Computer Methods and Programs in Biomedicine 255 (2024) 108348

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Alzheimer's disease diagnosis in the metaverse

Jalal Safari Bazargani [a,1], Nasir Rahim [b,1], Abolghasem Sadeghi-Niaraki [a,1], Tamer Abuhmed [b], Houbing Song [c], Soo-Mi Choi [a,*]

[a] Department of Computer Science and Engineering and Convergence Engineering for Intelligent Drone, XR Research Center, Sejong University, Seoul, Korea
[b] College of Computing and Informatics, Sungkyunkwan University, Suwon, Korea
[c] Department of Information Systems, University of Maryland, Baltimore County (UMBC), Baltimore, MD, 21250, USA

ARTICLE INFO

Keywords:
Alzheimer's disease
Cognitive assessment
Multimodal data
The metaverse
Virtual reality
Artificial intelligence
3DCNN-ML

ABSTRACT

Background and Objective: The importance of early diagnosis of Alzheimer's Disease (AD) is by no means negligible because no cure has been recognized for it rather than some therapies only lowering the pace of progression. The research gap reveals information on the lack of an automatic non-invasive approach toward the diagnosis of AD, in particular with the help of Virtual Reality (VR) and Artificial Intelligence. Another perspective highlights that current VR studies fail to incorporate a comprehensive range of cognitive tests and consider design notes for elderlies, leading to unreliable results.
Methods: This paper tried to design a VR environment suitable for older adults in which three cognitive assessments namely: ADAS-Cog, Montreal Cognitive Assessment (MoCA), and Mini Mental State Exam (MMSE), are implemented. Moreover, a 3DCNN-ML model was trained based on the corresponding cognitive tests and Magnetic Resonance Imaging (MRI) with different modalities using the Alzheimer's Disease Neuroimaging Initiative 2 (ADNI2) dataset and incorporated into the application to predict if the patient suffers from AD.
Results: The model has undergone three experiments with different modalities (Cognitive Scores (CS), MRI images, and CS-MRI). As for the CS-MRI experiment, the trained model achieved 97%, 95%, 95%, 96%, and 94% in terms of precision, recall, F1-score, AUC, and accuracy respectively. The considered design notes were also assessed using a new proposed questionnaire based on existing ones in terms of user experience, user interface, mechanics, in-env assistance, and VR induced symptoms and effects. The designed VR system provided an acceptable level of user experience, with participants reporting an enjoyable and immersive experience. While there were areas for improvement, including graphics and sound quality, as well as comfort issues with prolonged HMD use, the user interface and mechanics of the system were generally well-received.
Conclusions: The reported results state that our method's comprehensive analysis of 3D brain volumes and incorporation of cognitive scores enabled earlier detection of AD progression, potentially allowing for timely interventions and improved patient outcomes. The proposed integrated system provided us with promising insights for improvements in the diagnosis of AD using technologies.

1. Introduction

Cognition is the process through which a person uses the knowledge they have previously gathered to guide their behavior. As the brain ages, cognitive abilities often deteriorate progressively, eventually losing their independence and resulting in dementia. The term "dementia" can be considered an umbrella term to refer to a variety of symptoms, including a deterioration in memory, thinking, language, or perceptual interpretation. Alzheimer's Disease (AD), the most prevalent type of dementia, is brought on by a combination of some of the mentioned symptoms. The World Health Organization predicts that AD and associated dementias will be the cause of 1.37 percent of all deaths globally in 2030. When AD is identified, the neuronal damage has already advanced to the point where it cannot be repaired. In fact, damage cannot be repaired when neurons die since they do not proliferate and replenish themselves as other cells do. Simply put, no treatment has been found to cure dementia up to now [1]. Therefore, early diagnosis of Alzheimer's is of paramount importance. That is the reason behind the

* Corresponding author.
E-mail address: smchoi@sejong.ac.kr (S.-M. Choi).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cmpb.2024.108348
Received 5 March 2024; Received in revised form 29 June 2024; Accepted 20 July 2024
Available online 21 July 2024

fact that screening tests are being increasingly used by doctors to assess patients' conditions.

In comparison to earlier decades, the global population is aging at a faster pace. Although older people who use Virtual Reality (VR) technology are now in the minority, they could benefit from this technology in different ways [2] including daily activities [3]. Moreover, the aging population's health can be maintained through the use of VR as assessment, treatment, or other applications. At present, many neurologists and medical professionals are dedicating significant time to researching methods for early diagnosis of AD and encouraging results have been frequently achieved. A common approach toward that is the employment of Artificial Intelligence (AI) which has shown noticeable advantages including the possibility of having multi-criteria decision--making. Therefore, there is a significant contribution in exploring the employment of VR and other new technologies such as AI in AD diagnosis.

The present study aims to diagnose AD by employing cognitive tests in a VR setting and predicting the results using an AI model. Specifically, a wide range of design considerations is taken into account so that the designed virtual environment became suitable for older adults. Then, the cognitive tests namely: ADAS-Cog, Montreal Cognitive Assessment (MoCA), and Mini Mental State Exam (MMSE) are implemented inside the environment in which the patient can take the test with the help of an examiner in a multiplayer mode. Regarding the AI section, we presented a framework that integrates the capabilities of 3D Convolutional Neural Networks (3D CNN), feed-forward neural networks, and traditional Machine Learning (ML) classifiers, with the aim of outperforming existing methods for AD progression detection in the medical domain. The framework utilizes multimodal data, which consists of a 3D Magnetic Resonance Imagin (MRI) volume and cognitive scores, to predict the medical condition of a patient at 48 months (M48) after the initial assessment. The 3D CNN is designed as a lightweight feature extractor that captures intra-slice features of the MRI volume by analyzing the most critical brain regions and producing a latent representation of the current brain status. Additionally, we evaluated the impact of fusing other critical modalities, such as clinical score biomarkers, with the extracted deep features, and combinedly used them to evaluate five classifiers, including Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), and Deep Neural Network (DNN). It was concluded that each model improves the disease identification process when it sees multiple modalities of the same patient. The key contributions of this paper are as follows:

- The designed virtual environment was performed by considering an array of design notes in order to make the environment appropriate for older adults. The evaluation of the design environment indicated promising positive effects in the related area.
- We converted three important cognitive tests namely: ADAS-Cog, MoCA, and MMSE into a virtual setting.
- We proposed a novel and lightweight AD progression detection framework that addresses the limitations of the existing studies in the AD domain. The given framework predicts the advancement of AD at month 48, which refers to patients' health status after three years.
- This time frame is deemed appropriate for the patient's care providers and family members to attend to their required health preparations.
- To delve deeper into the impact of multimodal data in identifying the progression of AD, we combined the patients' cognitive scores with MRI and used it in the disease identification process. The fusion of modalities was then used to evaluate and improve the selection of classifiers namely RF, DT, LR, SVM, and DNN.
- We conducted various experiments to evaluate the proposed framework in different environments, such as AD progression detection using (1) MRI modality, (2) CSs biomarkers, and (3) combination of MRI + CSs. Results suggest that 3D CNN followed by DNN outperformed all other comparative models when tested with combined modalities. The performance reported were (precision:97%, recall:95%, F1-score:95%, AUC: 96%, and accuracy:94%).

The rest of the paper is as follows: the literature review is covered in Section II, and Section III provides information on the methodology. The results and discussion are provided in Sections IV and V respectively. Finally, Section VI provides us with the conclusion of the paper.

2. Literature review

One way to categorize AD detection approaches is to consider them as two classes namely invasive and non-invasive. With invasive approaches, information must be gathered from the patient's internal organs via procedures such as lumbar punctures or blood draws. While some of these tests are excruciatingly uncomfortable, they are not necessarily safe or comfortable for the patient. On the other hand, the second class includes methods that are harmless and more convenient for the patient. The involvement of new technologies provides opportunities to enhance the current non-invasive cognitive assessments. That is to say, over the past decade, the paper-based tests have been polished and currently, the new technologies make it possible not only to have more enhanced non-invasive tests but also to propose new ones. Additionally, it is stated that diagnosing AD is costly and it is recommended to incorporate e-health concepts in order to obtain a cost-efficient approach toward AD diagnosis [4].

Cognitive tests, which fall into the non-invasive category, are highly accurate at detecting Alzheimer's. In fact, programs targeting cognitive aspects include a variety of exercises that stimulate and/or test cognitive flexibility, executive processes, and spatial memory [5]. Having said that, paper-based tests used in diagnosing AD have been found to lack ecological validity, which means that a person's performance on the exam does not accurately reflect how they operate in everyday life [6]. On the other hand, the potentialities of Virtual Reality (VR) in improving ecological validity and offering a more accurate evaluation of a patient's cognitive performance in a real-world situation have been proved [7]. That is to say, the patient's mind can be manipulated in terms of being misled and tricked as a result of being fully immersed in a virtual environment. Through the use of realistic stimuli, VR creates virtual worlds that simulate the feelings associated with cognitive and physical processes. These qualities play a key role in the portrayal of real-world circumstances depending on the amount of immersion and engagement.

A novel cognitive test using VR is proposed in [8] with the purpose of early AD diagnosis. In the proposed solution, considerations such as targeting at least one of the cognitive domains and benefiting from virtual environments in terms of flexibility have been taken into account. The findings of the paper proved the efficiency enhancement of the current approaches when VR was employed. However, the paper lacks research in designing the environment appropriate to the targeted age group, and the proposed test requires more modifications to be more adaptive to patients.

Tan et al. [9] investigated the performance of VR in cognitive assessment among 35–74-year-old adults classified into four age groups. The evaluation was based on the score of the Montreal Cognitive Assessment (MoCA) and completion time spent performing several daily activities such as shopping. The six cognitive domains namely: perceptual-motor, executive function, complex attention, learning and memory, social cognition and language [10] were targeted in the paper. The findings highlighted noticeable differences between the performance of members in age groups, however, further enhancements were needed to define performance indices for each age group.

Apart from the ecological validity mentioned earlier, VR holds other benefits such as being self-administered, requiring little training, offering joy, and lessening the psychological anguish brought on by utilizing standard assessment methods [11]. Another benefit resulting from virtual settings would be the availability of more data compared to traditional assessment methods. Bourrelier et al. [12] highlighted this

capability by retrieving data such as body movement trajectory, completion time, and speed. The findings of the paper proved that VR-based daily activity assessments outperform traditional ones. Moreover, VR can help explore the role of emotion in MCI rehabilitation systems as a VR-based approach was examined in [13] and it was proved that emotion can be complementary to current treatment systems.

