International Journal of Neural Systems, Vol. 0, No. 0 (2022) 1–15

ENHANCING MULTIMODAL PATTERNS IN NEUROIMAGING BY SIAMESE NEURAL NETWORKS WITH SELF-ATTENTION MECHANISM

JUAN E ARCO[1,2,3], ANDRÉS ORTIZ[2,3], JUAN M GÓRRIZ[1,3], JAVIER RAMÍREZ[1,3]

[1] *Department of Signal Theory, Networking and Communications, University of Granada, 18010 Spain*

[2] *Department of Communications Engineering, University of Malaga, 29010 Spain*

[3] *Andalusian Research Institute in Data Science and Computational Intelligence, Spain*

E-mail: jearco@ugr.es

The combination of different sources of information is currently one of the most relevant aspects in the diagnostic process of several diseases. In the field of neurological disorders, different imaging modalities providing structural and functional information are frequently available. Those modalities are usually analyzed separately, although a joint of the features extracted from both sources can improve the classification performance of Computer-aided diagnosis (CAD) tools. Previous studies have computed independent models from each individual modality and combined then in a subsequent stage, which is not an optimum solution. In this work, we propose a method based on the principles of siamese neural networks to fuse information from Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET). This framework quantifies the similarities between both modalities and relates them with the diagnostic label during the training process. The resulting latent space at the output of this network is then entered into an attention module in order to evaluate the relevance of each brain region and modality at different stages of the development of Alzheimer's disease. The excellent results obtained and the high flexibility of the method proposed allows fusing more than two modalities, leading to a scalable methodology that can be used in a wide range of contexts.

Keywords: Multimodal combination; siamese neural network; self-attention; Deep learning; medical imaging.

1. Introduction

Current neuroimaging techniques provide very useful information regarding the structure or functional state of the brain. Thus, features extracted from those modalities are usually exploited in Computer Aided Diagnosis (CAD) tools, since they figure out discriminative information which is highly valuable to diagnose different neurological and neurodegenerative diseases. Both image modalities provide different but complementary information: while structural Magnetic Resonance Imaging (MRI) informs about the distribution of the different tissues in the brain (mainly gray and white matter), functional imaging such as Positron Emission Tomography (PET) inform about changes in cerebral blood flow and in cerebral glucose metabolism. This is an indicator of neuronal activity that can be figured out by means of different radiotracers. For instance, 18F-FDG-PET imaging have been extensively used for the diagnosis of neurodegenerative pathologies such as Alzheimer disease (AD).[1–5] In a similar way, changes in different brain structures related to the progression of a

neurodegenerative proccess can be revealed by structural MRI, and further used in differential diagnosis tasks.[2,6–10] These works use GM or WM images obtained by segmentation of MRI to classify controls and AD patients[9,10] or to compute Regions of Interest (ROI), searching for common patterns in Controls (CTL) and AD subjects. Moreover, the construction of neurodegeneration models to study the progression of the disease usually requires the use of both, functional and structural information. This is generally addressed by combining features computed from functional and structural images.[2,7]

Deep Learning (DL) architectures are particularly effective in these scenarios, given their ability to learn hierarchical representations from the input data. One crucial aspect is how information from different sources is combined, but it is also highly important when features from different modalities are fused. In early fusion, data from different sources are concatenated in the first stage of the processing pipeline. This means that the DL model treats them as if they belong to the same modality, eliminating any identification about their origins.[11–13] In intermediate fusion, the concatenation of the features is not done in the input of the network, but in a middle layer of the DL architecture. For example, data from one modality is processed by a fully convolutional layer, combining the output of this layer with the input data from another modality. This strategy is particularly flexible, allowing the fusion of the second modality once the first one has been conveniently processed.[14–16] Finally, in late fusion frameworks different modalities are combined at the end of the classification stage.[17–19] Specifically, one model is trained individually by each modality of the data, resulting in a number of models equal to the different modalities. Then, the classification decisions are combined according to a specific rule. Although the use of these fusion techniques can improve performance compared to using a single modality, they probably do not take advantage of the complementarity of the different modalities. In fact, these methods have demonstrated their effectiveness in differential diagnosis tasks, but their use in exploratory analysis is limited since it is difficult to link structural and functional features. Besides, they rely on the combination of different models (one for each modality) instead of fusing information from different modalities into the same model, which is a much more optimum solution.

To address this issue, we propose a classification framework formed by a siamese neural network and an attention module in order to combine MRI and PET imaging. First, the siamese architecture is employed to fuse structural and functional images by quantifying the similarity between them. The two modalities are independently entered into the network, but its training is simultaneously performed with data from both sources. The basis of this architecture is to relate the distance between the two individual inputs (structural and functional information) and the diagnostic label. The resulting latent space obtained at the output of the siamese network is then entered into an attention module that evaluates the relevance of each brain region and image modality. The proposed methodology has been evaluated using images from the Alzheimer Disease Neuroimaging Initiative (ADNI)[20] in order to explore structural and functional changes during the development of Alzheimer's disease (AD). The rest of the paper is organized as follows. Section 2 provides a detailed description of previous works that have employed data fusion and siamese neural networks. Section 3 describes the methodology proposed in this work, especially focused in the architecture for combining different imaging modalities. The discriminative power of our method is evaluated in Section 5, a complete discussion is presented in Section 6, whereas Section 7 summarizes future works and conclusions.

2. Related works

Recent developments of intelligent systems have demonstrated that fusion of different modalities leads to a boost in classification performance. This allows not only a more accurate classification system, but a better understanding about the relevance of each individual source. It is worth noting that this process is particularly effective when the information provided by one source complements the information extracted from a different one. Although the way the different modalities are combined is crucial for obtaining a robust and accurate system, this task is not particularly simple. A high number of studies have developed frameworks for combining information from different modalities by using machine learning[21–24] or deep learning.[25–28] Fusion of multimodal data has been successfully used both for segmentation and classification purposes,

especially when applied to medical imaging. Ref.[29] developed a method based on a recurrent framework for tumor segmentation based on PET and Computed Tomography (CT) images. Specifically, they combined features from these modalities with the intermediary segmentation results, which was previously estimated from multiple recurrent fusion phases. Their results demonstrated the generalizability of the method, and its flexibility to be applied in scenarios with limited computational resources. Ref.[30] introduced a method based on the features learned by a 2D convolutional autoencoder to create a 3D network able to segment spatial and volumetric information in a more efficient way. According to their findings, this method obtained a superior performance than common architectures (3D-UNet, 3D-MultiResUNet) while guaranteeing a much more reduced computational cost. Information contained in different sources has also been fused for classification tasks. Ref.[31] developed a Searchlight analysis that explored structural MRI in order to identify the brain region that leads to the maximum separability between the different classes. Briefly, they combined the voxels contained in small spherical regions with scores from psychological tests to determine the regions more affected by AD. In a similar context, Ref.[32] proposed an ensemble framework to combine information from MR images of different sessions of the longitudinal study. The contribution of each individual source to the global classification was estimated within a nested cross-validation scheme, which means that weights were derived from the scenario of maximum performance.

