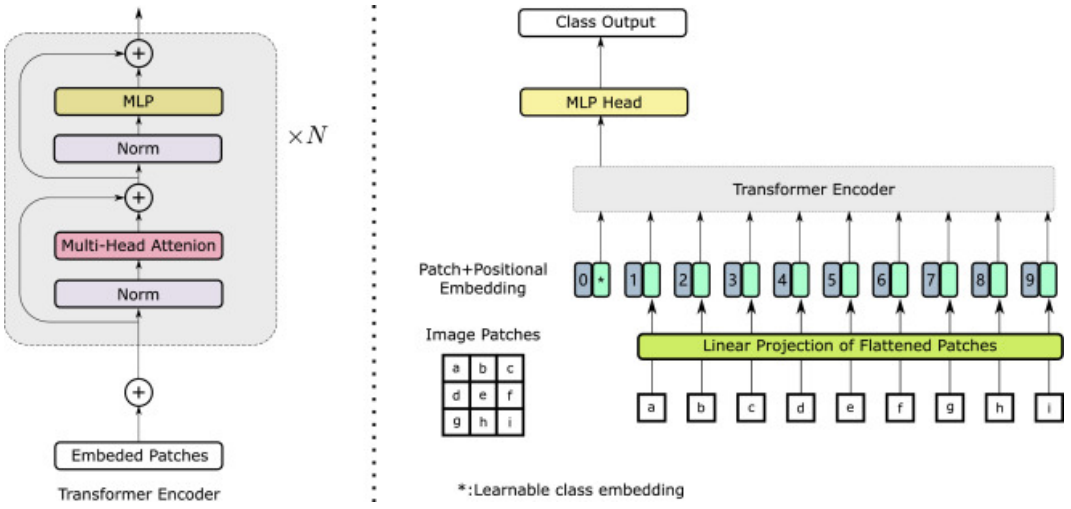

Figure 1.9: Architecture of ViT global[16]

As described in fig(1.9) We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard

Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable classification token to the sequence.

DeiT.

DeiT was created to alleviate the issue of ViT's need on large training datasets by ensuring effective performance on smaller datasets. The answer was a distillation token—a transformer term—that followed the input sequence to learn from the teacher model's output using a teacher-student arrangement and a knowledge distillation framework. Additionally, it was hypothesized that utilizing a CNN as the teacher model could make it simpler to train the transformer to inherit the beneficial inductive bias in the student network.

1.8 Conclusion

Convolutional neural networks (CNNs) and transformer-based deep learning models have considerably advanced the area of artificial intelligence and helped achieve cutting-edge performance in a variety of domains.

Chapter 2

Deep learning in medical images

2.1 Introduction

Medical imaging plays a crucial role in modern healthcare, enabling the diagnosis, monitoring, and treatment of various medical conditions. With the advancement of artificial intelligence (AI) and deep learning techniques, a promising approach called transformers has emerged as a powerful tool for analyzing and interpreting medical images. Transformers, originally introduced for natural language processing tasks, have demonstrated their efficacy in a wide range of applications, including computer vision tasks such as object recognition and image classification.

2.2 Medical Image

Technology advancements have made it possible for medical professionals to diagnose and treat patients without the risk of adverse effects. One of the best ways to achieve that objective is to be able to observe what is going on inside the body without undergoing any kind of invasive surgery. The process of creating visual representations of specific locations within the body for the purpose of diagnosing and monitoring medical conditions is known as medical imaging.

The process has significantly improved public health. Medical imaging is one of the most effective tools at the patients' disposal and can be used for both therapeutic and diagnostic purposes.

Medical image refers to any visual representation or picture of the human body or its parts, obtained through various imaging techniques. These images are used by healthcare professionals, such as radiologists and physicians, to diagnose and monitor medical conditions, as well as to plan and guide treatments[20].

2.2.1 Types of medical images

There are many different types of medical imaging, and as technology progresses, more methods are being created. Each type uses a different method to create visuals of what is happening inside the body.

X-rays(Radiography):

The imaging procedure takes pictures of the body is interior using electromagnetic radiation. X-ray is the most common and widely used method of radiography. For this imaging technique, a x-beam machine radiates high-energy waves onto the body. These waves are not absorbed by soft tissues like organs or skin, but are absorbed by hard tissues like bones. The x-ray results are transferred onto an image by the machine, which highlights the body parts that absorbed the waves in white and the unabsorbed material in black.

CT (Computed Tomography) scans:

These are an X-ray kind that creates 3D images to aid in diagnosis. It also goes by the name Computed Axial Tomography (CAT), and it creates cross-sectional images of the human body using X-rays. The patient can recline on a motorized table inside the enormous circular opening of the scanner. A narrow, "fan-shaped" X-ray beam is then created by rotating the detector and X-ray around the patient, passing through a portion of the patient's body to create an image. Compared to traditional x-rays, CT scans provide clearer, more accurate images of the body's bones, blood arteries, internal organs, and soft tissue. The use of CT scans typically eliminates the need for exploratory surgery.

PET (Positron Emission Tomography) scans:

PET scans use a small amount of radioactive material to produce images of organs and tissues.

SPECT (Single Photon Emission Computed Tomography) scans:

SPECT scans use radioactive material to produce images of organs and tissues.

Fluoroscopy:

Fluoroscopy uses X-rays to produce real-time images of internal structures and is often used in minimally invasive procedures.

Nuclear Medicine scans:

Small amounts of radioactive material are used in nuclear medicine scans to provide images of the organs and tissues.

MRI (Magnetic Resonance Imaging) scans:

Radio waves and magnetic fields are used in this particular imaging procedure to examine the human body's organs and other structures. The cycle needs a X-ray scanner, which is an immense cylinder that contains a huge roundabout magnet. The body's protons of hydrogen are aligned by this magnet's magnetic field. The protons are then presented to radio waves, making the protons pivot. The protons relax and realign themselves when the radio waves are turned off, releasing radio waves that the machine can sense to create an image.

Ultrasound:

Ultrasound, a medical imaging technique, utilises high-frequency sound waves that bounce off tissues to create visual representations of joints, muscles, organs, and soft tissues. It

operates similar to illuminating the interior of the body, but instead of light, the waves penetrate through the layers of skin and can only be captured using electronic sensors. Widely considered as a safe and cost-effective method of medical imaging, ultrasound has a broad range of applications and is recognized for its absence of harmful effects.

2.2.2 characteristics of medical images

Medical images can have several important characteristics that can impact their diagnostic utility and interpretation. Some of the key characteristics of medical images include[21]:

Resolution:

The quality of detail captured in a medical image is known as its resolution, which is usually quantified in pixels per inch or pixels per millimeter. Images with higher resolution contain greater levels of detail, making them more effective in identifying small abnormalities or lesions.

Contrast:

The contrast of a medical image pertains to the variation in brightness or color observed among different sections of the image. Images with high contrast exhibit a more pronounced distinction between various structures and tissues, facilitating easier differentiation. On the other hand, low-contrast images can pose challenges in discerning crucial details.

Noise:

Noise in a medical image refers to unpredictable fluctuations that can hinder the visualization of significant structures or patterns. Techniques in image processing can be employed to minimize noise, although excessive processing can potentially compromise the image's resolution and other vital characteristics.

Motion:

During the image acquisition process, motion artifacts may arise if there is movement in the object being captured. These artifacts can impede the visibility of significant structures or patterns within the image. To mitigate motion artifacts, various image processing techniques can be employed, or alternative imaging modalities that are less susceptible to motion can be utilised.

Overlapping structures:

When different parts of an image are obscured by other structures, overlapping structures can occur, making it difficult to see important details. This can be an issue in pictures

with unfortunate difference, or when imaging inward designs that are somewhat clouded by different designs.

Brightness and color:

Color and brightness can be used to bring out particular patterns or structures in an image, making it easier to see important details. However, image processing techniques can have an impact on brightness and color, which can make it difficult to see fine details.

Modality-specific features: :

Each imaging technique has advantages and limitations. For instance, X-beams give great bone detail, while X-ray examines are better at imagining delicate tissues. These are just a few of the important aspects of medical images that can affect how well they can be used for diagnosis and how they are interpreted.