Moving forward with further literature review in terms of AI, in order to clinically diagnose AD at an early stage, a variety of imaging biomarkers are widely investigated such as MRI, Positron Emission Tomography (PET), Functional Magnetic Resonance Imaging (fMRI), Single Photon Emission Computed Tomography (SPECT), and Magnetic Resonance Spectral Imaging (MRSI) [14]. MRI is a highly effective method for brain imaging which does not require invasiveness. It offers a more precise depiction of the brain's size and structure in comparison to Computed Tomography (CT), SPECT and PET. It provides a distinct advantage in terms of marking soft tissues, capturing precise spatial resolution or even identifying very minor abnormalities in the brain tissues. Furthermore, the diagnostic abilities of MRI modality has experienced a significant improvement thanks to the systematic and accurate labeling of MR images. This labeling process is a crucial factor in determining AD and healthy controls [15].

AD is a severe decline in cognitive abilities that is typically diagnosed by specialists using multiple forms of evaluation. Many studies focused on AD diagnosis use only a single modality of medical data, such as PET, MRI, or CS. For example, Bron et al. [16] proposed a computer-aided diagnostic system where they compared a wide array of traditional ML classifiers for the diagnosis of AD patients. This study was conducted based on 29 ML classifiers trained on an MRI modality collected at a single visit of a patient. The highest accuracy achieved was 63% for classifying three categories: CN vs MCI vs AD. In another study conducted by Jiang et al. [17], a lightweight deep CNN model was proposed to classify AD patients. The published architecture comprised an eight-layer deep network incorporating batch normalization and dropout layers to tackle the issue of overfitting. Their proposed framework was trained for binary classification tasks by implementing an MRI modality of CN and AD patients. Zhang et al. [18] proposed a 3D CNN model to separate cognitively normal from diseased patients using whole brain MRI volume. They first extracted all axial slices from the MRI volume and then processed them through a 3D CNN model in volumetric format to analyze a large amount of data simultaneously, thus predicting the patient's health status. They reported an increase in disease diagnostic accuracy when a raw MRI is processed through a standard preprocessing pipeline before applying ML algorithms. The preprocessing steps include inhomogeneity correction, extraction of brain tissues, and calculation of the wavelet entropy of the MRI. Their proposed 3D CNN was composed of a single layered network that was optimized using particle swarm optimizer resulting in an accuracy of 93%.

The diagnosis of AD commonly employs the use of a single modality, such as MRI, PET, or CS. However, recent research has shown that combining multiple modalities, or "multimodal data" can improve the overall accuracy of ML models used in medical diagnosis [19-21]. While MRI is a crucial modality for AD detection, combining it with other forms of data can provide a more comprehensive understanding of a patient's condition and potentially enhance the accuracy of models' prediction. Other modalities such as a patient's medical history, cognitive scores, and neuropsychological features have also been found crucial biomarkers in disease identification. By combining multiple sources of data, it is possible to reduce noise and attain more reliable and accurate results, which are more likely to be accepted by the medical community. Overall, multimodal studies offer a more complete picture of the disease due to the effect of AD on multiple biological processes [22]. Examples of studies that have fused multiple modalities include MRI, cerebrospinal fluid (CSF) tests, PET, and genetic features. Furthermore, utilizing multiple forms of data in medical research has been shown to provide more comprehensive and precise results. This approach also tends to be more widely accepted by the medical community. For example, in a study conducted by Zhang et al. [18], a neural network was proposed that combined data from MRI, CSF, and fluorodeoxyglucose FDG-PET to distinguish between healthy individuals, cognitively impaired, and AD patients. Other studies, such as those by Xu et al. [23] and Huang et al. [24], also employed a multimodal approach, using various forms of imaging and laboratory data to classify individuals with cognitive impairment. Gray et al. [25] integrated data from FDG-PET, MRI, CSF, and genetic analysis in a random forest model for the diagnosis of AD.

Table 1 provides a summary of the literature review in terms of employed methods and technologies. A general perspective reveals information on a lack of study using the combination of VR and AI. Moreover, most VR studies did not cover different cognitive tests which contributed to less reliable results obtained from their proposed solutions. The conducted literature review indicates no similar study to the present paper in terms of considering multiple cognitive tests, design notes for older people, and the integration of VR and AI.

3. Methodology

According to the goal of the paper, which is to diagnose AD in the Metaverse, the general framework can be explained in three main sections: VR, AI, and Evaluation.

3.1. The design of the environment

Despite the improvements made in VR, this technology still poses a wide range of challenges for older adults, cybersickness and discomforts being examples [2]. Generally, age-related characteristics such as the unfamiliarity with VR, vision and hearing influence designing the interface and can result in other challenges [32,33]. Older adults frequently face obstacles as a result of their restricted social, economical, physical, and cognitive resources, and they have to make more effort compared to younger people to grasp new technology [34]. In this regard, it is stated that researchers must correctly recognize older adults as a complex group rather than a homogeneous population [35]. Therefore, several design considerations were taken into account regarding the design of the environment suitable for older adults. These considerations can be categorized into seven groups: visuals, audio, onboarding and assistance, safety, minimizing side effects, usability, and realism. As for visual aspects, it is recommended to have familiar materials in the environment [36] while keeping the number of objects to a low number [37], see Fig. 1.

Moreover, simple, vibrant, and contrasting colors are highlighted to have positive effects on the suitability of the scene for elderlies [38]. Fig. 2 indicates the designed environment by focusing on visual-related considerations such as using simple and contrasting colors for familiar objects. Regarding audio, spatial [39] and ambient [40] sound is considered to be beneficial in virtual environments for elderlies. This

Table 1
The comparison of the related papers and this paper.

Paper	Cog domain	VR	AI	Detection method	Approach
[8]	EF	✓	×	Turning Test	–
[9]	All	✓	×	MoCA	Daily activity
[26]	EF – PM	✓	×	Free-Recall	Daily activity
[27]	All	✓	×	ACE-R	Navigation
[28]	PM	✓	×	ALTODIA-iADL	Obj Placement-finding
[29]	–	×	✓	MRI	RESU-Net
[30]	–	×	✓	MRI	FDN-ADNet
[31]	All	×	✓	Hippocampus features and CS	RNN
Ours	All	✓	✓	MRI and CS(ADAS-Cog, MMSE, MoCA)	3DCNN-ML and VR Version of ADNI2 Manual

Fig. 1. Familiar objects and low number of objects in the designed environment.

Fig. 2. Using simple and contrasting colors in designing the environment.

feature will, simply put, make older adults feel more familiar with their surroundings. Thus, a fireplace, as shown in Fig. 3, was incorporated into the scene to act as background ambient sound, and all sounds were turned into the spatial mode in order to provide the participants with some orientation information.

In order to target safety and onboarding and assistance, the researcher's presence is recommended [38,41]. For this purpose, the multiplayer option was added to the environment so that both the patient and the researcher could be present at the same time. As for other considerations in onboarding and assistance, introduction to the equipment is stated useful for designing a VR environment for older adults [39] which was met at the beginning of the experiment. Furthermore, the introduction session involved explanations of how to use hands and controllers for interacting with the objects [39]. As for the next consideration group, which is minimizing side effects, three notes were taken into account: minimizing stress [42], continuous communication [43], and clear instructions [44]. With the help of multiplayer mode in the environment, the patient and the doctor can communicate using voice. This, in turn, supports not only continuous communication during the experiment but also clear instructions since the doctor can clarify the required tasks to the patient to an extent that is allowed according to the experiment manual. Furthermore, an appropriate user interface that offers simplicity and high readability using larger fonts [37] was designed to facilitate the interaction of older adults with the features implemented in the environment. In addition, minimizing the use of controllers was considered to increase the usability aspect of the design [35]. In this regard, all features were implemented in a way that can be triggered using controllers and bare hands. The engine will automatically switch to hand tracking if no movement in controllers is detected. As for the locomotion system of the application, three types of navigation were implemented: teleportation through hands and continuous movement and snap turn through controllers. Lastly, avatars were selected to be the representative of the users in the environment. The human-like avatars enhance the realism aspect of the design which is necessary for our target audience [39].

Fig. 3. The fireplace offers ambient sound for the background.

3.2. The conversion of cognitive tests to VR

During the literature review, several cognitive tests were examined in terms of their suitability to be implemented in a VR setting, and it resulted in the selection of Alzheimer's Disease Neuroimaging Initiative 2 (ADNI2) as the reference manual. This manual covers different characteristics of the entire spectrum of AD to be monitored and examined. The cognitive assessment section of the manual includes a thorough list of cognitive tests categorized into different stages such as baseline, month-6 visit, and annual visit. By studying the procedure of taking cognitive tests in each stage, three main cognitive tests were chosen to be implemented inside the VR environment: ADAS-Cog, Montreal Cognitive Assessment (MoCA), and Mini Mental State Exam (MMSE). The reasons behind choosing these three tests were the fact that they were almost repetitive in all ADNI2 stages, and they covered the entire cognitive domain.

Among five clinical stages in ADNI2, ADAS-Cog, MMSE, and MoCA were employed in four stages which is a testimony to the importance of these cognitive tests. ADAS-Cog was implemented partially, while the other two cognitive tests were incorporated completely into the environment. Constructional praxis, naming task, orientation, word recognition, and number cancellation in the ADAS-Cog were implemented according to the manual. Regarding the ideational praxis stage, since it includes tasks such as folding a paper and putting it into an envelope, the authors decided to manipulate this stage of the test and replace it with other commands. That was done mainly because the folding of a paper in a virtual setting might not be as realistic as it is in the real world, so in order to avoid this negative point, modification was conducted. The other two cognitive tests were implemented completely according to the manual. The MMSE has 10 stages namely: orientation, immediate recall, attention, delayed recall, language naming, command, repetition, reading, writing, and construction. The MoCA covers a total of 11 stages namely: alternating trail making, visuconstructional skills (cube), visuconstructional skills (clock), naming, memory, attention, sentence repetition, verbal fluency, abstraction, delayed recall, and orientation. The high number of stages supports covering different cognitive domains so that the final screening output will be more reliable.

