Proceedings of the 4th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/49/20241286

A novel interpretative deep neural network with Grad-CAM's heatmap for the early diagnosis of Alzheimer's disease

Annabelle Yao

The Lawrenceville School, New Jersey, USA

ayhk2017@gmail.com

Abstract. In recent years, Alzheimer's disease, a disease which targets the patient's cognitive abilities and causes dementia, has become increasingly prevalent among the older population. Clinical practice today diagnoses the disease through MRI imaging and cognitive tests based on individual doctor's personal experience. This reliance on varying doctoral experiences poses a huge challenge in during its early stages, when the symptoms are subtle and hard for clinicians to discern. As a result, many elderly today are only diagnosed when the disease has already progressed into a later, much obvious stage, significantly reducing their chance of receiving effective treatments. In our work, we applied a novel method in Alzheimer's disease early diagnosis. Our unprecedented method uses a state-of-the-art image recognition model, a Vision Transformer (ViT), combined with a saliency map, Grad-CAM, to increase interpretability and alleviate the black box issue in the predicted results. We trained the ViT with ADNI imaging datasets and conducted similar experiments with other popular models for machine learning as baselines for performance comparison. The results show that our model has delivered the best performance out of all traditional methods, with an astonishing 99% accuracy. Compared to an average accuracy of 96% by other commonly used models for data prediction, our method has improved the final prediction accuracy to 3 standard deviation range, thereby accounting for almost all the possible outliers. These breakthroughs, paired with the website we created to increase the model's reach, will prove to be a revolutionary contribution to the field of Alzheimer's disease diagnosis.

Keywords: Alzheimer, Vision Transformer, Deep Learning, Neural Networks, Interpretability.

1. Introduction

In the field of using deep learning methods for Alzheimer's disease diagnosis, some methods have been introduced for sophisticated data modeling. Two particularly prominent techniques have gained traction in recent times: the 3D Deep Convolutional Neural Networks (3D-CNN) and the Long Short-Term Memory (LSTM) neural network models. Both models possess unique capabilities. For instance, 3D-CNN is adept at processing spatial hierarchies in data, making it a preferred choice for tasks involving volumetric images. On the other hand, LSTM, with its memory cells, excels in identifying and understanding temporal sequences, ensuring patterns over time aren't overlooked. Our method is the increasingly popular model, Vision Transformer, a model that can easily interconnect modelled relationships for every portion of the input. Yet, as the medical field gradually increases its reliance on

DOI: 10.54254/2753-8818/49/20241286

model outcomes for critical decisions, the interpretability of these models becomes crucial for upscale application.

In this work, we combined the advantages from both worlds. We implemented a state-of-the-art image recognition model, the Vision Transformer, and we added a post-hoc interpretation module, Grad-CAM, to provide model explanations.

The Vision transformer [1] is a model that has only started to rise in popularity in recent times due to ChatGPT's transformer structure. The model is highly computationally efficient and uses self-attention, an ability to draw information from the whole input, different to the most often used convolutions with other deep learning models that were previously used to do disease prediction. ViT is also previously trained with the ImageNet and ImageNet-21k datasets and process input data in parallel, allowing us to minimize computational resources needed to train, and do faster training and inferences. However, the largest advantage is that they can be interpreted very easily through post-hoc modules.

In our research endeavors, we recognized that achieving a high-accuracy model wasn't sufficient for full acceptance in the medical domain. There remained concerns about potential undetected errors in the model's reasoning that could compromise patient safety. As a response, we further explored methods to enhance the interpretability of deep learning models. Among the possible machine learning interpretation methods to choose from, Grad-CAM stood out. It functions by calculating the average weights of pixels across all layers, producing a detailed class activation map. This map provides a visual representation, illuminating the weight distribution of each pixel, making it easier to pinpoint areas of significance within the data. The added interpretability by Grad-CAM has significantly improved the transparency and trustworthiness of our model.

In sum, through this research, we have made the following contributions:

- We identified a new significant problem within the medical field
- We applied a novel method using a state-of-the-art deep learning model (ViT) in Alzheimer's disease prediction and analyzed its ability in diagnosing early Alzheimer's given MRI scan images.
- We solved the challenge of lack of interpretability through applying the Grad-CAM method, allowing the model to assist doctors to make decisions in real world diagnosis and applications.
- We modelled and compared other traditional methods of disease prediction, showing the different results and drawing conclusions for applicability.

2. Related works

2.1. Deep learning models for disease prediction

3D deep convolutional neural networks are a type of neural network that uses a 3D filter to perform convolutions. The kernel is able to move in 3 directions versus 2 directions in the 2D neural network, and all the input and output data is 4 dimensional. The deep CNNs can be traced back to the initial image classification 2D CNN architectures such as AlexNet, VGGNet, and GoogLeNet. In 2014, Tran et al. introduced concepts of 3D CNNs for action recognition in videos, applying CNNs to spatiotemporal data. Their work, called C3D [2], helped with the advancement of 3D CNNs in multiple areas, especially ones that were extensively used in medical imaging recognition such as volumetric segmentation [3], and brain imaging analysis [4] etc. In this work [5], the 3D deep convolutional neural network was used to find the most common and reoccuring pattern in Alzheimer prediction. After training, the model was able to conclude and identify imaging biomarkers that are predictive and indicative of Alzheimer's disease, pairing them with other given information to accurately early detect the disease.

Another method of data prediction is using Long Short-Term Memory (LSTM) neural networks. They were originally invented in 1995 by Sepp Hochreiter to solve the vanishing gradient problem [6]. This neural network uses feedback connections instead of feedforward neural network connections, resulting in a long short-term memory gradient compared to the short term memory vanishing gradient of a normal neural network. LSTMs preserve previous information from earlier sequences and carry it forward, making it ideal for predicting data and processing large amounts of data [7] . In this work [8],

DOI: 10.54254/2753-8818/49/20241286

the LSTM was used to predict the trajectory of an individual's Alzheimer'S disease progression. The use of LSTM instead of normal neural networks allowed the accuracy of the prediction to increase to 95 for both fast and short progression patients.

