frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 18 January 2023
DOI 10.3389/fnins.2022.1043626

Check for updates

OPEN ACCESS

EDITED BY
Enrico Capobianco,
Jackson Laboratory, United States

REVIEWED BY
Marco Diego Dominietto,
Paul Scherrer Institut (PSI), Switzerland
Medicharla Venkata Jagannadham,
University of Hyderabad, India

*CORRESPONDENCE
Jing Shen
✉ 15804257313@163.com

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 13 September 2022
ACCEPTED 12 December 2022
PUBLISHED 18 January 2023

CITATION
Sun X, Guo W and Shen J (2023) Toward attention-based learning to predict the risk of brain degeneration with multimodal medical data. *Front. Neurosci.* 16:1043626. doi: 10.3389/fnins.2022.1043626

Toward attention-based learning to predict the risk of brain degeneration with multimodal medical data

Xiaofei Sun[1], Weiwei Guo[2] and Jing Shen[3]*

[1]Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, Hong Kong SAR, China, [2]EchoX Technology Limited, Hong Kong, Hong Kong SAR, China, [3]Department of Radiology, Affiliated Zhongshan Hospital of Dalian University, Dalian, Liaoning, China

Introduction: Brain degeneration is commonly caused by some chronic diseases, such as Alzheimer's disease (AD) and diabetes mellitus (DM). The risk prediction of brain degeneration aims to forecast the situation of disease progression of patients in the near future based on their historical health records. It is beneficial for patients to make an accurate clinical diagnosis and early prevention of disease. Current risk predictions of brain degeneration mainly rely on single-modality medical data, such as Electronic Health Records (EHR) or magnetic resonance imaging (MRI). However, only leveraging EHR or MRI data for the pertinent and accurate prediction is insufficient because of single-modality information (e.g., pixel or volume information of image data or clinical context information of non-image data).

Methods: Several deep learning-based methods have used multimodal data to predict the risks of specified diseases. However, most of them simply integrate different modalities in an early, intermediate, or late fusion structure and do not care about the intra-modal and intermodal dependencies. A lack of these dependencies would lead to sub-optimal prediction performance. Thus, we propose an encoder-decoder framework for better risk prediction of brain degeneration by using MRI and EHR. An encoder module is one of the key components and mainly focuses on feature extraction of input data. Specifically, we introduce an encoder module, which integrates intra-modal and inter-modal dependencies with the spatial-temporal attention and cross-attention mechanism. The corresponding decoder module is another key component and mainly parses the features from the encoder. In the decoder module, a disease-oriented module is used to extract the most relevant disease representation features. We take advantage of a multi-head attention module followed by a fully connected layer to produce the predicted results.

Results: As different types of AD and DM influence the nature and severity of brain degeneration, we evaluate the proposed method for three-class prediction of AD and three-class prediction of DM. Our results show that the proposed method with integrated MRI and EHR data achieves an accuracy of 0.859 and 0.899 for the risk prediction of AD and DM, respectively.

Discussion: The prediction performance is significantly better than the benchmarks, including MRI-only, EHR-only, and state-of-the-art multimodal fusion methods.

KEYWORDS

risk prediction of brain degeneration, multimodal medical data, multimodal learning, self-attention mechanism, cross-attention mechanism

1. Introduction

With the advent of artificial intelligence (AI), many deep learning–based methods (Schlemper et al., 2019; Zhang et al., 2019; Ye et al., 2021) using medical data have emerged as essential tools for aiding the early identification of disease severity. Commonly, medical data can be divided into two broad modalities: image data, such as magnetic resonance imaging (MRI) and computed tomography (CT), and non-image data, such as Electronic Health Records (EHR).

Brain degeneration is a chronic brain disease that disturbs the brain's normal functioning and further brings a huge threat to public health (Pratico, 2008). Several research studies (Nicolls, 2004; Xu et al., 2009; Stanciu et al., 2020; Cheung et al., 2022) have revealed that adults with chronic diabetes mellitus (DM), including type 1 diabetes and type 2 diabetes, have a higher risk of developing AD. The severity and duration of DM could contribute to brain degeneration (Pruzin et al., 2018). Thus, AD becomes the most common cause of brain degeneration and typically begins with impairments in cognitive functions (Li and Hölscher, 2007). According to the different development of cognitive degradation, AD is divided into three stages, including the pre-clinical (e.g., cognitively normal) stage, mild cognitive impairment (MCI) stage, and dementia stage (Pratico, 2008). MCI is key to diagnosing the early stage of AD. Similarly, DM is classified as type 1 diabetes mellitus (T1DM) and type 2 diabetes mellitus (T2DM) depending on differences in diabetes mechanisms. Patients with T1DM and T2DM would present brain degeneration at different levels.

Many deep learning methods (Escott-Price et al., 2015; Moeskops et al., 2018; Li and Fan, 2019; Xu et al., 2020; Yang and Liu, 2020; Zhu et al., 2020) have been developed to predict the risk of brain degeneration from various aspects, e.g., the transition from MCI to AD in advance, and the cognitive impairment in patients with T1DM and T2DM. These risk prediction methods can effectively reduce the incidence rate of concurrent brain degeneration diseases. Because of a huge data domain gap between medical images and EHR, the difference in prediction accuracy is significant when using medical images or EHR, respectively. The medical images (e.g., MRI) present the vital anatomical information that non-image data (e.g., EHR) lack. EHR is regarded as an important auxiliary for accurate medical image interpretation, particularly for DM diagnosis (Biessels and Reijmer, 2014). Therefore, the fusion of medical images and EHR could provide sufficient information and improve prediction accuracy. Most deep learning–based methods (Li et al., 2019; Ljubic et al., 2020; Yang and Liu, 2020; Yiğit and Işik, 2020; Alexander et al., 2021; Zhang et al., 2021) for predicting the risk of brain degeneration from some chronic diseases only utilize single-modal data. The learnable features from single-modal data may suffer from serious biases of the learning model, which lack imaging or clinical context information. Several learning-based methods (Spasov et al., 2018; Li and Fan, 2019; Zhou et al., 2021) using medical images and EHR data have attempted to predict disease risk by a multimodal data fusion model. However, very few deep learning–based methods account for the inter-modal and intra-modal relationships and have been explored for better accurate risk prediction of brain degeneration.

Medical imaging datasets account for anatomical information and are insufficient to train a network alone. The main reason is the lack of clinical information that is embedded in the EHR dataset. It may lead to unbalanced classes and inaccurate prediction (Huang et al., 2020). EHR is a kind of hierarchical data that stores the historical health status of a patient in temporal sequences formed by multiple visits (Shickel et al., 2017). EHR data of a patient are usually represented by a sparse binary matrix. Only encoding a sparse vector in the deep learning–based method may cause a lack of diversity for potential embedding space, thus increasing the challenge for network training without large volumes of image data (Ye et al., 2020). Therefore, only leveraging EHR data for the risk prediction of brain degeneration is also insufficient.

To solve the above limitations, combining medical imaging with EHR data is necessary for compensating patients' more detailed historical health status. More specifically, medical images, such as MRI, could offer more complex interpretations of a patient's health status, thus leading to a more elaborate embedding space for potential risk-generation tasks. However, most deep learning–based methods (Shickel et al., 2017; Xu et al., 2020; Zhang et al., 2021) using multimodal data only integrate the medical data from different modalities in a simple manner, such as an early, intermediate, or late fusion structure.

A lack of deep exploration of the intra-modal and inter-modal dependencies leads to sub-optimal prediction performance.