3.3. The training of the AI model

3.3.1. Material and methods

The technique described in this section aims to monitor the progressive patterns of AD through a combination of MRI and CSs. The

framework, displayed in Fig. 5, incorporates necessary preprocessing steps for each MRI scan which entails eliminating artifacts and converting the data into a standard format. The refined MRI scans are then processed through a 3D CNN followed by classical ML algorithms to diagnose the current state of AD. A 3D CNN is capable of extracting high-level, representational features including both spatial and temporal aspects from multiple MRI slices, while the ML classifiers use the extracted features set to detect the health status of a patient by delivering binary outcomes. These outcomes signify whether the individual will experience AD or will stay cognitively normal.

3.3.2. Dataset

The dataset used in this article was acquired from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [45]. The ADNI was established in 2003 under the direction of the principal investigator Michael W. Weiner. The purpose of ADNI is to examine if a combination of MRI, PET, biological markers, clinical assessments, and neuropsychological evaluations can be utilized to track the advancement of MCI and AD patients. Furthermore, detecting significant biomarkers in the early stages of AD would aid medical professionals and researchers in creating new therapies, evaluating their effectiveness, and minimizing the cost clinical of trials.

This study was conducted using 564 MRI images along with cognitive scores. No private information or patient identities were disclosed during this study, as the data were anonymized/de-identified and provided by the ADNI, ensuring compliance with data protection and privacy regulations. The obtained MRIs were 3T T1-weighted anatomical sequences captured using volumetric 3D MPRAGE protocol with a voxel resolution size of $1 \times 1 \times 1$ mm. We extracted multiple coronal slices from the preprocessed MRI volume, since coronal planes have shown better representative features of the crucial brain regions [46] (i.e., hippocampus and subfields of the amygdala) that are highly vulnerable to neurodegenerative disease. We collected 110 slices from the coronal plane by determining the upper and lower slices corresponding to the middle slice of the MRI volume. Through this way, the most important information is taken from the 3D scan.

The criteria we followed to determine the eligibility of a subject for participation in this study were as follows:

- The AD progression of a patient must be diagnosed within 2.5 years by a physician. This ensured the enrollment of participants with early and active disease progression.
- The CN subject had:
  - Mini-Mental State Examination (MMSE) scores between 24 and 30, inclusive. These scores indicate normal cognitive function.
  - A Clinical Dementia Rating (CDR) of 0, indicating no signs of dementia.
  - No history of depression, Mild Cognitive Impairment (MCI), or any form of dementia.
- The AD patients had:
  - MMSE scores between 20 and 26, inclusive, reflecting mild to moderate cognitive impairment.
  - A CDR of 0.5 or 1.0, which corresponds to very mild or mild dementia.
  - Conformity to the National Institute of Neurological and Communicative Disorders and Stroke/Alzheimer's Disease and Related Disorders Association (NINCDS/ADRDA) criteria for probable AD.
- Subjects with 3D MRI scans using a 3T scanner were selected. This high-resolution imaging technique is crucial for accurately capturing brain structures and changes related to AD.

Our proposed 3D CNN leverages the 3D nature of the MRI volume consisting of the most crucial 2D slices and detects the key features relevant to the progressive patterns of AD. Our training dataset consists of 564 individuals, with 282 people being cognitively normal and the remaining 282 being AD patients. All of the MRI scans were processed through a standard preprocessing steps as depicted in Fig. 5.

3.3.3. Image preprocessing

The preprocessing of MRI scans involves removing irrelevant information from the raw data, allowing for easier comparison of multiple brain scans. In this study, all MRI volumes were processed before any experimental steps were taken. This included registering 2D slices to a standard template space, correcting inhomogeneities in 3D scans, separating skull from brain tissue, and aligning 2D slices to a standard template space.

In the first step, we used FreeSurser's Freeview tool [47] to visualize the data and found that some of the raw MRI scans were rotated by 180 ° during the formation phase. This step is essential for ensuring uniformity in the processing and analysis of voxels across various platforms. To reorient MRI scans in the correct view space, we used the fslreorient2std software package to align flipped slices to the standard format. In the next step, all MRI scans were passed through N4 bias field correction algorithm available in Advanced Normalization Tools (ANT) [48] to remove inhomogeneities. Inhomogeneities refer to low-frequency signals within the MRI scanner which can negatively impact the overall quality of the images and must be corrected in the first place. In the third step, we performed skull stripping to distinguish skull from non-skull regions. Non skull regions refer to residual neck voxels and can interfere with the classification task by adding noise and increasing the dimensionality of the training data. We used the brain extraction tool (BET) [49] from FSL package to perform skull stripping. Finally, all MRI volumes were registered to a standard reference space, the MNI152 template [50]. The registration process involved transforming the MRI scans through affine transformations such as scaling, rotating, translating, and shearing. In this study we used FLIRT tool from FSL to register MRI scans to MNI152 template and correlation ratio was used as the similarity metric during the registration process.

3.3.4. Experimental setup

The proposed approach improves the diagnostic process of AD progression by combining MRI and clinical scores. Experiments were conducted on a workstation equipped with NVIDIA GeForce GTX 1060 and 16 GB of RAM. We tuned several hyperparameters that include learning rate, training batch size, number of epochs, and best performing optimizer. 5-fold stratified cross-validation was used to balance subject distribution during training. The learning rate was set to 0.00015 with momentum and weight decay values of 0.99 and 0.1e-5 respectively. Using the Adam optimizer, the model reached the lowest training-testing loss. The input batch size was set between 5–8 with no significant improvement seen above or below this range. The number of epochs was set to 120, with the training loss reaching its lowest point at 100 epochs and not decreasing further.

Evaluation metrics: To determine the model's performance, the generalization ability was evaluated using stratified k-fold cross-validation. This method, with 5 folds, helps maintain the class distribution balance and minimize the chance of biased learning. To measure the model's goodness-of-fit, we compute five standard evaluation metrics including precision, recall, F1-score, AUC and accuracy. These metrics are frequently used in bioinformatics research and serve as a comparison tool for different studies [14,51]. The accuracy of a model is calculated as the percentage of correctly classified instances among all predicted data instances. Precision measures the accuracy of positive predictions, while recall, also known as sensitivity, computes the ability to correctly identify actual positive instances. The F1-score combines two metrics i.e., precision and recall and calculates the harmonic mean of these two measures. Another widely used evaluation metric is the area under the curve (AUC), which takes into consideration both true positive and false positive rates. An AUC value of 0.5 would indicate random guessing, while a value of 1.0 would indicate a perfect classifier. The mean AUC is determined by comparing the average likelihood of

correctly classified elements from one category compared to another.

Model Architecture: In this study, we employed a grid-search based hyperparameter tuning technique to optimize the performance of the underlined framework. We conducted a thorough series of experiments to refine the critical parameters of the model. We adjusted various parameters in the proposed deep convolutional framework that includes number of 3D conv layers, kernel size in each layer (e.g., 3×3, 5×5), the regularization coefficient applied to each layer, the number of dense layers, and the number of dense units in each layer. The backbone network is a lightweight 3D CNN that by nature can capture spatial and temporal features simultaneously from 3D volumetric data. In this study, we used 3D MRI volume composed of 110 2D middle slices that cover the most important brain regions highly prone to AD. The architectural design of the proposed 3D CNN is composed of convolutional, max-pooling and dropout layers. After highly optimizing the proposed 3D CNN, we came up using kernel size $3 \times 3 \times 3$, and pooling size $2 \times 2 \times 2$ throughout the network. Our proposed CNN is composed of five convolutions and three max-pooing layers. The first two convolution layers utilize 32 kernels followed by a max-pooling and dropout layer. The output of the initial two layers was regularized by a dropout threshold of 30%. In the third and fourth layers, 32 and 64 3D kernels were applied to the output of the previous layers respectively. Afterward, a dropout layers of 30% were applied to the output feature maps. In the fifth layer, 64 kernels followed by a 20% dropout coefficient were applied to the output feature maps. The ReLU activation function was applied to all hidden layers due to its simpler design that helps train a network faster. It is important to note that, we do not use maxpooling layer after every convolution layer but after two layers in order to maintain the spatial features of 2D slices in the deeper layers. On the other hand, this approach may lead to a longer training time or even models' overfitting. To address this problem, we used dropout and weight regularization techniques to prevent overfitting and reduce training time. This strategy helps stabilize the training process and avoid the corruption of the feature space. The output from the fifth convolution layers is flattened to 1D feature vector which represents the compact features set of the entire 3D MRI volume. This features vector is further fused with CSs and used for evaluating a group of ML classifiers including RF, DT, LR, SVM, and DNN. The motivation behind various ML classifiers with different modalities was to select the most informative and discriminative feature set in the disease identification process.

3.4. The design of the questionnaire

As mentioned earlier, a wide range of notes was taken into account in designing the VR environment for older adults. As for evaluating the suitability of the environment, a new questionnaire was produced using existing questionnaires covering different aspects of the design based on a 5 Likert scale. Utilizing questionnaires to test a product's usability is an affordable and effective approach. The designed questionnaire contains a total of 27 questions categorized into five sections namely: User Experience (UX), User Interface (UI), mechanics, in-env assistance, and VR induced symptoms and effects (VRISE). These sections were created by modifying and integrating three questionnaires namely: Virtual Reality Neuroscience Questionnaire (VRNQ) [52], User Experience Questionnaire (UEQ) [53], and User Interface Questionnaire (UIQ) [54]. The VRNQ includes 20 questions providing some information about the overall level of the VR software quality as well as subcategories about UX, game mechanics, in-game assistance, and VRISE. The primary objective of the UEQ is to quickly and accurately quantify UX in terms of usability. Lastly, the UIQ evaluates the interface in terms of Perceived Usefulness (PU), Perceived Ease of Use (PEU), Perceived Performance (PP), Expectations, Confirmation, Satisfaction, Continuance Intention (CI), and Interface Quality (IQ).