2.2. Machine learning interpretability

In this work, we focus on interpretability methods for computer vision.

Interpretability for computer vision can be split into two major methods, post-hoc method and intrinsic method. The post hoc method is when the deep learning model has a relatively simplistic structure that results in the easy interpretability of the generated predictions. Conversely, the intrinsic method refers to the interpretability measures implemented after the model is trained.

A recent work [9] used a Grad-CAM posthoc interpretability method to "to visualize the evidence on which a classifier bases its decisions," which doesn't constrain model architecture. The deep learning model does this by first transforming the image into a saliency map, an image that the computer highlights with varying degrees of color according to which region it chooses to focus on. This feature can help identify the most important parts of an image, and helps with segmentation of key portions [10] to help accurately determine which factors the computer thinks most affects the final conclusion and result, thereby increasing interpretability of the model.

The intrinsic Gaussian mixture model (GMM) is a type of mixture model that assumes all the underlying components use Gaussian distribution, also called a normal distribution, and uses a bell shape curve and has recently been used to help increase interpretability in deep learning models [11] . The Gaussian mixture model was around long before 1846, however in 1977, Arthur Dempster, Nan Laird, and Donald Rubin made significant adjustments to the model by creating the Expectation-Maximization algorithm for estimating parameters within the model accurately. The GMMs, with their simplicity, interpretability and ability to model complex data distributions quickly spread, being used across speech recognition, speech acoustics, anomaly detection and computer vision fields. This paper [12] used the Gaussian mixture model to "model intracacies of healthcare data" and "benchmark casual effect and policy estimators" with the advantage of it being interpretable.

3. Methods

3.1. Vision Transformer Architecture

For our training model, we used a Vision Transformer [1] to make predictions. Vision Transformers (ViT) is a more efficient alternative to the traditional Convolutional Neural Networks (CNNs) for image recognition and computer vision tasks. The overall structure is shown in Figure 1.

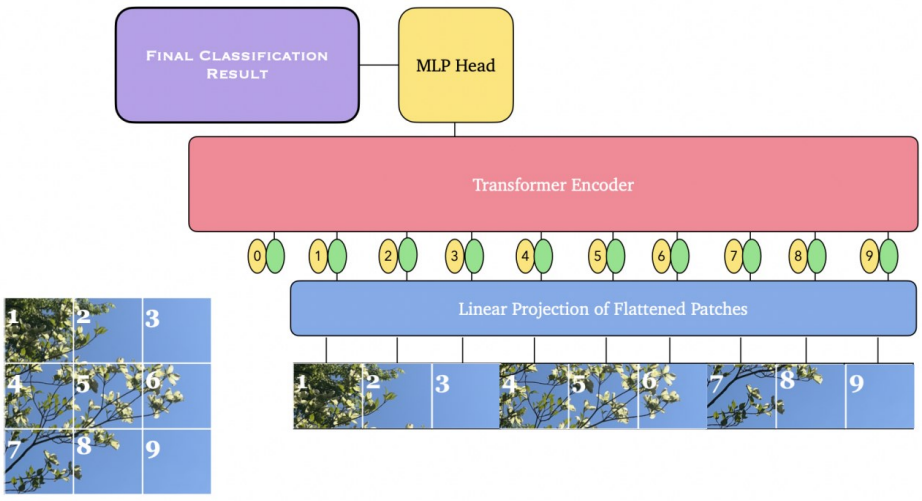

Figure 1. Vision Transformer Architecture

DOI: 10.54254/2753-8818/49/20241286

The inputs are images of height H, width W, and number of channels C. These images are cut up into smaller two dimensional square patches, resulting in $(HW)/P^2$ patches. Each patch has a resolution of (P, P) pixels and forms a sequence to makes up an image. These patches are then flattened into a singular vectors and input into a single feed forward layer that linearly projects the concatenation of the channels of all pixels in the image patch onto an input dimension. Then, a learnable class embedding and 1D positional embedding are attached to the end of the image's sequence as the classification output and to retain each patch's positional information throughout the training process. The final sequence of embedded image patches is then fed into a standard transformer encoder with 24 layers and 16 attention heads. Within the encoder, each layer's input is the previous layer's output. The output is calculated through a series of different equations and decisions as explained below in Figure 2.