Other previous works have introduced the reliability of a classification prediction for weighting each modality within an ensemble architecture. Ref.[33] employed a Bayesian deep learning approach for maximizing performance while quantifying the uncertainty of a model's prediction. Specifically, they replaced the deterministic weights along the neural network by a distribution over these parameters. The informativeness of different regions of an image and their influence in a classification have also been evaluated by a similar approach. Ref.[34] used a probabilistic version of a Support Vector Machines (SVM) for providing information about the uncertainty of the classification. The weight of each individual was not derived from the performance of the classifier, but from its reliability, leading to a system in which reliable predictions contribute more than those with a higher uncertainty.

Regarding the siamese networks, they were introduced in the 1990s within a signature verification system.[35] The idea behind these architectures is to check the similarity between two different samples, which is measured according to a specific distance (Euclidean, cosine, etc). One of the most clear applications of this technique is to quantify differences based on distance metrics. Ref.[36] focused on the problem of content-based retrieval in audio signals. The siamese neural network was employed for encoding the audio into a representation of lower dimensionality. They showed that the output of the siamese architecture extracted the semantic information associated with each individual event, leading to an effective tool for retrieving semantically similar instances. Another example of the applicability of siamese networks in audio can be found in Ref.[37] Authors developed a framework based on two convolutional neural networks (CNN) to extract features from two different samples: an original sound and an imitation. They proposed a semi-siamese alternative in which the two encoders were asymmetric and previously trained in other similar tasks like speech recognition. Results showed that the inclusion of transfer learning within the siamese framework significantly improved the performance of the system.

These architectures have also been evaluated in biological works, such as identification of chromosomes,[38] characterization of cellular heterogeneity,[39] metagenome interpretation[40] or drug response.[41] However, image processing is one of the fields where siamese networks are more widely used. Ref.[42] explored a method in a one-shot learning context, where a unique exemplar of each possible class was available. The high performance obtained in a 20-way one-shot classification demonstrated the powerful discriminative ability of the siamese networks, being able to generalize even with new classes whose distributions were unknown. In a similar context, Ref.[43] developed a siamese neural network for the identification of heavy mineral images. The main contribution of this work was the inclusion of an adversarial training to discard domain-related information, which allowed the identification of the target features in unseen scenarios. Ref.[44] designed a regularized siamese neural network in a context of

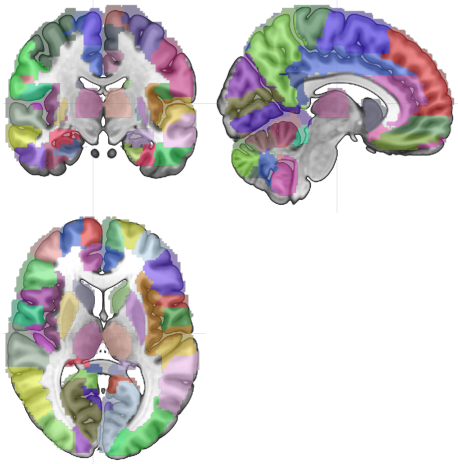

Figure 1. Brain parcellations provided by the AAL atlas.

anomaly detection, i.e. when the number of samples of a specific context is much lower than others. Specifically, authors proposed an architecture formed by stacked convolutional autoencoders for feature extraction and unsupervised deep siamese networks for learning the representational space generated by the distance between two samples. The resulting latent representations were then entered into a one-class SVM to detect subtle lesions in epileptic patients, outperforming in terms of sensitivity the most relevant studies in this field. Another medical application of siamese networks is given in Ref.[45] This work presented a model for the detection of malaria parasites from microscopic images. First, features from images of different classes (infected cell *vs* uninfected cell) were extracted by using a fully connected convolutional block. Then, the similarity of the resulting features associated with each initial image was evaluated in the siamese block. Finally, the output was feed forwarded to the last linear layer, assigning the corresponding label based on the activation of the neurons of that final layer.

3. Methodology

3.1. *Brain parcellation*

The method proposed in this work relies on measuring differences between structural and functional information in the brain. Although severe diseases can cause damage across the whole brain, it is likely that the level of the atrophy varies for different regions. For this reason, it is crucial to detect the presence of abnormalities, but even more important to identify where they are. One possible solution is to employ an atlas in order to delimitate the different anatomical brain regions. An important point is to choose a proper atlas, since there is a high number of them differing in the complexity and detail of the parcellations. Dividing the brain into a high number of regions improves spatial precision, but reduces the probability of identifying relevant patterns.[46] On the other hand, the use of too large regions could lead to mark as informative voxels that they are not, in case only a small part of the region is affected by a specific pathology. In order to provide a sensitive method while controlling Type I errors, we employed the 116 regions contained in the Automated Anatomical Labeling (AAL[47]) atlas since it strikes a balance between the number of regions and their size. Figure 1 depicts the brain subdivisions contained in this atlas.

3.2. *Siamese architecture*

Each region defined in the AAL atlas is iteratively extracted both in the MRI and PET images. Regarding the first one, we only focused on gray matter (GM) because atrophy in this tissue has been proven to be related to AD.[48–51] From the original MRI and PET images, all the regions except the target one are automatically discarded by setting the values of the voxels within these regions to zero. Thus, the resulting images contain only informative values in a small region, corresponding to the one that has not set to

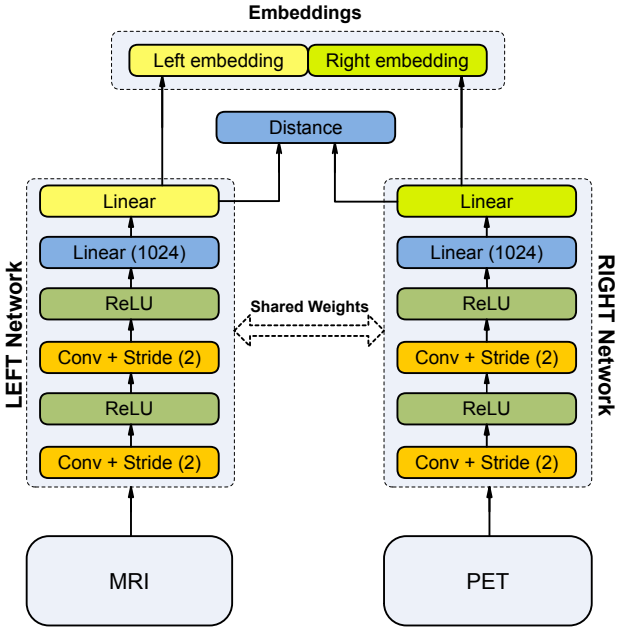

Figure 2. Both branches have exactly the same configuration, sharing the weights that are updated during the training process. The main difference is that the left branch receives as input the MRI images, whereas the right one receives the PET images.

zero. To reduce the computational burden associated with the processing of this non-relevant information, we cropped the images by automatically selecting a rectangular area that only contains the voxels of the target region. This considerably reduces the size of the images and the subsequent mathematical operations. After that, all the images of the database are standardized by removing the mean and scaling to unit variance, as follows:

$$\mathbf{z} = \frac{\mathbf{x} - \mu}{\sigma} \tag{1}$$

where $\mathbf{x}$ is a matrix containing the target region of all the patients in the dataset, μ and σ are the mean and standard deviation of this matrix, respectively, and $\mathbf{z}$ denotes the resulting scores.