Concerning the new questionnaire in this paper, the UX involves six questions, five of which were obtained from modifying corresponding questions in the VRNQ monitoring the degree of immersion, quality of experience, graphics, sound, and VR hardware factors. The sixth question is about stimulation and is obtained from UEQ. As for the UI section, a total of eight questions were used according to Ref. The implemented features in the VR environment were assessed in the mechanics section, in which there are six questions about the navigation system (teleportation without controllers and continuous movement with controllers) and interactions with the objects. Questions 1 to 5 and question 6 were obtained and modified respectively from VRNQ and UEQ. Regarding the assistance provided for the patient inside the environment, three questions about audial instructions, visual instructions, and prompts were employed from VRNQ. For the last section of the questionnaire, patients were asked about four induced symptoms and effects namely: nausea, disorientation, dizziness, and instability.

3.5. Procedure

The procedure of the experiment began with some initial introduction sessions for the participants after they had approved the consent form. Then, they were asked to mimic the responses of the ADNI2 dataset so that the responses could be sent to the AI model as input features. Fig. 4 indicates two stages from MoCA and MMSE cognitive tests as examples. As can be seen, two avatars – a patient and an examiner – were present in the environment, performing the procedure. At the final stage, a total of 12 participants were selected randomly and asked to fill out the designed questionnaire.

4. Results

4.1. AD diagnosis

We evaluated our model by conducting three experiments. The aim was to investigate models' performance using a single modality (i.e., MRI or CSs) or multimodality (MRI + CSs) in training input data. The experiments were carried out in three steps: 1) MRI-based AD progression detection, 2) CS-based AD progression detection, and 3) progression detection using a combination of MRI and CSs. The role of single and multimodal data in AD diagnosis is depicted in Fig. 6. The main purpose of designing the experiments in this way is to show the impact of each modality in the overall performance of the studied models. To enhance the reliability of a model, a stratified 5-fold cross-validation technique was used for the models' evaluation. To prevent data leakage, the MRI scans used in training were not repeated in the testing process. The performance of the classifiers was assessed by computing and comparing the average performance of each model using five commonly used evaluation metrics: precision, recall, F1 score, AUC, and accuracy.

Experiment 1. MRI based AD progression detection

In this experiment, we evaluated each ML classifier using MRI modality only. Then the results of multiple evaluation metrics were recorded to assess the performance of each model. To determine the performance and stability of each network, we also compared the AUC score of each network.

Table 2 shows the results of the experiment that employed various ML classifiers utilizing deep features obtained from a 3D CNN. The performance of five models: RF, DT, LR, SVM, and DNN were compared. The implementation of the four classifiers was done using scikit-learn 1.2.0 and Python 3.8. The DNN was implemented using PyTorch 1.1.2. We randomly split the training data into 80% training set and 20% testing set at runtime using a stratified cross-validation technique at each training fold. We trained the DNN part of the proposed framework in an end-to-end manner and the testing results for each fold were collected and averaged. In the case of training ML classifiers, we employed a grid search technique to perform hyperparameter tuning using deep features obtained from the proposed lightweight 3D CNN. We repeated each experiment five times and reported the average results for each metric along with the standard deviation value.

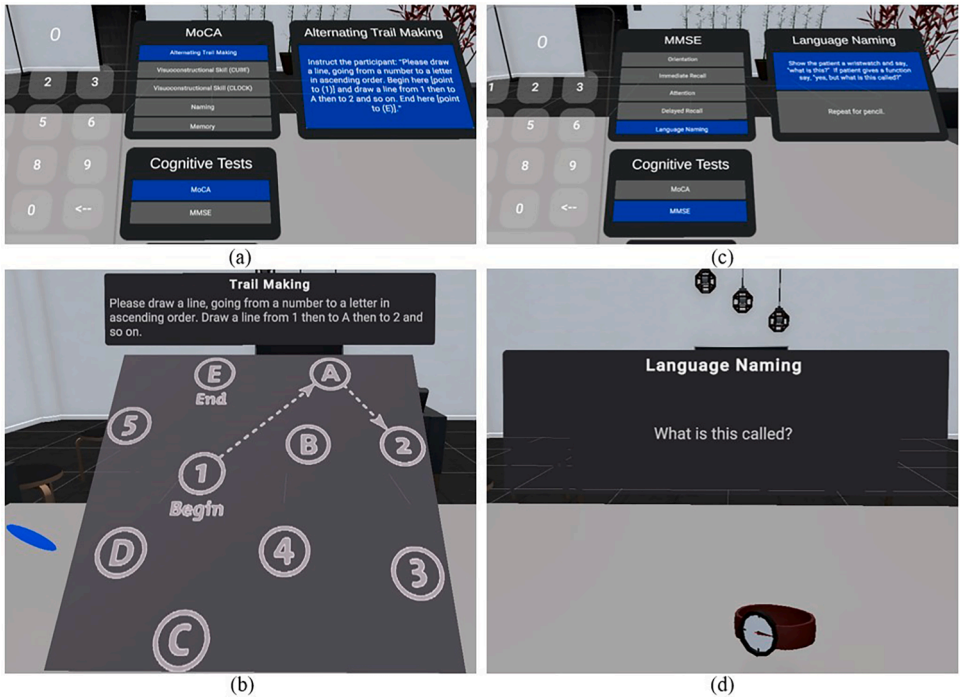

Fig. 4. Two examples of the implemented cognitive tests, (a) Trail Making in MoCA - Examiner's View, (b) Trail Making in MoCA - Patient's View, (c) Language Naming in MMSE - Examiner's View, (d) Language Naming in MMSE - Patient's View.

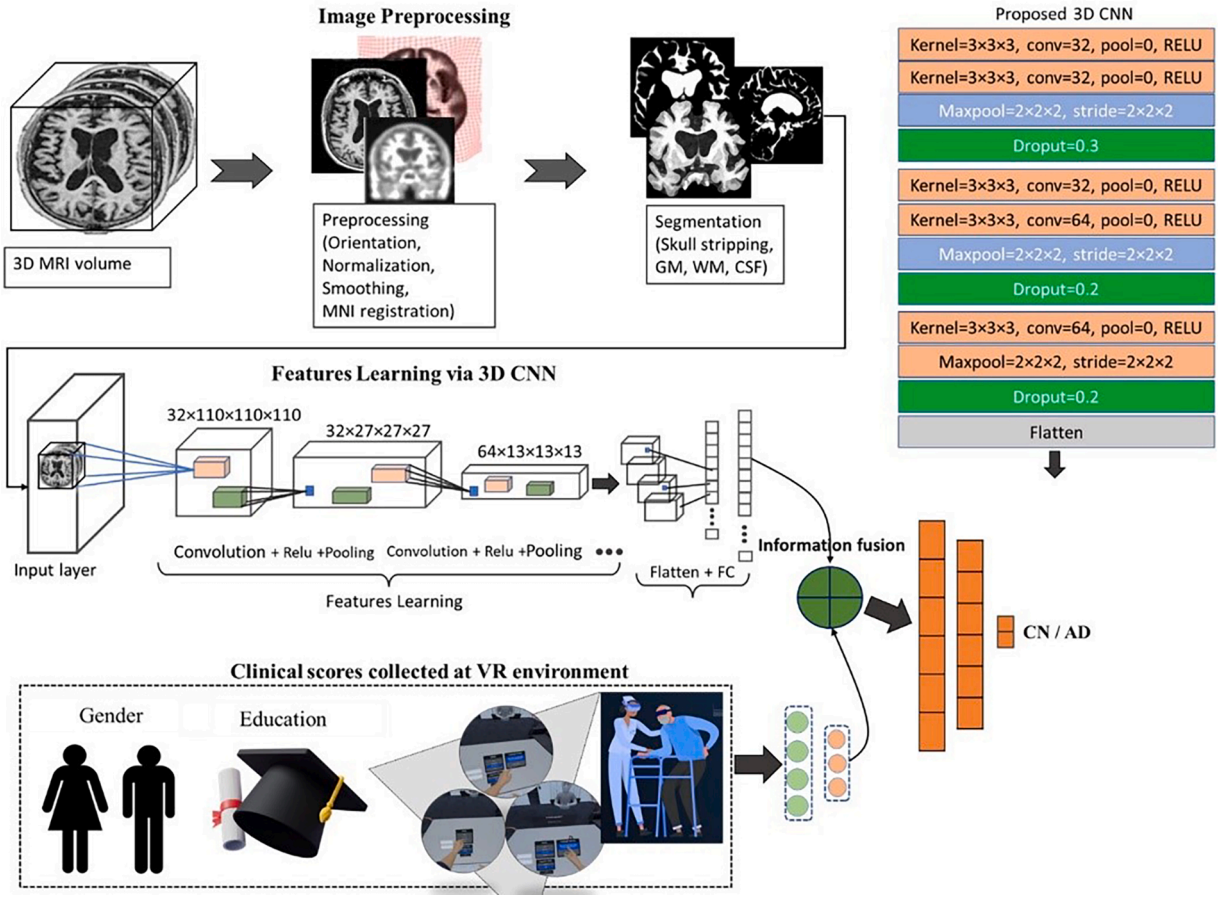

Fig. 5. Proposed framework of 3D CNN fused with clinical data for AD progression detection.

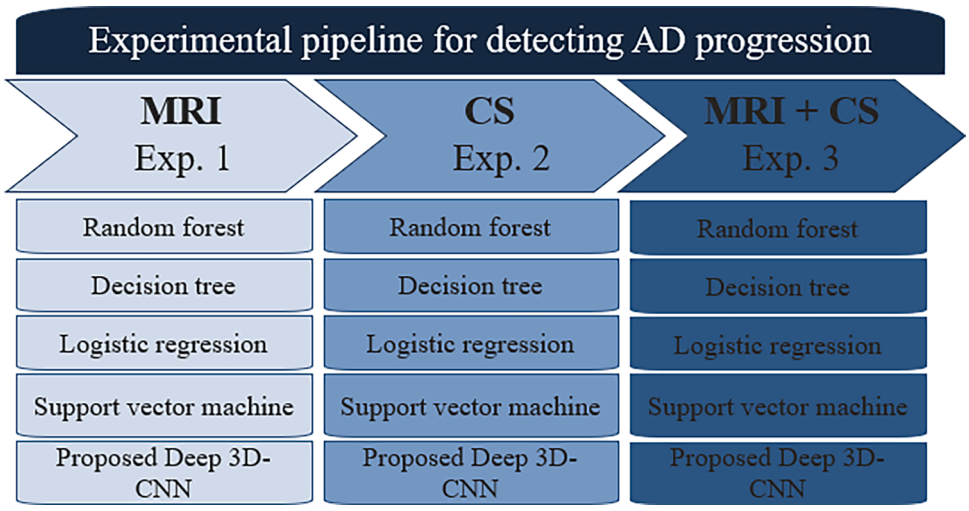

Fig. 6. An experimental route map featuring single (MRI, CS) and multimodal (MRI + CS) medical data.

