A Multimodal Artificial Intelligence Approach to Early Alzheimer's Diagnosis

Dr. Öğr. Üyesi Gülay Çiçek.

Yusuf Talha Yılmaz[1]
[1]Department of Software Engineering, Faculty of Engineering and Architecture
Istanbul Beykent University, Sarıyer, İstanbul, Türkiye
[1]2103013281@student.beykent.edu.tr,

I. Literature review

This section of the study examines AI-based research on early diagnosis of Alzheimer's disease, focusing particularly on image data-focused studies. Existing studies are grouped according to the data types used, model configurations, and performance criteria, and the strengths and weaknesses of each method are thoroughly evaluated. However, the literature generally utilizes only image data such as MRI or PET, and it is noteworthy that the number of multimodal approaches based on tabular or clinical data integration is limited.

After summarizing the research's key findings, data sources, and application areas, the focus is on the unique contributions and potential benefits of using multimodal data. The literature review demonstrates that this study, with its multimodal approach combining both image and tabular data, offers an innovative contribution to the literature and demonstrates its potential to provide significant improvements in terms of accuracy and generalizability in early diagnosis compared to other studies.

A. Deep Learning

This section focuses on the use of deep learning techniques in the early diagnosis of Alzheimer's disease. The importance of deep learning structures such as artificial neural networks and convolutional neural networks, particularly in extracting meaningful features from image data, is thoroughly examined. Studies in the literature are grouped according to different layer structures, activation functions, and optimization methods, and the effectiveness and limitations of each method in disease diagnosis are examined. It also emphasizes that deep learning approaches are not limited to a single dataset; they also explore the possibilities of integration with tabular and clinical data, demonstrating the advantages of reliability and accuracy in early diagnosis. This review highlights not only the effective performance of deep learning techniques in the early diagnosis of Alzheimer's disease but also the importance of model transparency, interpretability, and data diversity.

Doaa and colleagues (2023) propose a deep learning system using MRI data for the early diagnosis of Alzheimer's disease. In their study, they adopted a methodology that includes data processing, data augmentation, and classification stages; they tested CNN models developed from scratch and VGG16 models based on transfer learning. The CNN model was built with three convolutional layers and max-pooling, dropout was used to prevent overfitting, and binary classification was performed using a sigmoid activation function. In the VGG16 model, the first layers were kept constant, the upper layers were retrained, and different optimization methods were tested; the best result was achieved with Adam optimization with 97.44% accuracy. The performance of the CNN designed from scratch was in the 99.95% – 99.99% accuracy range. The authors state that the limitations of the dataset and the limitations of binary classification are obstacles that need to be addressed in the future [1].

Janani and his team (2021) developed multimodal deep learning models that combine image, genetic (SNP), and clinical (EHR) data to determine Alzheimer's disease stages. A 3D CNN was used for image data, and intermediate features were extracted using a stacked denoising auto-encoder for clinical and genetic data, and combined during the classification phase. The model was trained to distinguish between Alzheimer's disease, MCI, and healthy control (CN) categories, achieving higher accuracy, precision, recall, and F1 scores than traditional machine learning methods. Furthermore, masking and clustering-based methods were used to investigate the impact of biological markers such as the hippocampus, amygdala, and RAVLT test on model decision-making [2].

In the study presented by Seung et al. (2024), a middle-fusion multimodal deep learning model is proposed that fuses MRI and PET images for the early diagnosis of Alzheimer's disease. This model extracts features using depthwise separable convolution (DS-Conv) blocks without using an activation function, and then learns complex relationships between different modalities through mix skip connection convolution (MSC-Conv) and sharing weight convolution (SW-Conv) blocks. The research evaluated the model using T1-w MRI, FDG PET, AB PET, and tau PET images on the ADNI1, ADNI2, and ADNI3 datasets. The balanced accuracy rate was 1.00 (i.e., 100%) in the classification between Alzheimer's and control groups, and 0.76 in the comparison between mild cognitive impairment (MCI) and control groups. In addition, a new ROI extraction technique including hippocampus, middle temporal and inferior temporal regions was also proposed in the study [3].

Suriya and team (2021) designed a four-category deep learning model for the early diagnosis of Alzheimer's and dementia. The CNN structure, built using information obtained from MRI images, includes DEMNET blocks, dropout, and dense layers. Class imbalance in the dataset was corrected using the SMOTE method, and the training accuracy rate was determined to be 99% and the validation accuracy rate was 94%. In the testing phase, an accuracy rate of 95.23% and an AUC value of approximately 97% were achieved. This study demonstrated high success in multiple

