International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 14 (2022) pp.313-325

Submitted: 21 Feb, 2022; Accepted: 1 Apr, 2022; Publish: 17 May, 2022

Multimodal deep learning based on the combination of EfficientNetV2 and ViT for Alzheimer's disease early diagnosis enhanced by SAGAN data augmentation

Rahma Kadri[1,2], Bassem Bouaziz[1],Mohamed Tmar[1], and Faiez Gargouri[1]

[1]Multimedia, InfoRmation systems and Advanced Computing Laboratory,
MIRACL,University of Sfax,Tunisa

[2]Faculty of Economics and Management of Sfax ,
FSEG Sfax ,University of Sfax Tunisia
elkadrirahma@gmail.com

Abstract: **The digitalization of health data and the innovative eHealth technologies has created a new paradigm shift from traditional medicine methods to a new predictable, individualized medicine based on patient-centric approaches. The emerging fields of predictive and precision medicine are evolutionary methods to treat the disease based on the patient's characteristics such as his lifestyle, genetic profile ,and environment to understand the disease. Alzheimer's (AD) early detection is still a challenging task. Researchers adopt advanced imaging techniques such as Magnetic Resonance Imaging (MRI) and fluorodeoxyglucose (FDG)-positron emission tomography (PET) to ensure a relevant understanding of AD disease. Extracting insights from these data is the key step towards disease early prediction and preventing its progression. Recently, deep learning methods have shown unparalleled success and have made a significant headway on brain diseases detection. Convolution neural network is a type of deep learning that has shown a state-of-the-art performance on the AD detection and early diagnosis. However its application has many limitations. The recent architectures such as transformers ensure an efficient image recognition and feature extraction with less complexity. In this study we investigate and evaluate the application of the different CNN and transformers models on Alzheimer's disease early diagnosis. Further, we introduce a multi-modal method based on the MRI and PET modality for Alzheimer's disease detection using the combination of the Efficientnet V2 and the vision transformer enhanced by a new data augmentation based on the self attention generative adversarial networks(SAGAN). We validated the proposed method using the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Open Access Series of Imaging Studies (OASIS). Our proposed method combines the main advantages of the vision transformer and Efficientnet V2 achieving a 96% accuracy rate. This new method outperforms different CNN models and transformer methods and ensures a robust feature extraction and representation.**

Keywords: Alzheimer's early diagnosis,CNN,Transformer, Vision transformer,Self attention

I. Introduction

Traditional healthcare was based on reactive clinical methods for disease diagnosis. The reactive approach in healthcare is expensive and doesn't meet the individual characteristics in disease diagnosis. Recent technological advancements in healthcare ensure unprecedented opportunities that shift the emphasis in medicine from reaction to proactive and prevention based on personalized diagnosis. These new approaches tailor tremantes to patient specific disease detection. The main objective of personalized medicine is to create an optimized and optimal pathway of an accurate and early prediction for disease prevention. This is based on interpreting a huge amount of multidimensional datasets that capture genetic information of the patient, clinical history, his lifestyles and his personal data. Accurate and early diagnosis of disease relies on high quality healthcare data collection,integration and analysis. These data enable an effective characterization of the disease based on individual personal data. The main challenge is how to interpret and extract knowledge from these scattered and heterogeneous data to quantify the individual's risk for such disease. The key driver for an effective diagnosis is to collect personalized data that characterize the patient. Brain disease early detection for prevention is an open challenge within personalized medicine. Within this context, Alzheimer's disease is a brain disorder that destroys brain cells and affects the indi-

vidual's ability to carry out daily activities. This disease is a progressive disorder that causes a deterioration in cognitive and behavioral function over time.

There are different stages of the AD. Preclinical stages is the first stage, the prodromal AD is the second stage that is characterized by the brain dysfunction and the AD dementia is the final stage.Alzheimer's disease early prediction is a crucial and challenging task to prevent its progression. Researchers adopt various biomarkers for AD early detection such as neurological tests, clinical data and brain imaging techniques to cover and track the brain change and state. Analysing these brain imaging modalities is a common method and practice for Alzheimer's disease detection. It shrinks brain regions such as the hippocampus and cerebral cortex of the brain. The Hippocampus is the brain region that is responsible for spatial memory. It is the earliest affected brain region in AD. Shrinking Hippocampus causes short-term memory due to the damage of neuron connection. Hippocampal volume is one of the relevant AD biomarkers. MRI modality is a powerful tool that image and provide details about the hippocampal volume. It is the commonly used modality that detects subtle changes on the brain tissue volumes and state. This can foster the predictive aspect of the diagnosis and provide a potential opportunity for an early intervention for disease prevention before significant pathological changes. MRI modality has proven a substantial utility for AD diagnosis due to its capability to cover the brain atrophy and other static tissue abnormalities. FDG-PET is another powerful brain modality for the AD early detection. This modality can measure the brain's metabolic activity which is a valuable bio-marker of AD. Moreover this neuroimaging technique image the functional brain changes and tracks the A tracers to predict the conversion from MCI due to AD. Deep learning methods has made a considerable breakthroughs to empower an effective decision-making for the diagnosis of diseases and shown a significant potential for clinical decision within the Alzheimer's disease [1],[3], [25], [2]. The deep learning architecture consists of various layers that proceed the input data at different levels such as multiple nonlinear processing layers. The network extracts different levels of features from the data. The main advantage of deep learning is to automate the image feature extraction using an hierarchical learning process. The most commonly used deep learning architecture is the convolution neural network in various computer vision tasks. It is a type of deep network that applies a filter as a feature detector to extract the main features within the input image. CNN models are mainly applied on image classification, recognition and segmentation within the AD disease [[4],[5],[6] [7]].

However the feature representation mechanism in CNN doesn't allow to encode the multi-level dependencies within the input image to enhance its representational capacity. The only solution is to increase the size of the convolution filters used to detect features. This solution is to increase the complexity of the network. Transformers are state-of-the-art type of network used for NLP also recently for computer vision and outperforms both the RNN and CNN models in many tasks. The building block of the transformer is the self attention mechanism that improves the capability of the model to capture the main dependencies within the input data and extract much more relevant information. Vision transformer is a subtype of transformer designed for computer vision tasks and overcomes the mail limitation of the CNN models. In this study we proposed:

- Evaluation and examination the application of such CNN networks on AD diagnosis
- Investigation the application of transformers networks for AD diagnosis
- A new data augmentation based on the self attention generative adversarial neural network to enhance the training and tackle the problem of overfitting and lack of data.
- A multi-modal method based on the MRI and PET modality for Alzheimer's disease detection using the combination of the Efficientnet V2 and the vision transformer to combine the main advantages of the CNN and the transformer.

II. Related Work

In this section,we investigate the recent studies for Alzheimer's detection and prediction during the last three years. We divided the current methods into Multimodal that adopted different brain modalities and the unimodal methods that used only brain modality. The aim objective of this classification of the AD detection methods is to choose the optimal method for AD classification.

A. Multimodal methods

Multimodal methods for AD detection are characterized by the application of different brain modalities. [8]adopt a multi fusion modality that combines a magnetic resonance imaging (MRI)), genetic single nucleotide polymorphisms (SNPs)), and clinical test data to classify patient state into three classes,CN,AD and MCI. They used a stacked denoising auto-encoders for feature extraction from the genetic data and clinical data. Further they applied a 3D deep CNN for feature extraction from the imaging data. The main extracted features for AD detection in this study are the hippocampus, amygdala brain areas, and the Rey Auditory Verbal Learning Test (RAVLT). Data collection is carried out from the ADNI dataset. They used MRI data as imagining data. The clinical data consists of demographic data,neurological exams,medication data and imaging summary scores. They extracted the whole genome sequencing (WGS) data from 808 ADNI participants as genetic data. They combined different features from the different data(EHR features,imaging features and SNP features) for AD detection. The three features are combined and passed through a classification layer to classify AD satages. The main advantage of this study is the multi fusion of the different data which enhance the AD detection. However this model is slow to train. They demonstrated that their model outperforms the traditional machine learning techniques. [9]proposed novel image fusion method based on the MRI and PET modality. In this study,they extract GM information from MRI and PET and combine it into a single image. They acquired data from the ADNI dataset T1-weighted MRI and FDG-PET captured