The attention mechanism (Vaswani et al., 2017) has emerged with the coming of transformer architecture. It is an input processing technique for neural networks that allows the network to focus on specific parts of a complex input, one at a time until the entire dataset is processed. Attention can provide the ability to highlight vital information and suppress irrelevant information. In the tasks of medical imaging analysis, the spatial–temporal self-attention mechanism (Schlemper et al., 2019; Chen and Shi, 2020; Chen et al., 2020; Plizzari et al., 2021; Yu et al., 2021; Mehta et al., 2022) is often used to capture the spatial and temporal correlations of the same image sequences. The cross-attention mechanism (Hou et al., 2019; Huang et al., 2019; Yu et al., 2021) can capture the interdependent relationship between two sequences of single-modal or multimodal data by integrating two separate embedding sequences with the same dimension asymmetrically. The attention has been effectively applied to medical image analysis to achieve promising results. Some deep learning–based studies (Wang et al., 2018; Jiang et al., 2021) only use simple concatenation for the combination of multimodal features after a feed of medical images (e.g., MRI, CT, or X-ray) and clinical context features (e.g., EHR). The attention mechanism can provide the ability to emphasis on important information and suppress irrelevant counterparts of multimodal features. However, the attention mechanism is scarcely adopted to capture the correlations between medical images and non-image data. The goal of this study is to solve the abovementioned problems. We thus develop a novel attention–based framework for predicting the risk of brain degeneration by making better use of medical images and EHR data. First, a spatial and temporal attention encoder is composed of a set of self-attention blocks. The encoder is employed to extract the complementary features information based on multimodal data to achieve the intra-modal dependencies. This encoder often helps extract the critical pixel information of MRI. Then, for gaining the inter-modal dependencies between MRI and EHR data, a cross-attention mechanism is used to extract the cross-correlation from these two modalities. After two attention encoders, we also propose to adopt the multi-head attention decoder for combining the features of different modalities before the final fully connected (FC) layer. The decoder can guarantee an optimal global feature representation depending on its powerful combination ability in different subspaces.

To sum up, the contribution of this study is two-fold. First, different from the previous multimodal fusion methods of varying medical data modalities (Arevalo et al., 2017; Huang et al., 2020; Jiang et al., 2021; Nagrani et al., 2021), we focus on extracting the critical complementary information between MRI and EHR data with the attention mechanisms for the prediction of brain degeneration. Second, multi-head attention as a disease-oriented decoder is used to improve the prediction performance to avoid sub-optimal issues. We perform the experiments on an available publicly Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and an internally collected diabetes mellitus (DM) dataset to evaluate the performance of our proposed method.

2. Materials and methods

2.1. Materials

2.1.1. Internally collected datasets

All internal data used in the study are collected from Zhongshan Hospital Affiliated with Dalian University. The protocol for this retrospective study was approved by the Ethics Committee of Zhongshan Hospital Affiliated with Dalian University. The requirement for written informed consent from study participants was waived.

The dataset includes 396 subjects with T1-weighted MRI and the corresponding EHR. A patient's diagnosis in the internal data is classified as normal control (NC), T1DM, and T2DM. This study includes 99 NC cases, 135 T1DM cases, and 162 T2DM cases. The EHR data contain a total of 17 features [demographic information: age, gender, years in education; fasting glucose; glycated hemoglobin (HbA1c); triglyceride (TG); cholesterol (CHO); low-density lipoprotein (LDL); high-density lipoprotein (HDL); C Peptide; Montreal Cognitive Assessment (MoCA); clock drawing test (CDT); verbal fluency test (VFT); trial marking test A (TMT-A); anxiety level; depression level; and sleep quality]. The MRI data are directly used for all following experiments in this study to avoid information loss because of preprocessing operation. All 17 features of EHR data are considered to use in the following experiments.

2.1.2. Public datasets

The data used in the evaluation of this study are obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Jack et al., 2008) for analyzing the progression of Alzheimer's disease (AD). An essential goal of ADNI database is to evaluate whether medical images, including MRI and PET, and other modality EHR data including biological markers and clinical and neuropsychological assessment information, can be integrated to predict the AD progression from MCI or pre-clinical stage for accurate diagnosis and early prevention.

We select the training data according to the following rules (Jiang et al., 2021). For each patient, the first scanned MRI with description information "multiplanar reconstruction (MPR); GradWarp; B1 Correction; N3." A patient's diagnosis in the ADNI is typically classified as AD, MCI, and cognitively normal (CN). In this study, we select the whole data from 969 subjects, containing 288 AD cases, 365 MCI cases, and 316 CN cases. For each patient, one MRI sequence is accompanied by corresponding EHR data. The MRI data are also directly used in this study. The selected EHR data contain a total of 11 features

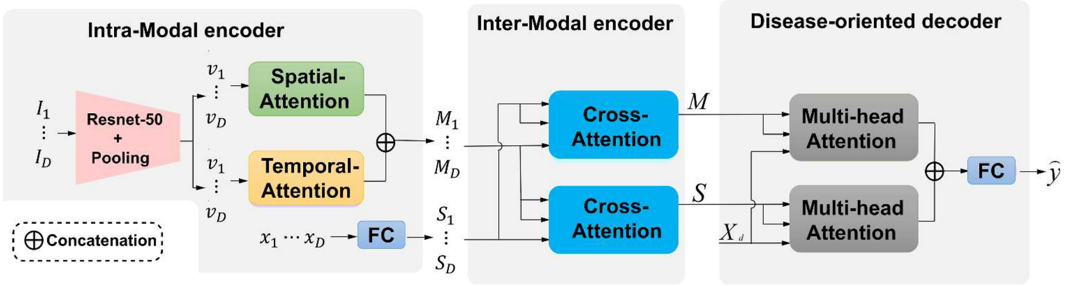

FIGURE 1
An illustration of the proposed framework for the risk prediction of brain degeneration. Our framework has two attention encoders and one decoder. $\{\boldsymbol{I}_i \in \mathbb{R}^{\{H\times W\}}|i \in \{1,..., D\}\}$ is an MRI sequence $\boldsymbol{I}$.$\{\boldsymbol{v}_i \in \mathbb{R}^{\{H\times W\}}|\boldsymbol{v} \in \{1,..., D\}\}$ is an MRI features representation. $\{\boldsymbol{x}_i \in \mathbb{R}^{\{H\times W\}}|x \in \{1,..., D\}\}$ is the corespongding Electronic Health Records (EHR) data of MRI. $\boldsymbol{S}$ and $\boldsymbol{M}$ are the feature representations across spatial-attention and temporal-attention, respectively. X_d represents the classification query. $\hat{\boldsymbol{y}}$ is the final prediction label according to the categories of diseases.

[demographic information: age, gender, years in education, and ethnic and racial categories; biofluids: APOe4 genotyping; cerebrospinal fluid (CSF) levels; behavioral assessments: clinical dementia rating (CDRSB); Alzheimer's disease assessment scale (ADAS13); the episodic memory evaluations in the Rey Auditory Verbal Learning Test (RAVLT_immediate); and The Mini-Mental State Examination (MMSE)]. All 11 features of EHR data are considered to use in the following experiments.

2.2. Methods

This study develops an end-to-end framework for predicting the risk of brain degeneration by taking in the complementary features between MRI and EHR data. The input data of the network are the paired MRI and EHR data. 3D ResNet-50 (Yu et al., 2021; Mehta et al., 2022) is the backbone network in the initial stage. Other deeper networks, such as DenseNet (Huang et al., 2017), also work with our proposed framework. The output is the prediction result, which is represented as binary values. To address the issues of the intra-modal and inter-modal dependencies, two attention mechanisms are deployed in the two-level encoder module. To be specific, self-attention as the first-level encoder, which includes spatial and temporal attention, is utilized to extract the spatial–temporal feature information for the internal-slice dependencies of the same MRI sequence. The EHR data and disease representations from the self-attention output are passed into the second-level cross-attention encoder. This encoder considers the inter-modal dependencies by extracting the correlations between the features from MRI and EHR data. After the encoder, the multi-head attention mechanism as a decoder aggregates the information from all dimensions for producing the final prediction. The overall network architecture of risk prediction of brain degeneration is shown in Figure 1.