Table 2
Comparison of different ML classifiers using MRI input data.

ML Model	Mean Precision	Mean Recall	Mean F1-score	Mean AUC	Mean Accuracy
RF	84 ± 0.05	80 ± 0.06	81±0.04	81 ± 0.05	79 ± 0.03
DT	81 ± 0.04	79 ± 0.05	80±0.05	79 ± 0.06	79 ± 0.05
LR	89 ± 0.05	86 ± 0.04	87 ± 0.05	86 ± 0.04	85 ± 0.05
SVM	86 ± 0.04	87 ± 0.05	87±0.06	85 ± 0.05	86 ± 0.03
DNN	89 ± 0.04	88± 0.04	88±0.04	87 ± 0.04	87 ± 0.04

As depicted in Table 2, the highest accuracy was achieved by the DNN model, with a precision of 89 ± 0.04, recall of 88 ± 0.04, F1-score of 88 ± 0.04, AUC of 87 ± 0.04, and accuracy of 77 ± 0.04. The LR and SVM models reported the second-highest accuracy, with a precision of 89 ± 0.05, recall of 86 ± 0.05, F1-score of 87 ± 0.05, AUC of 86 ± 0.04, and accuracy of 85 ± 0.05 for LR, and a precision of 86 ± 0.04, recall of 87 ± 0.05, F1-score of 87 ± 0.06, AUC of 85 ± 0.05, and accuracy of 86 ± 0.0 for SVM, with slight variations in different evaluation metrics. The RF model reported better performance than the DT, but less than all other classifiers, with a precision of 84 ± 0.05, recall of 80 ± 0.06, F1-score of 81 ± 0.04, AUC of 81 ± 0.05, and accuracy of 79 ± 0.03.

Fig. 7 illustrates a performance comparison of five ML models using the MRI modality, with the AUC as the evaluation metric. The LR, SVM, and DNN models achieved an AUC score above 80% when using the MRI modality alone. In particular, the LR model achieved the second-highest AUC score, at 86%, while the DNN model outperformed all other comparative models by achieving the highest AUC score of 87%. The SVM model showed a better AUC score than the DT and RF models, at 85%, but lower than the LR and DNN models. When using the MRI modality alone, the DT model achieved the lowest AUC among all classifiers, at 79%.

Experiment 2. AD progression using CS modality

Experiment 2 shows the impact of incorporating clinical scores on the detection of AD progression. We employed the same set of ML classifiers as discussed in Experiment 1 and incorporated clinical scores in the analysis. The primary objective was to assess the significance of clinical scores in the disease identification process.

As shown in Table 3, all ML classifiers demonstrated a significant improvement in disease identification compared to using only MRI modality. LR, SVM, and DNN models achieved performance scores above 90% in all metrics. This outcome highlights the importance of clinical scores in disease identification. The DNN model outperformed the performance of all other classifiers, achieving precision of 95±0.05, recall of 94±0.02, F1-score of 95±0.04, AUC of 94±0.04, and accuracy of 94±0.03. SVM also reported a significant improvement in overall

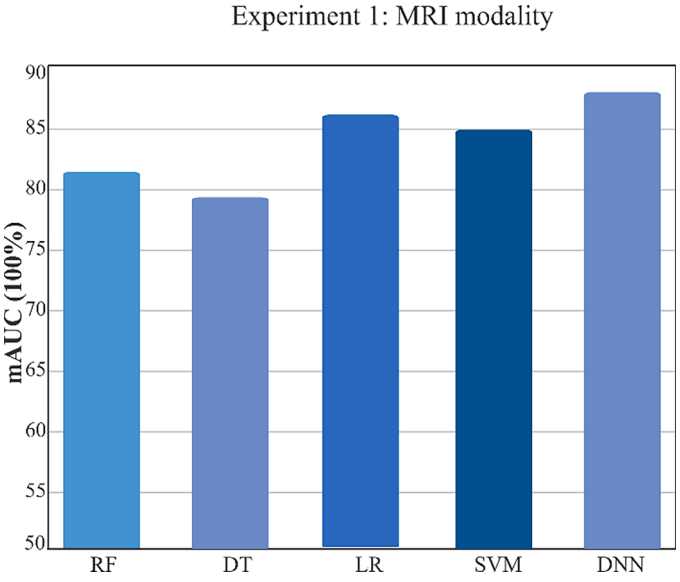

Fig. 7. Performance comparison of the different ML classifiers using mAUC metric and MRI modality.

Table 3
Performance comparison of different ML classifiers using mAUC metric and clinical scores.

ML Model	Mean Precision	Mean Recall	Mean F1-score	Mean AUC	Mean Accuracy
RF	86 ± 0.04	86 ± 0.05	86 ± 0.04	85 ± 0.05	86 ± 0.04
DT	89 ± 0.05	88 ± 0.04	89 ± 0.04	87 ± 0.03	89 ± 0.02
LR	93 ± 0.03	92 ± 0.03	91 ± 0.04	90 ± 0.04	92 ± 0.04
SVM	93 ± 0.05	93 ± 0.04	93 ± 0.04	92 ± 0.04	94 ± 0.03
DNN	95 ± 0.05	94 ± 0.02	95 ± 0.04	94 ± 0.04	94 ± 0.03

accuracy compared to the MRI modality alone, with precision of 93 ±0.05, recall of 93±0.04, F1-score of 93±0.04, AUC of 92±0.04, and accuracy of 94±0.03. LR also crossed 90% milestone in each metric but the accuracy was lower compared to SVM and DNN. Other classifiers, such as DT, also exhibited significant improvements in overall performance compared to using MRI modality only and reported performance in the range 85–90% in each metric.

The numerical features utilized for this study offer significant

insights into the detection and diagnosis of AD from both medical and ML perspectives. Clinical assessments, including medical history, physical examination, cognitive tests like the MMSE or MoCA, and behavioral assessments, are the most essential biomarkers for detecting and monitoring AD by providing a comprehensive view of a patient's cognitive and functional status. Biomarkers, such as amyloid beta and tau proteins detected through CSF analysis, PET imaging, and blood-based assessment, offer objective, quantifiable data for early AD detection and progression tracking. Other features such as neuropsychological evaluations assess cognitive functions like memory, executive functioning, attention, language, and visuospatial skills, crucial for differentiating AD from other cognitive disorders and tailoring intervention strategies. For ML algorithms, these features collectively enable accurate and robust models for early detection, diagnosis, and monitoring of AD, outperforming models based on neuroimaging modalities only i.e., MRI.

Fig. 8 shows the performance comparison of five ML classifiers using AUC metric and CSs as sources of training data. As compared to using the MRI modality, each model achieved a significant improvement in the disease identification process in terms of achieved accuracy. The DNN model, in particular, outperformed all other models in terms of achieved AUC scores, with an 8% improvement in mean AUC scores compared to using the MRI modality alone. Other models also showed a 5–10% improvement in mean AUC scores in the achieved results. The significant improvement of each model with the inclusion of cognitive scores highlights the importance of chosen clinical scores in the disease identification process. Furthermore, each model was able to consider important features during the training process that better represent patients' health status during the progression of AD. While the achieved mAUC scores using CS improved overall accuracy compared to the MRI modality, we observed unstable behavior characterized by large variance in reported accuracies in Experiment 1 and Experiment 2. Literature studies indicate that using multimodality in disease diagnosis enhances the diagnostic process compared to single data modality usage. Therefore, in Experiment 3, we explored the role of multimodality AD diagnostic process.

Experiment 3. AD progression using multimodal data (MRI + CS)

The primary objective of Experiment 3 was to optimize all comparative models using multimodal data particularly combining MRI and CS features. The goal was to investigate AD's progression detection by merging information about patients' cognitive abilities with MRI features, which in turn was used to improve the overall accuracy of AD diagnosis. In order to examine the progressive patterns of AD using multimodal data, we trained each model using CSs fused with MRI modality. In this way, the model was able to identify the key patterns from each modality that is commonly used in the disease identification process. The performance of different classifiers was further evaluated using the AUC metric as discussed previously.

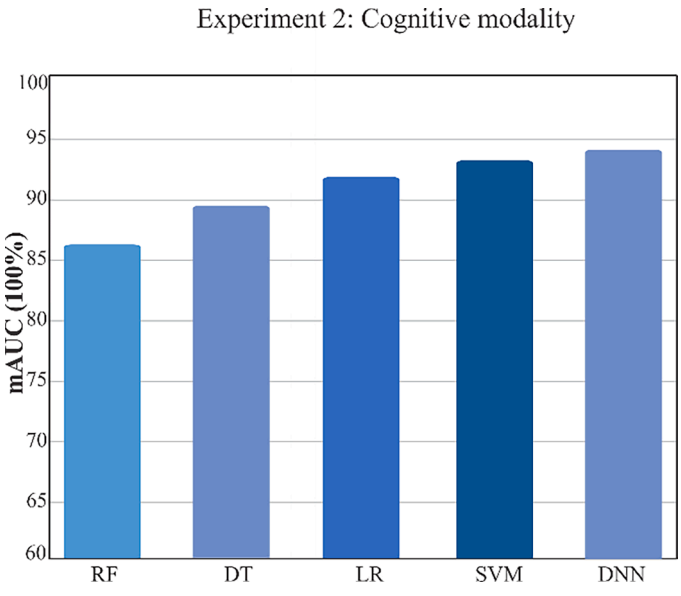

Fig. 8. Comparison of different ML classifiers using clinical scores.