At this point, information from structural and functional images is available to be fused. To do so, we developed a neural network based on the principles of siamese architectures.[35,42,52] This framework relies on the use of a convolutional neural network with two branches sharing the same architecture and weights. Each individual branch receives as input one of the modalities to be fused: GM MRI images in the left branch and PET images in the right one. During the training, each branch processes the information as a common feed-forward network: each neuron of a specific layer receives an input, processes it and sends the output to the next layer. Given the shared nature of this architecture, weights of both branches are simultaneously updated. The structure of the siamese network is shown in Figure 2, which is based on convolutional and linear layers in each branch. Specifically, we employed two convolutional layers whose outputs were modified according to a ReLU activation function. After that, the outputs were entered into two linear layers containing 1024 and 20 neurons, respectively. The training process was guided by a loss function that evaluated the similarity of the outputs of the last linear layers of both branches. The mathematical expression for computing the distance measure was based on the Hinge function,[53] as follows:

$$L(y) = max(0, 1 - t \cdot y) \tag{2}$$

where y refers to the outputs of the linear layers of the two branches and $t = \{-1, 1\}$ denotes the actual label.

The network was trained during 200 epochs, in-

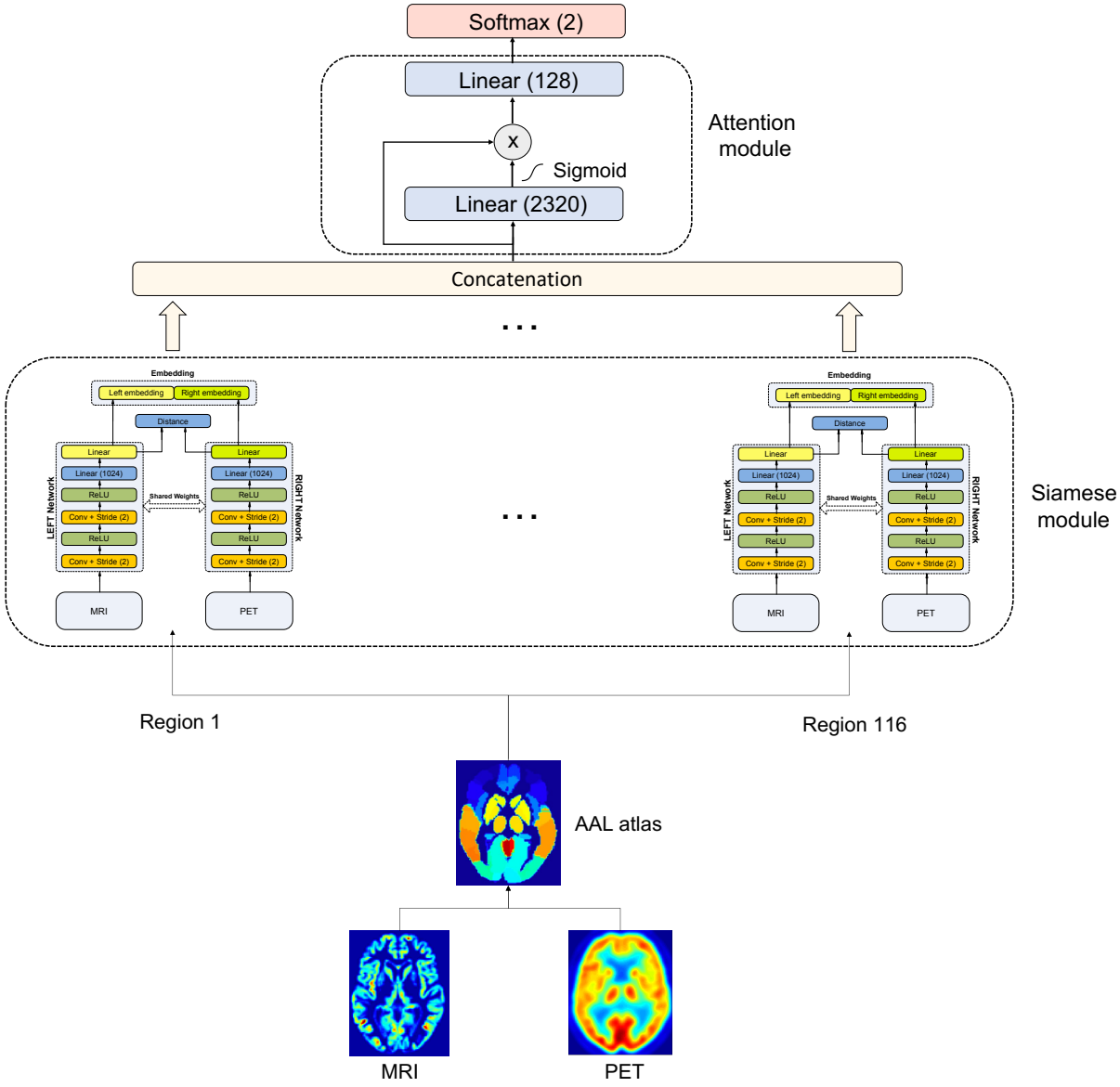

Figure 3. Scheme of the method proposed. Brain parcellations are extracted from MRI and PET images. Then, each individual region is entered into a siamese network which is trained with the aim of finding a relationship between the class of each sample and the distance between structural and functional information. The resulting embeddings are used as inputs of an attention module that evaluated the importance of each individual region and image modality into the final classification decision.

cluding an early stopping to finish this procedure when the validation loss was lower than $l = 0.01$. Then, the outputs of both branches were concatenated into an embedding that can be interpreted as a representation of lower dimensionality of the similarity between the MRI and PET images of the brain region. This initial network worked as a feature extractor, identifying the most relevant aspects that characterize the combination of both image modalities.

3.3. *Attention module*

The resulting embeddings for each region of the AAL atlas were then combined to have a global fusion of the structural and functional information of the brain. These features were used as input of a classi-

Table 1. Demographics of the subjects of the database according to their diagnosis: Alzheimer's disease (AD, Mild cognitive impairment converter (MCIc) or stable (MCIs), and normal controls (CTL).