Given the observed history of patient health status in multiple visits, an available visit is represented by $\{\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_D, \boldsymbol{x}\}$, where $\{I_i \in \mathbb{R}^{\{H\,\times\,W\}}|i \in \{1, \ldots, D\}\}$ represents the i-th slice from an MRI sequence, H and W denote the height and width, respectively. Binary vector set $\boldsymbol{x} \in \mathbb{R}^D$ is EHR data of each MRI sequence, each element in $\boldsymbol{x}$ belongs to $\{0, 1\}$, where 1 denotes the presence of the corresponding AD and vice visa. The task needs to predict the risks of getting K categories of diseases, which could be represented as $\hat{\boldsymbol{y}} \in [0, 1]^K$. Our framework consists of two encoders that integrate intra-modal and inter-modal dependencies in a spatial–temporal manner and a disease-oriented decoder with multi-head attention to extract the most relevant disease representations.

2.2.1. Intra-modal encoder

Given medical images, intra-modal dependencies are first generated by capturing the spatial–temporal relations of MRI modality in an independent module. Considering the MRI sequence $\{\boldsymbol{I_1},\ \boldsymbol{I_2}, \ldots, \boldsymbol{I_D}\}$, where D is the number of slices from one MRI sequence, a ResNet-50 and a spatial average pooling layer are adopted to extract the disease features representation $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_D}\}$, where each element is a C-dimensional vector with shape $(1, C)$. After repeating the above operations for all MR slices of one visit, $C \times D$ vectors are separately processed by two blocks from spatial and temporal domains. As shown in Figure 1, one disease representation v_i, which stands for the i-th slice, interacts with other representations in the spatial-attention block to capture the intra-slice relations. v_i interacts with other representations in the temporal block to compute the inter-slice variations from the same MR sequence. Based on the MRI sequence, the relations between two continuous slices are retrieved with temporal attention, and the relations of pixels in one slice are retrieved with spatial attention. Both the two attention mechanisms follow the spatial and temporal structure as described in Mehta et al. (2022).

As shown in Figure 1, spatial attention is used to capture intra-slice dependencies. The relationships between each pixel and other pixels in the slice are computed. These relations are passed with dominant intra-frame dependencies. The illustration of spatial attention is shown in Figure 2A and mathematically expressed by the following equation:

$$S_{j,i} = \frac{\exp\left(K(v_i)^T Q(v_j)\right)}{\sum_i^{H\times W}\exp\left(K(v_i)^T Q(v_j)\right)}. \tag{1}$$

The disease representation v through ResNet-50 and spatial average pooling layer is transformed to the key $K(v_i)$, query $Q(v_j)$, and value $V(v_i)$ by using $1 \times 1 \times 1$ convolution filter. The relationships between pixels are represented by the spatial dimension $(H \times W) \times (H \times W)$. $\boldsymbol{S_{j,i}} \in \mathbb{R}^{\{C\times D\times H\times W\times H\times W\}}$ is spatial correlation matrix for computing the impact of i-th position on j-th position and obtained by softmax of the inner product of $K(v_i)$ and $Q(v_j)$. Here, C is the number of channels. The output attention features across spatial dimensions are written as:

$$\widehat{\boldsymbol{M}}_{\boldsymbol{S}} = \sum_{i=1}^{H\times W} V(v_i)\, S_{j,i}. \tag{2}$$

Then, $\widehat{\boldsymbol{M}}_{\boldsymbol{S}} \in \mathbb{R}^{\{C\times H\times W\times D\}}$ is fed into $1 \times 1 \times 1$ convolution filter, which results in the final spatial-attention features $\boldsymbol{M_S}$ with C channels.

The temporal attention captures an MRI sequence's inter-slice dependencies and relates the global features between two slices of the same MRI sequence in the temporal domain. The illustration of temporal attention is shown in Figure 2B and mathematically expressed by the following equation:

$$\boldsymbol{T_{j,i}} = \frac{\exp\left(K(v_i)^T Q(v_j)\right)}{\sum_i^{D}\exp\left(K(v_i)^T Q(v_j)\right)}. \tag{3}$$

The relationships between pixels are represented by the depth dimension $D \times D$. $\boldsymbol{T_{j,i}} \in \mathbb{R}^{\{C\times D\times H\times W\times D\times D\}}$ is a dimensional temporal correlation matrix for computing the impact of i-th slice on j-th slice. The output attention features across temporal dimension are written as:

$$\widehat{\boldsymbol{M}}_{\boldsymbol{T}} = \sum_{i=1}^{D} V(v_i)\, T_{j,i}. \tag{4}$$

Then, $\widehat{\boldsymbol{M}}_{\boldsymbol{T}} \in \mathbb{R}^{\{C\times D\times H\times W\}}$ is fed into $1 \times 1 \times 1$ convolution filter, which results in the final temporal-attention features $\boldsymbol{M_T}$ with C channels.

For each spatial and temporal attention block, the final output is then concatenated along with the spatial dimension to form D matrices where each one owns the shape of (D, C). Finally, disease representations of medical images $\{\boldsymbol{M}_i \in \mathbb{R}^{D\times C} | i \in \{1, 2, \ldots, D\}\}$ are generated by summing matrices with the same visit index from different attention blocks. For the EHR vector sequence $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D\}$ comprise of D time points for one MRI sequence, a fully connected layer is adopted to embed each EHR vector into a C-dimensional space to capture the overall health information by producing a vector with shape $(1, C)$, which results in disease representations $\{\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_D\}$ of EHR data.

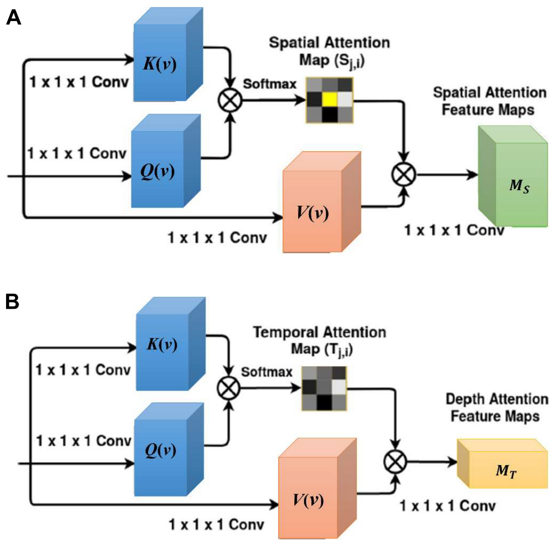

FIGURE 2
(A) 3D spatial-attention architecture; **(B)** 3D temporal-attention architecture. $K(v)$ is the key, $Q(v)$ is the query and $V(v)$ is the value. ⊗ is defined as an element-wise multiplication operation. Noted that spatial or depth is an interchangeable term.

2.2.2. Inter-modal encoder

Inter-modal dependencies between MRI and EHR are captured through a cross-attention mechanism, which exchanges the global health status from EHR data and detailed disease information from MRI in a parallel manner. Given disease representation $\{M_t, S_t\}$ for t-th slice, two cross-attention modules as shown in Figure 3 are leveraged to compute the cross-correlation of multimodal features by taking queries from their own modalities while key and value matrices from opposite modalities, which results in $\{M_t', S_t'\}$.

To be specific, disease representation *via* two $1 \times 1 \times 1$ convolution filter produces two feature maps Q and K, respectively, where $\{Q, K\} \in \mathbb{R}^{\{C\times H\times W\}}$. After obtaining Q and K, the feature attention maps are generated *via* affinity operation (Huang et al., 2019) and softmax.