The results, shown in Table 4, demonstrate that by combining CSs with MRI data, the DNN model outperformed all other models, with a precision of 97 ± 0.02, recall of 95 ± 0.02, F1-score of 95 ± 0.02, AUC of 96 ± 0.02, and accuracy of 94±0.03. The SVM model also showed consistent results across all metrics, with a precision of 94 ± 0.03, recall of 95 ± 0.02, F1-score of 94 ± 0.03, AUC of 93 ± 0.04, and accuracy of 94 ± 0.03. Both the LR and DT models achieved almost equal results when using multiple modalities, and overall showed significant improvement compared to using single modalities of MRI or clinical scores alone. In general, combining multiple modalities represents more complex features than using a single modality which makes the task of identifying the disease more challenging. However, all models were able to utilize the fused datasets to improve their performance, indicating that they are robust to noise in the data and can effectively avoid its effects. In addition, Table 4 highlights that using multiple modalities of data is crucial in improving the overall performance and stability of ML models, as well as helping reduce the variance compared to using a single modality. From a medical perspective, it is logical to investigate various modalities to achieve precise diagnosis. In terms of machine learning, it indicates that the addition of multimodal data provided supplementary information to the resulting feature set, thereby aiding the models in refining their decision boundaries. The reported results in this study are in line with earlier studies that supported the positive aspect of multimodal data in enhancing the performance, stability, and effectiveness of a trained model. By incorporating multimodal data, the suggested model gains insight into the fundamental patterns linked to disease progression from each modality throughout the training phase.

Fig. 9 demonstrates the impact of using multiple modalities on the detection of AD progression by displaying the mAUC scores achieved with multimodal data, specifically, the combination of MRI and CSs as training input. The results indicate that utilizing multimodal training data significantly improves the disease identification process for all models. The DNN model outperforms all other models, achieving the highest accuracy. The reported precision in the DNN model improved by 3% and 8% when compared to using cognitive scores or MRI modality alone. Similar improvements were observed in other classifiers, such as SVM, which reported a 6% improvement, and LR, which reported a 5% improvement in precision when compared to using cognitive scores alone. DT and RF also showed 4% and 6% improvements in precision scores respectively, when compared to using cognitive scores alone. Through the experiments, it was concluded that combining cognitive scores with MRI features significantly boosted the stability and performance of disease diagnosis models [55]. Biomarkers offer higher accuracy for early detection due to their sensitivity to biochemical changes,

Table 4
Comparison of ML classifiers using multimodal data (MRI + clinical scores).

ML Model	Mean Precision	Mean Recall	Mean F1-score	Mean AUC	Mean Accuracy
RF	92 ± 0.02	90 ± 0.02	91 ± 0.02	90 ± 0.02	90 ± 0.03
DT	94 ± 0.03	93 ± 0.02	93 ± 0.03	94 ± 0.04	92 ± 0.03
LR	93 ± 0.04	94 ± 0.03	93 ± 0.03	94 ± 0.02	93 ± 0.03
SVM	94 ± 0.03	95 ± 0.02&	94 ± 0.03	93 ± 0.03	94 ± 0.02
DNN	97 ± 0.02	95 ± 0.02	95 ± 0.02	96 ± 0.02	94 ± 0.03

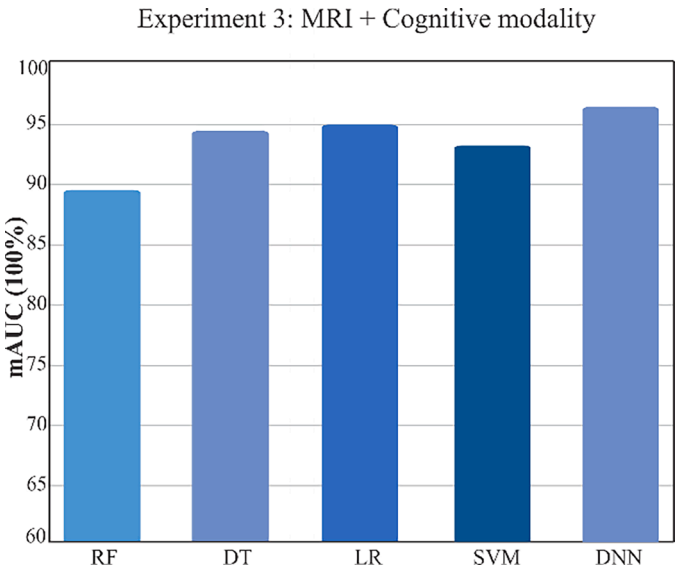

Fig. 9. Performance comparison of different ML classifiers using mAUC metric and multimodality (MRI + clinical scores).

while MRI provides important structural information. The integration of clinical assessments, biomarkers, and neuropsychological evaluations with the structural details of brain tissues offers a comprehensive approach, enabling precise diagnosis and effective tracking of AD progression. For the most accurate diagnosis, a multimodal approach that includes both biomarkers and MRI is recommended.

4.2. Designed environment

The design notes mentioned earlier were evaluated and the results are indicated in Fig. 10. The responses to the UX section of the questionnaire highlight an acceptable level of experience. In fact, it can be understood that the system performed well in motivating the participants, offering them an enjoyable experience as well as a feeling of immersion. However, the quality of graphics and sound seems to be not completely satisfying. Moreover, using the HMD for long periods appears to receive different feedback, mostly feeling almost comfortable. The last three mentioned results indicate a need to explore graphics, sound, and comfort aspects for offering a VR system for older people. The next box plot, displayed in Fig. 10, is about the UI. A general overview states the success of the designed UI. The results illustrate that aspects namely: PU, PEU, Expectations, Confirmation, and IQ received more positive feedback compared to other aspects. It seems the participants found the interface a bit challenging to navigate through, a negative impact on PP. Satisfaction and confirmation were also influenced probably because of the navigation and received less positive feedback. Moving forward with the mechanics of the proposed VR system, all aspects received a score of three to five which can be a testimony to the considered design notes. Regarding the navigation system in the environment, participants found continuous movement using controllers easier than teleportation using hands. On the other hand, they took the idea that manipulating objects was easier by hand rather than by controllers. Furthermore, it seems the users could easily use the items in the environment. Having said that, according to the last box plot of the mechanics section, it can be inferred that most users stated a need for additional help in order to perform the tasks.

In terms of in-env assistance, all three aspects namely audial and visual assistance along with prompts received almost similar feedback. That said, the users found prompts as the most effective assistance approach, followed by visual and audial instructional assistance respectively. Finally, as shown in Fig. 10(e), the designed VR environment performed acceptably in terms of inducing symptoms and effects, nausea, disorientation, dizziness, and instability. The system prevented the users from experiencing nausea and instability almost to an equal level. However, dizziness seemed to happen among the users while experiencing the environment, and disorientation was also another induced effect.

5. Discussion

In this study, we proposed a lightweight 3D CNN that extracts high-level, spatiotemporal features from 3D MRI volumes. These features were then used to evaluate the performance of various ML classifiers. The 3D CNN extracts features from both the MRI volume and the CSs and combines them to create a feature vector, which is then fed into the ML classifiers for the detection of AD progression. Our framework was evaluated using MRI, CS, and a combination of MRI and CS data, and it was found that the combination of both data sources produced the most

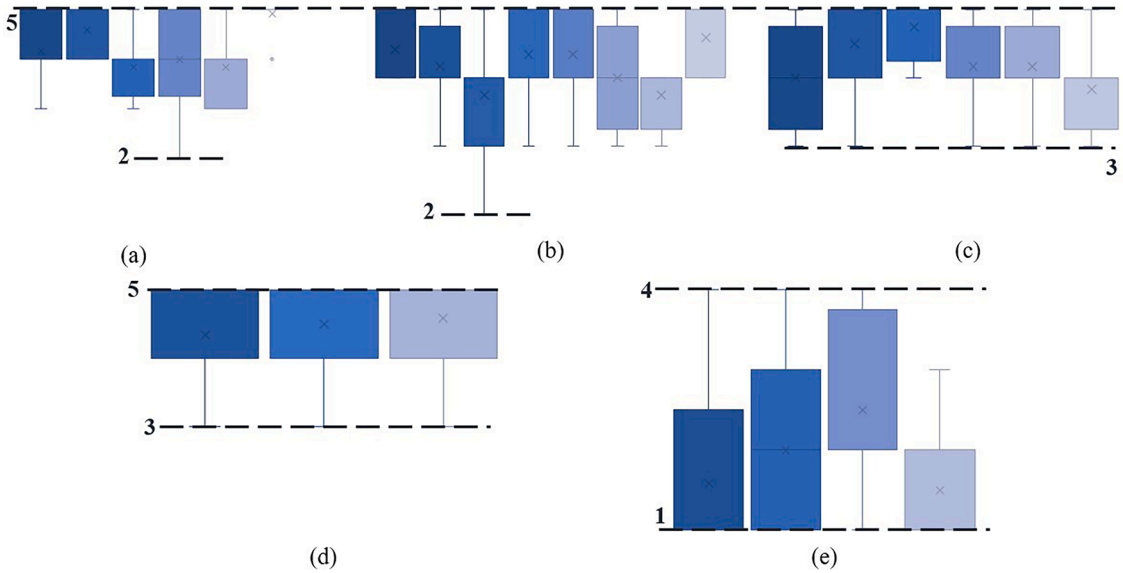

Fig. 10. Responses to the questionnaire in the form of box plot, (a) User Experience, (b) User Interface, (c) Mechanics, (d) In-env Assistance, (e) VR Induced Symptoms and Effects.

accurate and stable results using a DNN model.

Table 5 compares our method to the latest methods developed for AD progression detection. Our findings suggest that many existing studies in the AD domain only use a single slice or focus on a specific region of the entire MRI volume, which is a limitation of most current techniques. This leads to a significant loss of crucial information, which is essential for maintaining the stability of the model in predicting diseases. For example, Uysal et al. [56] employed a selective approach by utilizing only data from the hippocampal region of the brain, which was extracted using the ITK-SNAP tool. They utilized this data only to train ML models for detecting AD. Nguyen. [31] and their team developed an AD progression detection model where they proposed minimal RNN to impute large-scale missing values in the training data before evaluating a classifier. They reported 91% test accuracy with their proposed approach. El-Sappagh and their team [14] employed a longitudinal dataset including data collected at 15 different time points for progression detection. However, their approach failed to consider the interval between the last available data and the time when the AD prediction was performed. Abuhmed and their colleagues [20] developed a method for AD progression detection that integrates deep models with several other modalities collected at various time steps. The input data used for this study included MRI, PET, neuropsychological tests, clinical scores, and demographic data. They employed a variety of features, obtained from the ADNI database, to perform disease diagnostic tasks. El-Sappagh et al. [57] reported 98% precision in detecting AD progression using multimodal longitudinal data gathered at four distinct time steps. Their research relied on large-scale data of 1029 subjects, comprising various types of information such as demographics, cognitive scores, medications for brain disorders, non-brain disorder medications, and coexisting disorders. They also tested several ML classifiers, including DT, LR, SVM, and KNN.