Author(s) (Year)	Sample Size	Language	Platform	Method	Results	Deficiencies	Future Studies
Doaa et al. (2023) [1].	6400	English	MRI images	CNN and transfer learning framework based on deep learning	High success; demonstrated competence in Alzheimer's analysis with few labeled data	General validity has not been tested on different data sets.	Testing with different datasets, conducting multicenter studies, and integration with different imaging modalities have been suggested.
Janani et al. (2021) [2].	2,004 patients	English	Clinical data, SNP genetic data, MRI image data	Multimodal deep learning: 3D CNN for vision; stacked denoising auto-encoder for genetics + clinical; classification by fusion of data modules	Deep models outperformed traditional models. Multimodal data fusion was found to be superior to single-modality models in terms of accuracy, precision, recall, and F1-mean.	Only CN and AD classification was made in the imaging module; MCI/AD distinction was not made for genetic data; some modules worked with data that may be incomplete.	Working with datasets consisting of larger and more diverse modules; better addressing the MCI stage from a genetic and imaging perspective; improving the performance of the model in missing data modules.
Seung et al. (2024) [3].	-	English	Alzheimer's Disease Neuroimaging Initiative (ADNI) series	Multi-modal 3D deep learning model: a 3D CNN-based approach combining image/modal data with intermediate fusion structure	It has been found that multimodal data perform better than single modality	There are uncertainties such as sample size, missing data across modalities, and generalizability.	Testing the model with larger data sets, data from different centers, and missing-mode cases
Suriya et al. (2021) [4].	6400	English	MRI images	With the deep learning-based CNN framework ("DEMNET"): Structured multi-stage classification.	Classical: 95.23% accuracy; AUC 97%; Cohen's Kappa ≈0.93	Sample size was not detailed in full numbers on a class basis; the problem of imbalance between classes was stated.	Eliminating the problem of inter-class imbalance; testing generalizability with different datasets.
Aristidis et al. (2023) [5].	-	English	Non-invasive biomarker data platforms	Integration of Artificial Intelligence and Deep Learning techniques to analyze non-invasive biomarker data	It has been stated that processing large and heterogeneous data obtained by non-invasive approaches with AI and DL offers high potential for early diagnosis of Alzheimer's.	The article is not a primary experimental study; although it makes comparisons between various technologies and methods, it has limitations in terms of practical application and clinical prevalence.	Larger scale collection of diverse non-invasive data sources; integration of AI/DL models into clinical practice; and greater use of explainable artificial intelligence (xAI) techniques are recommended.
Doaa et al. (2022) [6].	Compilation study	English	Literature on imaging data and deep learning studies in the clinical/research environment	The use of deep learning approaches (CNN, DNN, AE, DBN etc.) for early detection of Alzheimer's Disease in the literature; image pre-processing, classification methods, literature review.	Many studies in the literature have shown high performance (e.g., accuracies above 95% are reported for three-class classifications).	Limited access to large datasets, Heterogeneity among imaging modalities (MRI, PET, fMRI), difficulty of multimodality integration, Review focused only on the most common DL methods, the full methodological diversity may not be comprehensive	The combined use of different data modalities has been suggested, and the development of deep learning models with transfer learning, automatic feature extraction, and lighter network architectures has been suggested.

TABLE I: Deep Learning

classification despite using only image data [4].

Aristidis and his team (2023) investigated the use of non-invasive biomarkers instead of invasive methods in the early diagnosis of Alzheimer's disease and the application of artificial intelligence and deep learning to this data. The study evaluated biomarker development platforms using data obtained from blood components, wearable sensors, and imaging systems. The study emphasized the importance of integrating artificial intelligence and deep learning due to the large scale, diversity, and analysis required for this data. While no specific model or specific success rates were provided, the study noted that non-invasive methods offer benefits in terms of patient comfort, cost, and risk. Furthermore, it noted that issues such as data management, interpretability, and model transparency remain significant challenges. The study demonstrates that non-invasive data sources and artificial intelligence techniques could significantly transform the early diagnosis of Alzheimer's [5].

Doaa et al. (2022) conducted a systematic review of current deep learning methods for the timely diagnosis and classification of Alzheimer's disease. The study comprehensively examined imaging techniques such as structural and functional MRI and PET, as well as data preprocessing stages (denoising, brain region extraction, intensity correction), and classification approaches (CNN, AE, DNN, etc.). Of the 159 studies reviewed, 110 focused on the early diagnosis of Alzheimer's disease, and it was noted that multimodal data provided better classification results compared to single-modal data. Furthermore, existing issues such as dataset inadequacies, modality integration, class imbalances, and model understandability were addressed. This systematic review identifies gaps in the literature and research areas for developing deep learning-based systems for early diagnosis of Alzheimer's disease, laying a solid foundation for future multimodal research [6].

B. Machine Learning

This section discusses machine learning-based methods for detecting Alzheimer's disease at an early stage. Studies typically evaluate algorithms such as support vector machines (SVM), decision trees, random forests (RF), k-nearest neighbors (k-NN), and logistic regression in classification tasks using single or multiple data sources such as MRI, PET, and clinical information. Data preprocessing, dimensionality reduction, and feature selection techniques (PCA, RFE) aim to improve model efficiency, while class imbalances are corrected using methods such as SMOTE. Findings in the literature indicate that machine learning models can achieve high accuracy with single-modal data, but when used with multimodal data, they provide more reliable and generalizable results for early diagnosis. Furthermore, model explainability and data diversity are significant challenges for the applicability of machine learning methods.

Vasco et al. (2022) developed a multi-diagnostic and extensible machine learning method using structured MRI data for the diagnosis of Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI). In the study, morphometric features were obtained from the ADNI (n=570) and OASIS (n=531) datasets. These features were trained with a 5-layer cross-validation method using l-SVM, decision trees, random forest, extreme random trees, LDA, logistic regression, and logistic regression with SGD algorithms. Feature selection was performed using mutual information, ANOVA F-values, and chi-square statistics; the best feature percentages and hyperparameters were determined using evolutionary algorithms. In the "Control vs AD" task in binary classification, the balanced accuracy was 90.6% and the MCC value was 0.811 on the combined ADNI + OASIS dataset. In the multi-class classification task, using only the ADNI dataset for the "Control vs. MCI vs. AD" task, balanced accuracy of 62.1% and an MCC of 0.438 were achieved. Based on the importance of the features, the hippocampus contributed 25-45%, the temporal areas 13%, and the cingulate and frontal regions 8-13%. Furthermore, some limitations were highlighted, such as the fact that graph theory metrics did not improve classification performance, that MCI patients were only available in one dataset, and that cognitive test scores were not included in the model [7].