At each position j in the spatial dimension of feature map Q, a vector $Q_j \in \mathbb{R}^C$ is obtained. For the total features set $\boldsymbol{\Omega}_j \in \mathbb{R}^{(H\times W-1)\times C}$ also can be obtained by capturing the spatial features vectors from feature map K, which are in the same row with position j. Here, $\boldsymbol{\Omega}_{i,j} \in \mathbb{R}^C$ represents the i-th element of $\boldsymbol{\Omega}_j$. The affinity operation is formulated as follows:

$$\boldsymbol{Aff}_{\boldsymbol{i,j}} = \boldsymbol{Q_j}\boldsymbol{\Omega}_{\boldsymbol{i,j}}{}^T \tag{5}$$

where $\boldsymbol{Aff}_{\boldsymbol{i,j}}$ is the correlation degree between $\boldsymbol{Q_j}$ and $\boldsymbol{\Omega}_{\boldsymbol{i,j}}$. Then, a softmax layer is applied on $\boldsymbol{Aff}_{\boldsymbol{i,j}}$ across each channel to calculate the attention map $\boldsymbol{A}$ from affinity operation.

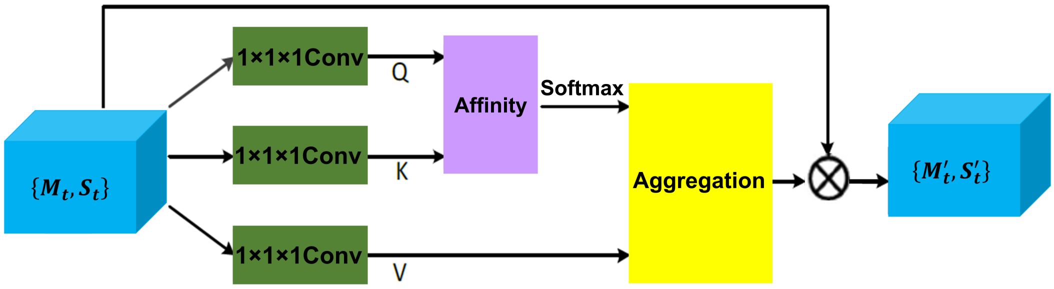
1×1×1Conv
Q
1×1×1Conv
K
1×1×1Conv
V
Affinity
Softmax
Aggregation
$\{\boldsymbol{M}_t, \boldsymbol{S}_t\}$
$\{\boldsymbol{M}_t', \boldsymbol{S}_t'\}$
$\otimes$

FIGURE 3
The details of cross-attention architecture to extract the cross-correlation of multimodal features. Q, K, and V represent all extracted feature maps. $\{\boldsymbol{M}_t, \boldsymbol{S}_t\}$ is the disease feature representation, and $\{\boldsymbol{M}_t', \boldsymbol{S}_t'\}$ is the final representation.

Another $1 \times 1 \times 1$ convolution filter is applied to disease representation $H \in \{\boldsymbol{M}_t, \boldsymbol{S}_t\}$ to produce feature map V, the final representations $\{\boldsymbol{M}_t^{'}, \boldsymbol{S}_t^{'}\}$ is obtained by aggregation operation (Huang et al., 2019) for achieving the mutual feature gains from MRI and EHR.

Similarly, at each position j in the spatial dimension of feature map V, a vector $V_j \in \mathbb{R}^C$ and the total features set $\widehat{\Omega}_j \in \mathbb{R}^{(H \times W-1) \times C}$ are obtained. Here, $\widehat{\Omega}_{i,j} \in \mathbb{R}^C$ represents the i-th element of $\widehat{\Omega}_j$. The aggregation operation is formulated as follows:

$$Agg_j = \sum_{i \in |\widehat{\Omega}_j|} A_{i,j} \widehat{\Omega}_{i,j} + H_j \tag{6}$$

where Agg_j is a feature vector at position j. $A_{i,j}$ is scalar data, which belongs to affinity feature map A. The most relevant contextual information is added to local disease representation H to enhance the local features and augment the disease representation. Thus, these disease feature representations achieve mutual gains between MRI and EHR data.

After repeating the operations for each slice corresponding to an independent time point, D updated vectors of EHR are concatenated into $\boldsymbol{S} \in \mathbb{R}^{D \times C}$, and a compressed disease representation of medical images $\boldsymbol{M} \in \mathbb{R}^{D \times C}$ is produced by concatenating and pooling the $\{\boldsymbol{M}_1^{'}, \boldsymbol{M}_2^{'}, \ldots, \boldsymbol{M}_D^{'}\}$ across the temporal dimension.

2.3. Disease-oriented decoder

Disease-oriented decoder seeks the most relevant information in two different modalities for predicting the risk of brain degeneration. The right part of Figure 1 shows that the decoder includes two multi-head attention layers and a fully connected layer. The multi-head attention layer expects disease representations $\boldsymbol{M}$, $\boldsymbol{S}$, and a classification query $\boldsymbol{X_d} \in \mathbb{R}^{K \times C}$ as input, where $\boldsymbol{K}$ is the number of disease risk categories included in the task. By conducting the multi-head attention mechanism, which follows the multi-head attention of the transformer (Vaswani et al., 2017), the most relevant clinical contextual information for brain degeneration is updated and stored in the query. Finally, the outputs of two multi-head attention layers are added together and transmitted into a fully connected layer to produce the final prediction result $\hat{\boldsymbol{y}} \in \mathbb{R}^{\boldsymbol{K}}$. Actually, the prediction risk of brain degeneration is a classification task, and the cross-entropy loss function is applied at the training stage to train the model.

3. Experiments and results

3.1. Implementation details

We implement our proposed method on Pytorch to classify three stages of AD progression, including CN, MCI, and AD. For the training stage, four Nvidia Tesla V100 GPUs with 32GB memory are used. We employ a polynomial learning rate policy where the initial learning rate is multiplied by $1-(\frac{iter}{total_{iter}})^{power}$ with $power = 0.9$. The initial learning rate we used is set to 0.01. Momentum and weight decay coefficients are 0.9 and 0.0001, respectively. The input size of MRI is $256 \times 256 \times 170$, the batch size is set to 32. Five-fold cross-validation is performed to split the training data. We perform 100 epochs of training for all settings. All the intensities of input MRI images are normalized to [0,1].

3.2. Results

3.2.1. Evaluation metrics

Four evaluation metrics are calculated to evaluate the risk prediction performance on the test cases of internally collected DM datasets and ADNI datasets. These metrics include sensitivity, accuracy, specificity, and area under the receiver operating characteristic curve (AUROC). All the evaluation metrics are reported in the following ablation and comparison experiments.

3.2.2. Ablation study for intra-modal and inter-modal encoders

We employ self-attention mechanisms, including a spatial-attention mechanism (SAM) and a temporal-attention mechanism (TAM) for the intra-modal encoder and a cross-attention mechanism (CAM) for the inter-modal encoder. The addition of these two encoders can contribute to capturing the intra-modal and inter-modal dependencies for better prediction. To verify the encoder module's performance and analyze each component's actual contribution, we conduct ablation experiments with different settings on both DM and ADNI datasets in **Tables 1**, **2**.

As shown in **Tables 1**, **2**, the intra-modal and inter-modal encoders remarkably improve the prediction performance on internally collected DM and public ADNI datasets. The baseline method only uses the multi-head attention mechanism, as shown in the first row of **Tables 1**, **2**. Compared with the baseline method, employing SAM and TAM in the intra-modal encoder achieved a significant prediction improvement with an accuracy of 0.762 on DM datasets and an accuracy of 0.742 on ADNI datasets. The visual attention maps in **Figure 4** with SAM and TAM showed that the attention mechanism in the intra-modal encoder could capture the critical area (around the location of the hippocampus) features, which are quite relevant to brain degeneration. Only employing the CAM in the inter-modal encoder yields an accuracy of 0.784 on DM datasets and 0.852 on ADNI datasets, which are higher than the accuracies of only employing the SAM and TAM in the intra-modal encoder. Then, in our proposed method, we further combine the SAM and the TAM in the intra-modal encoder with the CAM in the inter-modal encoder, and the highest accuracies of 0.859 on DM datasets and 0.899 on ADNI datasets are achieved. In particular, on DM datasets, the proposed method outperforms the method with only an intra-modal encoder and the method with only an inter-modal encoder by 16.4 and 16.1%, respectively. We also observe that our proposed method achieves the best results for other evaluation metrics for both DM and ADNI datasets. Similarly, results substantiated that multimodal encoders considering intra-modal and inter-modal dependencies greatly benefit the risk prediction of brain degeneration based on different disease datasets (e.g., DM datasets and ADNI datasets).