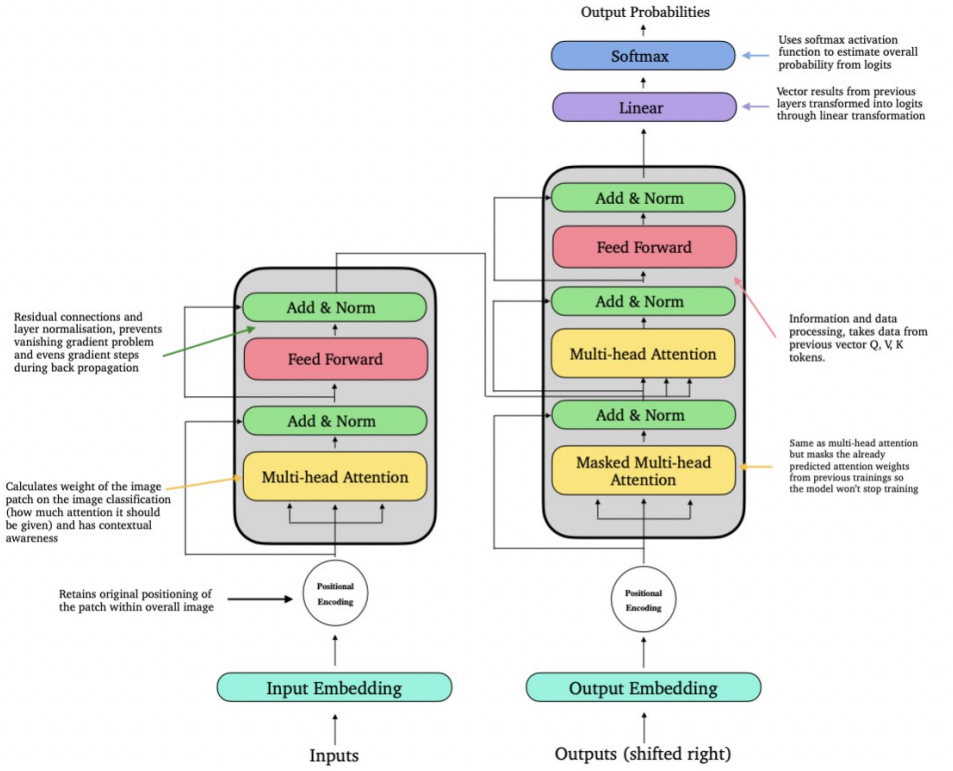

Figure 2. Vision Transformer Encoder Decoder Architecture

First, the input embedding is fed into the encoder and positional encodings are added to the sequence. Each patch is split into individual Q (the set of vectors we calculate attention for), K (the set of vectors we're given, ie. the image sequence's vector), and V (the encoder embedding vector) vectors. Then, the sequence has its self-attention weight, the amount of importance the patch of the image has on the overall classification of the image, calculated through the equation below, with d standing for the size of the patch, Q standing for the Query token, K standing for the Key, and V standing for Value.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) * V \tag{1}$$

The vector results from the self-attention have contextual awareness, meaning that the vector representation of the patch can take into consideration the patches around it in the original image. The vector results are then sent into a multi-head attention layer. This layer goes through the same process as the self-attention layer, except when initially splitting the patch into Q, K, and V vectors, it will further split the Q, K, and Vvectors into 8 smaller Q, K, and Vvectors. After computing the attention matrices for each case, the resulting vectors have a higher contextual awareness compared to the vectors

DOI: 10.54254/2753-8818/49/20241286

after the self-attention layer. The output from this layer is then put through the add & norm layer, where it goes through two individual steps, add and norm. The add step is a residual connection. This means that it sums the result from the positional encoding layer with the input, and it's done to prevent the vanishing gradient problem during the back propagation of neural networks (the model will keep back propagating until the gradient becomes 0 and it will stop learning). The norm step stands for layer normalisation. It gathers all the neurons for the neural network a smaller range centered around 0, and it allows for the gradient steps in the back propagation phase to be more even, and thus increase stability of the overall training. Through a series of more add & norm, feed forward, and multi-head attention layers, it is input into a linear layer, where the vectors are transformed into a smaller dimensional representation, logits, through a linear transformation and the application of an activation function. After obtaining the logits, the softmax layer applies the softmax function to turn the logits into a probability for each class. The class with the highest probability is the predicted class for our initial input. The output is then fed into the next cycle of layers where it goes through the same process, using a masked multi-head attention layer instead of a multi-head attention layer. This is because in the process, self-attention allows the transformer decoder to look at future predictions if they're not masked. The mask is done through multiplying the original I matrix, with $I = QK^T$, by

$$M = \begin{pmatrix} m_{1,1} & m_{1,2} & \dots & m_{1,L} \\ m_{2,1} & m_{2,2} & \dots & m_{2,L} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ m_{L,1} & m_{L,2} & \dots & m_{L,L} \end{pmatrix}$$

where $m_{i,j}$ is an element of the mask matrix. The value of $m_{i,j}$ is 1 if position j is allowed to know position i, and 0 if it isn't. After multiplying the original I matrix by the mask matrix to get I', the attention equation becomes

$$Attention(Q, K, V) = softmax\left(\frac{I'}{\sqrt{d_k}}\right) * V \tag{2}$$

and will prevent attention from future positions from affecting this patch's attention. After going through the entire process of more add & norm, feed forwards, and multi-head attention layers, the output of probabilities from the decoder go through an extra linear layer classification in the MLP head layer, where the most likely classification is output, and the other classifications are ignored.

3.2. Interpretability module Grad-CAM Architecture

To help solve the challenge of interpretability, we used Grad-CAM to help draw the saliency maps.

Grad-CAM, Gradient weighted class activation mapping, is an explanation method used for neural networks with gradients. Throughout the training sequence of the model, in the transformer encoder, the Grad-CAM back propagates the gradient to find the output class's probability score in respect to the self-attention layer's query embeddings. The gradients are aggregated via taking the mean or sum of each self-attention head with the equation below to obtain a single importance weight for each embedding. l is the number of self-attention layers, H is the number of attention heads, the query embedding is $q_{l,h}$, and the gradient with respect to query embedding is $\frac{\partial logits_c}{\partial q_{l,h}}$.

$$\alpha_l = \frac{1}{H}\sum_{h=1}^{H} \frac{\partial logits_c}{\partial q_{l,h}} \tag{3}$$

For each attention layer, Grad-CAM takes the weighted sum of each query embedding using the importance weights α_l. This allows the most important query to be emphasized. The weighted sums are then passed through the Softmax activation function to keep only the positive contributions. Then, the activations are converted into a heatmap through the equation below.

DOI: 10.54254/2753-8818/49/20241286

$$\text{Heatmaps}_l = \sum_{h=1}^{H} \alpha_1 \odot q_{l,h} \tag{4}$$

They are then upsampled to the size of the original image through an interpolation technique called bilinear interpolation. This technique uses estimates the values of pixels at non-integer coordinates based on the values of neighboring pixels, and resizes the heatmap to the desired size, maintaining its clear resolution. It uses the equation below for each spatial position (i, j) in the output heatmap [13]. Heatmap(u, v) is the value at position (u, v) in the original heatmap, $w(x)$ is the interpolation weight for position x, and is determined by finding the distance between the target position (i, j) and its neighboring positions (u, v).

$$\text{Heatmaps}_{\text{upsampled}}(i, j) = \sum_u \sum_v \text{heatmap}\,(u, v) \times w(i - u) \times w(j - v) \tag{5}$$

with the interpolation weight $w(x)$ determined by the equation below

$$w(x) = \begin{cases} 1 - |x|, & \text{if} |x| \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

The now input image sized heatmap in Figure 3 has highlighted red/yellow regions being the parts that the training model paid attention to when making its classification decision, explaining the result as in Figure 3.