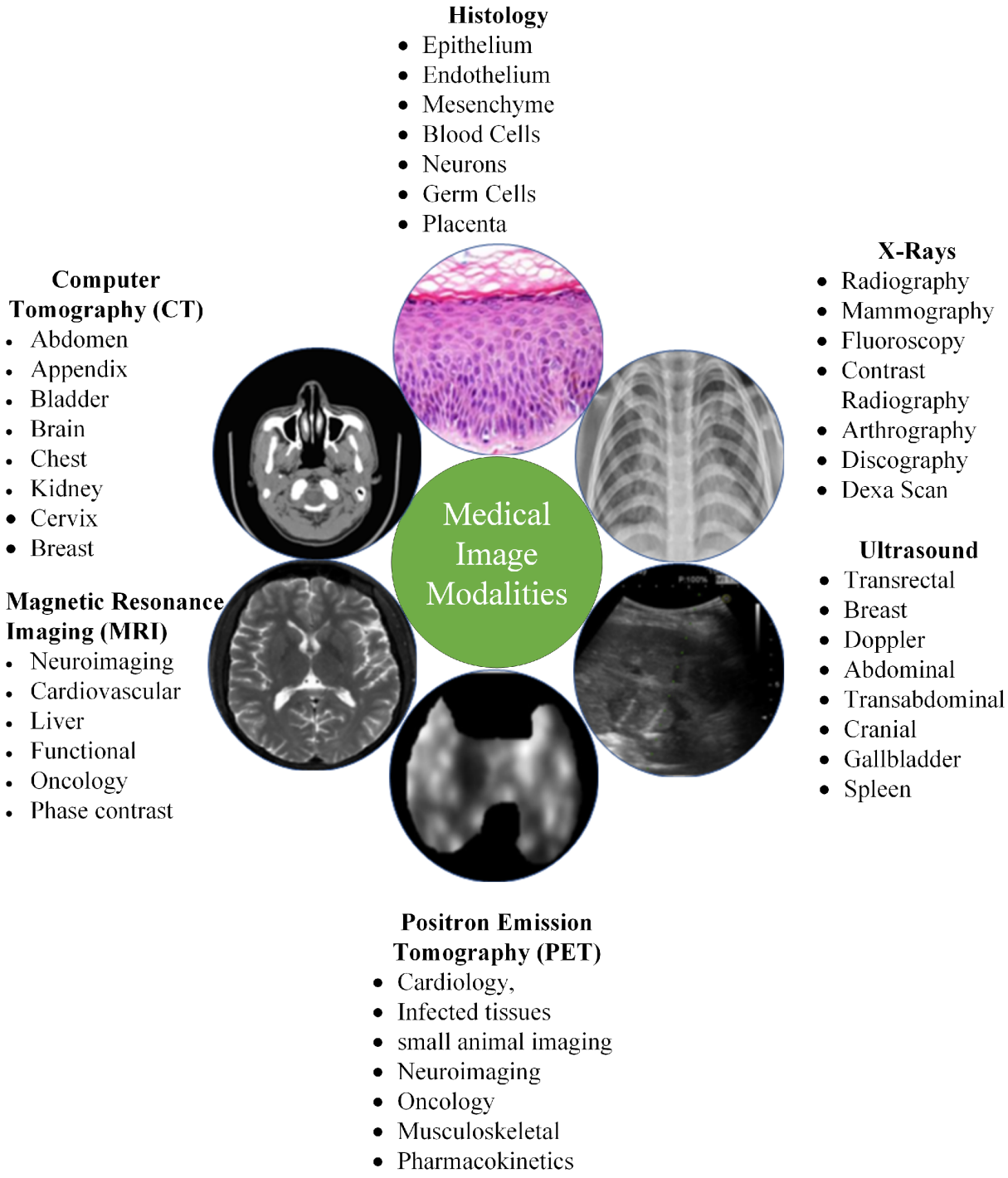

Figure 2.1: Typology of medical imaging modalities[22].

2.3 Deep learning applications in medical image

Deep learning has made significant strides in various medical image analysis tasks, offering improved accuracy and efficiency in diagnosis, and treatment planning. Some notable applications of deep learning in medical image analysis include[23]:

2.3.1 Classification

Deep learning models can classify medical images into different categories, aiding in disease diagnosis. For example:

- Using CNN diagnosing diabetic retinopathy by classifying fundus images into different severity levels.
- Using ViTs to classify different types of skin lesions in dermatology images.

2.3.2 Detection and Localization

Deep learning can locate and identify specific objects within medical images. This is useful for detecting and localizing abnormalities like tumors in radiological images for example:

- Using CNN Detecting and localizing lung nodules in CT scans for early lung cancer detection.
- Using ViTs Detecting abnormalities in X-ray images, such as fractures or foreign objects, using ViTs.

2.3.3 Segmentation

Deep learning models can segment medical images into different regions, allowing for precise identification of structures and abnormalities. Forexample :

- Using CNN Segmenting brain tumors in MRI scans to assist surgeons in accurate tumor removal.
- Using ViTs Segmenting and analyzing retinal layers in optical coherence tomography (OCT) scans for diagnosing retinal diseases.

2.3.4 Registration

Deep learning can help align and fuse different medical images from various modalities, aiding in multi-modal diagnosis and treatment planning.

2.4 Conclusion

Deep learning have emerged as a promising approach for analyzing medical images, offering the ability to capture global context and long-range dependencies. Their applications span various tasks, including image classification, segmentation, and generation.

Chapter 3

COVID19 Classification Using ViT-Models and CNN: comparative study

3.1 Introduction

Classification of medical images is a critical task in healthcare, enabling accurate diagnosis, treatment planning, and monitoring of various diseases. Computed tomography (CT) is a widely used imaging method that provides detailed cross-sectional images of the body. As devoted in the previous chapter, our comparative study consists in investigating the performance of the most popular deep learning techniques, namely CNN and Transformers, in the classification of COVID-19 CT images, as a case study.

3.2 COVID-19

COVID-19, a highly contagious illness, has resulted in approximately 386,000 fatalities worldwide, with 6.5 million confirmed cases as of June 3rd, 2020. The scarcity of testing poses a significant challenge in controlling the spread of this disease. The primary testing method, reverse transcription polymerase chain reaction (RT-PCR), faced a severe shortage of test kits during the peak of the COVID-19 outbreak. Consequently, numerous suspected cases could not undergo timely testing, unwittingly contributing to further transmission. To address the scarcity of RT-PCR test kits, healthcare facilities have resorted to alternative diagnostic approaches. One such method involves utilizing computed tomography (CT) scans for screening and diagnosing COVID-19. China's National Health Commission and State Administration of Traditional Chinese Medicine have recognized CT scans and other chest imaging techniques as important tools for diagnosing COVID-19 in their Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia[24].

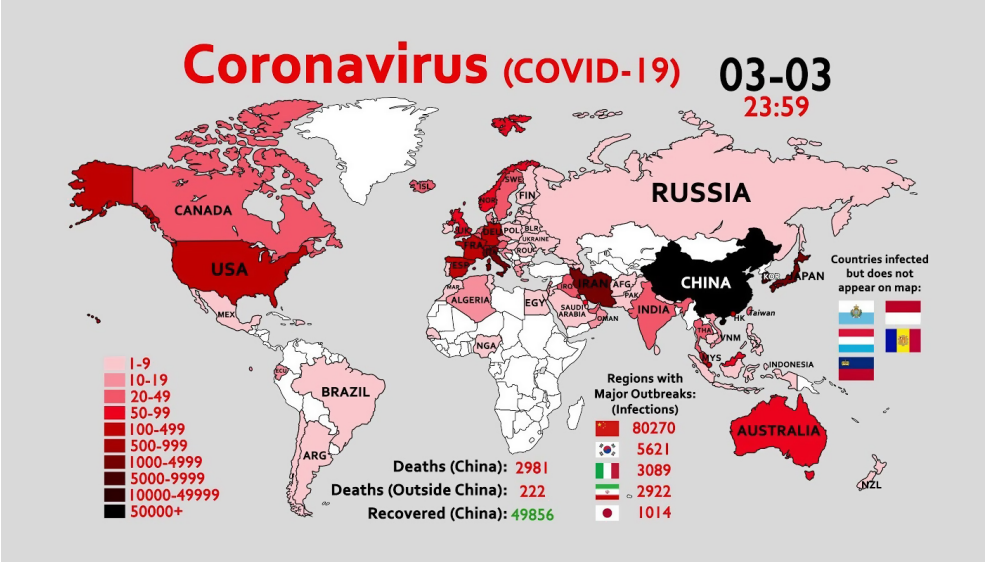

Figure 3.1: Distribution of Covid-19 cases worldwide[25].

The coronavirus, a member of the Coronavirus (CoV) virus family, was first known as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), then in February 2020, the World Health Organization (WHO) officially designated it as COVID-19.[26] Without a doubt, this dangerous virus has affected people in more than 210 countries throughout the world. A person who contracts COVID-19 may exhibit a variety of signs and symptoms of infection, including fever, coughing, and respiratory conditions that are similar to influenza. Additionally, some people may have bodily aches, weariness, nasal obstruction, and loss of taste. When the infection is severe, it may cause pneumonia, breathing problems, multiple organ failure, and, unfortunately, even death. Droplets released by infected people during coughing and sneezing are the main means of virus transmission.[27].

Several COVID-19 strains were discovered in 2021 in a number of countries, including the United Kingdom, Brazil, and India. Compared to earlier strains, the transmission and death rates linked to these new variants have become more significant[28].

The coronavirus has the capacity to spread infection to people in different waves. A recent uptick in cases is been reported in some of the countries that had successfully stopped the COVID-19 pandemic's spread.

A further difficulty with the virus that researchers have discovered is its propensity to evolve structurally over time. Finding a reliable and practical solution consequently becomes difficult. Researchers have confirmed that the virus exhibits dynamic patterns, which makes it challenging for an antidote created for a particular strain to successfully tackle subsequent versions. Even though vaccines are widely distributed worldwide, several nations continue to record new cases and daily mortality[29].

3.2.1 lung diseases

Pneumonia[30] is a condition in which the air sacs in the lungs enlarge due to a variety of causes, including viruses, bacteria, fungus, and more. Methods including lung x-rays, blood cultures, sputum analysis, computed tomography scans, and similar techniques are primarily used to diagnose pneumonia. The common signs of pneumonia include coughing, fever, and breathing problems. The best course of action for treating pneumonia depends on the type of illness, whether it is bacterial, viral, or fungal. Numerous viral infections, from mild ones like the common cold to more serious ones like SARS and MERS, can result in pneumonia. The main factor contributing to pneumonia-related deaths is usually respiratory failure.

The standard means for diagnosing pneumonia, such as chest x-rays and CT scans, may be overloaded by the surge of patients in pandemic-like conditions, overwhelming hospital imaging departments. The scenario becomes more complex because interpreting x-ray pictures calls for skill and ideal circumstances. It follows that it is necessary to analyze these photos, and technologies based on artificial intelligence and machine learning (AI/ML) are well suited for doing so. It might be difficult for radiologists to distinguish between distinct pulmonary abnormalities such pneumothorax, pleural effusion, pneumo-

nia, TB pulmonary, pulmonary scarring, and more since lung infections frequently appear as opaque patches on radiographs.