Diagnosis	Number	Gender (M/F)	Age	MMSE
AD	70	46/24	75.26 ±7.53	22.49 ±2.91
MCIc	39	25/14	77.77 ±7.41	26.00 ±2.97
MCIs	64	42/22	76.49 ±6.85	27.18 ±2.53
CTL	68	43/25	75.93 ±4.98	28.98 ±0.98

fication block that allowed the automatic distinction between the different classes. Given the flexibility of the model proposed, this block could be based on a simple linear classifier receiving as input features the embeddings of the different regions of the atlas. Although performance would be possibly excellent with this simple solution, we designed an attention module in order to refine the embeddings assuring an optimum classification solution. This block weights the contribution of each brain region and modality in the final classification task. Attention modules are commonly used to force a CNN to dismiss non-relevant information and focus only on the most important one, which leads to a boost in the discriminative power.[54] They are also beneficial for preventing gradients from non-informative regions of the images.[55] The attention module proposed in this work consists on an MLP and a sigmoid function at the end to generate a mask of the input feature map. Specifically, the embeddings computed by the siamese network were then entered into an MLP. Then, its output was multiplied by the embeddings before entering the result into a final linear layer. The architecture of the attention module, as well as its connections to the siamese block is shown in Figure 3.

4. Evaluation of Alzheimer's disease progression

The performance of the proposed method is evaluated by studying differences between structural and functional damage of the brain, with the aim of identifying changes associated with the different phases in the development of Alzheimer's disease. Specifically, we evaluated our proposal with 18F-FDG PET and MRI images, which provide functional and structural information, respectively. The following subsections contain information about the source of the database employed, as well as the demographics of the subjects included in the study and the preprocessing applied to each individual image modality.

4.1. *Database description*

The data used in the preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55-90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2, and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-

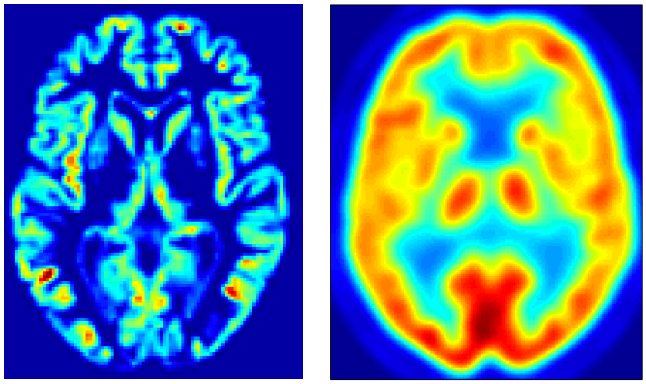

Figure 4. Gray Matter structural image and PET functional image for a control subject.

date information, see www.adni-info.org.

From the high amount of information contained in the ADNI database, we focused on the analysis of PET and MRI images. Given the longitudinal nature of the study, MRI scans were acquired at different sessions. However, PET images were not collected in all of these moments. For this reason, we selected only those patients with at least one image of each modality. Thus, the dataset used comprises data from 251 patients, consisting on 68 normal controls (CTL), 70 AD patients, 39 suffering from Mild Cognitive Impairment that converts at certain session to AD (MCIc) and 64 diagnosed with MCI that remains stable along the different sessions (MCIs). Table 1 summarizes demographics of the dataset in terms of age, gender and Mini Mental State Examination (MMSE) scores. Finally, Figure 4 shows an example of the MRI and PET images of a control patient.

4.2. *Images preprocessing*

4.2.1. *MRI images*

Images were firstly registered following a spatial transformation based on exponential Lie algebra.[56] The idea behind this process is to assure the structural correspondence between all the voxels across the images, so that the location of one specific voxel is the same in images of different subjects. After that, the images were resized to 121 x 145 x 121, with a voxel size of 1.5 mm in the sagittal, coronal and axial planes. Then, we performed a segmentation of the images into gray (GM) and white (WM) tissues, using the algorithms available in SPM12.[57] This tool provides the distribution of the intensities of the voxels of the T1-weighted MRI according to the tissue probability maps, containing values ranging from 0 to 1 as they reflect the probability that a specific voxel belongs to GM, WM or cerebrospinal fluid (CSF). These maps provided by the International Consortium for Brain Mapping (ICBM) are derived from 452 T1-weighted scans, which were aligned with an atlas space and corrected for scan inhomogeneities. Finally, a non-linear deformation field was computed to find the one that best fits the tissue probability maps of each patient.

4.2.2. *PET images*

We also employed SPM12 in this kind of images to spatially normalize them according to a specific template. Then, a normalization in intensity was applied to each individual image to guarantee that comparisons between them were properly done. Briefly, the normalization value was computed as the average of the 0.1% voxels with the highest intensities.[58] Besides, voxels whose intensity was lower than the 10% of the normalization value were automatically discarded for subsequent analysis. Specifically, they were considered as background, which means that they do not contain relevant information but can introduce different artifacts and noise.[59]

4.3. *Experimental setup*

The *small sample size problem* is widely present in most biomedical studies, and it is caused by the limited number of samples that this kind of works usually have. For this reason, it is necessary to assure that samples employed for training the model are independent from those used to test its perfor-

mance, guaranteeing the generalization ability of our method. Thus, we employed a resampling approach based on *k-fold* cross-validation ($k = 5$) to estimate the prediction error of the method proposed.

5. Results

The first classification context using the methodology described in previous sections consists on the classification between Controls (CTL) and AD patients. In this scenario, our method led to an accuracy of 96%. Additionally, classification experiments between CTL and MCIc were carried out, yielding 94% of accuracy. Finally, we also evaluated the performance of the proposed methodology when distinguishing between CTL and MCIs. In this case, we obtained an accuracy of 86%. Table 2 summarizes the results of the different experiments based on additional performance metrics. Figure 5 depicts the ROC curve for the three classification experiments carried out in this work. The ROC curve shows the trade-off between sensitivity and specificity, computed using the probability predictions derived from the neurons of the output layer of the DL architecture. In addition to the ROC plot, we included the area under the ROC curve (AUC) for the different experiments as a measure of the discrimination ability between two diagnostic groups. We can see that our method yielded a high performance for all classification contexts, but they are gradually ordered according to their difficulty.