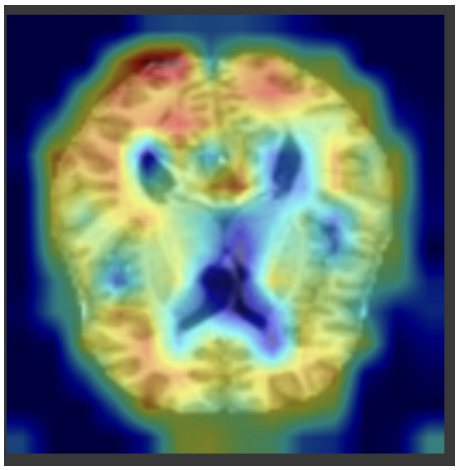

Figure 3. An example of the final heatmap created

4. Experiments

4.1. Datasets

For this research, we conduct a classification task and make predictions based on the Alzheimer's dataset using machine learning. Alzheimer's is a progressive disease that targets the brain, and causes dementia, destroying cognitive abilities and resulting in memory loss. We extract a patient dataset from Kaggle and public hospital repositories[1234] that are available upon request and application. This dataset includes a total of 6400 pre-processed MRI (Magnetic Resonance Imaging) Images, split into four classes of images: Very Mild Demented, Mild Demented, Moderate Demented, Non Demented, and are all resized into an image consisting of 128 x 128 pixels as shown in Figure4 below. The precise data statistics can be found in the Table 1.

[1] https://adni.loni.usc.edu/
[2] https://www.alzheimers.net/
[3] https://catalog.data.gov/dataset/alzheimers-disease-and-healthy-aging-data
[4] https://www.nature.com/articles/s41598-020-79243-9

DOI: 10.54254/2753-8818/49/20241286

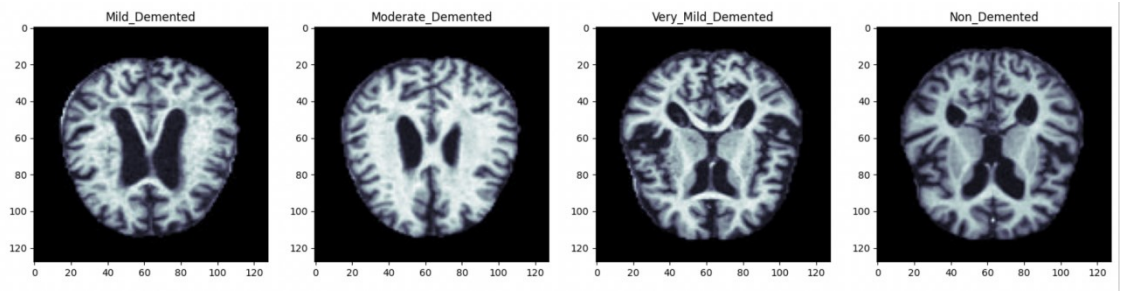

Figure 4. Images of a sample from each category in the Dataset

Table 1. Data statistics of Alzheimer's dataset

Dataset	Number of Cases
Very Mild Demented	2240
Mild Demented	896
Moderate Demented	64
Non Demented	3200

4.2. Baselines and Variants

We compare ours with the following baseline models. For fair comparison, we use all the data in the dataset given above for all the models. For our proposed ViT model, **we provide the code in Github**[5].

(a) Traditional Models

- SVM [14]: SVM (Support Vector Machine) is a linear model for regression and classification. It separates two classes of a dataset with a linear line or a hyperplane that optimises the precision of the split and the maximum margin of the split. In our task, we use SVM for classification.
- XGBoost [15]: XGBoost is a machine learning model that aims to scalably boost gradient enhancement. It trains multiple weak learning decision tree models on a subset of the total data and combines all the results to create a more accurate final prediction. In our task, we use XGBoost for classification.
- MLP: MLP classifier (Multi-layer Perceptron classifier) is a feed forward type neural network model used in machine learning that uses back propagation to classify classes within a dataset. It is commonly used in supervised learning and in our task, we use the MLP classifier for classification.

(b) Deep Learning Models

- AlexNet 16]: AlexNet is a convolutional neural network model that uses consecutive convolutional layers with a feedback and feed forward structure to perform image classification by matching and classifying an image according to pretrained models. In our task, we use AlexNet for classification.
- VGG-16 [17]: VGG-16 (Visual Geometry Group) is a convolutional neural network model that uses 13 convolutional layers, hidden layers, 3 fully connected layers, and a ReLu (Rectified linear unit activation function) unit to perform image classification and recognition. In our task, we use VGG for classification.

(c) Our Proposed Model

- ViT [6]: ViT (Vision Transformer) is a model for image classification and recognition that uses a transformer architecture to extract key portions of images by giving each subsection of an

[5] https://github.com/Annabelleyao/Interpretative-Deep-Neural-Network-with-Grad-CAM-s-Heatmap-Early-Diagnosis-of-Alzheimer-s-Disease

DOI: 10.54254/2753-8818/49/20241286

image different amounts of attention and sending the flattened images through an encoder which will classify the image. In our task, we use ViT for classification.

4.3. Implementation Details

For the Alzheimer's diagnosis task, we randomly split the dataset into training and testing sets in a 80:20 ratio. For data, we use MRI dataset taken from Kaggle and ADNI database. The precise dataset details can be found in the "Datasets" section. We train all models with a batch size of 64 and a learning rate of $1e$-4. We train all the models until convergence. All models are trained on CoLab T4 GPU with 25.5 GB memory.

We use a pre-trained Vision Transformer (ViT) model on ImageNet-21k (14 million images, 21,843 classes) at resolution 224x224. To embed patient Alzheimer's disease data, we use 12 blocks of transformers, each block applies a multi-head self-attention layer, a normalization layer, a feed forward layer and another normalization layer. For the transformer we set the depth as 12 and the number of heads as 12.

Table 2. Exp 1. Performance Comparison.