at the same period for three AD stages AD,CN and MCI. Preprocessing includes a Gradwarp, B1 non-uniformity, and N3. Gradwarp corrects non-uniformity using B1 calibration scans. For the PET modality Co-Registered dynamic: six 5-min FDG-PET frames are acquired within 30–60 min post-injection, co-registeretaion, Standardization of image , intensity normalization and Uniform resolution. Feature extraction is based on a 3D Multi-Scale CNN that merges multi-scale features flowed by a FC layer and a softmax layer for AD prediction. The main advantage of this model is the feature fusion method into a single GM image which ensures a relevant feature representation and requires few parameters.This proposed fusion method enhances the AD detection. [10]introduced an hybrid method for AD prediction progression (stable MCI (sMCI) and progressive MCI (pMCI)) based multitask multiple deep bidirectional long short-term memory (BiLSTM) using different data from the ADNI dataset such as time-series modalities and baseline data. The first LSTM extracted the low level feature and the second LSTM extracted the high level features. In this study they proposed 2 main architectures.The first architecture is a multitask regression model that predicts seven crucial cognitive scores for the patient 2.5 years after their last observations. The predicted scores are used to build an interpretable clinical decision support system based on a glass-box model. This architecture aims to explore the role of multitasking models in producing more stable, robust, and accurate results. Data preprocessing inculdes missing data, data normalization. [12]applied 2 VGG19 network for the MRI and PET modality. Each network is composed by 19 layers.Then they applied a correlation analysis and they combined the output with the results of clinical neuropsychological diagnosis. The data is selected from the ADNI dataset.[31] pointed out that the single-nucleotide polymorphisms (SNPs) data are vital for AD prediction. They applied an ensemble model consisting of two sub-networks, an image processing network based on a AlexNet network and a multilayer perceptron (MLP) for SNP data processing. Authors in this study combined the result between 2 networks the CNN and the MLP for 2 different data MRI and SNP data for Alzheimer's classification into 2 classes AD and CN. Data collection is from the ADNI data set. This study combined the advanatges of machine learning technique (NLP) and AlexNet and achieved an accuracy of 93%. However the dataset used is very small. [4] proposed a lightweight 3D deep convolutional network model based on the dense network for AD classification. Here they used the hippocampus magnetic resonance imaging (MRI) as biomarker for AD detection. They incorporate the global shape representations with the hippocampus brain segmentations to enhance the feature representation. They use T1-weighted structural MRI from initial screening or baseline from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The combination of the hippocampus brain segmentations and the global features enhance the AD detection in this study. However they use a small data set and the CNN model feature representation is not relevant for a good AD detection. We noticed that there are a few methods that combine different brain modalities for AD diagnosis.

B. Unimodal Methods

The unimodal methods are based on a single modality for the AD detection. [11]adopted a transfer learning using a fine tuned ResNet18 for a binary classification of AD which include EMCI/LMCI, AD/CN, CN/EMCI, CN/LMCI, EMCI/AD, LMCI/AD, and MCI/EMCI. The study's data consist of MRI data acquired from the ADNI dataset. Data preparation includes random resize and cropping to 256 × 256, random rotation, random horizontal flip, center cropping to 224 × 224, conversion to PyTorch tensor, and normalization. They reshape the fully connected layer of the original Resenet18 to adapt it for AD classification. Transfer learning here improves the feature extraction.However the network cannot extract the main changes of the brain within the MRI data. [16] proposed a new concatenated deep features approach based on the concatenation of the pre-trained networks densenet121 and resnet18 network for AD classification into five classes AD, MCI, EMCI, LMCI, and NC. This study is validated using 138 MRI images from the ADNI dataset.They reach an accuracy of 97%. This hybrid method outperforms the method proposed in [24] which also proposed an AD classification into five stages using only three-Dimensional DenseNet network achieving 0.86 as value of accuracy.
[17] asserted that the F-18 Fluorodeoxyglucose positron emission tomography/computed tomography is an effective tool for AD early detection. They proposed a CNN network that consists of five convolution layers for feature extraction. They replace the fully connected layer by a global average pooling layer that enhances the object localization and minimizes the network parameters. Through this layer they extract the heatmap of input that selects the relevant region for making a classification. In this study, they used a custom dataset to classify AD and NC. [30] noticed that most studies are focused on AD detection based on CNN and there is a need for a learning object detection methods on Alzheimer's Disease detection. For this end, they proposed a new AD detection based on on common object detection methods such as faster R-CNN, SSD and YOLOv3. In addition, they created a custom dataset from the ADNI dataset using a collection of T1 weighted sagittal MRI dicom slices in MP-Rage series in DICOM 16-bit and PNG 16-bit image format annotated with their respective class label and bounding box in Pascal VOC format. They figure out the main advantage of the object detection within the AD diagnosis. Their study yields a good detection accuracy (0.998 for YOLOv3, 0.982 for SSD and 0.988 for Faster R-CNN). [26] proposed an hybrid method that combines VGG16 and CNN model for AD classification. [27] also proposed an hybrid method based on the 3D CNN and 3D convolutional long short-term memory for AD classification. They extract an informative features within the input MRI using the 3DCNN. The 3D CLSTM is used to extract the channel-wise higher-level information. Extracting the channel-wise higher-level information enhances the capability of the network to learn more regions related to the AD. They validated their study using the ADNI dataset. They achieved 94.19% as an accuracy of classification. [28] 3D multiscale convolutional neural network consists of 8 layers (Conv1-Conv8) for muliscale feature extraction. They proposed a feature fu-

sion method and enhancement layers that merge multiscale feature maps into a vector. They applied a pooling to each feature map. This produced feature maps in different scales. Then a feature fusion strategy is applied to fuse the multiscale features by concatenating feature vectors from all scale levels. This method enhances the CNN feature representation for effective AD detection. Experiments were validated on the ADNI dataset. This method achieved an accuracy of 93.53%. Transformers today are [29] introduced an optimized vision transformer method for AD stages prediction based on fMRI data. Through this review we noticed that most studies are based on the single modality which is the MRI modality and the CNN models. However CNN models have many limitations regarding the feature representation and the need for huge amounts of data. It does not capture and encode the long large relationships at pixel level within the input image. CNN ensures a good generalizability however it does not capture the main dependencies and the high global features in the input image. There are many recent studies that pointed out that attention-based networks can outperform convolutional Neural Networks (CNNs) on image classification and recognition. There are few studies on the AD detection that incorporate the attention mechanism to boost the capability of CNN for optimal feature representation. Further, there is no studies that investigate the use of the transformers on AD detection. In this study we deal with this issues.

III. Methods

In this section we describe the proposed method. Firstly, we proposed a data augmentation method based on the self attention generative adversarial neural network to improve the training process. Furthermore we evaluated some CNN models for AD detection and basing on this evaluation we investigate the application of different transformer architectures on the AD detection. Finally we describe the proposed multimodal based on a combination of the Efficient net B7 and the vision transformer using the MRI and PET modality.

A. Data Selection

In this study we acquired data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (https://adni.loni.usc.edu/). ADNI is a longitudinal multicenter study that aims to enhance the clinical trials for the prevention and treatment of Alzheimer's disease. This multisite study develops different types of AD biomarkers such as clinical, imaging and genetic data for the early diagnosis of AD. The main objective of the ADNI is to make all the data available for scientists worldwide to foster AD detection and diagnosis. In this study we select 2 types of modalities MRI and FDG-PET scans captured in the same period.

B. Data preprocessing

Data preprocessing is an integral step for deep learning models that enhance the ability of the model to learn relevant features from he input image. The selected MRI and PET scans undergone various processing steps. The preprocessing of MRI scans includes bias correction, noise removal, Gradwarp, B1 non-uniformity, corrupted image removal, smoothing, image resizing, conversion to Pytroch tensor and image normalization.

Table 1: Demographic information of samples from ADNI dataset.

Class	Age	Sex	Modality	Total number
CN	(60-100)	(F)and (M)	MRIand FDG-PET	610
MCI	(60-100)	(F)and (M)	MRIand FDG-PET	670
AD	(55-100)	(F)and (M)	MRI and FDG-PET	690

The Table 1 present the main demographic information of subjects selected from ADNI dataset used in this study.