These results demonstrate that the information fusion provided by our method is extremely beneficial for boosting the classification performance. However, our proposal provides additional information in terms of explainability that is relevant in the study of AD. Figure 6 shows a representation of the embeddings after the sigmoid activation of one subject according to his/her structural and functional information, as well as a combination of both modalities. These maps were generated by computing, for each patient, the number of features from the embeddings that were important during the evaluation of similarity between structural and functional images. Specifically, we counted all the features associated with each brain region and modality that surpassed a threshold of $thr = 0.9$, as a measure of the *a priori* relevance of the region/modality in the fusion of both modalities. We can see in Figure 6 that there is a subtle discrepancy between the importance of regions in structural images (top of the figure) and functional images (middle of the image). The reason for the emergence of these differences will be clearly explained in next section.

Figure 7 depicts the weights of the neurons at the output of the first fully connected layer of the attention module. Thus, they represent the importance of each brain region in the classification decision when distinguishing between controls and AD patients. This map demonstrates that our method is able to identify the localization of the most relevant structural-functional differences, i.e. those dissimilarities between both modalities that are relevant in the classification outcome. Besides, it is important to note the ability to identify the role of different brain regions in the development of AD, evidencing changes in the anatomy (such as atrophy) and functionality according the progression of this pathology. We will discuss about some of these regions in Section 6.

6. Discussion

In this work, we present a method for fusing structural and functional information from brain imaging based on a siamese architecture. This approach evaluates the similarity between both modalities according to a specific distance measure, which is then related to the diagnostic label of both samples. Once the network is trained, the latent space obtained as the output of the last linear layer contains the embeddings associated with the different classes. They are finally used to train an attention module that performs the classification task. The generalization ability of our method was evaluated in MRI and PET imaging from Alzheimer's disease patients at different stages of this pathology. Specifically, we focused on studying differences between structural and functional brain damage in classification contexts of incremental difficulty, from the simplest scenario where evaluating normal controls *vs* AD patients to a more complicated one in which differences between MCI and AD were studied.

The idea behind ensemble frameworks has been used in previous studies in order to decompose a difficult problem in multiple simpler ones. For example, these alternatives are employed to alleviate the computational burden associated with deep learning. Images are usually partitioned and processed instead of being processed as a whole since it would require

Table 2. Performance obtained by our method in terms of balanced accuracy, sensitivity, specificity, precision, F1-score and area under the ROC curve, respectively.

Controls *vs* Alzheimer's					
Bal Acc	Sens	Spec	Prec	F1-score	AUC
0.96±0.03	0.99±0.01	0.93±0.05	0.94±0.04	0.97±0.04	0.98±0.01
Controls *vs* MCI converters					
Bal Acc	Sens	Spec	Prec	F1-score	AUC
0.94±0.02	0.97±0.05	0.93±0.05	0.92±0.05	0.94±0.02	0.93±0.03
Controls *vs* MCI stable					
Bal Acc	Sens	Spec	Prec	F1-score	AUC
0.86±0.04	0.94±0.06	0.91±0.05	0.90±0.06	0.92 ±0.05	0.88±0.02

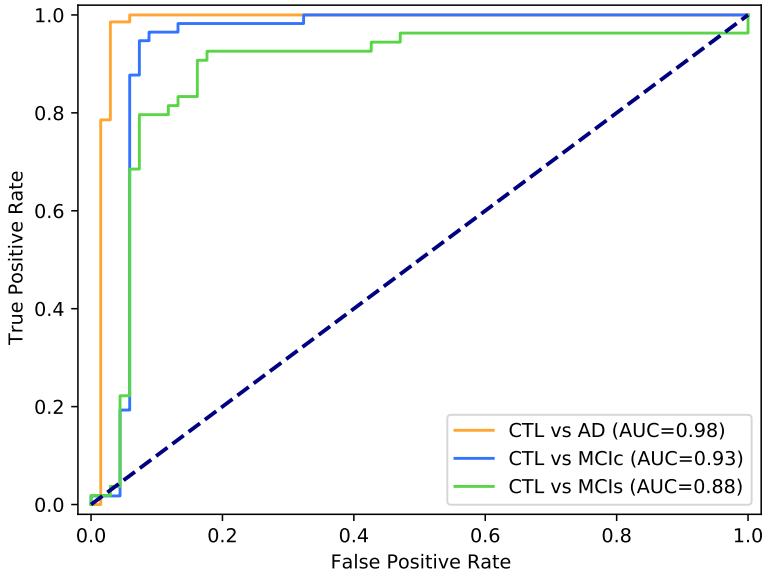

Figure 5. ROC curves obtained in the different classification scenarios.

high memory and computational resources that are not always met. Ensemble architectures have also been used for combining information from different modalities.[2,7] In these contexts, the aim is to extract information from different sources that can be relevant for classification. However, most of these works do not rely on an integrative model that simultaneously identifies informative patterns from different sources. On the contrary, they build one model for each modality, compute the relevance of each individual model and combine them according to their relevance. Although these alternatives can lead to a high performance, the interpretability of their results must be done carefully. Specifically, they provide information about how relevant a specific modality is according to its weight in the ensemble scheme. The higher the weight, the higher the importance, and viceversa. In the study of the development of AD, these techniques would allow to measure the importance of structural and functional damage in the prediction of the progression of this pathology. However, they would not inform about the relationship between both modalities, which could be crucial for a better understanding of this disease.

The method proposed in this work efficiently relates functional and structural information by computing the similarity between them. Thus, our approach exploits the complementarity between the two modalities of medical imaging to detect abnormalities in the brain. In healthy subjects, the basal

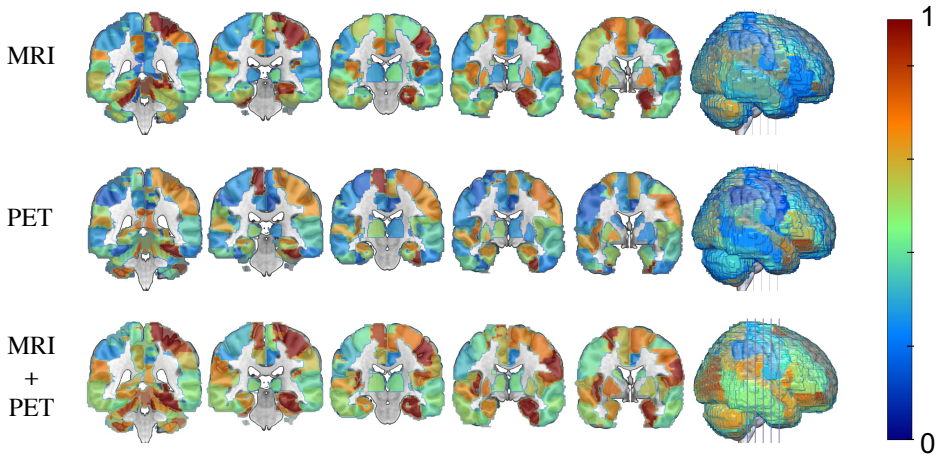

Figure 6. Example of the activation maps for different brain regions for structural and functional images, and after the fusion process in the CTL *vs* AD classification context.