3.3 COVID-19 Datasets

The datasets are divided into three main categories: speech data, text data, and medical imaging data. Medical image datasets, such as CT scans, X-rays, MRT, or ultrasound images, are largely used for screening and diagnosing COVID-19. We group medical datasets into these two groups since CT scans and X-rays make up the majority of COVID-19 datasets. A few datasets might include various kinds of medical images. Due to the scarcity and expensive cost of COVID-19 test kits, medical facilities in hospitals and clinics frequently use medical image-based diagnosis, which lessens the strain on traditional PCR-based screening. Datasets for medical images are frequently pre-processed utilizing methods like segmentation and augmentation. The process of segmentation entails dividing the image to pinpoint the area of interest [31].

The COVID-19 case reports have a number of uses, including (a) gathering data at regional levels (such as cities and counties), (b) visualizing cases and forecasting daily new cases, recoveries, etc., (c) examining the impact of non-pharmaceutical interventions (NPIs) on COVID-19 cases, and (d) studying the impact of mobility on the number of new cases. Cough and breath sounds are included in speech data sets, which can be used to diagnose COVID-19 and gauge the severity of the illness. These data sets can be examined using statistical, big data, and ML (machine learning) techniques for tasks involving prediction [31].

3.3.1 COVID-19 medical image datasets

The distribution of different samples of COVID-19 X-ray and CT images is shown in Fig(3.2) According to the distribution of datasets, X-ray images were classified in four categories: normal, bacterial pneumonia, viral pneumonia and COVID-19 pneumonia, and CT images were classified into two categories: normal and COVID-19 pneumonia. Figure 3a shows the X-ray image of normal lungs, (b) shows the X-ray image of COVID-19 infected lungs, (c) shows the X-ray image of the lung infected with virus, and (d) shows the X-ray image of the lung infected with bacteria. Figure 3e,f show CT sections of normal lung and COVID-19 virus-infected lung.

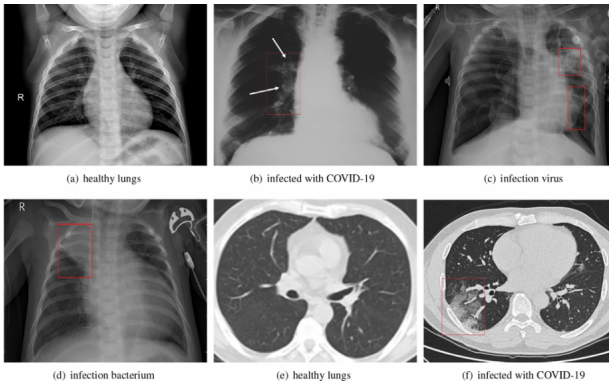

Figure 3.2: X-ray images (a–d) and CT images (e–f)[32].

3.3.2 COVID-19 textual datasets

various sources are used to gather data including: COVID-19 case reports and analysis, social media data, scholarly articles and NPI data sets.

3.3.3 Speech datasets

Utilizing three main techniques, namely cough sounds, speech analysis and speech-based stress detection, speech (audio) data sets are essential for the diagnosis and identification of COVID-19 positive cases using machine learning (ML) approaches. Meanwhile, telemedicine or smartphone applications can be used to collect datasets and provide speech-based COVID-19 diagnosis services remotely.

3.4 COVID-19 : Classification and Detection

DL algorithms such as, CNN and transformers use many imaging modalities to establish the main automatic COVID-19 diagnosis, namely COVID-19 classification and detection. the differences between the two tasks resides in outputs and performance evaluation. however, they share the generic pipeline for DL-based COVID-19 diagnosis ranging form pre-processing, segmentation, feature extraction, to classification, as described bellow.

3.4.1 Data Pre-processing

Pre-processing, the first step, entails transforming raw images into a format suitable for further analysis. Different equipment produce medical images with differences in size, slice thickness, and scan count (for example, 60 and 70 in CT). These elements increase the imaging data's heterogeneity, which leads to non-uniformity amongst datasets. So the pre-processing step mostly entails scaling, normalization, and occasionally RGB to grayscale conversion. When dealing with CT data, the voxel dimension is resampled to an isomorphic resolution in order to account for differences between datasets. Additionally, smoothing techniques are used to increase image quality by lowering noise and increasing

the signal-to-noise ratio.

Extraction or segmentation of particular regions of interest from a picture in order to aid the classification problem is a vital step in the pre-processing phase. When COVID-19 is diagnosed, the lungs are especially vulnerable to infection. Lung regions are therefore segregated from chest X-ray (CXR) and CT images for accurate diagnosis, and they are later used in the following processing processes. The manual segmentation of the lung region is a time-consuming, arduous job that primarily relies on the radiologists' knowledge and skill. To quickly analyze COVID-19 photos, DL-based segmentation approaches like few-shot segmentation and semantic segmentation allow automated identification of infected regions. For this task, commonly utilized segmentation models such fully convolutional networks are used. Pixel values may occasionally be thresholded.

3.4.2 Feature extraction and classification

Feature extraction and classification are the main steps in the DL based COVID-19 diagnosis phase. Automatic feature extraction and binary or multiclass classification are hallmarks of DL approaches. There are two techniques to extract features: using transfer learning using a trained model, or creating a unique CNN model from scratch. Many DL-based neural networks use CNN as their core building block to extract features from the input images. Additionally, there are extra layers for batch normalization and dropout as well as several convolutional and pooling layers.

In the following sections, we focus on the COVID-19 classification task, while attempting to introduce some theoretical points and evaluate the performance of a few new transformer models for comparison purposes.

3.5 COVID-19 Classification using CNN model

Chest computed tomography (CT) features are used to accurately categorize COVID-19-infected individuals and determine whether they fall into the infected class or not. Based on chest CT images, the COVID-19 disease classification is performed using the CNN model where the process of categorize patients with COVID-19 infections is depicted as describe in Fig(3.3):

- **Convolutional Layer**: from the input image, a spatial convolution operations is performed to provide feature maps organized in hierarchical manner. As feature space, such feature maps consist of low-level features which are extracted in initial layers, and high-level features which are extracted in deeper layers in ct scan image . The convolution operation is followed by an activation function, such as ReLU, that introduces non-linearity into the network.
- **Pooling Layer:** The Pooling Layer is responsible for reducing the dimensionality of the feature maps along the spatial dimension. The two primary pooling techniques used are Average-Pooling and Max-Pooling.

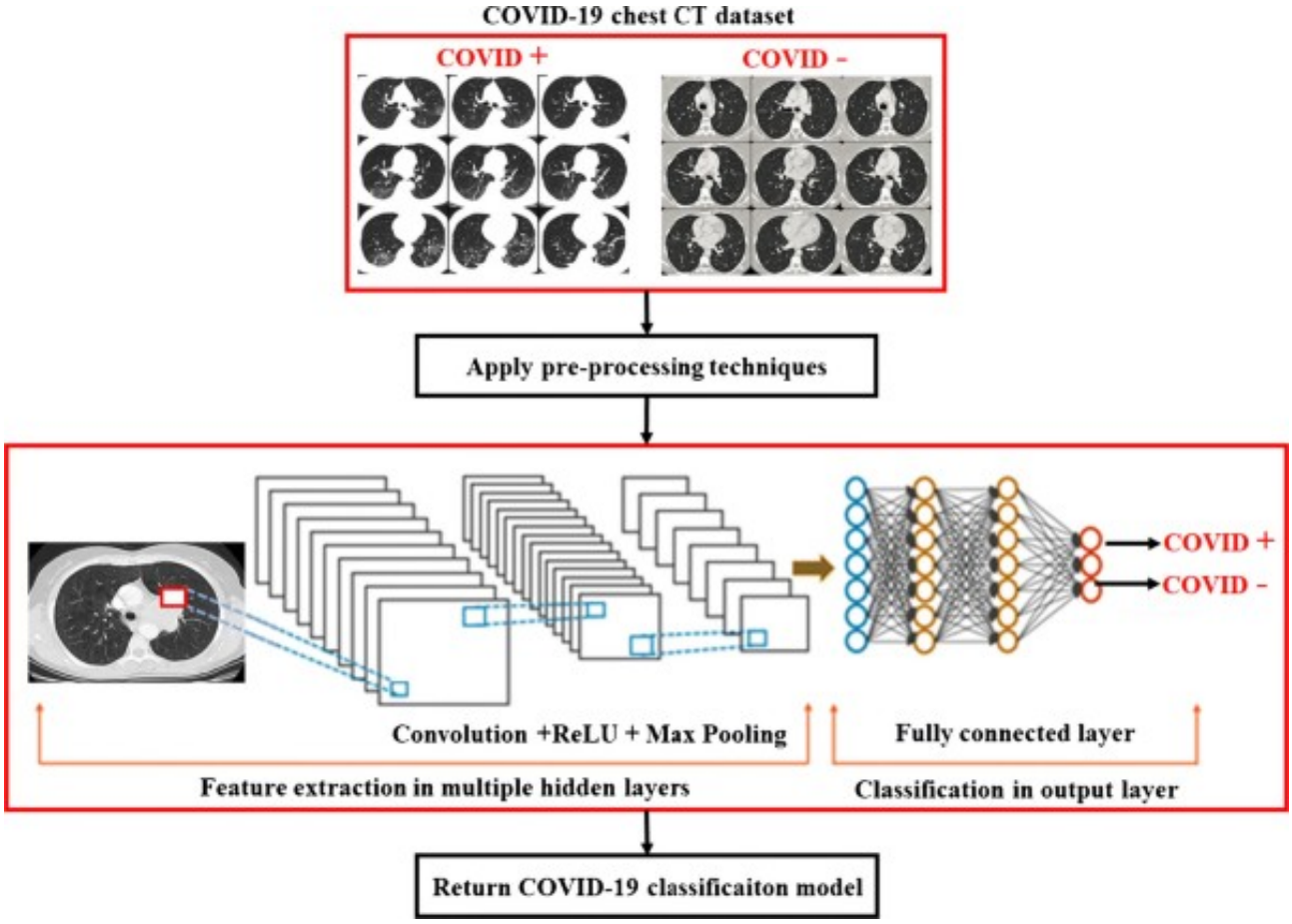

Figure 3.3: Diagram of the training process of the CNN-based COVID-19 classification model[33].