C. Data augmentation using self attention generative adversarial neural network

Deep learning networks such as convolution neural networks require a huge amount of data for effective training and for avoiding the overfitting problem. Traditional data augmentation are based on geometric transformation of the input images such as rotating,zooming, resizing,adding noise,image translation and image flipping. However this technique is not optimal for medical images. Recent data augmentation are based on the generative adversarial neural network(GAN). GAN has shown great potential in data augmentation.It is a class of deep learning technique that creates a real image from noise data to enlarge the size of the dataset in order to ensure a generalizable deep learning model. GAN consists of 2 networks. The first network is the generator that takes a noise data from latent space and create a synthetic data. The second network is the discriminator that classifies the generated images into real and fake or synthetic images. The generator tends to generate a realistic image that can be classified as real images by the discriminator.

GAN is an innovative technique for data augmentation. However it has many drawbacks. The main drawback is the convergence. There is no relevant method to find when to stop training the generator this leads to the unreliability in the training. The loss function is not a metric that can outlines the convergence of the GAN. Therefore we cannot figure out when the generator can create high quality synthetic images. The objective function of the GAN cannot outline the quality of the output images. Researchers proposed many solutions to mitigate and meet the main limitations of the GAN such as the realization of the loss function. Another solution is to replace the Jensen Shannon divergence of traditional GANs with the Earth Mover distance. We applied the traditional GAN however we noticed the unstable training of the 2 sub networks of the GAN due to the high resolution and the multidimensional brain images.

The traditional generative adversarial network as depicted in figure 1 are based on the CNN network which a ensure a relevant spatial locality information extraction. However within the CNN the receptive fields typically are not large enough to detect larger structures and long-range interactions within the input image.

Recently computer vision has been revolutionized by many innovative techniques. The most trendy technique today within deep learning is the attention mechanism. Researchers

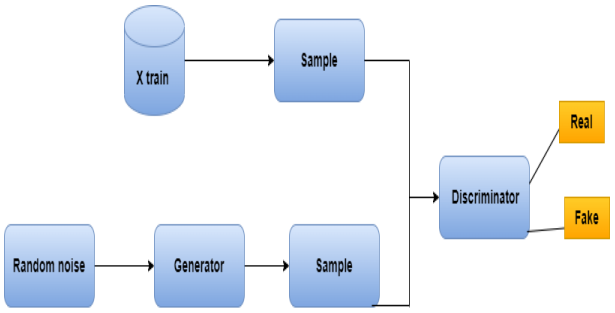


Figure. 1: Generative adversarial network architecture

foster and boost the CNN performance by the attention mechanism. This mechanism is the most relevant breakthrough in deep learning today which is inspired by the human perception mechanism. Based on this technique Self-Attention Generative Adversarial Networks ensure an attention-driven and long-range dependency for image synthesis.The SAGANs is a substantial improvement and extinction of GAN that integrate the self-attention mechanism into convolutional GANs to capture long-range, multi-level relationships within the input image.

1) Attention mechanism

Attention mechanism is inspired by the human cognitive process of selecting the relevant features of the image and ignoring the irrelevant features rather than concentrating on the whole image. The attention mechanism can defined as: $Attention = f(g(x), x)$
where $g(x)$ represent the process of concentrating on the most relevant and discriminative features within the image. $f(g(x), x)$ refers to the processing extracting the discriminative features of the input x based on the attention $g(x)$.

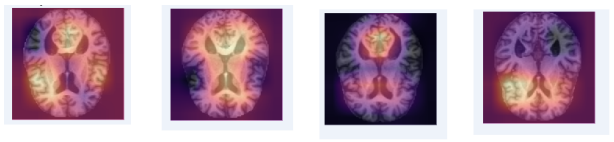


Figure. 2: Example of attention mechanism

2) Self attention

Self attention refers to the intra-attention which is a type of attention mechanism that takes in n inputs and returns n outputs. The self-attention mechanism ensures an interaction between the inputs ("self") and outlines who they should pay more attention to ("attention") [19] Figure 3.
The key component of the Self-Attention Generative Adversarial Network is the self-attention module that is incorporated with the convolution network. This module adopts the key-value-query model. The feature map created by the CNN network is fed into the self attention to transform it into three feature spaces named key $f(x)$, value $h(x)$, and query $g(x)$. These different feature maps are obtained by passing feature maps created by the CNN through three different 1*1 convolution maps. Then the multiplication between Key $f(x)$ and query $g(x)$ matrices is applied. In addition the softmax

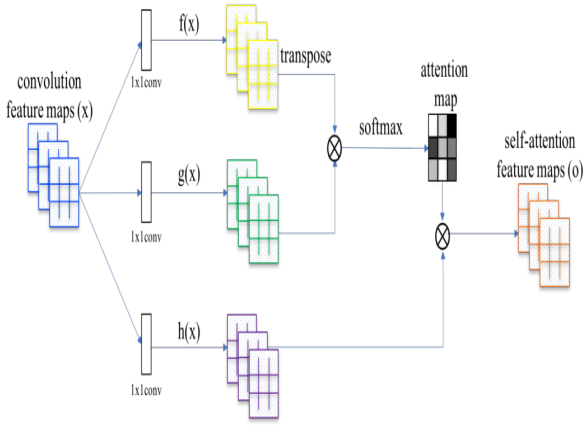


Figure. 3: Self attention mechanism adopted from [19]

is applied to each row operation of the multiplication result and produces an attention map. The main objective of the attention map from the softmax operation is to capture and select which regions of the image the model should attend to. Finally the attention map is multiplied with value $h(x)$ to create the self-attention feature map. The input feature map is added to the scaled attention map to get the output. Thanks to the self attention module the model focuses on the local and global features within the input image using a scaling parameter that is updated during the model training to focus on the relevant features of the image. Another motivation to adopt the SAGAN for data augmentation is to ensure stable training and avoid the instability in GAN training. The SAGAN uses 2 main techniques the Spectral normalization and the Two Time-scale Update Rule (TTUR) to deal with the instability in GAN training . The spectral normalization is applied in both the generator and the discriminator. This technique adjusts the Lipschitz constant without the network hyper-parameter tuning and avoids the gradients of the network. Furthemore, the TTUR accelerates the learning speed and avoid to the multiple discriminator updates for each generator update. The generator takes as input the noise data and creates a synthetic image. Firstly the input data is reshaped and then the model performs a stacking of transposed convolution layers for the input to generate the output. We applied the spectral batch normalization and ReLU For each of these layers as illustrated in the figure 4. We integrate the self attention module within the generator to ensure the long-range dependency modeling and allow the creation of high quality images.
The discriminator takes the generated image as input and applies a stacking of convolution layers flowed by the spectral batch normalization and Leaky ReLU layers as depicted in the figure 5.
During the training process, the discriminator takes real images from the training set and receives generated images from the generator network. The discriminator model is trained to maximize the probability of distinguishing between the real images from the training set and fake samples from the generator. We also add a self attention module for the discriminator to enhance its capability to capture and encode the long-range dependency within the image for an ef-