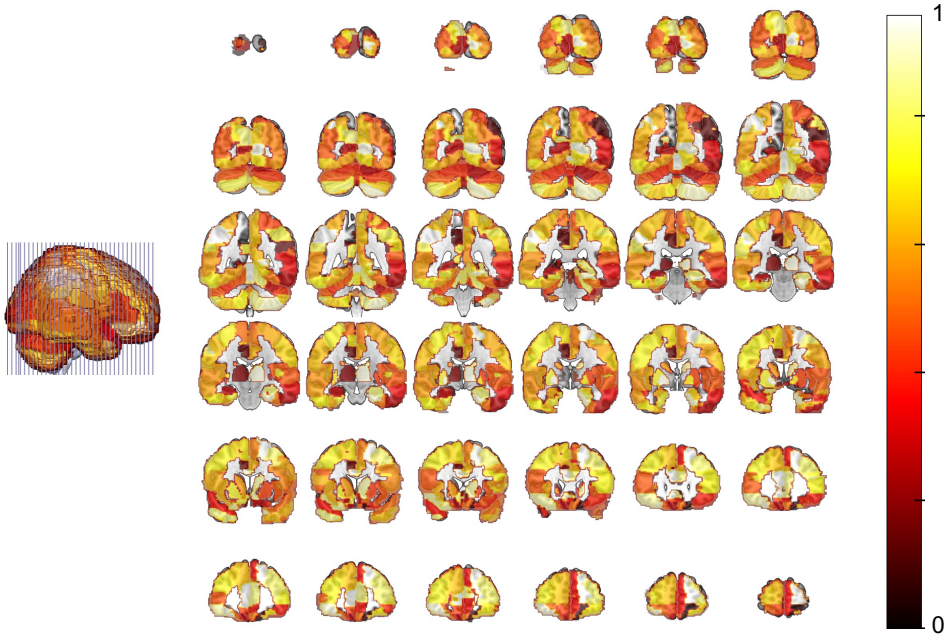

Figure 7. Weights of the neurons at the output of the first linear layer of the attention module in the CTL *vs* AD classification scenario. These values represent the relevance of each individual region in the classification outcome.

state of the brain would present a minimum atrophy, in addition to a correct functionality. This leads to a certain "distance", a specific measure that relates these two states. However, it is likely that patients with an incipient cognitive decline would present structural or functional changes in their brain. In this case, our method would detect a deviation in the relationship between the two imaging modalities. Moreover, people diagnosed with Alzheimer's disease will present a much higher atrophy in the brain, as well as a loss of brain function. Thus, it would be even easier for our tool to carry out a correct diagnosis. Nevertheless, there is not always a direct relationship between structural and functional damage in a

brain region, as Figure 6 shows. It is possible that atrophy is present in a brain region, but the functionality associated with this region is still preserved. Otherwise, it would not be necessary the acquisition of anatomical and functional imaging because they would provide exactly the same information. For this reason, we would like to highlight the importance of the contribution of this work, presenting a method that allows the intra-model combination of two imaging modalities by quantifying their differences.

It is important to note that our method evaluates the structural-functional relationship for individual brain regions. From the technical standpoint, this mitigates the curse of dimensionality problem that appears when the number of features (voxels in 3D images) is much higher than the number of samples (images in the database). From the clinicians point of view, it is quite interesting to predict the outcome of a patient in order to select a proper treatment that delay the brain damage caused by AD. Besides, it is highly important to identify the brain regions where abnormalities caused by AD are present. Our method shows a high sensitivity in the detection of these regions either for abnormalities in their structure or by a loss in their functionality. Specifically, we clearly identified alterations in the precuneus, hyppocampus, thalamus or the orbitofrontal cortex, regions with a crucial role in the development of AD.[60–63] This allows a better understanding of the affection and an early identification of the neural functions that will affect the patient in the future, leading to a tool for personalized medicine to improve the patient's health. Finally, the high performance obtained by our framework evidences its suitability in the identification of abnormal brain patterns. Results reveal the usefulness of our method not only for detecting AD, but for expanding our knowledge about the development of this disorder.

7. Conclusions and Future Work

In this work, we propose a method based on deep learning to combine medical images from different modalities. The method used relies on a siamese neural network, an architecture which computes differences between structural and functional data from the brain. This combined information is then entered into an attention module employed to identify the relevance of each brain region and modality in the development of Alzheimer's disease. The high performance shown by our method manifests its suitability for combining complementary information from different modalities. Besides, it paves the way for different applications not only in the medical field, but in imaging processing. Alternatively, the high flexibility of siamese architectures allows fusing more than two modalities by the inclusion of an additional branch for each new data type, leading to a scalable methodology that can be used in a wide range of scenarios.

8. Funding

This work was supported by projects PGC2018-098813-B-C32 and RTI2018-098913-B100 (Spanish "Ministerio de Ciencia, Innovación y Universidades"), UMA20-FEDERJA-086, A-TIC-080-UGR18 and P20 00525 (Consejería de economía y conocimiento, Junta de Andalucía) and by European Regional Development Funds (ERDF); and by Spanish "Ministerio de Universidades" through Margarita-Salas grant to J.E. Arco.

Bibliography

1. G. Chételat, J. Arbizu, H. Barthel, V. Garibotto, I. Law, S. Morbelli, E. van de Giessen, F. Agosta, F. Barkhof, D. J. Brooks, M. C. Carrillo, B. Dubois, A. M. Fjell, G. B. Frisoni, O. Hansson, K. Herholz, B. F. Hutton, C. R. Jack, A. A. Lammertsma, S. M. Landau, S. Minoshima, F. Nobili, A. Nordberg, R. Ossenkoppele, W. J. G. Oyen, D. Perani, G. D. Rabinovici, P. Scheltens, V. L. Villemagne, H. Zetterberg, and A. Drzezga, "Amyloid-pet and 18f-fdg-pet in the diagnostic investigation of alzheimer's disease and other dementias," *The Lancet Neurology*, vol. 19, no. 11, pp. 951–962, 2020.
2. A. Ortiz, J. Munilla, J. M. Górriz, and J. Ramírez, "Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease," *International Journal of Neural Systems*, vol. 26, no. 07, p. 1650025, 2016.
3. F. J Martinez-Murcia, J. M Górriz, J. Ramírez, A. Ortiz, s Disease Neuroimaging Initiative *et al.*, "A spherical brain mapping of mr images for the detection of alzheimer's disease," *Current Alzheimer Research*, vol. 13, no. 5, pp. 575–588, 2016.
4. D. C. Alsop, M. Casement, C. de Bazelaire, T. Fong, and D. Z. Press, "Hippocampal hyperperfusion in Alzheimer's disease," *Neuroimage*, vol. 42, no. 4, pp. 1267–1274, 2008.
5. F. J. Martínez-Murcia, A. Ortiz, J. Górriz, J. Ramírez, and D. Castillo-Barnes, "Studying the manifold structure of Alzheimer's disease: A deep learning approach using convolutional autoen-