In this study, conducted by Khandaker and his team (2023), a machine learning method was proposed for the early detection of Alzheimer's disease. The OASIS dataset was used; missing data were filled with mean values, then feature selection was performed using the Select KBest technique, and the data was scaled using the Standard Scaler. Classification algorithms used included Gaussian Naive Bayes, Decision Tree, Random Forest, XGBoost, Gradient Boost, and Voting Classifier. The highest accuracy level was 96% with the voting classifier. Limitations of the study include the dataset being from a single source and the scarcity of various data protocols [8].

A study by Chun-Hung et al. (2021) investigated the role of machine learning and novel biomarkers in the early diagnosis of Alzheimer's disease. The publication noted that current diagnostic methods (e.g., A and p-tau levels in spinal fluid) are invasive and expensive, and therefore, it was stated that less invasive biomarkers should be developed. In addition to traditional A/tau indicators, the study evaluated neurofilament light (NFL), a neuroinjury marker; neurogranin, BACE1, SNAP-25, GAP-43, and synaptophysin, as indicators of synaptic dysfunction; and sTREM2 and YKL-40, as markers of neuroinflammation. Among the machine learning methods used, support vector machine (SVM), logistic regression, random forest, and naive Bayes were reported to be useful in better understanding the differences between patients and healthy individuals. The article emphasizes that these methods have the potential to provide benefits in terms of sensitivity and specificity, but issues such as the size of data sets, clinical applicability of biomarkers and standardization of machine learning models are still among the challenges that need to be resolved [9].

Daniele and colleagues (2022) combined EEG signal analysis and supervised machine learning methods to facilitate the early diagnosis of Alzheimer's disease. A total of 105 individuals participated in the study. EEG data were normalized to 256 Hz and processed using a 1-30 Hz band-pass filter to remove interfering elements. Features were extracted using time-frequency analysis and dual digital FIR filters, and power densities in low- and high-frequency bands were used in the classification process. Decision trees, support vector machines, and k-nearest neighbor algorithms were applied, and the data were split into training and test sets at 70% and 30%. Accuracy rates for binary classifications were 97% for HC and AD, 95% for HC and MCI, and 83% for MCI and AD; accuracy was reported as 75% for three-class classification. The researchers emphasize that the correct selection of filter cutoff frequencies significantly affects classification success, and that the 7-16 Hz range increases the discriminatory power. Limitations of the study include the small data set, imbalance between groups, and the lack of use of different EEG recording protocols [10].

Gargi and his team (2023) proposed a machine learning-based

Author(s) (Year)	Sample Size	Language	Platform	Method	Results	Deficiencies	Future Studies
Vasco et al. (2022) [7].	ADNI dataset: 570 participants, OASIS dataset: 531 participants	English	Structural MRI data	Multi-class and two-class classification using machine learning algorithms + voting. Different MRI protocols and graph theory features are also evaluated.	In HC vs AD classification: Balanced accuracy 90.6% and Matthews correlation coefficient 0.811, in "HC vs MCI vs AD" classification: BAC 62.1%, MCC 0.438	The MCI category is small in number, does not fully align with clinical real-life decision processes, and does not include cognitive test scores or other biomarkers.	It was suggested that it should be tested with larger, multicenter and longitudinal data sets, and it was emphasized that real-time studies (prospective studies) should be carried out for clinical use.
Khandaker et al. (2023) [8].	150 individuals (64 demented, 72 non-demented) from the OASIS dataset	English	Structural MRI-based "Open Access Series of Imaging Studies (OASIS)" dataset	Feature selection (SelectKBest), standardization (StandardScaler), classifiers: GaussianNB, Decision Tree, Random Forest, XGBoost, GradientBoost, Voting Ensemble	Best accuracy 96% (Voting–Classifier or Random Forest combination)	Limited sample size; only one data source; multimodality (e.g., PET, biomarkers) not included.	Validation with larger, multi-center datasets; integration of multimodal data; prospective studies for clinical applicability.
Chun-Hung et al. (2021) [9].	Compilation / meta-analysis study	English	Clinical biomarkers, imaging, and other biomarker sources	Combining new biomarkers with machine learning and deep learning algorithms	The combined use of machine learning and new biomarkers has shown that it can increase sensitivity and specificity in the diagnosis of Alzheimer's disease.	Studies focusing on a single data source are limited; methods have limited applicability to clinical practice; some biomarkers lack standardization.	Testing with larger and multicenter datasets; more widespread use of blood-based biomarkers; and integration of deep learning models into clinical decision support systems have been suggested.
Daniele et al. (2022) [10].	105 records (48 AD, 37 MCI, 20 HC)	English	EEG recordings	High and low frequency band filtering with FIR dual time-domain filter, power density extraction: absolute square of high-low band difference and classification	HC vs AD: 97% accuracy, HC vs MCI: 95% accuracy, MCI vs AD: 83% accuracy, HC vs MCI vs AD: 75% accuracy	Sample size is relatively limited; only resting EEG, single-center data; accuracy drops significantly in three-class classification.	Scaling up of the dataset; use of different EEG protocols (e.g. under cognitive task); potential for real-time application on embedded devices.
Gargi et al. (2023) [11]	3,692 MRI images	English	ADNI dataset	Image preprocessing: selective clipping, grayscale conversion, histogram equalization and Classification algorithms: Random Forest, XGBoost, CNN	Accuracy: 97.57% Sensitivity: 97.60%	Single dataset (ADNI MRI only), only two-class discrimination (AD vs Normal), generalization ability of the model uncertain	Data augmentation (using GAN), cross-validation with different datasets, multi-class diagnosis (including MCI, for example) have been proposed.
Nitsa et al. (2021) [12]	750 participants	English	Alzheimer's Disease Neuroimaging Initiative (ADNI) MRI images	Detecting brain asymmetries (left-right hemisphere), statistical feature extraction (MSE, variance, etc.), SVM and CNN-based supervised learning	NC vs EMCI: SVM 92.5% accuracy; NC vs AD: 93.0% accuracy; NC vs AD for CNN: 90.5% accuracy	The number of participants was limited, and only MRI images were used. Variables such as individual handedness were not examined.	The effects of different handedness groups should be investigated; architecture optimization should be performed in deep learning; longitudinal studies are recommended.