Table 5 reports five aspects of each comparative study which includes the number of subjects used in this study, the modalities of data used, the level of achieved performance, and the design of the architecture. The uniqueness of this study lies in the following characteristics. This study leveraged 3D CNNs, which are particularly suited for analyzing volumetric data inherent in medical imaging. Unlike 2D CNNs that process images in two dimensions, 3D CNNs consider the depth aspect, allowing for a more holistic analysis of the brain's structure and enabling the detection of subtle changes over time. The unique advantage of using 3D CNNs in our study lies in their ability to capture the spatial hierarchies in three-dimensional data, which is crucial for identifying patterns indicative of AD progression. This is because AD-related changes are characterized by complex patterns in the brain's 3D structure, such as amyloid plaques, neurofibrillary tangles, brain atrophy, synaptic loss, and vascular changes. These pathological changes occur across the three-dimensional space of the brain, affecting its structure and connectivity. 3D CNN are particularly effective in recognizing these patterns because they capture the intricate details and spatial relationships inherent in the brain's anatomy, unlike 2D representations that may miss subtle volumetric changes. 3D CNNs can model the depth information, hierarchical feature extraction, and spatial dependencies essential for detecting the widespread and multifaceted nature of AD-related changes, providing a more comprehensive and realistic analysis of the brain's structure.

Furthermore, our 3D CNN model is designed to identify AD progression by learning from a large dataset of brain scans, which enables the model to discern intricate patterns that are often imperceptible to the human eye or missed by less sophisticated methods. The proposed approach not only enhances the accuracy of AD identification but also contributes to a more reliable prediction of the disease's trajectory. In contrast to other methods, such as traditional ML [22,56] that may require handcrafted features or 2D CNNs that only analyze single slices of the brain, our 3D CNN method processes a large portion of brain volume simultaneously. This comprehensive analysis ensures that no critical information is overlooked, leading to a more effective and early identification of AD progression. As highlighted in the table, the proposed framework demonstrated better results compared to many other studies. The proposed system's strong performance and stability make it a prime foundation for developing a clinical support system for detecting the progression of AD. However, there are few studies [20,51] that outperform our proposed framework, and we noticed that, they might own a large number of data instances in the training data or the dataset used in this study contains time series or longitudinal data about the patient. Despite the proposed model exhibiting superior performance compared to other leading DL models in AD management, further advancements are needed prior to its implementation for diagnosing real patients.

Regarding the VR part of the study, the research gap suggested implementing different cognitive tests in a virtual environment as well as designing the environments according to the needs of the target audience which are older people. In terms of cognitive assessment, the

Table 5
Comparison of the proposed framework with various literature studies.

Paper, Year	Data samples	Data modality	ML technique	Performance				
				ACC	PRE	Recall	F1-score	AUC
[22], (2020)	449	MRI biomarkers	Linear mixed effects	85.00	–	86.30	–	94.00
[31], (2020)	1677	Hippocampus features and CS	RNN	–	–	–	–	91.00
[56], (2020)	485	Segmented hippocampal regions	LR, KNN, SVM, DT, and RF	92.00	–	92.00	–	–
[20], (2020)	1536	SMRI, PET, CS, assessment data, and neuropathological data	CNN Bi-LSTM	92.62	94.02	98.82	92.56	–
[58], (2020)	216	MRI	3D DenseNet	88.90	–	–	–	92.50
[59], (2021)	151	MRI	Temporally Structured SVM	90.00	–	–	–	96.20
[60], (2021)	492	Segmented hippocampus	DeepAtrophy	–	–	–	–	88.00
[14], (2021)	1371	MRI, PET, CS, and Comorbidities	Bidirectional LSTM	74.55	84.68	84.80	–	–
[61], (2022)	1371	NS, MRI, CS, and CSF	2-staged AD progression detection	93.87	94.07	94.07	94.07	–
[29], (2022)	1500	MRI	RESU-Net	94.34	–	–	–	–
[30], (2022)	400	MRI	FDN-ADNet	90.83	–	95.00	–	–
[62], (2022)	809	Multiple neuroimaging modalities	Multi-Classification Framework	–	–	–	–	96.00
[57], (2022)	559	MRI, PET, CS, and medication data	Ensemble Classifier	98.56	98.56	98.56	98.56	98.56
[63], (2023)	823	SMRI	VGG-TSwinformer	77.20	–	–	–	81.53
[51], (2023)	1682	CNN, Demographics, and CS	3D-CNN-BRNN	94.00	97.00	95.00	–	96.90
Ours, (2024)	560	MRI, CS	3DCNN-ML	97.00	95.00	95.00	96.00	94.00

system incorporates a range of cognitive tasks tailored to the elderly population, allowing for a more accurate and comprehensive assessment of cognitive functions. As for the design of the environment, the findings of the questionnaire indicated that the considered design notes performed well in terms of suitability for the elderly. Simply put, the system provides an immersive and engaging experience for the elderly, assisting them with their motivation and participation in cognitive assessment and training exercises. This, along with the results obtained from the AI model, can support the idea of the paper which is the diagnosis of AD using VR and AI. In fact, the claim that cognitive assessments can be employed in a safe and regulated situation in VR has shown to be of great value in the treatment and diagnosis of medical conditions, in this case, AD.

6. Conclusion

The most serious type of dementia is Alzheimer's disease, for which there is presently no recognized medical treatment. The objective of this paper was to evaluate a VR-AI-based system capable of AD diagnosis. A virtual environment based on a wide range of design notes for older adults was designed, followed by incorporating three important cognitive tests: ADAS-Cog, MoCA, and MMSE. Afterward, a 3DCNN-ML model was trained using ADNI2 CS and MRI and added to the virtual environment so that the responses to the tasks can be fed to the model to diagnose AD. The evaluation of the selected design notes and the way to implement them showed a high level of suitability for the elderly. Moreover, the trained AI performed well in the diagnosis of AD. On a final note, the collected results show that the suggested integration of VR and AI along with the implemented cognitive tests and design notes can deliver precise indications of the presence of AD.

7. Limitations

While the present study demonstrates the promise for a VR-AI-based system for AD diagnosis, there are a number of limitations from both VR and AI sides that should be addressed in future research.

Regarding the VR related aspects, one of the primary challenges is the potential barriers to adoption, including the relatively high cost of VR technology and the need for technical expertise, which could limit access to this diagnostic tool. To overcome this limitation, future research should focus on developing more affordable and user-friendly VR systems that can be easily integrated into clinical settings. Another limitation is the potential gap between the virtual environment and real-life scenarios, which may affect the accuracy of cognitive test results. As mentioned in the literature review, paper-based tests used to diagnose AD have been criticized for their limited ability to accurately capture real-world cognitive abilities, as they often fail to reflect how individuals function in their daily lives. This can be addressed with the help of VR, even though there may still be differences between real and virtual worlds. In this regard, further research is needed to ensure that VR-based tests accurately reflect real-world cognitive abilities. Finally, there are concerns over data security and privacy raised by the use of AI and VR technology in healthcare. As the use of these technologies continues to grow, it is essential to develop and implement robust protocols to protect sensitive medical information.

As for the AI aspect, our model has some limitations that we plan to address in future work. Currently, the study focuses primarily on performance. However, in real medical environments, domain experts require interpretable results in addition to accurate models. Therefore, we will extend this study to evaluate and enhance our model's interpretability features. Moreover, this study is based on the baseline visit of patients, which sometimes does not adequately capture the progressive deterioration of brain tissues. To address this issue, we will include follow-up visits to better understand disease progression. In future studies, we also aim to incorporate additional modalities such as medications, comorbidities, and PET scans, alongside MRI and cognitive scores, to provide a more comprehensive analysis.

CRediT authorship contribution statement

Jalal Safari Bazargani: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Nasir Rahim:** Writing – original draft, Visualization, Methodology, Investigation. **Abolghasem Sadeghi-Niaraki:** Writing – review & editing, Supervision, Conceptualization. **Tamer Abuhmed:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Houbing Song:** Supervision, Conceptualization. **Soo-Mi Choi:** Writing – review & editing, Supervision, Resources, Funding acquisition.

Declaration of competing interest

This manuscript has not been published or presented elsewhere in part or in entirety and is not under consideration by another journal. We have read and understood your journal's policies, and we believe that neither the manuscript nor the study violates any of these. There are no conflicts of interest to declare.

Acknowledgement

This work was supported in part by the ITRC Support Program under Grant IITP-2024-RS-2022–00156354 and in part by the Metaverse Support Program to Nurture the Best Talents under Grant IITP-2024-RS-2023–00254529 funded by the Ministry of Science and ICT of Korea and the Institute of Information and Communications Technology Planning and Evaluation (IITP).