3.2.3. Evaluation of multi-head attention decoder

After two encoders, we employ the two multi-head attention layers as a disease-oriented decoder. The multi-head attention mechanism with multiple head numbers can focus on the most relevant features from multimodal representation subspaces to reach an optimal global representation. We evaluate the multi-attention decoder in our method with varying head numbers for a comprehensive comparison. We evaluate the impact of the head number on the multi-head attention mechanism. As shown in **Figure 5**, the accuracy performance of multi-head attention with head numbers from 1 to 12 is evaluated on both DM and ADNI datasets. From the observation of **Figure 5**, when the head number reaches the optimal head number, the performance decreases with increasing head number values. It is observed that the head number is set to six for DM datasets, and the highest accuracy of risk prediction of brain degeneration is demonstrated. Similarly, as shown in **Figure 5B**, the head number is set to five for our used dataset from the ADNI database, and the highest accuracy is observed. It implies that the optimal head number may vary for different data domains due to the data domain gap (Liu et al., 2021).

3.2.4. Comparison

We compare our method with MRI-only method-3D DenseNet (Ruiz et al., 2020), EHR-only method-ElasticNet (Zou and Hastie, 2005), and three typical learning-based multimodal fusion methods.

The MRI-only method only depends on the pixel information from MRI data for predicting the outcome. For our MRI-only method, we use the 3D DenseNet model (Ruiz et al., 2020), which utilizes MRI and is capable of considerable risk prediction of brain degeneration. The 3D DenseNet primarily consists of layers of 3D convolutions with skip connections.

The EHR-only method only depends on parsing the EHR data through preprocessing step. More precisely, the EHR data of a patient are usually denoted by a sparse binary matrix where each element is an International Classification Disease code (ICD-9) (Benesch et al., 1997) in a specified visit. Several learning-based methods (Ma et al., 2018; Zhang et al., 2019; Luo et al., 2020; Ahuja et al., 2021) have put effort into encoding the potential temporal relations, especially between distinct visits of EHR and output the risk prediction of disease through a multi-task paradigm. For our EHR-only method, we use an ElasticNet (Zou and Hastie, 2005) model, which takes in a concatenation of all EHR features.

In clinical practice, pertinent clinical information is vital for providing accurate diagnostic decisions during medical imaging interpretations (Boonn and Langlotz, 2009; Zhou et al., 2021). The fused feature maps from MRI and EHR data in our compared multimodal fusion methods are performed by (1) Early fusion (Spasov et al., 2018) based on concatenation; (2) intermediate fusion (Jiang et al., 2021) based on linear layers and (3) late fusion (Arevalo et al., 2017) based on single-head attention strategies.

The concatenation method is implemented by concatenating the pooled image feature and EHR feature at the input level. Different from the concatenation method, linear layers of a conventional neural network (CNN) mainly adopt a linear transformation for each modality data to obtain the transformed features with the same size for multimodal data. These two transformed features from medical image

TABLE 1 Quantitative results on internally collected diabetes mellitus (DM) datasets for the proposed method with or without the specified components.

Intra-modal encoder		Inter-modal encoder	Sensitivity	Accuracy	Specificity	AUROC
SAM	TAM	CAM				
			0.562 ± 0.016	0.596 ± 0.012	0.742 ± 0.012	0.714 ± 0.012
√			0.601 ± 0.012	0.634 ± 0.015	0.762 ± 0.011	0.772 ± 0.012
	√		0.719 ± 0.012	0.716 ± 0.013	0.766 ± 0.012	0.802 ± 0.012
√	√		0.762 ± 0.012	0.752 ± 0.016	0.771 ± 0.012	0.839 ± 0.011
		√	0.764 ± 0.012	0.784 ± 0.014	0.778 ± 0.012	0.842 ± 0.014
√		√	0.771 ± 0.016	0.801 ± 0.016	0.792 ± 0.011	0.861 ± 0.013
	√	√	0.834 ± 0.015	0.823 ± 0.015	0.816 ± 0.012	0.887 ± 0.012
√	√	√	**0.887 ± 0.016**	**0.859 ± 0.012**	**0.867 ± 0.012**	**0.916 ± 0.012**

The '√' symbol represents the inclusion of components. The results from the proposed method with SAM, TAM, and CAM are highlighted in bold. SAM represents spatial-attention mechanism, TAM represents temporal-attention mechanism, and CAM represents cross-attention mechanism.

TABLE 2 Quantitative results on Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets for the proposed method with or without the specified components.

Intra-modal encoder		Inter-modal encoder	Sensitivity	Accuracy	Specificity	AUROC
SAM	TAM	CAM				
			0.536 ± 0.011	0.626 ± 0.013	0.755 ± 0.012	0.732 ± 0.012
√			0.588 ± 0.012	0.726 ± 0.015	0.772 ± 0.011	0.791 ± 0.013
	√		0.708 ± 0.013	0.826 ± 0.015	0.864 ± 0.013	0.897 ± 0.012
√	√		0.742 ± 0.012	0.833 ± 0.014	0.872 ± 0.012	0.909 ± 0.011
		√	0.802 ± 0.013	0.852 ± 0.015	0.878 ± 0.013	0.913 ± 0.013
√		√	0.841 ± 0.014	0.866 ± 0.016	0.893 ± 0.010	0.931 ± 0.012
	√	√	0.886 ± 0.015	0.885 ± 0.016	0.884 ± 0.013	0.936 ± 0.012
√	√	√	**0.901 ± 0.014**	**0.899 ± 0.013**	**0.892 ± 0.012**	**0.953 ± 0.013**

The '√' symbol represents the inclusion of components. The results from the proposed method with spatial-attention mechanism (SAM), temporal-attention mechanism (TAM), and cross-attention mechanism (CAM) are highlighted in bold.

and EHR are added up to a fused feature. The fusion based on single-head attention is performed by employing standard attention as an aggregation strategy before the FC layer.

We use the ResNet-50 as the backbone for all methods and the same datasets to guarantee a fair comparison. We benchmark the performance of different methods on the entire test data using four different evaluation metrics. The results of the metrics are reported in Tables 3, 4 on DM and ADNI datasets. For both DM and ADNI datasets, we observe that the EHR-only method can achieve better performance than the MRI-only method for the risk prediction of brain degeneration on all the evaluation metrics. It means that EHR data could provide informative data for the clinical diagnosis of brain degeneration. When combining MRI and EHR data, the three multimodal fusion methods further enhance the prediction performance compared with the MRI-only and EHR-only methods. It proves that EHR is crucial for the complementary interpretation of MR images. Given the results of prediction performance from Tables 3, 4, late fusion works better for fusing MRI and EHR data to predict the risk of brain degeneration than early fusion and intermediate fusion. Unlike these three typical fusion methods, the proposed method considers the intra-modal and inter-modal dependencies for learning more modality-aware mutual and complementary features. These enhanced features lead to noticeable performance improvement on DM and ADNI datasets. Thus, the proposed method achieves the best results on all four evaluation metrics. Especially on ADNI datasets, the accuracy of 0.899 in our method is much higher than the accuracy of 0.757 in the worst MRI-only method, with a significant improvement of 18.7%.