TABLE II: Machine learning

method for the early diagnosis of Alzheimer's disease (AD). In the study, four-dimensional MRI images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset were converted to two-dimensional, and preprocessing operations such as selective cropping, grayscale transformation, and histogram equalization were performed on the images. Three different algorithms (Random Forest, XGBoost, and Convolutional Neural Network (CNN)) were then used for classification. According to the findings obtained after preprocessing, the success rate was 97.57% and the sensitivity was 97.60%. Significant limitations of the study include the use of only a single data source and the inability to accommodate data diversity across different protocols [11].

Nitsa and his team (2021) developed a machine learning method that identifies brain asymmetries for the diagnosis of early dementia and Alzheimer's disease. T1-weighted MRI images obtained from the ADNI dataset were standardized to 256×256×3 pixels, subjected to segmentation and projection operations, and then 10 statistical features (mean, variance, RMS, entropy, skewness, kurtosis, etc.) were obtained. Classification was performed using Naive Bayes, LDA, SVM, k-NN, and AlexNet-based CNN algorithms. Accuracy was found to be between 92.5% and 93.0% with SVM and between 75.0% and 90.5% with CNN for tasks between NC and EMCI and NC and AD. Limitations include not considering hand dominance, not performing long-term follow-up analyses, and not using different MRI protocols [12].

C. Hybrid Approaches

This section examines research based on hybrid models for the early diagnosis of Alzheimer's disease. Studies typically utilize multiple data sources, such as MRI, PET, demographic data, and psychometric tests, to create hybrid models that combine deep learning and machine learning methods. These methods aim to achieve high accuracy and generalizability by combining feature extraction, data augmentation, segmentation, and classification under a single framework. Current literature indicates that hybrid models are particularly effective in distinguishing MCI from AD compared to single-modal methods, but their clinical applicability remains limited due to limitations such as data limitations, single-center studies, and class imbalances.

In a systematic review conducted by Akhilesh et al. (2023), 47 studies using MRI and PET data were evaluated; traditional machine learning techniques achieved 85% accuracy, deep learning achieved 96–98% accuracy, and multimodal approaches were found to perform better than single-modality approaches. However, most studies were dependent on the ADNI dataset, which limited translation to real-world clinical applications [13].

Badiea et al. (2021) achieved accuracy around 94%–99% with CNN and hybrid AlexNet + SVM models after preprocessing with SMOTE and t-SNE using OASIS and MRI datasets. The study emphasized the need for multimodal data integration and clinical application because the data was single-centered and limited [14].

Anuradha et al. (2023) achieved 99% accuracy with a hybrid CNN-SoftMax model using image preprocessing and segmentation techniques on the Kaggle dataset. However, the number of samples in some classes was very small and the data came from a single source. The importance of validation with larger and multi-center datasets is emphasized. [15].

Ibrahim and his team (2022) developed a hybrid approach using four-class MRI data with local feature extraction and pre-trained deep models, achieving 99.8% accuracy and 100% specificity. The limitations include insufficient sample details and the use of a single data source [16].

Balaji et al. (2023) analyzed MRI and PET data using a CNN + LSTM architecture and achieved 98.5% accuracy in classifying cognitively normal and EMCI. The limitations of the study include the limited dataset and the absence of some classes [17].

Afreen and his team (2022) developed a three-stage hybrid machine learning algorithm using demographic and psychometric test data; accuracy rates ranged from 89.6% to 95.1%. The study emphasized the need for multimodal analyses because imaging or biomarker data were not included [18].

These findings indicate that hybrid approaches can provide high accuracy in early Alzheimer's diagnosis with the use of multimodal data, but there is still significant room for improvement in terms of data diversity, multicenter validation, and clinical validity.

D. Yöntemlerin Karşılaştırılmalı Değerlendirmesi

Artificial intelligence-supported techniques used to identify the preliminary stages of Alzheimer's disease can be classified into three main groups: deep learning, traditional machine learning, and mixed methods. The studies in Tables I, II and III provide important comparison opportunities in terms of various data types, model designs, success criteria and limitations.

Deep learning-based techniques are particularly notable for their ability to automatically extract meaningful features from visual data, particularly MRI and PET. For example, a study by Doaa et al. (2023) reported that their scratch-built CNN and transfer-learning-based VGG16 models achieved accuracy between 99.95% and 99.99% in binary classification tasks, demonstrating that high performance can be achieved with a small amount of labeled data [1]. Similarly, Janani et al.'s (2021) multimodal deep learning method offered a distinct advantage over traditional methods in three-class classification by combining image, genetic, and clinical information, demonstrating the potential of integrating multimodal data to increase early diagnosis accuracy [2]. While the benefits of deep learning methods include high accuracy, automatic feature extraction, and effective representation of visual-based features, their limitations include the fact that data sets are generally single-centered and limited, the model's explainability is low, and difficulties in managing missing data across different modalities.

Machine learning-based methods generally provide faster and more comprehensive solutions for limited datasets. The method developed by Vasco et al. (2022) based on multiple classification and voting achieved 90.6% balanced accuracy in classifying HC from AD, but encountered issues such as decreased performance in multi-class tasks and limited clinical data usage [7]. Similarly, Khandaker et al (2023) achieved 96% accuracy with the single-center OASIS dataset, but the lack of multimodal data and the small sample size negatively affected the generalizability of the model [8]. While the advantages of machine learning methods are the clarity and understandability of the model and the ability to work in data sparseness, their shortcomings include the limitation of multimodality integration, loss of performance on high-dimensional features and limited data generalizability.