	Precision	Recall	Macro F1 Score	Weighted F1 Score	ROC AUC
Traditional Methods					
SVM	0.25	0.50	0.17	0.33	0.50
XGBoost	0.99	0.94	0.96	0.98	0.99
MLP	0.82	0.78	0.77	0.78	0.96
Deep Learning Methods					
AlexNet	0.47	0.42	0.43	0.63	0.85
VGG	0.93	0.95	0.88	0.92	0.96
Our Method					
ViT	0.99	0.98	0.99	0.98	0.99

4.4. Metrics

Since the datasets in the experiment are imbalanced, we choose metrics that are fit for measuring model performance under data imbalance setting.

1. **Precision**: The precision metric quantifies the number of correct positive predictions that the model made. It calculates the prediction accuracy for positive predictions, and is shown in the equation below. True positives stand for correctly predicted positive values, and false positives stand for incorrectly predicted positive values.

$$\text{Precision}=\frac{\text{True Positives}}{\text{True Positives + False Positives}} \tag{7}$$

2. **Recall**: The recall metric quantifies the number of predicted true positives out of the amount of total positives that could have been predicted. It provides an insight into missed positive predictions. False negatives stand for incorrectly predicted negatives.

$$\text{Recall}=\frac{\text{True Positives}}{\text{True Positives + False Negatives}} \tag{8}$$

3. **F1 Score**: The F1 Score is the weighted average of precision and recall, and considers the false positives and false negatives in both, constituting to an overall accuracy of the model. It is calulated with the equation below.

$$\text{F1 Score}=\frac{2(\text{Precision}\times\text{Recall})}{\text{Precision}+\text{Recall}} \tag{9}$$

Since our dataset is imbalanced, the reported averages include macro average (averaging the unweighted mean per label), weighted average (averaging the support-weighted mean per label).

DOI: 10.54254/2753-8818/49/20241286

4. **ROC AUC** The ROC AUC metric is calculated by finding the area under (AUC) the Receiver Operator Characteristic (ROC) curve. The ROC curve is an evaluation metric for binary classification. It's a probability curve on the graph of TPR against FPR and helps show the performance of the model at all classification threshold points. TPR stands for True Positive Rate and is the same as the Recall metric. FPR stands for False positive rate, and is calculated by the equation below.

$$\text{False Positive Rate} = \frac{\text{False Positives}}{\text{False Positives + True Negatives}} \tag{10}$$

The AUC is used as a summary of the ROC curve, and it shows the model's ability to distinguish between positive and negative classes [18]. The equation of the ROC AUC metric is below.

$$\text{ROC AUC} = \int_0^1 \text{TPR}\left(\text{FPR}^{-1}(x)\right)dx \tag{11}$$

5. Results & Analysis

5.1. Exp 1. Performance Comparison

Looking at the Performance Comparison Table 2, we see that our Vision Transformer model had the highest scores across all evaluation metrics, followed by the XGBoost model, the VGG model, the MLP model, the AlexNet model, and lastly the SVM model. Our Vision Transformer model performed better than the other models, having an overall ROC AUC, Macro F1 Score, and precision of 0.99, and a weighted F1 Score, and recall of 0.98. We also see that XGBoost had a high ROC AUC and precision compared to the SVM, MLP, and AlexNet methods. With VGG, a more recent CNN model, we see that the results obtained also come close the results of our Vision Transformer model.

Reasons behind XGBoost and VGG's outstanding performances could be because of their creation for classification. XGBoost uses predefined models to reach a classification result, and VGG uses CNNs and ReLu to reach a result. When comparing the two, we see that XGBoost did better than VGG. When performing classification, XGBoost tries to fit the training data with a predefined model made up of trees or linear combination. On the other hand, VGG tries to find coefficients that minimize error functions on the training data. This means that the VGG might converge at a local minimum rather than a global minimum which XGBoost may reach on first try, resulting in a better performance from XGBoost rather than VGG.

However, when we compare both to our Vision Transformer model, we see that XGBoost and VGG are still weaker in image classification and prediction. Looking at the Grad-CAM comparison (Exp 2. Leveraging Grad-CAM's heatmap to Improve Interpretability), we see that VGG and our ViT pay attention to different areas of the same image, potentially resulting in the observed performance difference between the two classification models.

After our Vision Transformer and the XGBoost model, VGG did the best with a ROC AUC of 0.96, a weighted F1 score of 0.92, a Macro F1 Score of 0.88, a recall of 0.95, and a precision of 0.93. Then it was the MLP odel which had a ROC AUC of 0.96, a weighted F1 Score of 0.78, a Macro F1 Score of 0.77, a Recall of 0.78 and a Precision of 0.82. Then it was the AlexNet CNN with a ROC AUC of 0.85, a weighted F1 Score of 0.63, a Macro F1 Score of 0.43, a Recall of 0.42 and a Precision of 0.47. The SVM model did the worst out of all the models tested, with a ROC AUC of 0.50, a weighted F1 score of 0.33, a macro F1 score of 0.17, a recall of 0.50 and a precision of 0.25.

From these results, we can see that our innovative method of using a Vision Transformer model delivered a better overall performance compared to all the other traditional disease diagnosis models, with an astonishing 99% accuracy. Compared to the highest accuracy of 96% amongst the other previously used models for data prediction, our method has improved the final prediction accuracy to 3 standard deviation range. This means that our model appropriately accounts for all the possible errors that may occur in its final prediction, and resulting in a more trustworthy and reliable result.

DOI: 10.54254/2753-8818/49/20241286

5.2. Exp 2. Leveraging Grad-CAM's heatmap to Improve Interpretability

In our paper, we also used a Grad-CAM heatmap with our Alzheimer's early prediction ViT model to show the reasoning behind each early-prediction's classificiation. The highlighted portions of the heatmap according to the spectrum in Figure 5.

Grad-CAM's Heatmap Relative Intensity

Figure 5. Relative intensity of Grad-CAM's heatmap, with the right colors being the most intense and crucial areas in the model's decision making process, and the left being the least crucial areas

Below in Figure 6 are some samples from the ViT's Grad-CAM model. Having Grad-CAM will increase interpretability of our machine learning model, and can help doctors' understand our model's decision making process, thus helping solve the black-box problem seen across all machine learning models.