References

[1] World Health Organisation, Dementia, 2023 [cited 2023 Jan 16, 2023]Available from, https://www.who.int/news-room/fact-sheets/detail/dementia.
[2] K. Ijaz, et al., Design considerations for immersive virtual reality applications for older adults: a scoping review, Multimodal. Technol. Interact. 6 (7) (2022) 60.
[3] A.-I. Corregidor-Sánchez, et al., Effectiveness of virtual reality systems to improve the activities of daily life in older people, Int. J. Environ. Res. Public Health 17 (17) (2020) 6283.
[4] T. Lewis, et al., E-health in low-and middle-income countries: findings from the Center for Health Market Innovations, Bull. World Health Organ. 90 (2012) 332–340.
[5] R.I. García-Betances, et al., A succinct overview of virtual reality technology use in Alzheimer's disease, Front. Aging Neurosci. 7 (2015) 80.
[6] D.M. Spooner, N.A. Pachana, Ecological validity in neuropsychological assessment: a case for greater consideration in research with neurologically intact populations, Arch clin neuropsychol. 21 (4) (2006) 327–337.
[7] T.D. Parsons, Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences, Front. Hum. Neurosci. 9 (2015) 660.
[8] J.M.F. Montenegro, V. Argyriou, Cognitive evaluation for the diagnosis of Alzheimer's disease based on turing test and virtual environments, Physiol. Behav. 173 (2017) 42–51.
[9] N.C. Tan, et al., Age-related performance in using a fully immersive and automated virtual reality system to assess cognitive function, Front. Psychol. 13 (2022) 847590.
[10] American Psychiatric Association, D. and A.P. Association, Diagnostic and Statistical Manual of Mental disorders: DSM-5, 5, American psychiatric association, Washington, DC, 2013.
[11] R. Jin, A. Pilozzi, X. Huang, Current cognition tests, potential virtual reality applications, and serious games in cognitive assessment and non-pharmacological therapy for neurocognitive disorders, J. Clin. Med. 9 (10) (2020) 3287.
[12] J. Bourrelier, et al., Use of a virtual environment to engage motor and postural abilities in elderly subjects with and without mild cognitive impairment (MAAMI Project), Irbm 37 (2) (2016) 75–80.
[13] Y. Zhao, et al., Psychodynamic-based virtual reality cognitive training system with personalized emotional arousal elements for mild cognitive impairment patients, Comput. Methods Programs Biomed. 241 (2023) 107779.
[14] S. El-Sappagh, et al., Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data, Fut. Gener. Comput Syst. 115 (2021) 680–699.
[15] Y. Zhang, et al., Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion, Inform Fusion 66 (2021) 170–183.

[16] E.E. Bron, et al., Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge, Neuroimage 111 (2015) 562–579.
[17] X. Jiang, L. Chang, Y.-D. Zhang, Classification of Alzheimer's disease via eight-layer convolutional neural network with batch normalization and dropout techniques, J. Med. ImAging Health Inform. 10 (5) (2020) 1040–1048.
[18] Y. Zhang, et al., Multivariate approach for Alzheimer's disease detection using stationary wavelet entropy and predator-prey particle swarm optimization, Journal of Alzheimer's Disease 65 (3) (2018) 855–869.
[19] G. Muhammad, et al., A comprehensive survey on multimodal medical signals fusion for smart healthcare systems, Inf. Fusion 76 (2021) 355–375.
[20] S. El-Sappagh, T. Abuhmed, K.S. Kwak, Alzheimer disease prediction model based on decision fusion of CNN-BiLSTM deep neural networks, in: Proceedings of the Intelligent Systems and Applications: Proceedings of the 2020 Intelligent Systems Conference (IntelliSys) 3, Springer, 2021.
[21] N. Rahim, et al., Time-series visual explainability for Alzheimer's disease progression detection for smart healthcare, Alexandria Eng. J. 82 (2023) 484–502.
[22] G. Martí-Juan, G. Sanroma-Guell, G. Piella, A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease, Comput. Methods Programs Biomed. 189 (2020) 105348.
[23] L. Xu, et al., Multi-modality sparse representation-based classification for Alzheimer's disease and mild cognitive impairment, Comput. Methods Programs Biomed. 122 (2) (2015) 182–190.
[24] S. Huang, et al., Identifying Alzheimer's disease-related brain regions from multi-modality neuroimaging data using sparse composite linear discrimination analysis. Advances in Neural Information Processing Systems, 2011, p. 24.
[25] K.R. Gray, et al., Random forest-based similarity measures for multi-modal classification of Alzheimer's disease, Neuroimage 65 (2013) 167–175.
[26] K. Seo, et al., Virtual daily living test to screen for mild cognitive impairment using kinematic movement analysis, PLoS. One 12 (7) (2017) e0181883.
[27] D. Howett, et al., Differentiation of mild cognitive impairment using an entorhinal cortex-based test of virtual reality navigation, Brain 142 (6) (2019) 1751–1766.
[28] A. Tort-Merino, et al., ALTOIDA-iADL for the diagnosis of Mild Cognitive Impairment and early Alzheimer's disease, Alzheimer's & Dementia 17 (2021) e057982.
[29] H.A. Helaly, M. Badawy, A.Y. Haikal, Toward deep mri segmentation for alzheimer's disease detection, Neural Computing and Applications 34 (2) (2022) 1047–1063.
[30] R. Sharma, et al., FDN-ADNet: fuzzy LS-TWSVM based deep learning network for prognosis of the Alzheimer's disease using the sagittal plane of MRI scans, Appl. Soft. Comput. 115 (2022) 108099.
[31] M. Nguyen, et al., Predicting Alzheimer's disease progression using deep recurrent neural networks, Neuroimage 222 (2020) 117203.
[32] J. Johnson, K. Finn, Designing User Interfaces For an Aging population: Towards universal Design, Morgan Kaufmann, 2017.
[33] W. Boot, et al., Designing For Older adults: Case studies, methods, and Tools, CRC Press, 2020.
[34] T.L. Mitzner, et al., Technology adoption by older adults: findings from the PRISM trial, Gerontologist 59 (1) (2019) 34–44.
[35] S. Lindsay, et al., Engaging older people using participatory design, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2012.
[36] A.M. Mol, R.S. Silva, L. Ishitani, Design recommendations for the development of virtual reality focusing on the elderly, in: Proceedings of the 14th Iberian Conference on Information Systems and Technologies (CISTI), IEEE, 2019.
[37] K. Ijaz, et al., An immersive virtual reality platform for assessing spatial navigation memory in predementia screening: feasibility and usability study, JMIR. Ment. Health 6 (9) (2019) e13887.
[38] B. Ahmed, et al., Treatment of Alzheimer's, cognitive, chronic pain rehabilitation, depression and anxiety disorders in one system for elderly using VR, in: Proceedings of the 15th International Conference on Ubiquitous Robots (UR), IEEE, 2018.
[39] S. Baker, R.M. Kelly, J. Waycott, R. Carrasco, T. Hoang, F. Batchelor, F. Vetere, Interrogating social virtual reality as a communication medium for older adults, Proc. ACM. Hum. Comput. Interact. 3 (2019) 1–24. CSCW.
[40] N.C. Lecavalier, et al., Use of immersive virtual reality to assess episodic memory: a validation study in older adults, Neuropsychol. Rehabil. (2018).
[41] M. Eisapour, et al., Virtual reality exergames for people living with dementia based on exercise therapy best practices, in: Proceedings of the human factors and ergonomics society annual meeting, SAGE Publications Sage CA, Los Angeles, CA, 2018.
[42] M. Eisapour, et al., Participatory design of a virtual reality exercise for people with mild cognitive impairment, in: Proceedings of the Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, 2018.
[43] J.R. Bruun-Pedersen, S. Serafin, L.B. Kofoed, *Going outside while staying inside—Exercise motivation with immersive vs. non–immersive recreational virtual environment augmentation for older adult nursing home residents*, in: Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI), IEEE, 2016.
[44] I. Kovar, Use of virtual reality as a tool to overcome the post-traumatic stress disorder of pensioners, Int. J. Adv. Sci. Eng. Inf. Technol. (2019).
[45] R.C. Petersen, et al., Alzheimer's disease Neuroimaging Initiative (ADNI) clinical characterization, Neurology. 74 (3) (2010) 201–209.
[46] W. Kang, et al., Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer's disease diagnosis, Comput. Biol. Med. 136 (2021) 104678.
[47] B. Fischl, FreeSurfer, Neuroimage 62 (2) (2012) 774–781.
[48] *Advanced Normalization Tools*. Available from: https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BET.
[49] *BET - FslWiki - Skull Stripping*. Available from: https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BET.
[50] *MNI Atlases - FslWiki*. Available from: https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases.
[51] N. Rahim, et al., Prediction of Alzheimer's progression based on multimodal deep-learning-based fusion and visual explainability of time-series data, Information Fusion 92 (2023) 363–388.
[52] P. Kourtesis, et al., Validation of the virtual reality neuroscience questionnaire: maximum duration of immersive virtual reality sessions without the presence of pertinent adverse symptomatology, Front. Hum. Neurosci. 13 (2019) 417.
[53] M. Schrepp, J. Thomaschewski, A. Hinderks, Construction of a Benchmark For the User Experience Questionnaire (UEQ), 2017.
[54] A. Baharum, et al., Development of questionnaire to measure user acceptance towards user interface design, in: Advances in Visual Informatics: 5th International Visual Informatics Conference, IVIC 2017, Springer, Bangi, Malaysia, 2017. *November 28–30, 2017, Proceedings 5*.
[55] N. Rahim, et al., Information fusion-based Bayesian optimized heterogenous deep ensemble model based on longitudinal neuroimaging data, Appl. Soft. Comput. (2024) 111749.
[56] G. Uysal, M. Ozturk, Hippocampal atrophy based Alzheimer's disease diagnosis via machine learning methods, J. Neurosci. Methods 337 (2020) 108669.
[57] S. El-Sappagh, et al., Automatic detection of Alzheimer's disease progression: an efficient information fusion approach with heterogeneous ensemble classifiers, Neurocomputing. 512 (2022) 203–224.
[58] M. Liu, et al., A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease, Neuroimage 208 (2020) 116459.
[59] Y. Zhu, et al., Long range early diagnosis of Alzheimer's disease using longitudinal MR imaging data, Med. Image anal. 67 (2021) 101825.
[60] M. Dong, et al., DeepAtrophy: teaching a neural network to detect progressive changes in longitudinal MRI of the hippocampal region in Alzheimer's disease, Neuroimage 243 (2021) 118514.
[61] S. El-Sappagh, et al., Two-stage deep learning model for Alzheimer's disease detection and prediction of the mild cognitive impairment time, Neural Comput. Appl. 34 (17) (2022) 14487–14509.
[62] F. Nan, et al., A Multi-classification Accessment framework for reproducible evaluation of multimodal learning in Alzheimer's disease, IEEE/ACM. Trans. Comput. Biol. Bioinform. (2022).
[63] Z. Hu, et al., VGG-TSwinformer: transformer-based deep learning model for early Alzheimer's disease prediction, Comput. Methods Progr. Biomed. 229 (2023) 107291.