Hybrid methods offer significant benefits in multimodal data usage by combining the feature extraction capabilities of deep learning with the understandable and optimizable structure of machine learning. For example, Balaji et al. (2023) achieved 98.5% accuracy in classifying "cognitively normal" and "EMCI" with a CNN + LSTM-based hybrid model they developed by combining MRI and PET images with neuropsychological test

Author(s) (Year)	Sample Size	Language	Platform	Method	Results	Deficiencies	Future Studies
Akhilesh et al. (2023) [13].	Approximately 47 studies were selected for review.	English	MRI and PET imaging data, ADNI and OASIS datasets.	Systematic analysis of machine learning and deep learning techniques. Feature extraction and multi-modality utilization are evaluated.	Traditional ML techniques have achieved accuracy of up to 85%; deep learning has achieved accuracy of up to 96–98%. Multi-modality approaches have outperformed single-modality approaches.	Most studies are limited to MRI/PET imaging; data are largely dependent on ADNI; few studies have translated into actual clinical practice.	It is recommended to increase multimodality studies, integrate explainable Artificial Intelligence (XAI) techniques, and develop generalizable models with different data sources and data diversity.
Badiea et al. (2021) [14].	OASIS dataset: 373 records and MRI dataset: 6,400 images	English	OASIS medical record dataset and MRI image dataset	For OASIS: data balancing with SMOTE, dimensionality reduction with t-SNE, for classification and MRI: CNN + data augmentation, hybrid models	OASIS part: Random Forest with accuracy 94%; precision 93%; recall 98%; F1 96% and MRI part: Hybrid AlexNet + SVM model with accuracy 94.8%; sensitivity 93%; specificity 97.75%; AUC 99.70%.	Data are single-center and limited; classes on MRI images are few in number; limited to image and registration data only; other biomarkers or multimodality combinations are not included.	Multicenter testing with more data sources; multimodal data integration; and a shift toward prospective studies to transition hybrid models to clinical practice.
Anuradha et al. (2023) [15].	6,400 views	English	Kaggle "Alzheimer's dataset 4 class of images" dataset	Pre-processing: HSV and LAB transformations, K-means segmentation, image resizing, increased to 3,200 images for each class. Hybrid model, feature extraction in parallel, followed by CNN layer and SoftMax classification	The model using SMOTE achieved 99% accuracy.	Imbalance in data classes was studied with very few examples, especially in the "moderate dementia" class ($\leq$100 images). Data comes from only one source.	Validation with larger, multicenter datasets; integration of multimodal data (e.g., PET, biomarker); transition to real clinical application.
Ibrahim et al. (2022) [16].	-	English	Four-class classification of MRI images.	Local feature extraction, Pre-trained deep models, Hybrid model and Hybrid combine	99.8% accuracy; precision %99.9; sensitivity 99.75%; specificity %100; AUC 99.94%.	Sample details inadequate; only certain data sources may have been used; actual clinical validation limited.	It is recommended to work with a larger data set (multicenter); integrate different modalities and biomarkers; and conduct longitudinal studies for transition to clinical practice.
Balaji et al. (2023) [17].	512 MRI images and 112 PET images	English	MRI + PET images and neuropsychological test scores	CNN + LSTM architecture; ACO noise reduction; MFCM segmentation; Adam optimization	Accuracy in classifying "cognitively normal" and "EMCI" was achieved at 98.5%.	Dataset limited to MRI + PET only; some classes or real-world clinical data coverage is specified as limited.	Addition of different modalities (e.g., biomarkers) and validation with larger, multicenter datasets are recommended.
Afreen et al. (2022) [18].	ADNI data set	English	demographics + psychometric test results	Three-stage cognitive-hybrid machine learning algorithm and Feature selection, demographic + cognitive test data were used.	Tier 1: 89.63% accuracy Tier 2: 93.90% accuracy with Random Forest Tier 3: 95.12% accuracy	Class distributions are not comprehensive; only psychometric and demographic data are used. Imaging or biomarker data modalities are not included.	It is recommended to move to multimodal analyses with imaging (MRI/PET) and biomarker data; to test generalizability with data from different centers; and to perform prospective applications in real clinical environments.

TABLE III: Hybrid Approaches

results. Thus, it was observed that combining multimodal data increases classification accuracy [17]. In the research conducted by Afreen et al. (2022), they examined demographic and psychometric test data through a three-stage hybrid machine learning algorithm and achieved 95.12% accuracy at the Tier3 stage, demonstrating the effectiveness of tabular data in multimodal analysis [18]. The most important advantages of hybrid methods are the increased accuracy and generalizability by integrating various data sources; on the other hand, the disadvantages are the increased complexity of the model, the higher computational costs, and the expertise required for data preprocessing.

A general performance analysis reveals that while deep learning techniques based on a single modality offer high accuracy, they face issues with insufficient data diversity and interpretability. While machine learning techniques provide fast results with smaller, more structured datasets, they are limited in their ability to integrate multiple datasets. Hybrid approaches, on the other hand, enhance both accuracy and reliability in early Alzheimer's diagnosis by blending image, tabular, and clinical data; however, this increases model complexity and computational cost.

In summary, the comparison of studies demonstrates that multimodal and hybrid approaches offer superior potential in terms of both accuracy and generalizability in the early diagnosis of Alzheimer's disease. Furthermore, the fact that the datasets are largely single-center, small in size, and limited to specific modalities poses a significant obstacle to the implementation of these methods into clinical practice. In future studies, the use of multimodal datasets from various centers, management of incomplete data, model understandability, and improvement in clinical application are critical factors that will increase the success of AI methods in the early diagnosis of Alzheimer's disease.