4. Discussion

The main novelty of the proposed method is to incorporate the correlated features between MRI and EHR data into a global disease representation in a tightly coupled way, which

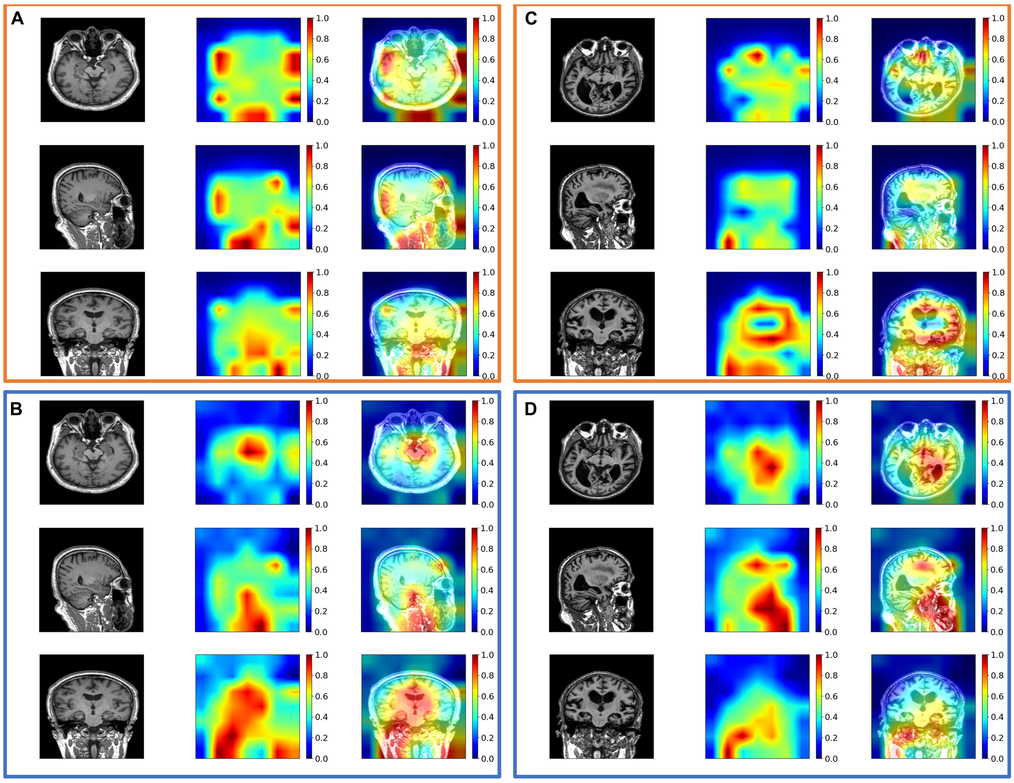

FIGURE 4
The exemplary attention maps **(A)** with spatial-attention mechanism (SAM) and temporal-attention mechanism (TAM) and **(B)** without SAM and TAM on diabetes mellitus (DM) datasets; **(C)** with SAM and TAM, and **(D)** without SAM and TAM on Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets. The views from the top row to the bottom row are axial, coronal, and sagittal views. The corresponding images from left to right are the original image, attention map, and image overlayed with the attention map. The value of the attention map from zero to one is assigned blue to red colors. Noted that attention maps without SAM and TAM may suffer from inaccurate feature extraction, such as high attention values close to 1 out of the head in panels **(A,C)**.

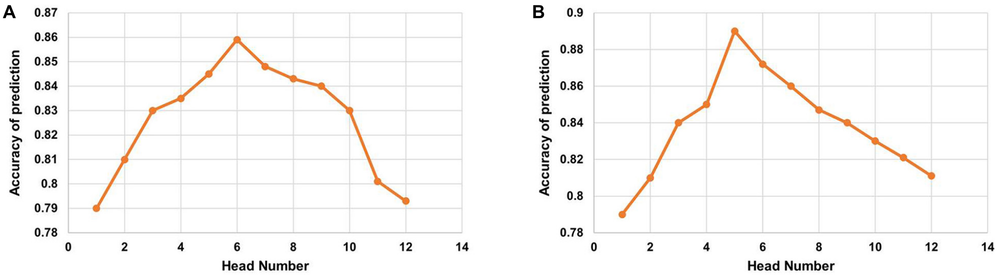

FIGURE 5
Accuracy of the multi-head attention with the varying head number on **(A)** diabetes mellitus (DM) datasets and **(B)** Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets.

depends on the attention mechanisms in intra-modal and inter-modal encoders. To further emphasize the impact of each attention component, the ablation experiments are performed by the single addition or the combined addition of different attention mechanisms to the baseline method. Our proposed method has the highest predictive ability to distinguish the three levels of brain degeneration progression, which occur in DM and AD patients, respectively. This is mainly because our method preserves the high correlation between MRI and EHR data by capturing intra-modal and inter-modal dependencies. Notably, our method adds spatial–temporal attention and cross-attention to capture the intra-modal dependencies of an MRI sequence. The intra-modal dependencies provide sufficient anatomical features and significantly improve the prediction.

TABLE 3 Performance comparison of the MRI-only method, the Electronic Health Records (EHR)-only method, the early fusion method, the intermediate fusion method, and the late fusion method on the test diabetes mellitus (DM) dataset.

Methods	Sensitivity	Accuracy	Specificity	AUROC
MRI-only	0.674 ± 0.012	0.763 ± 0.012	0.772 ± 0.013	0.742 ± 0.012
EHR-only	0.745 ± 0.012	0.818 ± 0.014	0.822 ± 0.012	0.832 ± 0.012
Early fusion	0.789 ± 0.013	0.825 ± 0.011	0.826 ± 0.012	0.841 ± 0.012
Intermediate fusion	0.827 ± 0.012	0.831 ± 0.014	0.839 ± 0.011	0.853 ± 0.012
Late fusion	0.841 ± 0.012	0.833 ± 0.012	0.851 ± 0.012	0.867 ± 0.011
Proposed	**0.887 ± 0.016**	**0.859 ± 0.012**	**0.867 ± 0.012**	**0.916 ± 0.012**

The bold values means the best performance among these methods.

TABLE 4 Performance comparison of the MRI-only method, the Electronic Health Records (EHR)-only method, the early fusion method, the intermediate fusion method, and the late fusion method on the test Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset.

Methods	Sensitivity	Accuracy	Specificity	AUROC
MRI-only	0.658 ± 0.013	0.757 ± 0.014	0.853 ± 0.011	0.843 ± 0.014
EHR-only	0.786 ± 0.012	0.829 ± 0.016	0.866 ± 0.013	0.903 ± 0.012
Early fusion	0.806 ± 0.014	0.852 ± 0.013	0.877 ± 0.012	0.913 ± 0.013
Intermediate fusion	0.815 ± 0.012	0.851 ± 0.015	0.875 ± 0.011	0.910 ± 0.012
Late fusion	0.873 ± 0.014	0.882 ± 0.014	0.886 ± 0.012	0.928 ± 0.011
Proposed	**0.901 ± 0.014**	**0.899 ± 0.013**	**0.892 ± 0.012**	**0.953 ± 0.013**

The bold values means the best performance among these methods.

The visualization of the different attention maps is shown in Figure 4. For the DM dataset, we can observe the SAM and TAM that can emphasize the critical brain area, which implies the features of the critical area are more relevant to the classification of DM patients. As for the ADNI dataset, the SAM and TAM can also focus on the critical brain area, such as the area around the hippocampus.

In addition to finding a method that can capture the intra-modal and inter-modal dependencies, there is an important need to seek the most relevant features to avoid sub-optimal prediction performance. Following that, we employed two multi-head attention layers to project the inputs into multiple different subspaces to a more elaborate embedding space for the final prediction. Because of different head numbers, the effectiveness of multi-head attention may vary. To reach the optimal performance, Figure 5 shows that larger head numbers do not bring a consistent increase in the prediction performance.