coders," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 17–26, 2020.
6. A. Ortiz, J. Gorriz, J. Ramírez, and F. Martínez-Murcia, "Automatic ROI selection in structural brain mri using som 3D projection," *PloS one*, vol. 9, p. e93851, 04 2014.
7. A. Ortiz, J. Munilla, I. Álvarez Illán, J. M. Górriz, J. Ramírez, and A. D. N. I. , "Exploratory graphical models of functional and structural connectivity patterns for alzheimer's disease diagnosis," *Frontiers in Computational Neuroscience*, vol. 9, p. 132, 2015.
8. A. Ortiz, J. M. Górriz, J. Ramírez, and F. J. Martínez-Murcia, "Lvq-SVM based CAD tool applied to structural MRI for the diagnosis of the alzheimer's disease," *Pattern Recognition Letters*, vol. 34, no. 14, pp. 1725–1733, Oct. 2013.
9. D. Chyzhyk, M. Graña, A. Savio, and J. Maiora, "Hybrid dendritic computing with kernel-lica applied to Alzheimer's disease detection in MRI," *Neurocomputing*, vol. 75, no. 1, p. 72–77, Jan. 2012.
10. R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M. Habert, M. Chupin, H. Benali, O. Colliot, and Alzheimer's Disease Neuroimaging Initiative, "Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database," *Neuroimage*, vol. 56(2), pp. 766–781, 2010.
11. S. El-Sappagh, H. Saleh, R. Sahal, T. Abuhmed, S. R. Islam, F. Ali, and E. Amer, "Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data," *Future Generation Computer Systems*, vol. 115, pp. 680–699, 2021.
12. V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Multi-level sensor fusion with deep learning," *IEEE Sensors Letters*, vol. PP, pp. 1–1, 10 2018.
13. K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets," *Vis. Comput.*, vol. 38, no. 8, p. 2939–2970, 2022.
14. W. Alahamade, I. Lake, C. E. Reeves, and B. De La Iglesia, "A multi-variate time series clustering approach based on intermediate fusion: A case study in air pollution data imputation," *Neurocomputing*, vol. 490, pp. 229–245, 2022.
15. S. Srivastava and S. Sadistap, "Data processing approaches and strategies for non-destructive fruits quality inspection and authentication: a review," *Journal of Food Measurement and Characterization*, vol. 12, pp. 2758–2794, 2018.
16. S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: a review," *Briefings in Bioinformatics*, vol. 23, no. 2, 01 2022.
17. B. Ding, T. Zhang, G. Liu, L. Kong, and Y. Geng, "Late fusion for acoustic scene classification using swarm intelligence," *Applied Acoustics*, vol. 192, p. 108698, 2022.
18. M. Hassan, S. Ali, H. Alquhayz, J. Y. Kim, and M. Sanaullah, "Developing liver cancer drug response prediction system using late fusion of reduced deep features," *Journal of King Saud University - Computer and Information Sciences*, 2022.
19. B. W.-Y. Hsu and V. S. Tseng, "Hierarchy-aware contrastive learning with late fusion for skin lesion classification," *Computer Methods and Programs in Biomedicine*, vol. 216, p. 106666, 2022.
20. Alzheimer's Disease Neuroimaging Initiative, "Available: http://adni.loni.ucla.edu/. Accessed 2021 Nov 5," 2021.
21. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 423–443, 2019.
22. S. Qiu, H. Zhao, N. Jiang, Z. Wang, L. Liu, Y. An, H. Zhao, X. Miao, R. Liu, and G. Fortino, "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges," *Information Fusion*, vol. 80, pp. 241–265, 2022.
23. A. Saadallah, F. Finkeldey, J. Buß, K. Morik, P. Wiederkehr, and W. Rhode, "Simulation and sensor data fusion for machine learning application," *Advanced Engineering Informatics*, vol. 52, p. 101600, 2022.
24. H. Zhou, T. Ma, H. Rong, Y. Qian, Y. Tian, and N. Al-Nabhan, "Mdmn: Multi-task and domain adaptation based multi-modal network for early rumor detection," *Expert Systems with Applications*, vol. 195, p. 116517, 2022.
25. A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowledge-Based Systems*, vol. 244, p. 108580, 2022.
26. F. Farahnakian and J. Heikkonen, "Deep learning based multi-modal fusion architectures for maritime vessel detection," *Remote Sensing*, vol. 12, no. 16, 2020.
27. R. Guo, D. Li, and Y. Han, "Deep multi-scale and multi-modal fusion for 3d object detection," *Pattern Recognition Letters*, vol. 151, pp. 236–242, 2021.
28. R. Hou, G. Chen, Y. Han, Z. Tang, and Q. Ru, "Multi-modal feature fusion for 3d object detection in the production workshop," *Applied Soft Computing*, vol. 115, p. 108245, 2022.
29. L. Bi, M. Fulham, N. Li, Q. Liu, S. Song, D. Dagan Feng, and J. Kim, "Recurrent feature fusion learning for multi-modality pet-ct tumor segmentation," *Computer Methods and Programs in Biomedicine*, vol. 203, p. 106043, 2021.
30. S. Najeeb and M. I. H. Bhuiyan, "Spatial feature fusion in 3d convolutional autoencoders for lung tumor segmentation from 3d ct images," *Biomedical Signal Processing and Control*, vol. 78, p. 103996, 2022.
31. J. E. Arco, J. Ramírez, C. G. Puntonet, J. M.