E. General Evaluation of the Literature

Deep learning, machine learning, and hybrid methods are gaining attention in research on early diagnosis of Alzheimer's disease. Deep learning techniques provide high accuracy by automatically extracting features, particularly from MRI and PET images, but they have some limitations regarding explainability. Machine learning algorithms can obtain more general results with clinical data and demographic information, but they are limited in analyzing high-dimensional images.

Hybrid models offer benefits in terms of both accuracy and generalizability by combining different types of data. Research has shown that multimodal and hybrid methods are more effective than single-modal methods, particularly in distinguishing between MCI and AD. However, data shortages and challenges to clinical implementation remain significant obstacles. In this context, current research demonstrates the potential of artificial intelligence in the early diagnosis of Alzheimer's disease and highlights the importance of multimodal approaches.

II. Experimental Results

A. Dataset and Methodological Summary

In this study, the open image dataset named "Alzheimer's Dataset (4 class of Images)", which covers four distinct stages of Alzheimer's disease, was used. This dataset contains a total of 6400 brain Magnetic Resonance (MRI) images, categorized into four groups: MildDemented, ModerateDemented, NonDemented, and VeryMildDemented. The original class distribution of the dataset was as follows: NonDemented (3200 samples), VeryMildDemented (2240 samples), MildDemented (896 samples), and ModerateDemented (64 samples). This distribution reveals a particularly significant imbalance in the ModerateDemented group.

To test the project's multimodal approach, clinical and demographic information not included in the original dataset was artificially generated based on known pathological characteristics of each image group. Each image was then supplemented with features such as age, Mini-Mental State Examination score, years of education, and gender.

To fairly and reliably evaluate the generalization ability of the models, a dataset containing a total of 6400 samples was initially separated by 20% (20%, test_size=0. 2) to create the Locked Final Test Set. This test set, consisting of 1280 samples, was never used in the model training or hyperparameter tuning stages, but was reserved only for comparative evaluation after all model training was completed. The remaining 80% (5120 samples) was evaluated during model training and in the 5-fold cross-validation stages. This separation method (using the stratify parameter) ensured that the class distribution in both the training and validation sets, as well as the locked test set, remained consistent with the main dataset.

B. Preprocessing and Feature Extraction

1) Tabular Data Preprocessing: When using synthetic tabular data in machine learning models, the categorical feature defined as gender is converted to a numerical form using the One-Hot Encoding technique. All numerical features, such as age, MMSE, and education level, are standardized using StandardScaler to prevent different scales from negatively impacting model performance. These processes aim to improve model performance by ensuring that algorithms assign equal importance to features.

2) Image Data Preprocessing: The MRI images used for the deep learning systems were resized to ensure the same input size (128x128) for all structures. Pixel values were normalized to a specific range. Furthermore, to increase the model's generalization capacity and reduce overfitting, data augmentation methods such as random rotation, panning, zooming, and horizontal flipping were applied to the images in the training set. These techniques are expected to enable the model to learn different variations in classes with limited examples, enabling it to develop more effective features.

C. Attribute Selection

In the machine learning phase of this study, only four synthetic features (Age, MMSE, Education, Gender) were used, reflecting basic demographic and clinical information related to the problem. Due to the limited and simple nature of the feature set, no feature selection method was required. Because each of these four features is hypothesized to be theoretically important in the diagnosis of Alzheimer's disease, we aim to understand the extent to which the model can learn based on this fundamental information. If a more complex and high-dimensional tabular feature set were used, feature selection methods could play a significant role in reducing model complexity and improving performance. However, in the current framework, this step is not implemented.

D. Experimental Setup and Model Groups

1) Hardware/Software Environment: All experiments in this research are conducted on Google Colaboratory (Colab), a cloud-based computing environment. To ensure reproducibility of the

study, the main software and hardware infrastructure used are described below:

- Hardware: NVIDIA Tesla T4 Graphics Processing Unit (GPU) provided by Google Colab was used to accelerate the training of deep learning models.
- Programming Language: Python 3.x
- Data Processing and Analysis: Pandas (for data frames), NumPy (for numerical operations)
- Machine Learning: Scikit-learn (for ML models, metrics, and K-Fold)
- Deep Learning: TensorFlow 2.x and Keras (for DL model architectures and training)
- Visualization: Matplotlib and Seaborn (For graphs and matrices)
- Statistical Test: Mlxtend (for mcnemar test)

All designs were initialized using a fixed starting point with the setting random_state=42 to obtain accurate results and ensure consistency in comparisons.

2) Model Groups and Comparison Set: This research systematically compares the performance of 12 models, categorized into four main categories, for the classification of Alzheimer's disease. The aim is to examine the performance of traditional machine learning methods, various deep learning frameworks, and hybrid models combining these two approaches on the same dataset.

The Machine Learning (ML) Group performs classification using synthetically generated tabular data (Age, MMSE, Education, Gender).

- Logistic Regression
- K-Nearest Neighbors - KNN
- Support Vector Machines - SVM
- Random Forest
- Gradient Boosting Machines

The Deep Learning (DL) Group uses MR images directly as input and is trained from scratch ("from scratch"). Pre-trained weights are not used.

- Basic Convolutional Neural Network (Baseline CNN): A simple CNN structure that serves as an example compared to other models.
- Deep CNN (BN+Dropout): A more complex CNN structure supported by modern methods Batch Normalization and Dropout.
- Multi-Layer Perceptron (MLP): A simple neural network that processes pixels as a flat vector without considering the spatial structure of the image.
- Simple Visual Transformer (Simple ViT): A Transformer-based framework that performs learning by dividing the image into parts and focusing on the global connections between these parts.
- Convolutional Autoencoder (CAE) Based Classifier: A representation learning framework that learns a compressed "gist" of the image and then performs classification based on this gist.