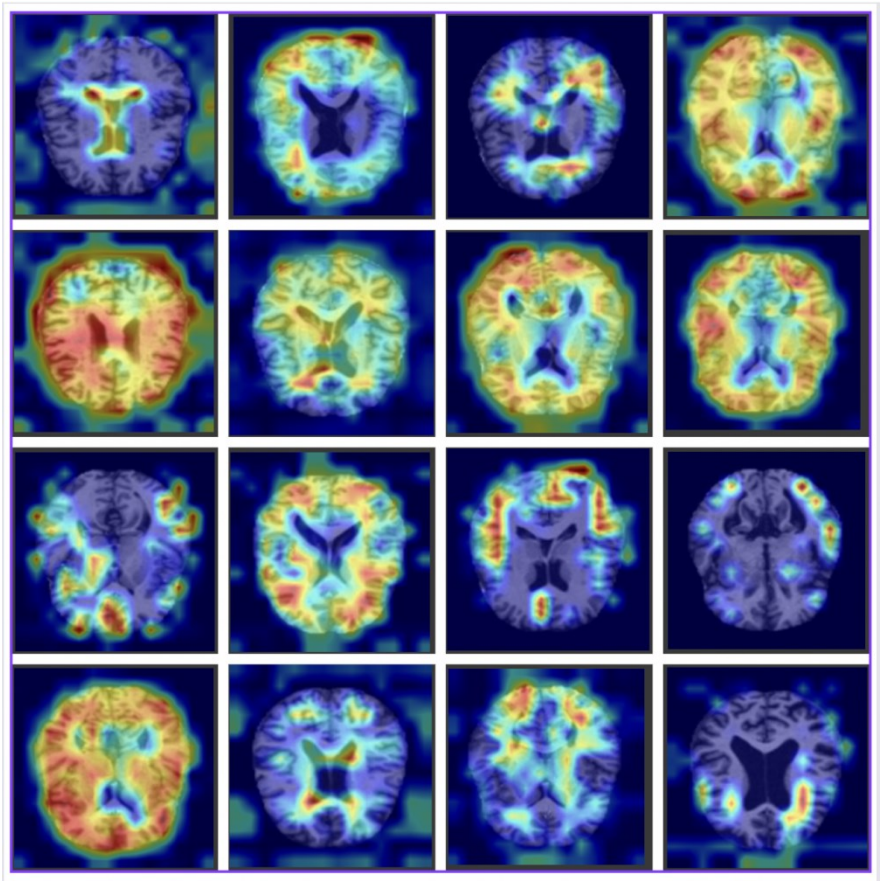

Figure 6. Multiple samples of the final heatmaps with ViT and Grad-CAM to aid doctors with diagnosis

The portions of the heatmap highlighted with colors nearer to the right of the spectrum show that the crucial portion of the decision making process is typically gathered around the center of the MRI's brain and the darker areas of the heatmap show the areas around the side of the image that are less important. These heatmaps will be extremely beneficial in guiding doctors to find possible related regions of threat of the disease when diagnosing early Alzheimer's.

DOI: 10.54254/2753-8818/49/20241286

When previously contrasting alternate methods of classification, we mentioned VGG, a deep CNN used for classification, which performed second best compared to the other models. Whilst VGG performed well, our ViT model performed much better. Below in Figure 7 are two test images taken from both VGG and ViT Grad-CAMs that show disparities in the models' areas of focus.

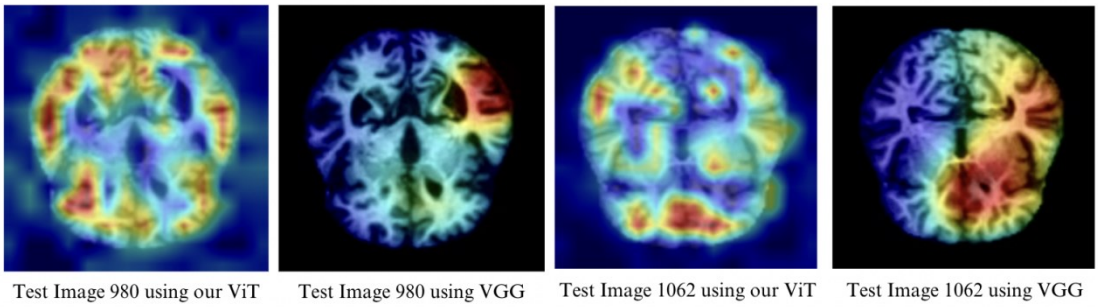

Figure 7. Two sample test cases from the Grad-CAM of VGG and ViT for performance comparison

In the comparison images above, we see that there are areas of attention weight overlap, proving our model's reliability. In Test Image 1062, both our ViT and the VGG model pay attention to the lower middle of the brain, showing that our ViT model is a very strong model (as proven by mostly overlapping results with the previously proven and heavily-relied on VGG model in data classification within science).

However there are also slight differences in each model's individual pixels' attention weights. As seen above with Test Image 980, the ViT pays more attention to the left region of the upper and lower brain whereas the VGG pays more attention to the upper right side of the brain. The ViT makes its classification decisions based on the entire image's individual pixels' attention weight in the convolutional layer. In comparison, the VGG model makes its classification decisions based on only each region's pixels' attention weight. This difference in regions of focus in making classifications causes the ViT to make decisions on a more global scale compared to the VGG, resulting in higher (ViT) and lower (VGG) performances in the models.

6. Conclusion

In this research we made 3 major contributions to the field of Alzheimer's disease early predictions. Firstly, we created a model that improved the accuracy of Alzheimer's disease early prediction by 3% to 99% prediction accuracy. Secondly, we created interpretability features for our model to solve the black-box problem commonly seen with machine learning. This can help explain the decision making process of our model and help with doctor diagnosis. Thirdly, we created a website with our model's code built in that people can upload their MRIs to for early predictions. It also contains information on Alzheimer's that people can reference to help prevent or slow the disease as well as information on treatment and related causes of the disease.