Górriz, and M. Ruz, "Improving short-term prediction from MCI to AD by applying Searchlight analysis," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016, pp. 10–13.
32. J. E. Arco, J. Ramírez, J. M. Górriz, and M. Ruz, "Data fusion based on Searchlight analysis for the prediction of Alzheimer's disease," *Expert Systems with Applications*, vol. 185, p. 115549, 2021.
33. J. E. Arco, A. Ortiz, J. Ramírez, F. J. Martínez-Murcia,
Y.-D. Zhang, and J. M. Górriz, "Uncertainty-driven ensembles of multi-scale deep architectures for image classification," *Information Fusion*, vol. 89, pp. 53–65, 2023.
34. J. E. Arco, A. Ortiz, J. Ramírez, F. J. Martínez-Murcia, Y.-D. Zhang, J. Broncano, M. Álvaro Berbís, J. R. del Val, A. Luna, and J. M. Górriz, "Probabilistic combination of non-linear eigenprojections for ensemble classification," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 1, pp. 1–12, 2023.
35. J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," in *Proceedings of the 6th International Conference on Neural Information Processing Systems*, ser. NIPS'93. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993, p. 737–744.
36. P. Manocha, R. Badlani, A. Kumar, A. Shah, B. Elizalde, and B. Raj, "Content-based representations of audio using siamese neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 04 2018, pp. 1–5.
37. Y. Zhang, B. Pardo, and Z. Duan, "Siamese style convolutional neural networks for sound search by vocal imitation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 429–441, 2019.
38. Swati, G. Gupta, M. Yadav, M. Sharma, and L. Vig, "Siamese networks for chromosome classification," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 72–81.
39. B. Szubert, J. Cole, C. Monaco, and I. Drozdov, "Structure-preserving visualisation of high dimensional single-cell datasets," *Scientific Reports*, vol. 9, p. 8914, 06 2019.
40. S. Pan, C. Zhu, X. Zhao, and L. P. Coelho, "A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments," *Nature Communications*, vol. 13, p. 2326, 04 2022.
41. M. Jeon, D. Park, J. Lee, H. Jeon, M. Ko, S. Kim, Y. Choi, A.-C. Tan, and J. Kang, "ReSimNet: drug response similarity prediction using Siamese neural networks," *Bioinformatics*, vol. 35, no. 24, pp. 5249–5256, 05 2019.
42. G. R. Koch, "Siamese neural networks for one-shot image recognition," 2015.
43. H. Hao, Z. Jiang, S. Ge, C. Wang, and Q. Gu, "Siamese adversarial network for image classification of heavy mineral grains," *Computers & Geosciences*, vol. 159, p. 105016, 2022.
44. Z. Alaverdyan, J. Jung, R. Bouet, and C. Lartizien, "Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: Application to epilepsy lesion screening," *Medical Image Analysis*, vol. 60, p. 101618, 2020.
45. G. Madhu, B. Lalith Bharadwaj, B. Rohit, K. Sai Vardhan, S. Kautish, and P. N, "Chapter 12 - convolutional siamese networks for one-shot malaria parasite recognition in microscopic images," in *Demystifying Big Data, Machine Learning, and Deep Learning for Healthcare Analytics*, P. N, S. Kautish, and S.-L. Peng, Eds. Academic Press, 2021, pp. 277–306.
46. J. E. Arco, P. Díaz-Gutiérrez, J. Ramírez, and M. Ruz, "Atlas-based classification algorithms for identification of informative brain regions in fMRI," *Neuroinformatics*, vol. 18, pp. 219–236, 2019.
47. N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273 – 289, 2002.
48. P. M. Thompson, K. M. Hayashi, G. de Zubicaray, A. L. Janke, S. E. Rose, J. Semple, D. Herman, M. S. Hong, S. S. Dittmer, D. M. Doddrell, and A. W. Toga, "Dynamics of gray matter loss in Alzheimer's disease," *Journal of Neuroscience*, vol. 23, no. 3, pp. 994–1005, 2003.
49. Z. Wu, Y. Peng, M. Hong, and Y. Zhang, "Gray matter deterioration pattern during Alzheimer's disease progression: A regions-of-interest based surface morphometry study," *Frontiers in Aging Neuroscience*, vol. 13, 2021.
50. G. B. Frisoni, C. Testa, A. Zorzan, F. Sabattoli, A. Beltramello, H. Soininen, and M. P. Laakso, "Detection of grey matter loss in mild Alzheimer's disease with voxel based morphometry," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 73, no. 6, pp. 657–664, 2002. [Online]. Available: https://jnnp.bmj.com/content/73/6/657
51. P. Weston, I. Simpson, N. Ryan, S. Ourselin, and N. Fox, "Diffusion imaging changes in grey matter in Alzheimer's disease: A potential marker of early neurodegeneration," *Alzheimer's research & therapy*, vol. 7, p. 47, 07 2015.
52. D. Chicco, "Siamese neural networks: An overview," *Methods in molecular biology*, vol. 2190, pp. 73–94, 2021.
53. K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *J. Mach. Learn. Res.*, vol. 2, p. 265–292, 2002.

54. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.
55. Z. Jiang, Y. Wang, C. Shi, Y. Wu, R. Hu, S. Chen, S. Hu, X. Wang, and B. Qiu, "Attention module improves both performance and interpretability of four-dimensional functional magnetic resonance imaging decoding neural network," *Human Brain Mapping*, vol. 43, no. 8, pp. 2683–2692, 2022.
56. J. Ashburner and K. J. Friston, "Unified segmentation," *NeuroImage*, vol. 26, no. 3, pp. 839 – 851, 2005.
57. Wellcome Centre for Human Neuroimaging, "Statistical Parametrical Mapping," https://www.fil.ion.ucl.ac.uk/spm/software/spm12, 2018.
58. I. A. Illán, J. M. Górriz, J. Ramírez, D. Salas-Gonzalez, M. M. López, F. Segovia, R. Chaves, M. Gómez-Rio, and C. G. Puntonet, "18F-FDG PET imaging analysis for computer aided Alzheimer's diagnosis," *Inf. Sci.*, vol. 181, no. 4, p. 903–916, 2011.
59. A. Ortiz, F. Lozano, J. Gorriz, J. Ramirez, F. Martinez-Murcia, and A. D. N. Initiative, "Discriminative sparse features for Alzheimer's disease diagnosis using multimodal image data," *Current Alzheimer Research*, vol. 15, no. 1, pp. 67–79, 2018.
60. T. Yokoi, H. Watanabe, H. Yamaguchi, E. Bagarinao, M. Masuda, K. Imai, A. Ogura, R. Ohdake, K. Kawabata, K. Hara, Y. Riku, S. Ishigaki, M. Katsuno, S. Miyao, K. Kato, S. Naganawa, R. Harada, N. Okamura, K. Yanai, M. Yoshida, and G. Sobue, "Involvement of the precuneus/posterior cingulate cortex is significant for the development of alzheimer's disease: A pet (thk5351, pib) and resting fmri study," *Frontiers in Aging Neuroscience*, vol. 10, 2018.
61. Y. Rao, B. Ganaraja, B. Murlimanju, T. Joy, A. Krishnamurthy, and A. Agrawal, "Hippocampus and its involvement in alzheimer's disease: a review," *3 Biotech*, vol. 12, no. 2, p. 55, Feb. 2022, © The Author(s) 2022.
62. Y. Wu, X. Wu, L. Gao, Y. Yan, Z. Geng, S. Zhou, W. Zhu, Y. Tian, Y. Yu, L. Wei, and K. Wang, "Abnormal functional connectivity of thalamic subdivisions in alzheimer's disease: A functional magnetic resonance imaging study," *Neuroscience*, vol. 496, pp. 73–82, 2022.
63. G. W. V. Hoesen, J. Parvizi, and C. C. Chu, "Orbitofrontal cortex pathology in alzheimer's disease." *Cerebral cortex*, vol. 10 3, pp. 243–51, 2000.