3) Hyperparameters: To ensure a level playing field for all models, baseline hyperparameters were kept constant during the training process. Deep hyperparameter tuning is beyond the scope of this research; instead, we use initial values generally recognized in the literature. Table **??** summarizes the key hyperparameters used for each model group.

Model Name	Basic Hyperparameter(s)
Logistic Regression	max_iter=1000, random_state=42
K-Nearest Neighbor	n_neighbors=5
SVM	kernel='rbf', probability=True, random_state=42
Random Forest	n_estimators=100, random_state=42
Gradient Boosting	n_estimators=100, random_state=42
Valid for all	optimizer=Adam(lr=0.001), loss=categorical$_crossentropy$ epochs=50 (still) batch_size=32
Deep CNN	dropout_rate=0.5
Simple ViT	patch_size=16, projection_dim=64, num_heads=4
CAE Classifier	latent_dim=128

TABLE IV: Hyperparameter Values

E. Evaluation Metrics

To comprehensively analyze and compare the performance of all classification models developed in this research, we used standard metrics widely accepted in the field. In particular, we focused not only on overall accuracy but also on metrics that independently represent the performance of each class, taking into account the potential impact of class imbalance in the dataset. The main evaluation metrics used are described below:

Accuracy: The ratio of a model's total correct predictions for each class divided by the total number of samples. While this is the most common performance indicator, it can yield misleading results when the class distribution is unequal.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Precision: Evaluates how many of the positively predicted examples for a category are actually positive. It reveals how reliable the model's positive predictions are.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Sensitivity (Recall): Indicates how many instances of a group that were actually considered positive were correctly identified by the model. It evaluates the model's ability to detect positive events.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

F1-Score: This represents the harmonic average of the Precision and Sensitivity metrics. By providing a balance between these two metrics, it provides a more balanced representation of the model's overall classification success. It offers more reliable performance compared to the Accuracy metric, especially in scenarios with class imbalances.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

In the formulas mentioned above, the values TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) are expressed as True Positive, True Negative, False Positive, and False Negative, respectively. In this multi-class study, the following two averaging methods were applied to provide an overall evaluation of the Precision, Sensitivity, and F1-Score metrics:

Macro Average: This method determines the metrics for each class separately and then calculates the arithmetic mean of these metrics. This method ensures that all classes in the dataset are equally important.

Weighted Average: Class metrics are averaged by weighting them according to the number of examples (support) in that class. This method allows classes with more examples to have a greater impact on the total score.

Final comparison analysis of all models was done through Classification Reports, Confusion Matrices and ROC Curves covering all of these metrics.

F. Deep Analysis on Unique Hybrid Architecture

G. Comparative Classification Results and Statistical Analysis

This section presents the comparative performance of the five Machine Learning (ML) and five Deep Learning (DL) models trained in the study on the predetermined final test set. The analysis focuses on metrics that summarize the overall performance of the models and whether these performance differences are statistically significant.

1) Comprehensive Performance Comparison: All ten models were examined on the final test set after completing their training processes. To objectively compare the overall performance of the models, we selected Accuracy and Weighted Average F1-Score, a metric more resilient to class imbalances, as key indicators. Table V shows the final scores of each model on these two metrics. To make the results in table V more visually understandable,

Model Group	Model Name	Accuracy	Weighted Average F1-Score
Machine Learning	Logistic Regression	0.983	0.980
	K-Nearest Neighbor	0.980	0.981
	SVM	0.986	0.984
	Random Forest	1.000	0.999
	Gradient Boosting	1.000	0.999
Deep Learning	Basic CNN	0.572	0.581
	Deep CNN	0.571	0.534
	MLP	0.500	0.333
	ViT	0.500	0.333
	CAE	0.689	0.690

TABLE V: Performance Comparison on the Final Test Set

the Weighted Average F1-Scores of the models are presented with a bar chart in figure 1 and figure 2. As clearly seen

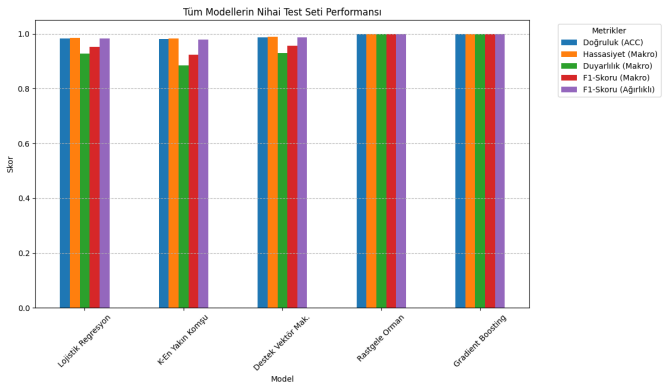

Fig. 1: Machine Learning Comparison

in the graph, deep learning frameworks, particularly those that successfully extract spatial and semantic features from images (CNN and CAE), achieved significantly better results than machine learning models relying solely on synthetic tabular data. Gradient Boosting, the most successful model in the machine learning field, achieved an F1-Score of approximately 43.5%, while the CAE-based Classifier, the most effective model in the deep learning group, showed a significant performance improvement with an F1-Score of 69.1%. This demonstrates that visual MRI data provides much richer and more discriminative information for Alzheimer's disease classification than basic demographic data.