- **Fully-connected Layer:** The Classification Layer is responsible for performing the actual task of classification. It is composed of multiple layers in a neural network. The configuration of these layers, including the number of layers and the number of nodes in each layer, are hyperparameters that need to be optimized. Following the layers, there is a softmax layer that assigns class scores to each class for the input CT scan image, representing the probabilities of belonging to two classes. The input image is then classified into COVID-19 or non-COVID-19.

In general, CNN models have been used in this context from two perspectives. Most studies have used pre-trained models that were initially trained on other large datasets such as ImageNet, which contains a vast collection of natural images [34]. Using transfer learning, the weights learned from the pre-trained deep learning architecture serve as the initial starting point for fine-tuning the weights of layers from the target COVID-19 dataset. So, the original classification layer should be replaced with a new layer with nodes corresponding to the specific classes for COVID-19 classification. Various pre-trained models are commonly used for COVID-19 diagnosis, including AlexNet[35], different versions of Visual Geometry Group (VGG) [36] and ResNet [37], Inception [38], Xception [39], InceptionResNet [40], DenseNet [41], and more.

In addition to pre-trained models, custom models are also popular for COVID-19 classification, involving training a model from scratch without utilizing any pre-trained model.

3.6 COVID-19 Classification using Vision Transformer model

The Vision Transformer (ViT) architecture can be used to classify computed tomography (CT) scans into several groups after analyzing the data. By taking particular actions, the ViT architecture, which was initially created for image classification tasks, may be modified to process CT scans (see Fig 3.4).

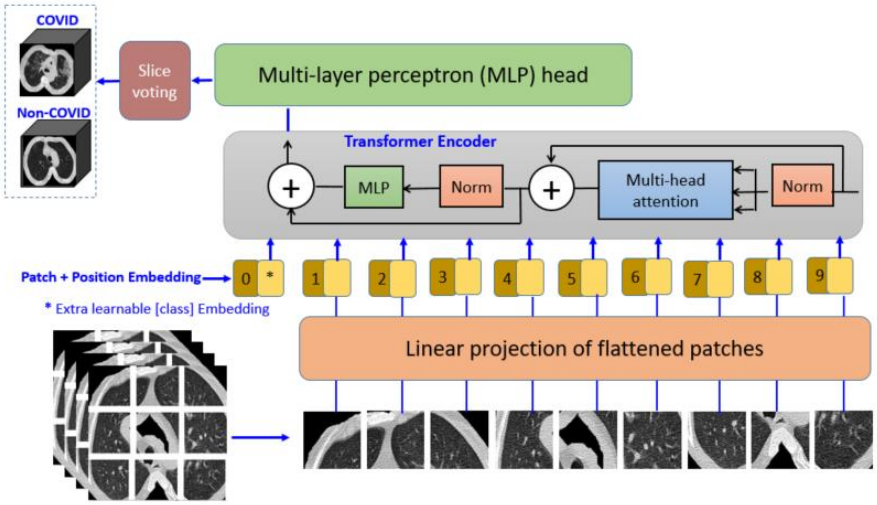

Figure 3.4: Typical ViT architecture used in COVID-19 classification[42].

As described in Fig(3.4), the classification pipeline of COVID-19 CT scans using simple ViT is detailed bellow [43, 44]:

- **Image Preprocessing**: CT scans are typically high-resolution images with multiple slices. Preprocessing steps may include resizing the images to a fixed size, normalizing pixel values, and converting the CT scans into a suitable format for input to the ViT model.

- **Patch Embedding:** Number of fixed patches are divided up into the input image. These patches are flattened and supplied to the transformer encoder with positional embedding. In a transformer encoder, classification is done using a multi-layer perceptron head. In comparison to a convolutional neural network, the use of a transformer enables more detailed and reliable predictions.

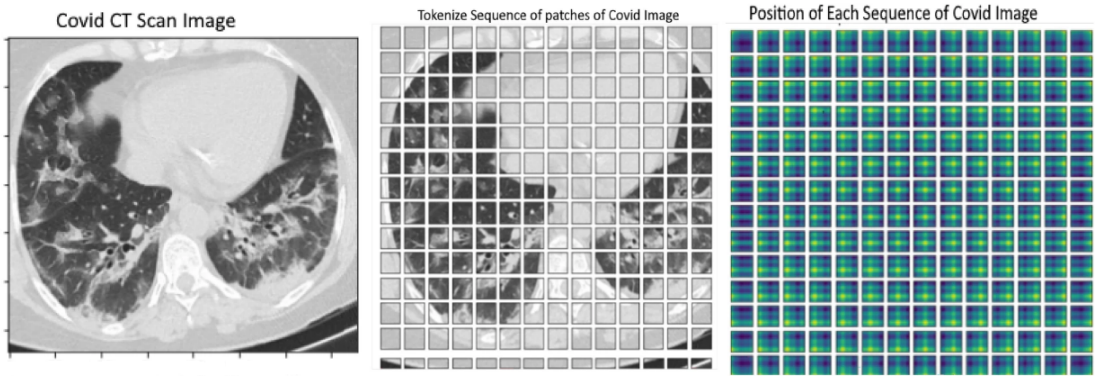

Figure 3.5: patch and positional embeddings[44]

A single image $xIH \times W \times C$ is split up into a number of patches $x = [x_1, ..., xn]$ I. Each image patch is converted into a vector and then linearly projected to produce a series of patch embeddings, where $x_0 = [Ex_1, ..., Ex_n]$ IND, where N is the total number of patches. Positional embeddings are added to the series of patches for having a tokenize input sequence in order to capture the positional information.

Positional encodings are added to the patch embeddings. These encodings convey the spatial relationship between the patches in the image.
This series of tokens is put through a transformer encoder to create a contextualized encoding with a wealth of semantic data. Transformer's encoder layers, which include layer normalization, multi-layer perceptrons, multi-head self-attention, and residual connections, are the same as those used in the regular Transformer. Three linear layers make up the self-attention mechanism, which maps tokens into intermediate representations, keys, queries, and values.

- **Transformer Encoder :** It patch embeddings are fed into a stack of transformer encoder layers along with the positional encodings. Multi-head self-attention techniques and feed-forward neural networks make up each transformer encoder layer. The feed-forward networks analyze the data locally within each patch, and the self-attention mechanism enables the model to capture global dependencies between the patches.

- **A linear layer** or a combination of linear layers and activation functions is attached to the output of the classification token. This Classification Head is responsible for mapping the features learned by the ViT model to the specific classification task at hand, producing class probabilities or predictions.

In order to enable the ViT model to perform image classification, a special token known as the classification token is prepended to the sequence of patch embeddings. The output of this token is used for classification tasks.

In some variations of the ViT architecture, such as Distillation and Tokens-to-Token, a different pieces can be used for extending the capabilities of the ViT model beyond image classification.

3.6.1 Distillation

In order to transmit knowledge from one neural network (the teacher or pre-trained model) to another (the student or smaller model), a common deep learning technique known as distillation modeling is used. This can assist reduce complex models into simpler ones that are both more effective and, to some extent, perform as well[19].

Soft Distillation

[45] limits the Kullback-Leibler uniqueness between the softmax of the educator and the softmax of the understudy model. Let Z_t be the logits of the teacher model, Z_s the logits of the student model. We denote by τ the temperature for the distillation,$lambda$ the softmax function ψ, and the Kullback–Leibler divergence loss (KL) and cross-entropy (LCE) coefficient on the ground truth labels y. The objective of distillation is:

$$\mathcal{L}_{global} = (1-\lambda)\mathcal{L}_C E(\psi(Z_s, y) + \lambda\tau^2 KL(\psi(Z_s/\tau), \psi(Z_t/\tau)). \tag{3.1}$$

Hard-label distillation:

It present another strategy for refining in which the educator's hard choice fills in as the genuine mark. The teacher's hard decision is $y_t = argmax_c Z_t(c)$, and the objective of this hard-label differentiation is:

$$\mathcal{L}_{global}^{hardDistil} = \frac{1}{2}\mathcal{L}_C E(\psi(Z_s, y) + \frac{1}{2}\mathcal{L}CE(\psi(Z_s), y_t) \tag{3.2}$$

Depending on the particular data augmentation, the teacher is hard label may shift for a given image. We'll see that this choice is better than the usual one because it doesn't have any restrictions and is a little easier: The instructor's prediction, y_t, assumes the same role as the actual name, y. Keep in mind that mark smoothing can also transform hard marks into delicate marks[46] , where the true label is considered to have a probability of $1-\varepsilon$, and the remaining ε is shared across the remaining classes. We fix this parameter to $\varepsilon = 0.1$ in our all experiments that use true labels[19].

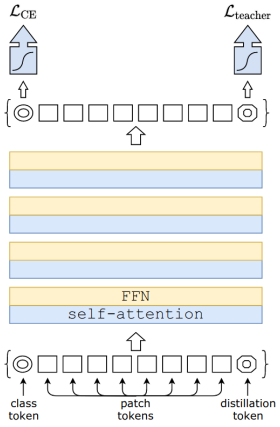

Figure 3.6: Distillation model architecture[47] .