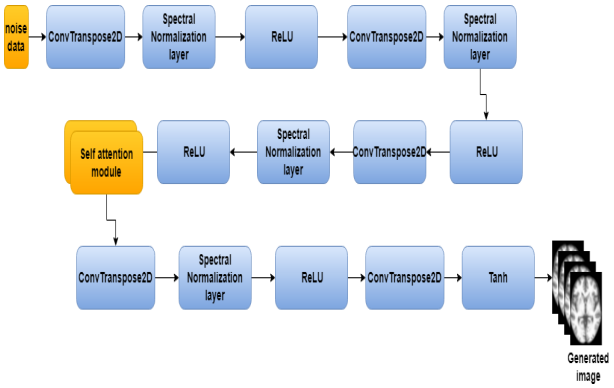


Figure. 4: Generator architecture

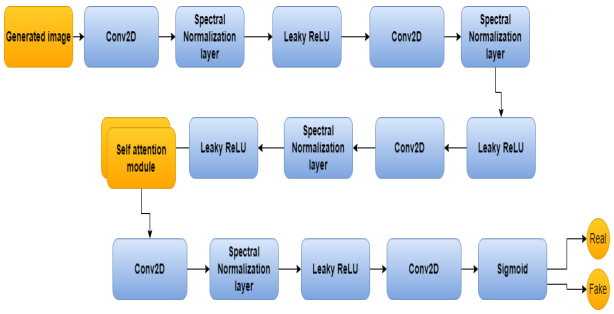


Figure. 5: Discriminator architecture

fective data classification. For the training we adopt the two-timescale update rule (TTUR) to deal with the slow learning in the regularization of the discriminator. The traditional regularization methods of the discriminator are based on multiple regularized discriminators. The discriminator is updated per generator is updated during training this slows the learning of the network.The separate learning rate (TTUR) specifically overcomes this issue.
The Spectral Normalization is added to the generator and the discriminator to enhance the stability of the GAN and improve the quality of the generated images. The Self-Attention Generative Adversarial Network ensures a high quality of image generation as depicted in the figure 6 thanks to the attention-driven, and long-range dependency. In our previous work "Deep Squeeze and Excitation-Densely Connected Convolutional Network with cGAN for Alzheimer's disease early detection" within our previous work [32] we adopted a cGAN for image generation. We noticed that the new method SGAN yields better results comparing to our previous work.

D. Alzheimer's disease early detection using CNN models

In the first step we evaluated different convolution neural networks such as VGG16,Resnet 152,Inception model and DenseNet network.
We figure out that the CNN models lack from a high feature representation capability and an effective information extraction from image. Researchers recently adopted the attention mechanism to boost the CNN capability for an effective feature representation and extraction.
Figure 7,8 and 9 illustrate the application of the Alexnet, Resnet 152 and VGG model. As shown in these figures, these models do not achieve a high accuracy for AD detection. Further, CNN models cannot capture the long relation between image pixels and cannot model the relative position of the feature within an image. CNN networks cannot encode long-range dependencies within the input image for an effective recognition. The solution is to increase the size of the filterers used to detect the image features. However this increases the computational cost of the model and produces the vanishing gradient problem.

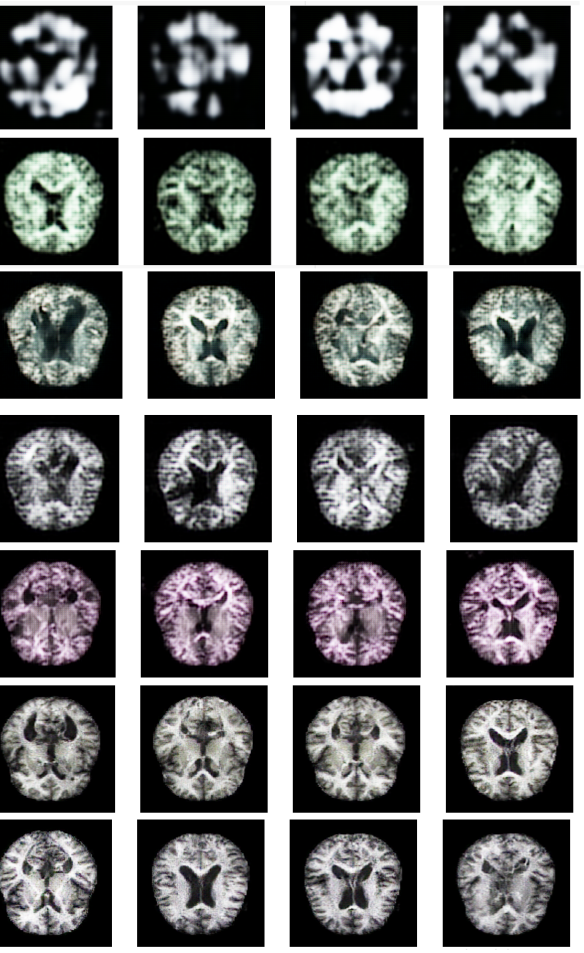
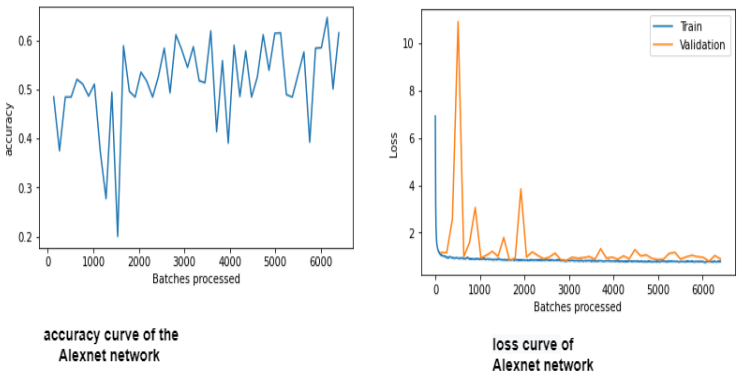


Figure. 6: Output of the Self Generative Adversarial Network for data augmentation



accuracy curve of the Alexnet network

loss curve of Alexnet network

Figure. 7: Alexnet application

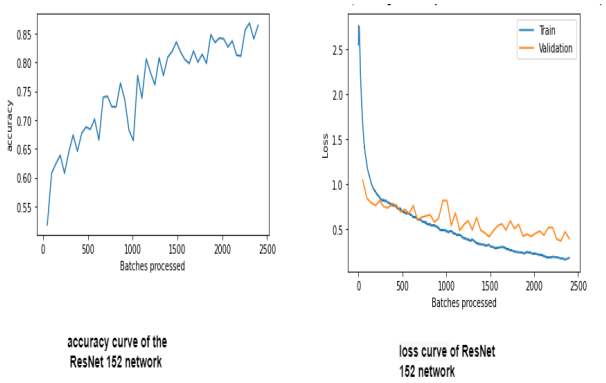


Figure. 8: Resnet 152 application

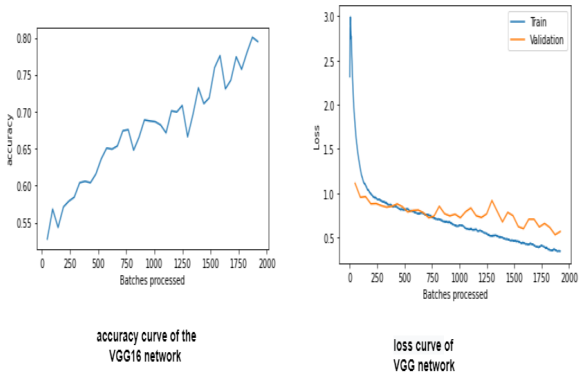


Figure. 9: VGG16 application

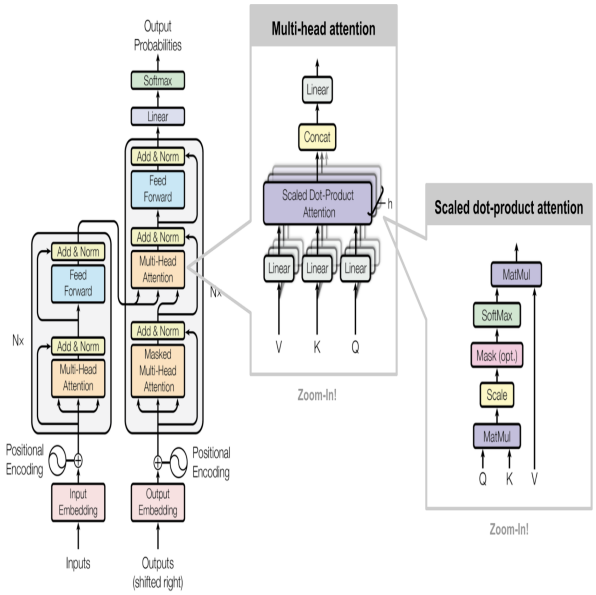
E. Alzheimer's disease early detection using Transformer architecture

Recently researches adopt the self attention mechanism to mitigate these limitations. This technique enables an effective feature representation by tracking and encoding the main long-range dependencies within the image. Transformers are based on the self attention mechanism motivated by this idea researchers applied the transformer on computer vision tasks such as image classification.