Although our results on DM and ADNI datasets demonstrate the great potential for integrating MRI and EHR data to improve the risk prediction performance of brain degeneration; however, there are some limitations of the proposed method.

Considering the inherent bias of DM and ADNI datasets (Pipitone et al., 2014), it is essential to investigate the performance of multimodal learning models on diversified data, such as more than two modalities of data, to generalize the prediction ability of our method in clinical applications. The number and diversity of datasets are still critical bottlenecks for the performance improvement of the proposed learning model. With a large number of diversified datasets, the prediction performance gain can be obtained by diversified feature enhancements. In addition, the internally collected DM datasets with different patient groups are not well balanced, which may impact the evaluation of the sensitivity gap. With limited DM datasets, the proposed method has improved the prediction of brain degeneration by classifying the three levels of DM patients. Therefore, more extensive studies will be necessary to validate the generalization ability of the proposed attention-based learning model despite our promising preliminary results from internal DM and public ADNI datasets.

In this study, we only select limited features (e.g., 17 features of DM patients and 11 features of ADNI patients) to create the EHR data, the extensive study to rely on MRI image features to guide the selection of more EHR features needs a deep exploration.

With the advent of deep transfer learning technology (Grassi et al., 2019; Bae et al., 2021; Alanazi et al., 2022), the performance of the proposed framework may be optimized by using other modalities of data, such as functional MRI and molecular imaging by mass spectrometry to provide more efficient and accurate predictions. Our method can aid the early diagnosis of brain degeneration and improve the diagnosis workflow. Meanwhile, our proposed method has great potential to be translated to predict the risk of other diseases. Based on other modalities of data, it incorporates more data properties to construct multimodal learning strategies for the prediction of other diseases, such as melanoma and multiple sclerosis (Huang et al., 2020).

5. Conclusion

In this study, we propose a novel attention–based learning framework by incorporating MRI images and EHR data, to improve the precision of brain degeneration diagnosis. Compared to the single-modal features, the optimal global feature representations extracted from MRI features and EHR features play an essential role in the final decisions of the learning model. Through the study, the proposed method is superior to the MRI-only, EHR-only, and typical multimodal fusion methods for predicting brain degeneration.

We deployed suitable attention mechanisms for each module of our framework to extract related information to improve the performance model, which may also be applied to other prediction tasks. Meanwhile, we should focus on the multi-head attention mechanism with different head numbers, which is usually valuable and practical to enhance the final elaborating representations from multimodal data. The designed encoder and decoder modules only depend on self-attention mechanisms, which are flexible to further applications and extensions.

In general, the proposed method provides an efficient aid for clinical diagnosis and early prevention of brain degeneration by extracting disease-oriented related information based on medical images and non-image clinical context information.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Affiliated Zhongshan Hospital of Dalian University, Department of Radiology. The patients/participants provided their written informed consent to participate in this study.

Author contributions

XS and WG contributed to the conception and design of the study. XS performed the data analysis and wrote the first draft of the manuscript. All authors contributed to the manuscript revision, read, and approved the submitted version.

Conflict of interest

WG was employed by EchoX Technology Limited.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahuja, Y., Kim, N., Liang, L., Cai, T., Dahal, K., Seyok, T., et al. (2021). Leveraging electronic health records data to predict multiple sclerosis disease activity. *Ann. Clin. Transl. Neurol.* 8, 800–810. doi: 10.1002/acn3.51324

Alanazi, M. F., Ali, M. U., Hussain, S. J., Zafar, A., Mohatram, M., Irfan, M., et al. (2022). Brain tumor/mass classification framework using magnetic-resonance-imaging-based isolated and developed transfer deep-learning model. *Sensors* 22:372. doi: 10.3390/s22010372

Alexander, N., Alexander, D. C., Barkhof, F., and Denaxas, S. (2021). Identifying and evaluating clinical subtypes of Alzheimer's disease in care electronic health records using unsupervised machine learning. *BMC Med. Inform. Decis. Mak.* 21:343. doi: 10.1186/s12911-021-01693-6

Arevalo, J., Solorio, T., Montes-y-Gómez, M., and González, F. A. (2017). Gated multimodal units for information fusion. *arXiv* [preprint] arXiv:170201992

Bae, J., Stocks, J., Heywood, A., Jung, Y., Jenkins, L., Hill, V., et al. (2021). Transfer learning for predicting conversion from mild cognitive impairment to dementia of Alzheimer's type based on a three-dimensional convolutional neural network. *Neurobiol. Aging* 99, 53–64. doi: 10.1016/j.neurobiolaging.2020.12.005

Benesch, C., Witter, D., Wilder, A., Duncan, P., Samsa, G., and Matchar, D. (1997). Inaccuracy of the international classification of diseases (Icd-9-Cm) in identifying the diagnosis of ischemic cerebrovascular disease. *Neurology* 49, 660–664. doi: 10.1212/WNL.49.3.660

Biessels, G. J., and Reijmer, Y. D. (2014). Brain changes underlying cognitive dysfunction in diabetes: What can we learn from mri? *Diabetes* 63, 2244–2252. doi: 10.2337/db14-0348

Boonn, W. W., and Langlotz, C. P. (2009). Radiologist use of and perceived need for patient data access. *J. Digit. Imaging* 22, 357–362. doi: 10.1007/s10278-008-9115-2

Chen, B., Zhang, Z., Liu, N., Tan, Y., Liu, X., and Chen, T. (2020). Spatiotemporal convolutional neural network with convolutional block attention module for micro-expression recognition. *Information* 11:380. doi: 10.3390/info11080380

Chen, H., and Shi, Z. (2020). A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens.* 12:1662. doi: 10.3390/rs12101662

Cheung, C. Y., Ran, A. R., Wang, S., Chan, V. T., Sham, K., Hilal, S., et al. (2022). A deep learning model for detection of Alzheimer's disease based on retinal photographs: A retrospective, multicentre case-control study. *Lancet Digit. Health* 4, e806–e815. doi: 10.1016/S2589-7500(22)00169-8

Escott-Price, V., Sims, R., Bannister, C., Harold, D., Vronskaya, M., Majounie, E., et al. (2015). Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain* 138, 3673–3684. doi: 10.1093/brain/awv268

Grassi, M., Loewenstein, D. A., Caldirola, D., Schruers, K., Duara, R., and Perna, G. (2019). A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion: Further evidence of its accuracy via a transfer learning approach. *Int. Psychogeriatr.* 31, 937–945. doi: 10.1017/S1041610218001618

Hou, R., Chang, H., Ma, B., Shan, S., and Chen, X. (2019). Cross attention network for few-shot classification. *Adv. Neural Inf. Process. Syst.* 32.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (Piscataway, NJ: IEEE). doi: 10.1109/CVPR.2017.243

Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., and Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digit. Med.* 3, 1–9. doi: 10.1038/s41746-020-00341-z

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., and Liu, W. (2019). "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, (Piscataway, NJ: IEEE). doi: 10.1109/ICCV.2019.00069

Jack, C. R. Jr., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The Alzheimer's disease neuroimaging initiative (Adni): Mri methods. *J. Magn. Reson. Imaging* 27, 685–691. doi: 10.1002/jmri.21049

Jiang, C., Chen, Y., Chang, J., Feng, M., Wang, R., and Yao, J. (2021). Fusion of medical imaging and electronic health records with attention and multi-head machanisms. *arXiv* [preprint] arXiv:211211710

Li, H., and Fan, Y. (2019). "Early prediction of Alzheimer's disease dementia based on baseline hippocampal Mri and 1-year follow-up cognitive measures using deep recurrent neural networks," in *Proceedings of the 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, (Piscataway, NJ: IEEE). doi: 10.1109/ISBI.2019.8759397

Li, H., Habes, M., Wolk, D. A., Fan, Y., and Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle Study of Aging (2019). A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. *Alzheimers Dement.* 15, 1059–1070. doi: 10.1016/j.jalz.2019.02.007