As described fig(3.6)The architecture of a distillation model includes the following components:

Teacher Model (Pre-trained Model): The teacher model is a well-established and pre-trained model that has already learned to perform a specific task effectively. It is usually a deep neural network with a large number of parameters. The architecture of the teacher model depends on the task, but it is typically designed to be highly accurate.
Student Model: The student model is a smaller neural network that is being trained to mimic the teacher model's behavior. It is designed to be more compact and computationally efficient than the teacher model. The student model's architecture may be a simplified version of the teacher's architecture or a different neural network architecture entirely.

Class Token: It's a special token or embedding that captures high-level features or information about the image as a whole. The class token aggregates information from different parts of the image and helps the model make a final prediction. It's like a summary of the image's content.

Patch Tokens: the input image is divided into smaller non overlapping patches, and each patch is treated as a token. These patch tokens are typically linearly embedded and processed by the model. Patch tokens capture local information within their respective patches. The model processes these patch tokens individually and then aggregates their representations to form a global understanding of the image, often using techniques like selfattention.

Distillation Token: The "distillation token" refers to a specific token or embedding within the student model (the smaller model) during the knowledge distillation process. It plays a crucial role in transferring knowledge from the teacher model (the larger model) to the student. The distillation token can be seen as a bridge between the teacher and student models. During training, the distillation token may be used to capture information from the teacher's model's predictions, intermediate representations, or other relevant aspects. This information is then used to guide the training of the student model to mimic the teacher.

Tokens-to-Token ViT :Training Vision Transformers from Scratch

Tokens-to-Token ViT

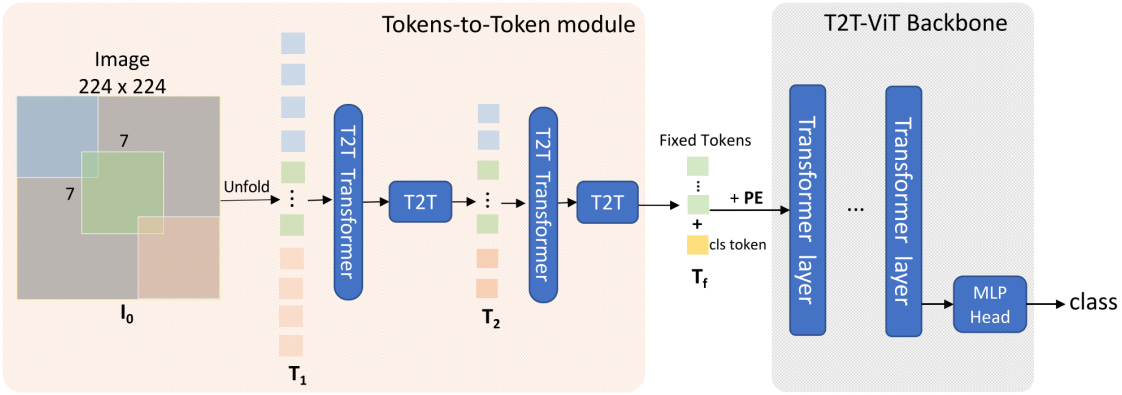

Figure 3.7: The overall network architecture of T2T-ViT[48].

Proposed Tokens-to-Token Vision Transformer (T2T-ViT) as a solution to get over the restrictions of basic tokenization and the ineffective backbone of ViT. T2T-ViT can gradually tokenize the image into tokens and has an effective backbone. Consequently, T2T-ViT contains two essential parts (Fig. 3.7)[48]:

- layer-wise "Tokens-to-Token module" (T2T module) to model the local structure information of the image and reduce the length of tokens progressively
- "T2T-ViT backbone" to draw the global attention relation on tokens from the T2T module. We adopt a deep-narrow structure for the backbone to reduce redundancy and improve the feature richness after exploring several CNN based architecture designs.

1.Tokens-to-Token: Progressive Tokenization

The Token-to-Token (T2T) module means to defeat the limit of basic tokenization in ViT. It continuously structurizes a picture to tokens and models the nearby construction data, and in this way the length of tokens can be diminished iteratively. There are two steps in each T2T process: Soft Split (SS) and restructuration (Fig. 3.7).

Soft Split:

they applied the soft split to the restructured image I to model local structure information and shorten token lengths. We specifically divided the restructured image into patches that overlapped in order to prevent information loss when creating tokens from it. Thusly, each fix is related with encompassing patches to lay out an earlier that there ought to be more grounded connections between's encompassing tokens. Each split patch's tokens are joined together to form a single token, making it possible to combine information from adjacent pixels and patches[48].

When conducting the soft split, the size of each patch is $k \times k$ with s overlapping and p padding on the image, where k_s is similar to the stride in convolution operation. So for the reconstructed image $I \in R^{h \times w \times c}$, the length of output tokens T_o after soft split is

$$l_0 = [\frac{h + 2p - K}{K - s} + 1] \times [\frac{w + 2p - k}{k - 1} + 1]. \tag{3.3}$$

Each split patch has size $k \times k \times c$. We flatten all patches in spatial dimensions to tokens $T_o \in R^{l_o \times k^2}$

After the soft split, the output tokens are fed for the next T2T process.

T2T module

The T2T module is able to gradually reduce the length of tokens and alter the image's spatial structure by carrying out the aforementioned Re-structurization and Soft Split iteratively. The iterative cycle in T2T module can be planned as

$$T_i^{'} = MLP(MSA(T_i), \tag{3.4}$$

$$I_i = Reshape(T_i^{'}) \tag{3.5}$$

$$T_{i+1} = SS(I_i), i = 1....(n - 1). \tag{3.6}$$

For the input image I_0 , we apply a soft split at first to split it to tokens: $T_1 = SS(I0)$. After the final iteration, the output tokens T_f of the T2T module has fixed length, so the backbone of T2T-ViT can model the global relation on Tf .

Additionally, the T2T module uses a lot of memory and MACs because the tokens are longer than in the typical ViT case (16 × 16). To address the constraints, in our T2T module, we set the channel aspect of the T2T layer little (32 or 64) to diminish Macintoshes, and alternatively take on a proficient Transformer like Entertainer [49] layer to cut down on memory usage when the GPU has limited memory. In our experiments, we conduct an ablation study to examine the difference between using the Performer layer and the standard Transformer layer[48].

Re-structurization:

As shown in (Fig 3.8) , given a sequence of tokens T from the preceding transformer layer, it will be transformed by the self-attention block :

$$T' = MLP(MSA(T)), \tag{3.7}$$

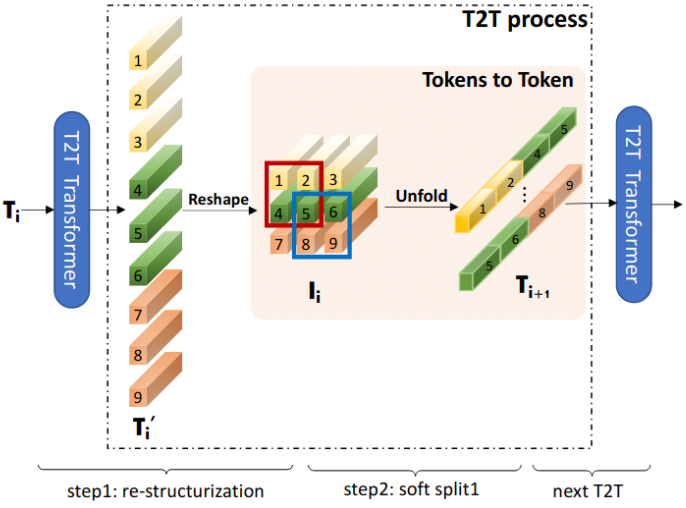

Figure 3.8: T2T model procces[48]

where "MLP" is the standard Transformer's multilayer perceptron with layer normalization, and "MSA" is the multihead self-attention operation with layer normalization[18] . Then the tokens$T^{'}$ will be reshaped as an image in the spatial dimension

$$I = Reshape(T^{'}). \tag{3.8}$$

Here "Reshape" re-organizes tokens $T^{'} \in R^{l \times c}$ to $I \in R^{h \times w \times c}$ to
where l is the length of $T^{'}, h, w, c$ are height, width and channel respectively, and $l = h \times w$.

2.T2T-ViT Backbone
backbone of ViT. We explore five architecture designs from CNNs to ViT:

- **. Efficient backbone for T2T-ViT**
  - Reduce the redundancy
  - Improve feature richness
- **.Architecture designs**
  - Dence connection as DenseNet
  - "Deep-narrow vs shallow wide " structures as wide -ResNet
  - Squeeze - an Excitation(SE)Networks
  - Split head in multi-head attention layer as ResNext
  - Ghost operations as GhostNet .

For tokens with fixed length Tf from the last layer of T2T module, we concatenate a class token to it and then add Sinusoidal Position Embedding (PE) to it, the same as ViT to do classification:

$$T_{f_0} = [t_{cls}; T_f] + E, E \in R(l+1) \times d \tag{3.9}$$

$$T_{f_i} = MLP(MSA(T_{f_i} - 1))), i = 1....b \tag{3.10}$$

$$y = fc(LN(Tf_b)) \tag{3.11}$$

where E is Sinusoidal Position Embedding, LN is layer normalization, fc is one fully-connected layer for classification and y is the output prediction[48].

3.7 Related Works

In the table that follows, we have outlined many studies publications that are associated with classification CT-Scans images using CNN and Transformers model together with their most impressive achievements. We make an effort to categorize them in accordance with the year of publication, technique used, dataset, and performance metric.