1) Transformer

The transformer is an attention-based encoder-decoder architecture. It is made up of a stack of encoder-decoders.
The encoder consists of six identical layers. Each layer involves 2 sub layers flowed by a normalization layer.The first layer is a multi-head self-attention mechanism, and the second is a simple position wise fully connected feed-forward network that applied transformations with Rectified Linear Unit (ReLU) activation.
$FFN(x) = ReLU(\mathbf{W}_1 x + b_1)\mathbf{W}_2 + b_2$

The input is passed through the self-attention layer and it takes each step of the encoding from the previous encoder as input and weighs their importance to each other in order to obtain the output encodings. The output is fed into the feed-forward neural network which processes each output encoding individually. These output encodings are then passed to the next encoder as its input, as well as to the decoders. Authors claimed that they used residual blocks within each



adapted from [20]

Figure. 10: Transformer architecture

two sub-layers. The output of each layer is defined as $LayerNorm(x + Sublayer(x))$ Where the $LayerNorm$ is a normalization layer and the $Sublayer(x)$ represent the function processed by the sublayer itself. The decoder is made up of six layers. Each layer consists of three sub layers: the self-attention mechanism, an attention mechanism over the encodings, and a feed-forward neural network.

2) The Transformer Attention

The transformer attention represents the mapping between a query and a set of key-value pairs, to an output. There are 2 main types of attention mechanisms named a scaled dot-product attention and a Multi-Head Attention.

The scaled dot-product attention takes as input a vectors of dimension that represent the queries and keys and a applied a dot product for each query with all the input keys.Then it defined as:
$attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
Where q adn k d_k are vectors of dimension that represent the queries and keys, respectively.

Multi-Head Attention The Multi-Head Attention is an attention module that capture a different representation of subspaces of queries, keys, and values by projecting linearly these values. Each step is an Attention Head. The Attention module splits its Query, Key, and Value parameters N-ways and passes each split independently through a separate Head. All of these similar Attention calculations are then combined together to produce a final Attention score. This is called Multi-head attention and gives the Transformer greater power to encode multiple relationships and nuances for each word. The multi-head attention is defined as follows:

$$MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = concat(head_1, \ldots, head_h)\mathbf{W}^O$$

Where each $head_i = attention(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V)$

3) The application of the vision transformer

The vision transformer is new approach based on the transformer that replace the convolutions by the transformer architecture. The transformer architecture process the input on sequences. Here we divide the image into patches and and flattening each patch to a vector as depicted in figure 11. We investigated the application of the vision transformer for

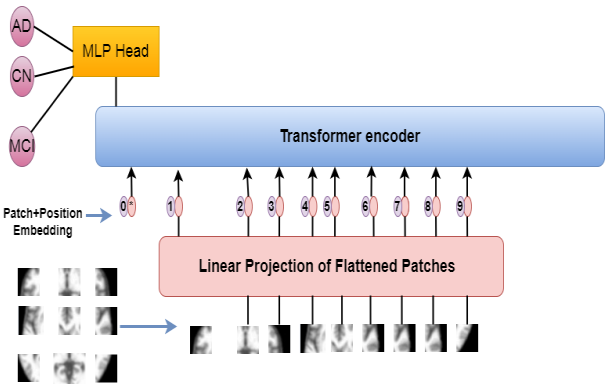


Figure. 11: ViT architecture

Alzheimer's classification using a uni-modal method based on the MRI data. The vision transformer splits the input image as a sequence of image patches with fixed size and flatten them. Then create a lower-dimensional linear embedding from these patches and apply a positional embedding as an input to the transformer encoder. The transformer consists of a Multi-Head Self Attention Layer (MSP). This layer ensures a training of the local and the global dependencies in an image by concatenating all the attention outputs linearly to the right dimensions. The Multi-Layer Perceptrons (MLP) Layer involves a two-layer with a Gaussian Error Linear Unit (GELU). The Layer Norm (LN) is a normalization layer added to each block to foster the model performance. We noticed that the vision transformer outperforms the CNN models such as VGG16. It has a good learning ability. Figure

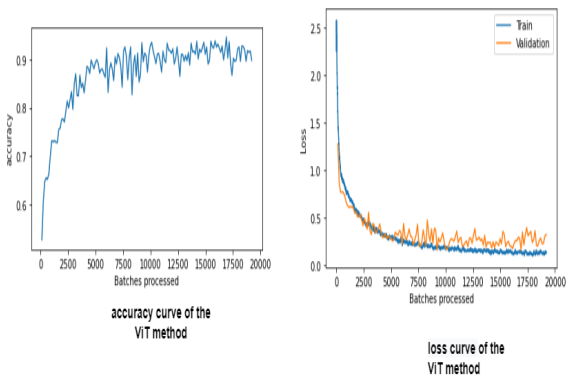
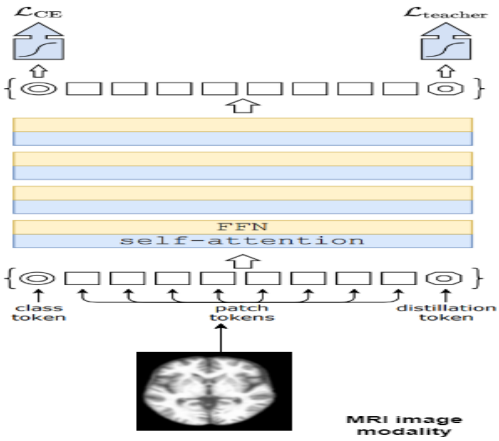


Figure. 12: Accuracy vs loss curve of the ViT method application using MRI modality

12 outlines the application of the ViT using the MRI modality for AD early detection. ViT ensures relevant and robust feature representation and extraction. It captures the local and the global features within the images. In addition, it captures and encodes the main long large dependencies within the input image which enhance the feature extraction. We achived a good accuracy using the ViT network.However, the performance of the ViT is depend on the size of the dataset due to low locality inductive bias. In addition the performance of the model depends on a sect of factors such as the network depth the optimizer and the network hyperparameters. ViT is difficult to optimize whereas CNN is easy to optimize. The ViT require a huge amount of data.

4) Data-efficient Image Transformer

Data-efficient Image Transformer is a subtype of vision transformer that uses a teacher-student strategy specific to transformers for training. This model is based on a distillation token that enables the student to learn from the teacher through attention. The DeiT can be considered as an extinction of the ViT architecture by adding to the original architecture an additional distillation token to the the input token as depicted in the figure 13. This new architecture processes the input images as a sequence of input tokens. We investi-



adapted from [22]

Figure. 13: Deit architecture

gated the use of the DeiT on Alzheimer's classification using MRI data. The Deit overcome the main problem of the ViT which is the need of huge amount of data for training.

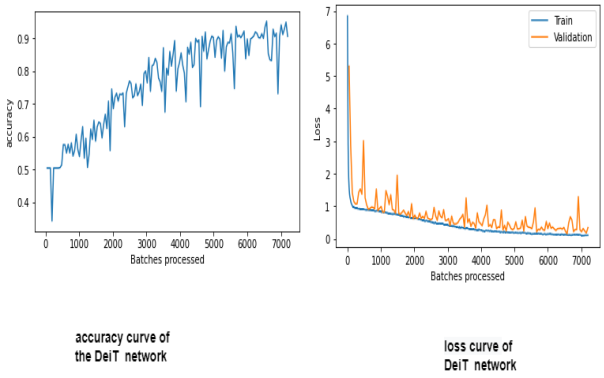
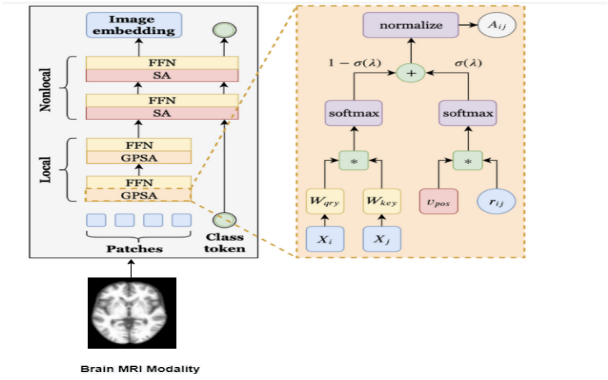


Figure. 14: Deit application

It requires less data than the ViT. Hence the DeiT yields optimal performance with less computing resources comparing to the ViT. The figure 14 illustrate the Deit performance on the Ad early detection.