Alzheimer's disease has become more prevalent amongst many seniors today, causing 33% of dementia cases and is one of the top 10 causes of death within adults. It is currently the 5th leading cause of death for adults over 65 worldwide, and will be responsible for over 13.6 million adult deaths by 2060 [19]. This disease is best mitigated and prevented from getting worse at an early stage. Our Vision Transformer Model has shown the best performance out of all the traditional and deep learning methods tested for Alzheimer's Early prediction based off on MRI Brain scans. MRI brain scan is currently one of the most common ways to identify Alzheimer's within a patient. By using these scans, our ViT model was able to reach a 0.99 ROC AUC and Precision. This will largely improve the overall accuracy of Alzheimer's disease early prediction as it can help doctors highlight key areas of problem for them to diagnose patients.

Previously Alzheimer's disease early prediction models, though with a convincing level of accuracy, have failed to populate many hospitals due to their being a black box neural network. End users, i.e.

DOI: 10.54254/2753-8818/49/20241286

doctors, don't understand those models' reasoning behind their decisions, making the models thereby undesired in high stake medical domains. Our Grad-CAM heatmap solves this issue through highlighting important regions of the MRI scan that indicate or largely contribute to the model prediction. End users will receive more interpretation and guidance regarding the diagnosis, thus greatly boosting trustworthiness in the predictions. Such improvement in trust will ultimately lead to an increased overall usage of Neural Network models such as ours, in the Alzheimer's early prediction field, benefiting millions of the elderly and their families around the world.

To help expand the positive impact our model will bring beyond just the scientific community, we also created a website linked with a server containing our model's code for users to upload their own MRI scans to for early prediction results. In addition, the website contains all the necessary information the user needs to know on Alzheimer's to help them prevent the disease, or get proper treatment when diagnosed. The link is attached here. https://s3.s100.vip:4794/. **The website's code is all provided in the GitHub.**[6]. The images of the website are below as reference. The images of the website are below as reference. The main page of our website that includes our model and method of early Alzheimer's Disease prediction is shown in Figure 8. The resources section is shown in Figure 9, and the background referencing information is shown in Figure 10.

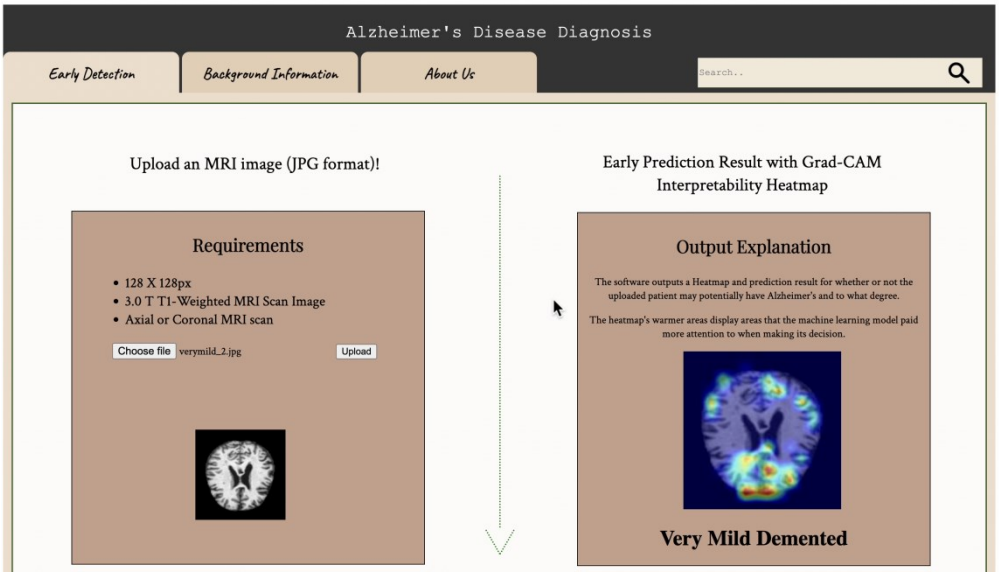

Figure 8. Image of our created Alzheimer's Disease Website User MRI Upload and Early Prediction Result page

[6] https://github.com/Annabelleyao/A-novel-interpretative-Deep-NN-with-Grad-CAM-s-heatmap-for-the-diagnosis-of-Alzheimers-disease

DOI: 10.54254/2753-8818/49/20241286

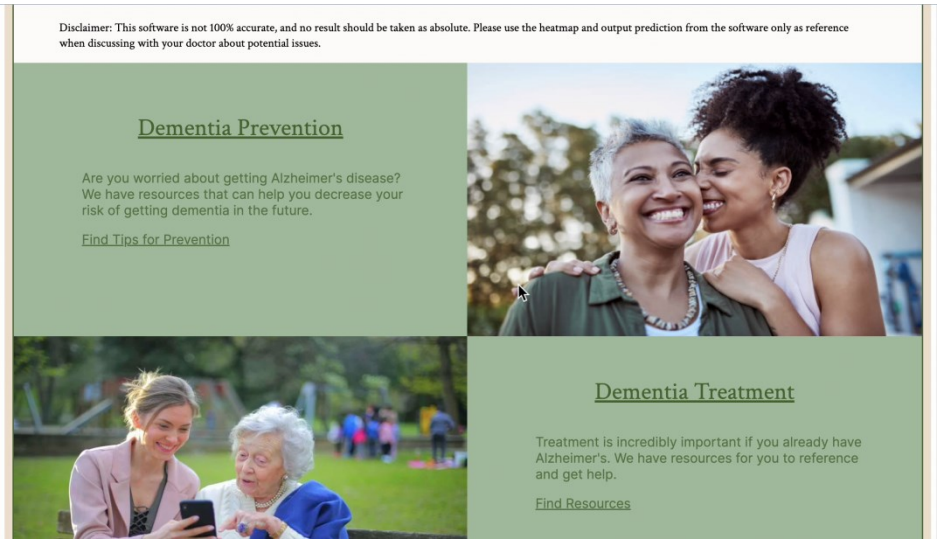

Figure 9. Image of our created Alzheimer's Disease Website treatment and prevention section