Ref.	**Year**	**Datasets**	**Method**	**Accuracy %**
[50]	2018	COVIDx	VGG-16	98.87%
[51]	2020	SARS-CoV-2 CT	CNN	95.8%
[52]	2020	UCSD COVID-CT	ResNet50	88.6%
[53]	2021	COVIDx	ViT with performer	91.0% 96.0%
[54]	2021	COV19-CT-DB	Use sub-volumes for 3D images	94.3%
[55]	2021	COV19-CT-DB	Feature aggregation by trans, CNN features, data resampling	-
[56]	2021	RSNA intracranial hemorrhage dataset	Multiple CNNs, GAN for domain alignment	98.0%
[57]	2021	-	Teacher-student model for knowledge distillation	-
[58]	2021	COVID-CTset	Research on patch size	95.4%
[59]	2021	-	Anatomy-aware transformer with localization Unet	-

3.8 Experimental setup

3.8.1 DataSet: COVID-CT

The most popular in literature, this dataset is a collection of chest computed tomography (CT) images from patients with confirmed or suspected COVID-19 infections is the COVID-CT dataset. From 216 patients, the dataset includes 349 CT scans, both positive and negative. The DICOM format is used to store the images, which have a resolution of 512 by 512 pixels. The figure(3.9) illustrates some positive examples[60].

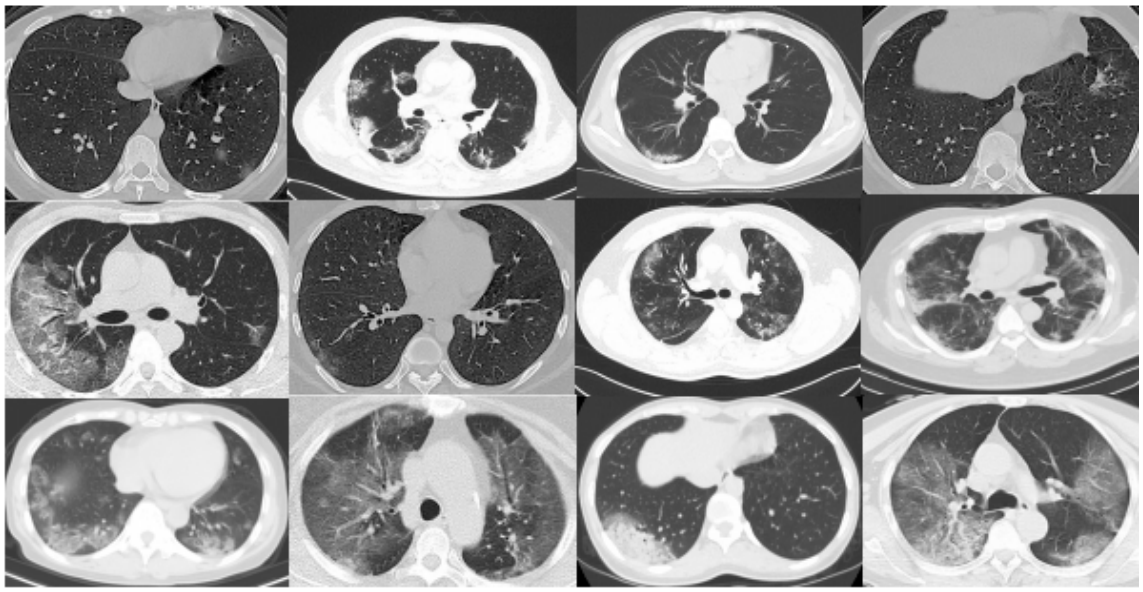

Figure 3.9: Some positive examples[60]

Data Preparation

Image preparation is an essential step in training, validating, and testing deep learning models that work with images. It involves transforming and enhancing the raw image data to make it more suitable for the specific task at hand .The datasets used in the experiment were collected from GitHub . the datasets composed tow categories Covid and non-Covid . it was splited for three subsets like this table :

	ViT(Vision Transformer)				CNN		
	Train	Validation	Testing	Total	Train	Validation	Total
	2D slice	2D slice	2D slice	2D slice	2D slice	2D slice	2D slice
percentage	58%	27%	15%	100%	80%	20%	100%
Images	425	208	118	751	485	108	593

Table 3.1: Dataset preparation

- **Train set:** Train set is used to train the deep learning model by feeding it input samples and their corresponding labels or targets. The model learns from this data to make predictions or perform the desired task.
- **Validation set:** the validation set is a subset of the datasets that is held out and used for evaluating the model is performance during training. It helps in assessing the model's generalization ability, identifying issues like overfitting, and guiding decisions on hyperparameter tuning.
- **Testing set:**Test data, also known as test sets , are used to assess the performance and generalization ability of a trained deep learning model. They serve as a benchmark to evaluate how well the model can make predictions or perform tasks on unseen data that it has not been exposed to during training or validation.

3.8.2 Tools

The following tools and libraries have been used during implementation of this project:

Kaggle

Kaggle is an online community for enthusiasts of machine learning and data science. Kaggle lets users work together, find and publish datasets, use notebooks with GPU integration, and compete with other data scientists to solve problems related to data science. With its powerful tools and resources, this online platform, founded in 2010 by Anthony Goldbloom and Jeremy Howard and acquired by Google in 2017, aims to assist professionals and students in achieving their data science journey goals[61].

Figure 3.10: Kaggle logo [62]

Python Language

Python is an object-oriented, high-level, interpreted language with dynamic semantics. Due to its significant level, in-fabricated information structures, dynamic composing, and dynamic restricting, this language is ideal for Quick Application Improvement and may likewise be utilized as a prearranging or stick language to associate dissimilar parts. The focus on readability in Python's simple syntax helps to keep maintenance costs low. Both modularity and code reuse in applications are made easier by Python's module and package system. Open-source, the Python programming language and its extensive standard library are freely available in binary or source form for all major platforms[63].

Figure 3.11: Python logo[64]

CUDA

is an equal figuring stage and programming model created by NVIDIA for gen- eral figuring on graphical handling units (GPUs). Using the power of GPUs, developers can dramatically accelerate computing applications with CUDA. The CUDA Tool compartment from NVIDIA gives all that you really want to foster GPU- sped up applications. The CUDA runtime, development tools, GPU-accelerated libraries, and a compiler are all included in the CUDA Toolkit. An easy-to-use distribution for CUDA-supported platforms and architectures is provided by the CUDA container images[65].

Figure 3.12: *cuda logo[66]*

Pytorch

Facebook developed the Open Source Deep Learning library Pytorch, which is compatible with the Python programming language. It is a tool that makes it simple for novice and experienced developers to create Deep Learning applications. The use of multi-dimensional matrices known as tensors to represent data is one of PyTorch's unique features. These keep the predictions, the parameters of the hidden layers, and the inputs of the Deep Learning architectures. PyTorch is a Python-based framework designed to be adaptable as a Deep Learning tool, as was previously mentioned. Numpy's workflow is so similar to PyTorch's that you might not even notice it[67].

Figure 3.13: *Pythorch logo[68]*

3.8.3 Performance evaluation: evaluation metrics

The next step is to use the test dataset to determine the model's performance using a variety of evaluation metrics. The most used in classification task are:

Confusion Matrix

It is a presentation estimation for AI characterization issue where result can be at least two classes. There are four distinct combinations of predicted and actual value in this table[69].

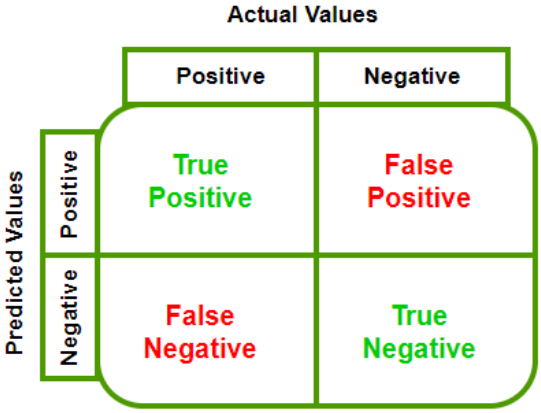

Figure 3.14: Confusion Matrix[70]

The terms used in defining a confusion matrix are :

True positive **(True Positive)**: Observation is predicted positive and is actually positive.

False positive**(False Positive)** : Observation is predicted positive and is actually negative.

True negative **(True Negative)**: Observation is predicted negative and is actually negative.

False negative **(False Negative)**: Observation is predicted negative and is actually positive.

Accuracy

Accuracy indicates the proportion of all forecasts that were correct. determine whether the ratio of patients with the illness to those without it (true negative, false positive) accurately reflects the disease's prevalence (true positive, false negative).

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{3.12}$$

Precision

The extent of positive qualities out of undeniably expected positive cases is known as accuracy or the positive prescient worth. In other words, precision is the percentage of positive values that have been correctly identified:

$$Precision = TP/TP + FP \tag{3.13}$$

Sensitivity

Sensitivity, recall, or the TP rate (TPR) is the percentage of positive values from all actual positive instances (the percentage of real positive cases that are correctly identified):

$$Sensitivity = TP/TP + FNF1 \tag{3.14}$$

F1 score

The F1 score, frequently known as the F score or F measure, is the consonant mean of accuracy and awareness.

$$F1score = 2 \times (Precision \times Sensitivity)/(Precision + Sensitivity) \tag{3.15}$$

3.8.4 Performance evaluation: evaluation model

ViT model parameters

For the comparison purpose, we attempt to evaluate the performance of two variant of simple ViT model, such as ViT Distillation and ViT Tokens-to-Token. the Table (3.2) presents the most parameters used in our study.