5) Covit

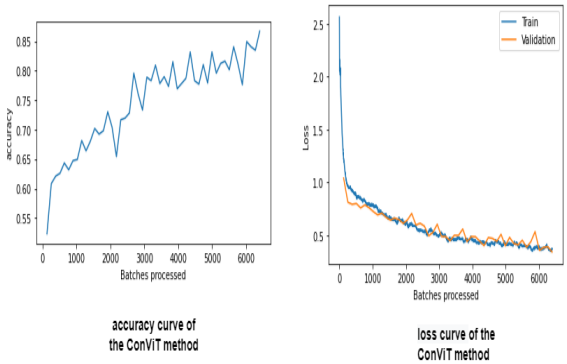
ConViT is a new type of vision transformer that is based on a gated positional self-attention module (GPSA). The GPA is type of positional self-attention that incorporates a "soft" convolutional inductive bias. The main objective of the GPSA layers is to ensure the locality of convolutional layers,



adapted from [21]

Figure. 15: ConViT application

The ConViT can be considered as a modification of the ViT architecture. It replaces the first 10 self-attention layer of the Vision Transformer with gated positional self-attention (GPSA) layers. Here there is a new type of self-attention layer named gated positional self-attention (GPSA) layer. As depicted in the figure 15 , the ConViT architecture adopted a gated position self-attention (GPSA) layers in the first part of the network followed by a self-attention (SA) layer in the second part of the network.



accuracy curve of the ConViT method

loss curve of the ConViT method

Figure. 16: ConViT application

The ConVit enhances the ViT architecture by gated positional self-attention to mimic the locality of convolutional layers. However it requires costly pretraining. Figure 16 present the performance of the ConVit on the AD early diagnosis. The Transformer based architectures outperform CNNs but this approach requires a huge amount of data. In this study, we find that the best solution is to combine the CNN model with the transformer. CNN has a relevant generalization ability thanks to its inductive bias, and transformer has stronger learning ability due to its global receptive field. We propose to combine these main advantages.

F. Alzheimer's classification using ViT and EfficientNet-V2

We proposed to combine the EfficientNet-V2 and the vision transformer for Alzheimer's detection based on the MRI data. The self attention within the transformer allow it to highlight the relevant features within the image and learn the crucial long-range dependencies between the image features. Furthermore,EfficientNetV2 ensures an optimal generalization. The proposed method is built upon 2 main networks, the EfficientNetV2 and the vision transformer as depicted in the figure 17. The first stage is the feature extraction using the EfficientNetV2 network in order to extract local features. The output of the CNN is then passed through the Vision transformer network.

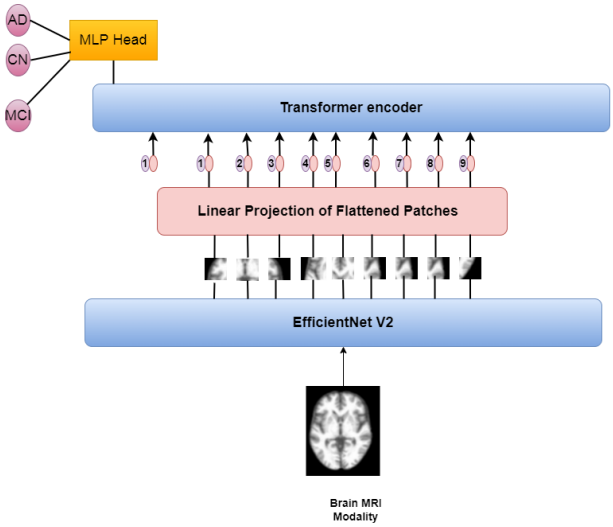


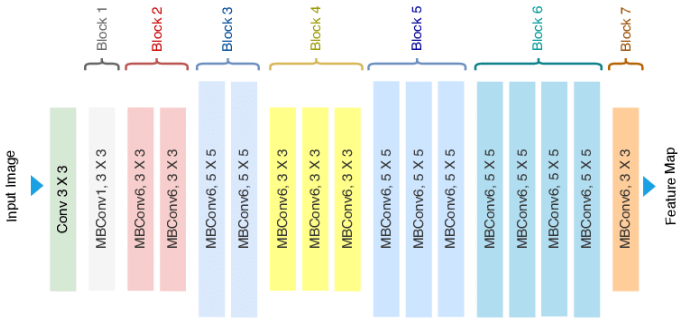
Figure. 17: ViT EfficientNet-V2 using MRI data

1) EfficientNet-V2 architecture

EfficientNet is a scaling method and lightweight NAS-based network that overcomes the main limitations of the traditional CNN methods. Traditional methods improve the performance of the CNN network based on the width of the network only or the depth or the resolution of the input image. However, scaling these three dimensions at the same time is the key solution for improving accuracy and training efficiency. Increasing the network width, depth and image resolution foster the network capability for extracting more complex features from the input image.
EfficientNet balances all the network parameters (depth,width,resolution) at the same time using a compound coefficient dimensions.

Compound Scaling The building block of the EfficientNet is the compound scaling method, which is based on a compound coefficient α to scale and balance the network parameters at the same time.

depth=$d = \alpha^{\phi}$
width=$w = \beta^{\phi}$
resolution=$r = \gamma^{\phi}$
$\alpha.\beta^2.\gamma^2$
$\alpha >= 1, \beta >= 1, \gamma >= 1$



adapted [23]

Figure. 18: Basic EfficientNet architecture

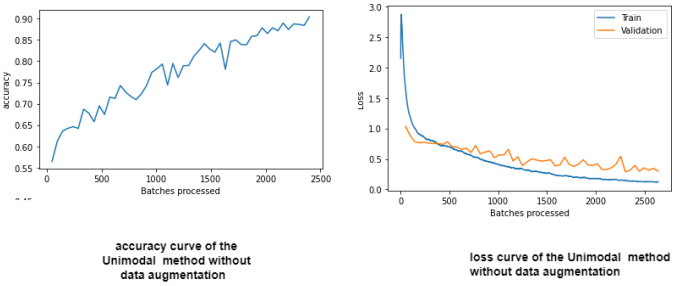


Figure. 20: ViT and EfficientNet-V2 unimodal application based the MRI modality

EfficientNetV2 is an extinction of the traditional EfficientNet in which the original MBConv in EfficientNetV1 are replaced by the Fused-MBConv. This new module replaces the depthwise 3x3 convolution and expansion 1*1 convolution in MBConv with 3*3 convolution. The EfficientNetV2 architecture adopted the MBConv and the Fused-MBConv modules in the early layers. It uses small 3*3 kernel sizes rather than 5x5 in EfficientNetV1. It removes the last stride-1 stage as in EfficientNetV1. EfficientNetV2 used a progressive learning strategy to enhance the training by increasing the image size and the magnitude of regularization simultaneously.

2) Vision transformer network ViT L-16

We adopted the ViT L-16 model for the second "Large" variant with a patch size of 16 × 16. This model involves 23 stacked transformer encoder layers.

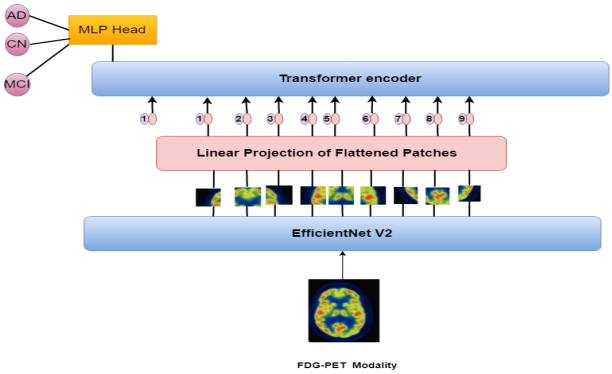


Figure. 19: ViT and EfficientNet-V2 unimodal application using FDG-PET modality

The ViT L-16 takes as input the feature maps produced by the EfficientNetV2 . The ViT applied the patch embedding projection to the patches extracted from the EfficientNetV2 feature map. As illustrated in the figure 20 the combination of the EfficientNet-V2 and ViT without the application of the proposed data augmentation and based only on the MRI modality achieves the best accuracy (91%) compared to the application of CNN models and the Vit,ConViT and the DeiT. We also applied our proposed method for the PET-FDG modality as depicted in the figure 19. We noticed that the unimodal based on MRI modality produce better accuracy than the unimodal based on the FDG-PET modality.