Li, L., and Hölscher, C. (2007). Common pathological processes in Alzheimer disease and type 2 diabetes: A review. *Brain Res. Rev.* 56, 384–402. doi: 10.1016/j.brainresrev.2007.09.001

Liu, L., Liu, J., and Han, J. (2021). Multi-head or single-head? An empirical comparison for transformer training. *arXiv* [preprint] arXiv:210609650

Ljubic, B., Roychoudhury, S., Cao, X. H., Pavlovski, M., Obradovic, S., Nair, R., et al. (2020). Influence of medical domain knowledge on deep learning for Alzheimer's disease prediction. *Comput. Methods Programs Biomed.* 197:105765. doi: 10.1016/j.cmpb.2020.105765

Luo, J., Ye, M., Xiao, C., and Ma, F. (2020). "Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, (New York, NY: Association for Computing Machinery). doi: 10.1145/3394486.3403107

Ma, F., Gao, J., Suo, Q., You, Q., Zhou, J., and Zhang, A. (2018). "Risk prediction on electronic health records with prior medical knowledge," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, (New York, NY: Association for Computing Machinery). doi: 10.1145/3219819.3220020

Mehta, N. K., Prasad, S. S., Saurav, S., Saini, R., and Singh, S. (2022). Three-dimensional densenet self-attention neural network for automatic detection of student's engagement. *Appl. Intell.* 52, 13803–13823. doi: 10.1007/s10489-022-03200-4

Moeskops, P., de Bresser, J., Kuijf, H. J., Mendrik, A. M., Biessels, G. J., Pluim, J. P., et al. (2018). Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in mri. *NeuroImage* 17, 251–262. doi: 10.1016/j.nicl.2017.10.007

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. (2021). Attention bottlenecks for multimodal fusion. *Adv. Neural Inform. Process. Syst.* 34, 14200–14213.

Nicolls, M. R. (2004). The clinical and biological relationship between type II diabetes mellitus and Alzheimer's disease. *Curr. Alzheimer Res.* 1, 47–54. doi: 10.2174/1567205043480555

Pipitone, J., Park, M. T. M., Winterburn, J., Lett, T. A., Lerch, J. P., Pruessner, J. C., et al. (2014). Multi-Atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. *Neuroimage* 101, 494–512. doi: 10.1016/j.neuroimage.2014.04.054

Plizzari, C., Cannici, M., and Matteucci, M. (2021). "Spatial temporal transformer network for skeleton-based action recognition," in *International conference on pattern recognition*, Vol. 12663, eds A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, et al. (Cham: Springer). doi: 10.1007/978-3-030-68796-0_50

Pratico, D. (2008). Evidence of oxidative stress in Alzheimer's disease brain and antioxidant therapy: Lights and shadows. *Ann. N. Y. Acad. Sci.* 1147, 70–78. doi: 10.1196/annals.1427.010

Pruzin, J. J., Nelson, P. T., Abner, E. L., and Arvanitakis, Z. (2018). Relationship of Type 2 diabetes to human brain pathology. *Neuropathol. Appl. Neurobiol.* 44, 347–362. doi: 10.1111/nan.12476

Ruiz, J., Mahmud, M., Modasshir, M., Shamim Kaiser, M., and Alzheimer's Disease Neuroimaging Initiative ft (2020). "3d densenet ensemble in 4-way classification of Alzheimer's disease," in *International conference on brain informatics*, eds M. Mahmud, S. Vassanelli, M. S. Kaiser, and N. Zhong (Cham: Springer), 85–96. doi: 10.1007/978-3-030-59277-6_8

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., et al. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* 53, 197–207. doi: 10.1016/j.media.2019.01.012

Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. (2017). Deep Ehr: A survey of recent advances in deep learning techniques for electronic health record (Ehr) analysis. *IEEE J. Biomed. Health Inf.* 22, 1589–1604. doi: 10.1109/JBHI.2017.2767063

Spasov, S. E., Passamonti, L., Duggento, A., Lio, P., and Toschi, N. (2018). "A multi-modal convolutional neural network framework for the prediction of Alzheimer's disease," in *Proceedings of the 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, (Piscataway, NJ: IEEE). doi: 10.1109/EMBC.2018.8512468

Stanciu, G. D., Bild, V., Ababei, D. C., Rusu, R. N., Cobzaru, A., Paduraru, L., et al. (2020). Link between diabetes and Alzheimer's disease due to the shared amyloid aggregation and deposition involving both neurodegenerative changes and neurovascular damages. *J. Clin. Med.* 9:1713. doi: 10.3390/jcm9061713

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inform. Process. Syst.* 30.

Wang, X., Peng, Y., Lu, L., Lu, Z., and Summers, R. M. (2018). "Tienet: Text-image embedding network for common thorax disease classification and reporting in Chest X-Rays," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (Piscataway, NJ: IEEE). doi: 10.1109/CVPR.2018.00943

Xu, J., Wang, F., Xu, Z., Adekkanattu, P., Brandt, P., Jiang, G., et al. (2020). Data-driven discovery of probable Alzheimer's disease and related dementia subphenotypes using electronic health records. *Learn. Health Syst.* 4:e10246. doi: 10.1002/lrh2.10246

Xu, W., von Strauss, E., Qiu, C., Winblad, B., and Fratiglioni, L. (2009). Uncontrolled diabetes increases the risk of Alzheimer's disease: A population-based cohort study. *Diabetologia* 52, 1031–1039. doi: 10.1007/s00125-009-1323-x

Yang, Z., and Liu, Z. (2020). The risk prediction of Alzheimer's disease based on the deep learning model of brain 18f-Fdg positron emission tomography. *Saudi J. Biol. Sci.* 27, 659–665. doi: 10.1016/j.sjbs.2019.12.004

Ye, M., Cui, S., Wang, Y., Luo, J., Xiao, C., and Ma, F. (2021). "Medretriever: Target-driven interpretable health risk prediction via retrieving unstructured medical text," in *Proceedings of the 30th ACM international conference on information & knowledge management*, (New York, NY: Association for Computing Machinery). doi: 10.1145/3459637.3482273

Ye, M., Luo, J., Xiao, C., and Ma, F. (2020). "Lsan: Modeling long-term dependencies and short-term correlations with hierarchical attention for risk prediction," in *Proceedings of the 29th ACM international conference* *on information & knowledge management*, (New York, NY: Association for Computing Machinery). doi: 10.1145/3340531.3411864

Yiğit, A., and Işik, Z. (2020). Applying deep learning models to structural mri for stage prediction of Alzheimer's disease. *Turk. J. Electr. Eng. Comput. Sci.* 28, 196–210. doi: 10.3906/elk-1904-172

Yu, K., Qin, X., Jia, Z., Du, Y., and Lin, M. (2021). Cross-attention fusion based spatial-temporal multi-graph convolutional network for traffic flow prediction. *Sensors* 21:8468. doi: 10.3390/s21248468

Zhang, S., Xu, S., Tan, L., Wang, H., and Meng, J. (2021). Stroke lesion detection and analysis in mri images based on deep learning. *J. Healthcare Eng.* 2021:5524769. doi: 10.1155/2021/5524769

Zhang, X. S., Tang, F., Dodge, H. H., Zhou, J., and Wang, F. (2019). "Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, (New York, NY: Association for Computing Machinery), 2487–2495. doi: 10.1145/3292500.3330779

Zhou, Y., Huang, S.-C., Fries, J. A., Youssef, A., Amrhein, T. J., Chang, M., et al. (2021). Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from Ct and Ehr. *arXiv* [preprint] arXiv:211111665

Zhu, T., Li, K., Herrero, P., and Georgiou, P. (2020). Deep learning for diabetes: A systematic review. *IEEE J. Biomed. Health Inform.* 25, 2744–2757. doi: 10.1109/JBHI.2020.3040225

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x