Parameters	Desription	Type	Default
Device	Computation device. ('cuda' or 'cpu')	str	cuda
Epochs	the number times to iterate over the dataset	int	40
image-size	Image size. If you have rectangular images, make sure your image size is the maximum of the width and height	int	224
patch-size	Number of patches. image-size must be divisible by patch-size.The number of patches is: $n = (image - size//patch - size) * *2$ and n must be greater than 16	int	32
num-classes	Number of classes to classify	int	2
dim	Last dimension of output tensor after linear transformation nn.Linear(..., dim)	int	12
depth	Number of Transformer blocks	int	12
heads	Number of heads in Multi-head Attention layer	int	8
mlp-dim	Dimension of the MLP (FeedForward) layer	int	40
channels	Number of image is channels	int	3
dropout	Dropout rate.	float between [0, 1]	0
emb-dropout	Embedding dropout rate	float between [0, 1]	0.1

Table 3.2: Arguments used in VITs Models training

CNN Model parameters

in our study, we have used a pre-trained CNN model (effcientNetB0), which was initially trained on ImageNet database. The Fig(3.5) (mettre la ref ici) illustrates the architecture of effcientNetB0 .

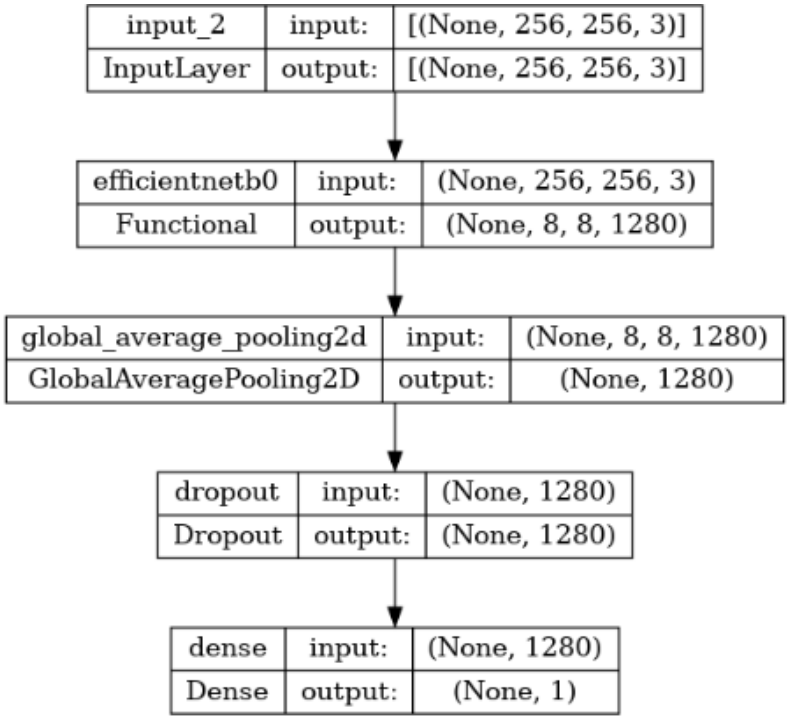

Figure 3.15: An input layer

- The input layer has the correct image dimension of (256, 256, 3).
- The None in output shape is a reserved place for the number of samples, which the model does not know yet.
- EfficientNetB0 sits right after the input layer.
- The last (classification) layer has output of dimension 1: NonCovid probability.
- Most importantly, since we've frozen the parameters of EfficientNetB0, the number of trainable parameters is only 1281.

3.9 Results and discussion

3.9.1 COVID-19 Classification using CNN model

Figure 3.16: Evaluation model CNN

3.9.2 COVID-19 Classification using ViT model and its variants

COVID-19 Classification using Simple VIT model

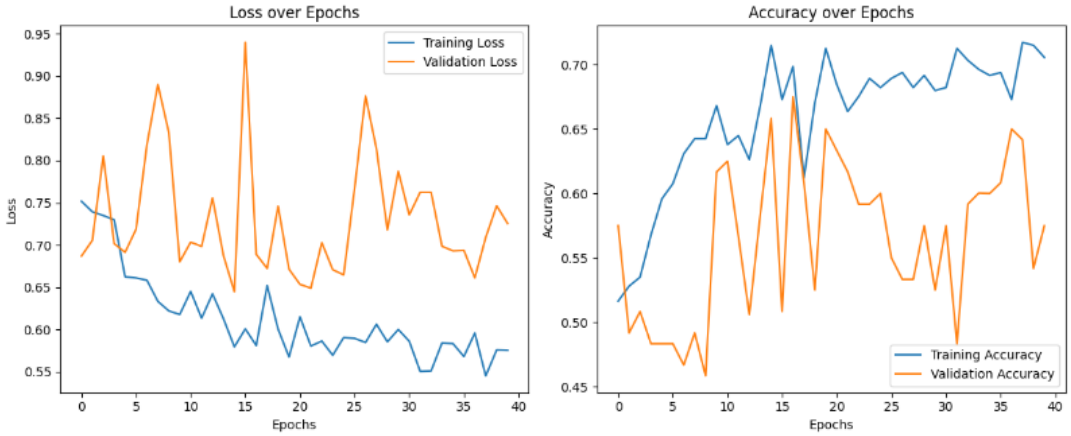

Figure 3.17: Evaluation curves of simple vit model

COVID-19 Classification using VIT Distillation model

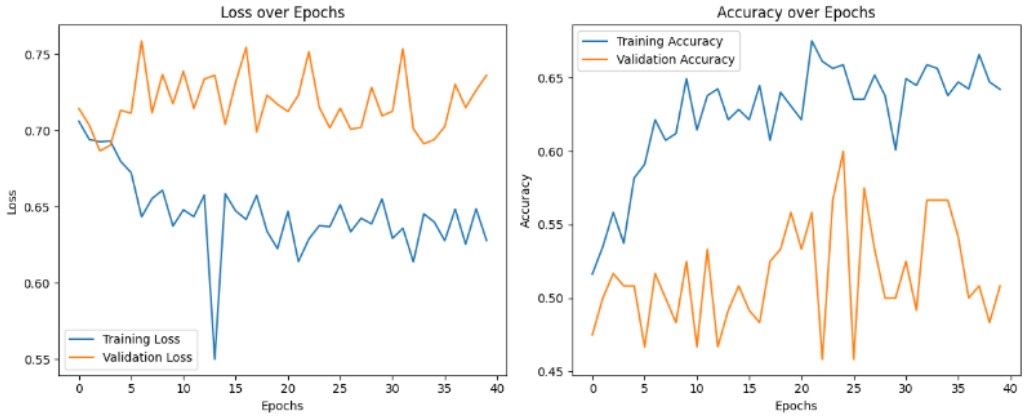

Figure 3.18: Evaluation curves of vit Distillation

COVID-19 Classification using VIT Tokens-to-Token model

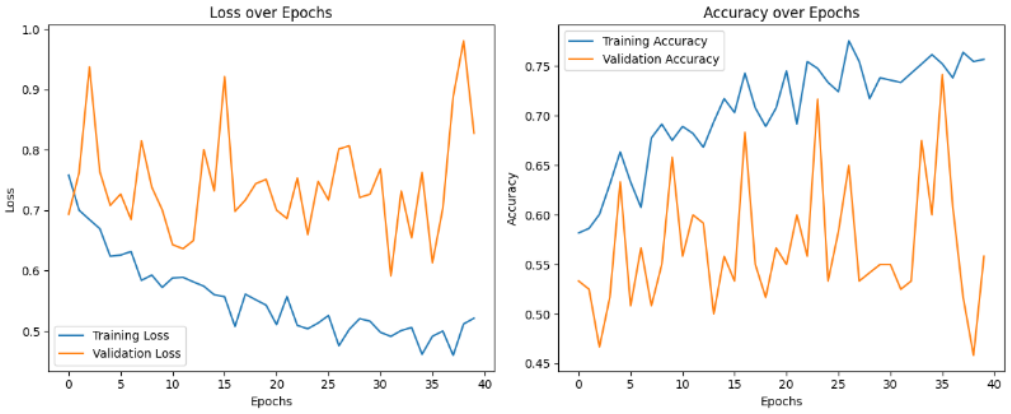

Figure 3.19: Evaluation curves of vit Tokens-to-Token model

	CNN			
	COVID	Nom COVID	Acc (%)	F1
COVID (prediet)	67	15		
Nom COVID (prediet)	3	63		
Average			87.84	0.881

Table 3.3: Evaluation results of CNN Model

	VIT-Simple			
	COVID	Nom COVID	Acc (%)	F1
COVID (prediet)	22	38		
Nom COVID (prediet)	12	46		
Average			57.5	0.467

Table 3.4: Evaluation results of VIT-Simple model

	VIT-Distillation			
	COVID	Nom COVID	Acc (%)	F1
COVID (prediet)	12	48		
Nom COVID (prediet)	8	52		
Average			50.83	0.3

Table 3.5: Evaluation results of Vit-Distillation

	VIT Tokens-to-Token			
	COVID	Nom COVID	Acc (%)	F1
COVID (prediet)	21	39		
Nom COVID (prediet)	17	41		
Average			52.5	0.55

Table 3.6: Evaluation results of Vit-Tokens-to-Token

3.9.3 Discussion

After conducting a comparative analysis between CNN and the three VIT models, the results show that the CNN model performed better than the VIT models. This is due to the size of the dataset and the architecture of the CNN and VIT models. In addition, one may say that VIT models are more sensitive to the dataset size compared to the CNN models (it requires large number of training data). However, this work did not consider the aspect of transfer learning that considers the pre-trained models for both CNN (VGG and RESNET) and VIT. Accordingly, this opens the door for a further investigation

3.10 Conclusion

In this chapter we present the definition of COVID-19, types of COVID-19 datasets, and our explanation Classification and detection of the Coronavirus (COVID-19) We have explained the classification of the Coronavirus (COVID-19) using the CNN model and the classification of the Coronavirus (COVID-19) using the Vision Transformer and also We summarized the related work and in the experimental setting, defined the database for this COVID-CT project, learned about the tools and libraries during the implementation of this project, and presented detailed results for each model, as well as comparisons between the results.