3) Multimodal Alzheimer's disease early detection

We adopt the image fusion method proposed by to [18].This method consists of extracting the GM area that is vital for the Alzheimer's detection diagnosis from the FDG-PET, using the MRI scan as an anatomical mask. The main objective of this method is to concatenate the relevant information from structural MRI information and functional PET information. Firstly the MRI image is passed through a prepossessing pipeline consisting of skull striping using a watershed" module in FreeSurfer 6.0. This method ensures the extraction of the brain volume without irrelevant information and preserves the intracranial tissue structure. Then the image registration is applied to the output image from the previous step through the FMRIB's Linear Image Registration Tool module within the FSL package. Finally the segmentation of the GM-MRI area using the FMRIB's Automated Segmentation Tool module in the FSL package. The preprocessing pipeline of the FDG-PET involved the image co-registering of the FDG-PET image to its respective MRI image using the FSL FLIRT module. The GM-MRI is then used as an anatomical mask to cover the full FDG-PET image obtained after the co-registering step through a mapping method. Finally the output image is co-registered to the corresponding Origin-PET image, using the FSL FLIRT module. As illustrated in the figure 21 the fused image is fed

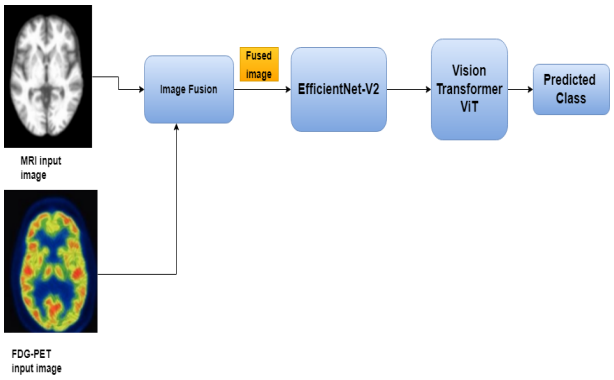


Figure. 21: Multimodal Alzheimer's disease early detection

into the EfficientNetV2 network to extract the local features and lower level features within the input image. EfficientNetV2 like the other CNN models suffer from the nonlocal features extraction. However it overcomes the inductive bias problem of the transformer. For this end, the output of the EfficientNetV2 is used as the encoding or embedding of the vision transformer.The ViT then extracts and encodes the main complex spatial relationships between high level image features which enhance the network learning and feature representation. Multimodal method or fusion of different brain modality ensure the extraction of different features and bring additional information that improve the feature representation and the learning process.

IV. Results and discussion

Our hybrid method has many advantages over the CNN models and the transformer methods because they are more computationally efficient and extract different global and local features. The figure 22 illustrated the performance of our multi-modal proposed model and we achieved an accuracy of 96%. of the network.

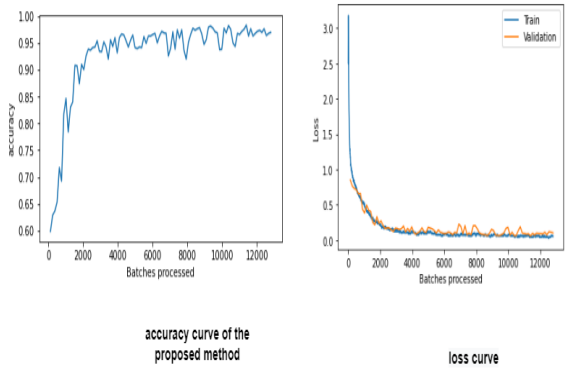


Figure. 22: The proposed method metrics

Our method performs well the ViT method, the CNN networks and the different current transformer networks. We evaluated different CNN models. The main advantages of these models are that they are easy to optimize and have a good generalization. However deep networks suffer from the vanishing gradient problem and over-fitting. Furthermore, CNN models don't encode and capture the relations at the pixel level within the input image. These models capture the local features but they don't capture the global features or the high level features within the image. Another limitation is that CNN doesn't pay attention to the relevant features within the image and doesn't encode the relative position of different features within the image. Another issue regarding the CNN is the receptive field size, increasing the size of the receptive field to capture more features can increase the model complexity. An effective feature extraction is based on a relevant feature representation that tracks the long dependencies within the image. The main limitation of CNN is that it doesn't encode the main long dependencies within an input image and doesn't hold the weighing of importance of each feature within the image. Vision transformer is a sub type of transformer for computer vision that incorporates the self attention mechanism that ensures a relevant weighing

Table 2: Comparative table

Method	Accuracy
AlexNet	61%
ResNet 152 model	85%
VGG16	79%
Densenet121	86%
ViT	89
DeiT	88%
ConViT	86%
Unimodal based on ViT and EfficientNet-V2	91%
Proposed method	96%

mechanism of importance of each feature within the image without compromising computational ability. ViT splits the input image into small patches which allow the model to process the local and global features within the image. The ViT demonstrated promising and good performance. In addition, it reduces the architecture complexity,ensures good scalability and features learning . However it requires a huge amount of data training and it is hard to optimize. Our method combines the main advantages of the ViT and the EfficientNetV2 to ensure global and local features and a robust feature representation that encodes the main long dependencies within the image. Our proposed method ensure an efficient feature extraction by combining the main advantage of ViT and EfficientNetV2. We noticed that the proposed data augmentation enhanced the model accuracy. The proposed model outperforms different models as depicted in the table 2 and ensure a remarkable improvement in accuracy compared to CNN and transformer models.

V. Conclusion

Brain disease prevention is a challenging task within personalized medicine. Alzheimer's disease (AD) is a progressive neurodegenerative disease that affects the person's ability to carry out daily tasks. The early diagnosis of this disease based on such biomarkers is the key step towards stopping its progression. Brain modality plays a critical role for understanding this disease. Deep learning opens new horizons and shows pertinent results within this context. CNN is one of the mods networks used for AD early detection. However its application brings some challenges such as the need for huge data for training, and the lack of a good mechanism to extract non local features. CNN does not ensure a valuable feature representation and does not track the relevant long range dependencies within the image. In addition, it does not hold a weighting mechanism that encodes the importance of each feature within the image. Vision transformers are emerging and promising networks that overcome the main issues of the CNN and ensure a robust feature representation and extraction with low complexity and computational cost. However these networks need huge amounts of data for training and they are hard to optimize. Furthermore, ViT has a low inductive bias compared to CNN. In this study, we proposed a hybrid method that overcomes these issues by combining the ViT and the EfficientNetV2. Our method combines the main advantages of the CNN and transformer at the same time and ensures a robust feature extraction and representation enhanced by a good data augmentation based on the self attention generative adversarial network. The proposed

method achieved the best accuracy compared to CNN models and transformer networks. The main advantage of our method is that we also combine the main advantages of the PET and MRI modality and our work is among the recent works that adopt transformers on the AD detection.

References

[1] Sethi, M., Ahuja, S., Rani, S., Bawa, P., and Zaguia, A. Classification of Alzheimer's Disease Using Gaussian-Based Bayesian Parameter Optimization for Deep Convolutional LSTM Network in *Computational And Mathematical Methods In Medicine*, pp. 1-16, 2021

[2] Yagis, E., Atnafu, S., Herrera, A., Marzi, C., Scheda, R., Giannelli, M., Tessa, C., Citi, L. & Diciotti, S. Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Scientific Reports*, 11 ,2021.

[3] Oh, K., Chung, Y., Kim, K., Kim, W. & Oh, I. Classification and Visualization of Alzheimer's Disease using Volumetric Convolutional Neural Network and Transfer Learning. *Scientific Reports*. 9, 2019.