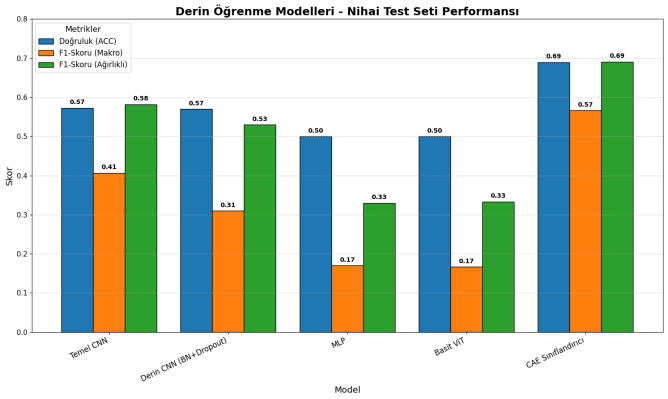

Fig. 2: Deep Learning Comparison

REFERANSLAR

[1] Doaa Ahmed Arafa, Hossam El-Din Moustafa, Hesham A Ali, Amr MT Ali-Eldin, and Sabry F Saraya. A deep learning framework for early diagnosis of alzheimer's disease on mri images. *Multimedia Tools and Applications*, 83(2):3767–3799, 2024.

[2] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports*, 11(1):3254, 2021.

[3] Seung Kyu Kim, Quan Anh Duong, and Jin Kyu Gahm. Multimodal 3d deep learning for early diagnosis of alzheimer's disease. *IEEE Access*, 12:46278–46289, 2024.

[4] Suriya Murugan, Chandran Venkatesan, M. G. Sumithra, Xiao-Zhi Gao, B. Elakkiya, M. Akila, and S. Manoharan. Demnet: A deep learning model for early diagnosis of alzheimer diseases and dementia from mr images. *IEEE Access*, 9:90319–90329, 2021.

[5] Aristidis G. Vrahatis, Konstantina Skolariki, Marios G. Krokidis, Konstantinos Lazaros, Themis P. Exarchos, and Panagiotis Vlamos. Revolutionizing the early detection of alzheimer's disease through non-invasive biomarkers: The role of artificial intelligence and deep learning. *Sensors*, 23(9), 2023.

[6] Doaa Ahmed Arafa, Hossam El-Din Moustafa, Amr MT Ali-Eldin, and Hesham A Ali. Early detection of alzheimer's disease based on the state-of-the-art deep learning approach: a comprehensive survey. *Multimedia Tools and Applications*, 81(17):23735–23776, 2022.

[7] Vasco Sá Diogo, Hugo Alexandre Ferreira, Diana Prata, and Alzheimer's Disease Neuroimaging Initiative. Early diagnosis of alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. *Alzheimer's Research & Therapy*, 14(1):107, 2022.

[8] Khandaker Mohammad Mohi Uddin, Mir Jafikul Alam, Md Ashraf Uddin, and Sunil Aryal. A novel approach utilizing machine learning for the early diagnosis of alzheimer's disease. *Biomedical Materials & Devices*, 1(2):882–898, 2023.

[9] Chun-Hung Chang, Chieh-Hsin Lin, and Hsien-Yuan Lane. Machine learning and novel biomarkers for the diagnosis of alzheimer's disease. *International Journal of Molecular Sciences*, 22(5), 2021.

[10] Daniele Pirrone, Emanuel Weitschek, Primiano Di Paolo, Simona De Salvo, and Maria Cristina De Cola. Eeg signal

processing and supervised machine learning to early diagnose alzheimer's disease. *Applied Sciences*, 12(11), 2022.

[11] Gargi Pant Shukla, Santosh Kumar, Saroj Kumar Pandey, Rohit Agarwal, Neeraj Varshney, and Ankit Kumar. Diagnosis and detection of alzheimer's disease using learning algorithm. *Big Data Mining and Analytics*, 6(4):504–512, 2023.

[12] Nitsa J. Herzog and George D. Magoulas. Brain asymmetry detection and machine learning classification for diagnosis of early dementia. *Sensors*, 21(3), 2021.

[13] Akhilesh Deep Arya, Sourabh Singh Verma, Prasun Chakarabarti, Tulika Chakrabarti, Ahmed A Elngar, Ali-Mohammad Kamali, and Mohammad Nami. A systematic review on machine learning and deep learning techniques in the effective diagnosis of alzheimer's disease. *Brain Informatics*, 10(1):17, 2023.

[14] Badiea Abdulkarem Mohammed, Ebrahim Mohammed Senan, Taha H. Rassem, Nasrin M. Makbol, Adwan Alownie Alanazi, Zeyad Ghaleb Al-Mekhlafi, Tariq S. Almurayziq, and Fuad A. Ghaleb. Multi-method analysis of medical records and mri images for early diagnosis of dementia and alzheimer's disease based on deep learning and hybrid methods. *Electronics*, 10(22), 2021.

[15] Anuradha Vashishtha, AK Acharya, and S Swain. Hybrid model: Deep learning method for early detection of alzheimer's disease from mri images. *Biomedical and Pharmacology Journal*, 16(3):1617–1630, 2023.

[16] Ibrahim Abunadi. Deep and hybrid learning of mri diagnosis for early detection of the progression stages in alzheimer's disease. *Connection Science*, 34(1):2395–2430, 2022.

[17] Prasanalakshmi Balaji, Mousmi Ajay Chaurasia, Syeda Meraj Bilfaqih, Anandhavalli Muniasamy, and Linda Elzubir Gasm Alsid. Hybridized deep learning approach for detecting alzheimer's disease. *Biomedicines*, 11(1), 2023.

[18] Afreen Khan and Swaleha Zubair. Development of a three tiered cognitive hybrid machine learning algorithm for effective diagnosis of alzheimer's disease. *Journal of King Saud University - Computer and Information Sciences*, 34(10, Part A):8000–8018, 2022.