General Conclusion

In this thesis, we tried to provide a solution to the problem of COVID-19 classification using several methods belonging to the field of machine learning. The main objective of this project is to compare different machine learning techniques in terms of learning and performance in order to solve the problem of covid19 classification.

This is done by CT-scans images this images medicale , it as a dataset for different technologies.

In this regard, we have presented tow different models , which are Convolutional Neural Network and Vision Transformers to apply them on a CT-scans datasets . After teaching and evaluating them, we get generally weak results for all models, especially during the testing phase, and this is due to the weak learning of models because of the small dataset.

CNN model is the bestperformanceand achieved the highest F1 score, which was 0.88. ViT-Pythorch models produced satisfactory results and achieved F1s score.

Throughout the time working on this project, we learned so much about artificial intelligence techniques specially machine learning models and how to apply these models to classification deseases.

Notation And Abbreviated Terms

AI : artificial intelligence
ML : Machine Learning
MLP : Multilayer Perceptron
CNN : Convolutional Neural Network
DNN : Deep neural network
ANNs : Artificial Neural Networks
DNN : Deep Neural Network
CV: Computer Vision
DETR: DEtection TRansformer
ViT : Vision Transformer
DeiT : Data-efficient Image Transformers
CT : Computed Tomography
MRI : Magnetic Resonance Imaging
PET : Positron Emission Tomography
SPECT: Single Photon Emission Computed Tomography
WHO : World Health Organization
CoV: CoronaVirus
CXR : Chest X-Ray
Covid-19: Coronavirus 2019
GGO: Ground-Glass Opacification
RT-PCR: Reverse Transcription Polymerase Chain Reaction
RBF : Radial basis function
SVM: Support Vector Machine.
CAD : vector supporting machine.
GPU: graphics processing unit.
JPG: Joint Photographic Experts Group.
T2T-ViT : Tokens-to-Token ViT.
DICOM : Digital Imaging and Communications

Bibliography

Bibliography

[1] URL: https://https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence.

[2] URL: https://www.ad-ins.com/7-examples-of-the-application-of-machine-learning-in-different-fields/.

[3] URL: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML.

[4] URL: https://chisoftware.medium.com/supervised-vs-unsupervised-machine-learning-7f26118d5ee6.

[5] URL: https://medium.com/dataseries/two-minutes-of-semi-supervised-learning-f0eb62729530.

[6] MOLLA YIGZAW KIBRET. "AGE ESTIMATION BY USING TRANSFER LEARNING IN A DEEP NEURAL NETWORK". PhD thesis. 2022.

[7] URL: https://www.javatpoint.com/computer-vision-applications.

[8] URL: https://www.deeplearning.ai/resources/natural-language-processing/.

[9] Xing Hao, Guigang Zhang, and Shang Ma. "Deep learning". In: *International Journal of Semantic Computing* 10.03 (2016), pp. 417–439.

[10] URL: https://www.ibm.com/cloud/learn/neural-networks.

[11] URL: https://www.whyofai.com/blog/ai-explained.

[12] URL: https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/.

[13] URL: https://vitalflux.com/different-types-of-cnn-architectures-explained-examples/.

[14] URL: https://medium.com/@gdenizbektass/cnns-building-blocks.

[15] URL: https://www.frontiersin.org/articles/10.3389/fnbot.2021.690220/full.

[16] K He et al. "Transformers in medical image analysis: A review. arXiv 2022". In: *arXiv preprint arXiv:2202.12165* ().

[17] Nicolas Carion et al. "End-to-end object detection with transformers". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer. 2020, pp. 213–229.

[18] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[19] Hugo Touvron et al. “Training data-efficient image transformers & distillation through attention”. In: *International conference on machine learning*. PMLR. 2021, pp. 10347–10357.

[20] URL: https://www.ahu.edu/blog/types-of-medical-imaging.

[21] URL: https://siim.org/page/archiving_chapter2.

[22] URL: https://link.springer.com/article/10.1007/s10916-018-1088-1.

[23] Justin Ker et al. “Deep learning applications in medical image analysis”. In: *Ieee Access* 6 (2017), pp. 9375–9389.

[24] Xingyi Yang et al. “COVID-CT-dataset: a CT scan dataset about COVID-19”. In: *arXiv preprint arXiv:2003.13865* (2020).

[25] URL: https://i.pinimg.com/736x/36/4a/1f/364a1f23739a3ce05a6b30a3781f1565.jpg.

[26] Asif Iqbal Khan, Junaid Latief Shah, and Mohammad Mudasir Bhat. “CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images”. In: *Computer methods and programs in biomedicine* 196 (2020), p. 105581.

[27] N Narayan Das et al. “Automated deep transfer learning-based approach for detection of COVID-19 infection in chest X-rays”. In: *Irbm* 43.2 (2022), pp. 114–119.

[28] Walid Hariri and Ali Narin. “Deep neural networks for COVID-19 detection and diagnosis using images and acoustic-based techniques: a recent review”. In: *Soft computing* 25.24 (2021), pp. 15345–15362.

[29] Temitope Emmanuel Komolafe et al. “Diagnostic test accuracy of deep learning detection of COVID-19: A systematic review and meta-analysis”. In: *Academic Radiology* 28.11 (2021), pp. 1507–1523.

[30] S Suba and Nita Parekh. “Machine Learning Approaches in Detection and Diagnosis of COVID-19”. In: *Artificial Intelligence and Machine Learning in Healthcare* (2021), pp. 113–145.

[31] Junaid Shuja et al. “COVID-19 open source data sets: a comprehensive survey”. In: *Applied Intelligence* 51 (2021), pp. 1296–1325.

[32] URL: https://www.nature.com/articles/s41598-023-32462-2/figures/3.

[33] URL: https://link.springer.com/article/10.1007/s10096-020-03901-z.

[34] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[36] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[37] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[38] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[39] François Chollet. "Xception: Deep learning with depthwise separable convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.

[40] Christian Szegedy et al. "Inception-v4, inception-resnet and the impact of residual connections on learning". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[41] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[42] Xiaohong Gao, Yu Qian, and Alice Gao. "COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models". In: *arXiv preprint arXiv:2107.01682* (2021).

[43] Fozia Mehboob et al. "Towards robust diagnosis of COVID-19 using vision self-attention transformer". In: *Scientific Reports* 12.1 (2022), p. 8922.

[44] URL: `https://pubmed.ncbi.nlm.nih.gov/35618740/#&gid=article-figures&pid=figure-2-uid-1`.

[45] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network". In: *arXiv preprint arXiv:1503.02531* (2015).

[46] Christian Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[47] URL: `https://keras.io/examples/vision/deit/`.

[48] Li Yuan et al. "Tokens-to-token vit: Training vision transformers from scratch on imagenet". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 558–567.

[49] Krzysztof Choromanski et al. "Rethinking attention with performers". In: *arXiv preprint arXiv:2009.14794* (2020).

[50] Mark Sandler et al. "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[51] Dina Kushenchirekova et al. "COVID-19 classification based on CNNs models in CT image datasets". In: *2021 16th International Conference on Electronics Computer and Computation (ICECCO)*. IEEE. 2021, pp. 1–4.

[52] Zharfan Zahisham, Chin Poo Lee, and Kian Ming Lim. "Food recognition with ResNet-50". In: *2020 IEEE 2nd international conference on artificial intelligence in engineering and technology (IICAIET)*. IEEE. 2020, pp. 1–5.

[53] Gabriel Sousa Silva Costa et al. "COVID-19 automatic diagnosis with ct images using the novel transformer architecture". In: *Anais do XXI simpósio brasileiro de computação aplicada à saúde*. SBC. 2021, pp. 293–301.

[54] Lei Zhang and Yan Wen. "MIA-COV19D: a transformer-based framework for COVID19 classification in chest CTs". In: *Proceeding of the IEEE/CVF International Conference on Computer Vision Workshops*. 2021, pp. 513–8.

[55] Shuang Liang, Weicun Zhang, and Yu Gu. "A hybrid and fast deep learning framework for Covid-19 detection via 3D Chest CT Images". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 508–512.

[56] Yassine Barhoumi and Rasool Ghulam. "Scopeformer: n-CNN-ViT hybrid model for intracranial hemorrhage classification". In: *arXiv preprint arXiv:2107.04575* (2021).

[57] Jingxing Li, Zhanglei Yang, and Yifan Yu. "A medical ai diagnosis platform based on vision transformer for coronavirus". In: *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*. IEEE. 2021, pp. 246–252.

[58] Joel CM Than et al. "Preliminary study on patch sizes in vision transformers (vit) for covid-19 and diseased lungs classification". In: *2021 IEEE National Biomedical Engineering Conference (NBEC)*. IEEE. 2021, pp. 146–150.

[59] Yingda Xia et al. "Effective pancreatic cancer screening on non-contrast CT scans via anatomy-aware transformers". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer. 2021, pp. 259–269.

[60] Yang Xingyi et al. "Covid-ct-dataset: a ct image dataset about covid-19". In: *arXiv preprint arXiv:2003.13865* (2020).

[61] URL: https://www.kaggle.com.

[62] URL: https://www.stickpng.com/img/icons-logos-emojis/tech-companies/kaggle-logo.

[63] URL: https://www.python.org/doc/essays/blurb.

[64] URL: https://www.python.org/community/logos/.

[65] URL: https://catalog.ngc.nvidia.com/orgs/nvidia/containers/cuda.

[66] URL: https://medium.com/geekculture/introduction-to-cuda-7bf6909ea57c.

[67] URL: https://www.kernix.com/pytorch-le-framework-a-la-croissance-la-plus-rapide-en-deep-learning/.

[68] URL: https://icon-icons.com/icon/pytorch-logo/169823.

[69] URL: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62.

[70] URL: https://www.geeksforgeeks.org/visualize-confusion-matrix-using-caret-package-in-r/.