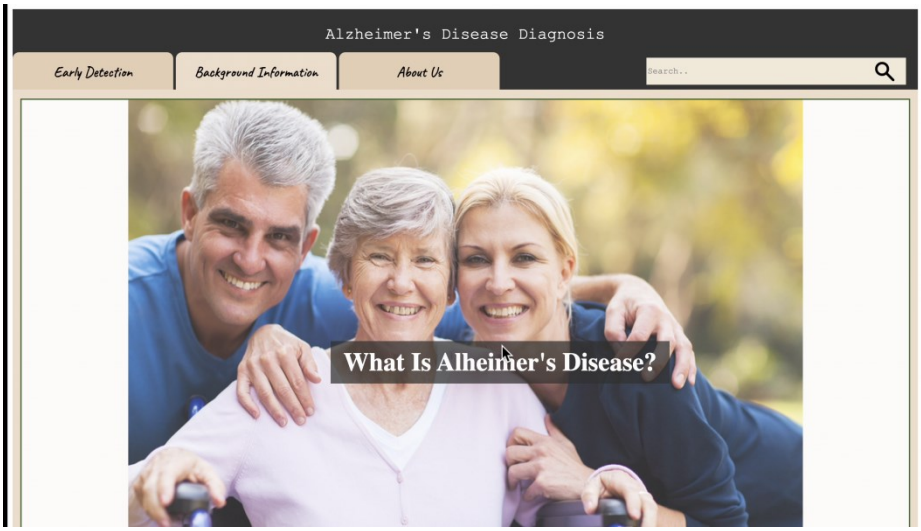

Figure 10. Image of our created Alzheimer's Disease Website Background Information page

In conclusion, our Vision Transformer model has delivered a better performance in Alzheimer's disease early prediction than the traditional methods and the other deep learning methods we found and tested. Our novel method also provides clear interpretation of the model's prediction through a Grad-CAM heat map, offering bigger chances for a wider adoption of Neural Networks technology in the high-stake fields. With our website, our model will able to expand to many different countries and regions with less medical and healthcare resources, helping reduce the differences in healthcare between countries, informing users of disease information and helping predict the disease so they can get early treatment, and aiding in alleviating the pressing issue of widespread Alzheimer's disease.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenbom, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010. 11929, 2020.

[2] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 4489-4497, 2015.
[3] Xin Yang, Lequan Yu, Shengli Li, Huaxuan Wen, Dandan Luo, Cheng Bian, Jing Qin, Dong Ni, and Pheng-Ann Heng. Towards automated semantic segmentation in prenatal volumetricultrasound. IEEE transactions on medical imaging, 38(1): 180-193, 2018.
[4] Wei Li, Xuefeng Lin, and Xi Chen. Detecting alzheimer's disease based on 4d fmri: An exploration under deep learning framework. Neurocomputing, 388: 280-287, 2020.
[5] Sheng Liu, Arjun V Masurkar, Henry Rusinek, Jingyun Chen, Ben Zhang, Weicheng Zhu, Carlos Fernandez-Granda, and Narges Razavian. Generalizable deep learning model for early alzheimer's disease detection from structural mris. Scientific reports, 12(1): 17106, 2022.
[6] Alex Graves and Alex Graves. Long short-term memory. Supervised sequence labelling with recurrent neural nervorks, pages 37-45, 2012.
[7] Alex Sherstinsky. Fundamentals of recurrent neural network (mn)and long short-term memory (lstm)network. Physica D: Nonlinear Phenomena, 404: 132306, 2020.
[8] Vipul K Satone, Rachneet Kaur, Anant Dadu, Hampton Leonard, Hirotaka Iwaki, Mary Makarious, Lana Sargent, Alzheimer's Disease Neuroimaging Initiative, Ali Daneshmand, Sonja W Scholz, et al. Predicting alzheimer's disease progression trajectory and clinical subtypes using machine learning. bioRxiv, page 792432, 2019.
[9] Merel Kuijs, Catherine R Jutzeler, Bastian Rieck, and Sarah C Brüningk. Interpretability aware model training to improve robustness against out-of-distribution magnetic resonance images in alzheimer's disease classification. arXiv preprint arXiv: 2111. 08701, 2021.
[10] Mengying Xiao, Liyuan Zhang, Weili Shi, Jianhua Liu, Wei He, and Zhengang Jiang. A visualization method based on the grad-cam for medical image segmentation model. In 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS), pages 242-247. IEEE, 2021.
[11] Nourah Alangari, Mohamed El Bachir Menai, Hassan Mathkour, and Ibrahim Almosallam. Intrinsically interpretable gaussian mixture model. Information, 14(3): 164, 2023.
[12] Newton Mwai Kinyanjui and Fredrik D Johansson. Adcb: An alzheimer's disease benchmark for evaluating observational estimators of causal effects. arXiv preprint arXiv: 2111. 06811, 2021.
[13] Rachel Draelos. Grad-cam: Visual explanations from deep networks, May 2020.
[14] Shihong Yue, Ping Li, and Peiyi Hao. Svm classification: Its contents and challenges. Applied Mathematics-A Joumal of Chinese Universities, 18: 332-342, 2003.
[15] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. R package version 0. 4-2, 1(4): 1-4, 2015.
[16] Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S Awwal, and Vijayan K Asari. The history began from alexnet: A comprehensive survey on deep learning approaches. arXiv preprint arXiv: 1803. 01164, 2018.
[17] Srikanth Tammina. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. International Joumal of Scientific and Research Publications(IJSRP), 9(10): 143-150, 2019.
[18] Aniruddha Bhandari. Guide to auc roc curve in machine leaming: What is specificity?, Jun 2020.
[19] Alzheimer’s disease facts and figures 2023. 19(4): 1598-1695, Mar 2023.