[4] Katabathula, S., Wang, Q. Xu, R. Predict Alzheimer's disease using hippocampus MRI data: a lightweight 3D deep convolutional network model with visual and global shape representations. in *Alzheimer's Research Therapy*, 13,2021.

[5] Zaabi, M., Smaoui, N., Derbel, and Hariri, W. Alzheimer's disease detection using convolutional neural networks and transfer learning based methods. *17th International Multi-Conference On Systems, Signals Devices (SSD)*, (2020,7).

[6] Chaihtra, D, and Vijaya Shetty, S. Alzheimer's disease detection from brain MRI data using deep learning techniques. *2021 2nd Global Conference For Advancement In Technology (GCAT)*, (2021,10).

[7] Alshammari, M. and Mezher, M. A modified convolutional neural networks for MRI-based images for detection and stage classification of Alzheimer disease. *2021 National Computing Colleges Conference (NCCC)*, (2021,3).

[8] Venugopalan, J., Tong, L., Hassanzadeh, H, and Wang, M. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific Reports*,11, (2021,2).

[9] Song, J., Zheng, J., Li, P., Lu, X., Zhu, G,and Shen, P. An Effective Multimodal Image Fusion Method Using MRI and PET for Alzheimer's Disease Diagnosis. in *Frontiers In Digital Health*. 3, (2021,2).

[10] Abuhmed, T., El-Sappagh, S. and Alonso, J. Robust hybrid deep learning models for Alzheimer's progression detection. in *Knowledge-Based Systems*. 213 ,pp. 106688, (2021,2).

[11] Odusami, M., Maskeliūnas, R., Damaševičius, R. & Krilavičius, T. Analysis of Features of Alzheimer's Disease: Detection of Early Stage from Functional Brain Changes in Magnetic Resonance Images Using a Fine-tuned ResNet18 Network, in *Diagnostics*, 11, 1071 (2021,6).

[12] Zhang, F., Li, Z., Zhang, B., Du, H., Wang, B, and Zhang, X. Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease. in *Neurocomputing*,361,pp. 185-195 (2019,10).

[13] Khagi, B, and Kwon, G. 3D CNN Design for the Classification of Alzheimer's Disease Using Brain MRI and PET. *i*n IEEE Access,8,pp. 217830-217847, 2020.

[14] Hirose, S., Wada, N., Katto, J, and Sun, H. ViT-GAN: Using Vision Transformer as Discriminator with Adaptive Data Augmentation, in *2021 3rd International Conference On Computer Communication And The Internet (ICCCI)*. (2021,6).

[15] Durall, R., Frolov, S., Hees, J., Raue, F., Pfreundt, F., Dengel, A., and Keuper, J. Combining Transformer Generators with Convolutional Discriminators. *KI 2021: Advances In Artificial Intelligence*, pp. 67-79, 2021.

[16] Odusami, M., Maskeliūnas, R., and Damaševičius, R. An Intelligent System for Early Recognition of Alzheimer's Disease Using Neuroimaging. in *Sensors*. 22, 740 (2022,1).

[17] Kim, H., Lee, H., Oh, K., Lee, S., Yun, M., and Yoo, S. Multi-slice representational learning of convolutional neural network for Alzheimer's disease classification using positron emission tomography. in *BioMedical Engineering OnLine*,19, (2020,9).

[18] Song, J., Zheng, J., Li, P., Lu, X., Zhu, G., and Shen, P. An Effective Multimodal Image Fusion Method Using MRI and PET for Alzheimer's Disease Diagnosis. in *Frontiers In Digital Health*,3,(2021,2).

[19] Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-Attention Generative Adversarial Networks. (2019). .

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin, I. Attention Is All You Need. (2017).

[21] D'Ascoli, S., Touvron, H., Leavitt, M., Morcos, A., Biroli, G.,and Sagun, L. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. (2021).

[22] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A.,and Jégou, H. Training data-efficient image transformers distillation through attention. (2021).

[23] Tan, M., and Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. (2020).

[24] Solano-Rojas, B., and Villalón-Fonseca, R. A Low-Cost Three-Dimensional DenseNet Neural Network for Alzheimer's Disease Early Discovery. in *Sensors*. 21, 1302, (2021,2).

[25] Sethi, M., Ahuja, S., Rani, S., Koundal, D., Zaguia, A., and Enbeyle, W. An Exploration: Alzheimer's Disease Classification Based on Convolutional Neural Network. *BioMed Research International*,pp. 1-19 (2022,1).

[26] C, N., and Kusuma Comparative study of detection and classification of Alzheimer's disease using Hybrid model and CNN, in *2021 International Conference On Disruptive Technologies For Multi-Disciplinary Research And Applications (CENTCON)*. (2021,11).

[27] Xia, Z., Yue, G., Xu, Y., Feng, C., Yang, M., Wang, T., and Lei, B. A novel end-to-end hybrid network for Alzheimer's disease detection using 3D CNN and 3D CLSTM,in *2020 IEEE 17th International Symposium On Biomedical Imaging (ISBI)*, (2020,4).

[28] Ge, Chenjie and Qu, Qixun and Gu, Irene Yu-Hua,Jakola, Asgeir, Multiscale deep convolutional networks for characterization and detection of Alzheimer's disease using MR images, in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019.

[29] Sarraf, S., Sarraf, A., DeSouza, D., Anderson, J., Kabia, M.,and ADNI, T. OViTAD: Optimized Vision Transformer to Predict Various Stages of Alzheimer's Disease Using Resting-State fMRI and Structural MRI Data. (Cold Spring Harbor Laboratory,2021,11).

[30] Fong, J., Shapiai, M., Tiew, Y., Batool, U,and Fauzi, H. Bypassing MRI pre-processing in Alzheimer's disease diagnosis using deep learning detection network. in *2020 16th IEEE International Colloquium On Signal Processing Its Applications (CSPA)*. (2020,2).

[31] Ying, Q., Xing, X., Liu, L., Lin, A., Jacobs, N., and Liang, G. Multi-Modal Data Analysis for Alzheimer's Disease Diagnosis: An Ensemble Model Using Imagery and Genetic Features. (Cold Spring Harbor Laboratory,2021,5).

[32] Kadri, R., Tmar, M., Bouaziz, B. & Gargouri, F. Deep Squeeze and Excitation-Densely Connected Convolutional Network with cGAN for Alzheimer's Disease Early Detection, in *Intelligent Systems Design And Applications*. pp. 441-451 (2022).

Author Biographies

Rahma Kadri received her Master Thesis degree in Computer Science from Higher Institute of Computer Science and Multimedia, Sfax University, Tunisia in 2018. She is currently a PhD student at Sfax University, Tunisia and a member of Multimedia, InfoRmation systems and Advanced Computing Laboratory (MIRACL). Her research areas includes Image Processing, Machine Learning and Deep Learning.

Bassem Bouaziz Hold a PhD in computer science from University of Sfax. He is currently University council member. He is senior member of the Digital Research Center of Sfax (CRNS). He coordinates several multinational projects on biodiversity informatics and digital health technologies using Artificial Intelligence in partnership with industry and academia. His research areas include Multimedia document indexing and processing, Computer Vision for video analysis, Deep learning, biomedical signals processing.

Mohamed Tmar Currently, he is Associate Professor at the Department of Computer Science of the Higher Institute of Computer Science and Multimedia at the University of Sfax, Tunisia. he is a member of the Multimedia, Information systems and Advanced Computing Laboratory, University of Sfax. His main research areas are Multimedia document indexing and processing, Computer Vision for video analysis, Deep learning, Objects recognition.

Faiez Gargouri Professor of Computer Science at University of Sfax, he is a member of the Multimedia, Information systems and Advanced Computing Laboratory and Vice-President of the University of Sfax. He was the head of the Higher Institute of Computer science and Multimedia from 2007 to 2011. He has got his maitrise diploma in computer management, faculty of economics and management of Sfax (1988), his master in computer science from the Paris 6 University (1990) and his PhD thesis, Paris 5 University (1995). He got his Habilitation Degree in Computer Science, Faculty of Sciences of Tunis (2002). His research interests include Business Information Systems, Business Intelligence, multimedia Information systems, Ontology, Deep learning...He supervises several theses.