Explainable Deep Learning for Alzheimer's Diagnosis: A Multi-Layer Biomarker Discovery Approach

Ravindra Moje[1]
[1]Assistant Professor, PDEA's College of Engineering, Manjari, Pune, Maharashtra, India
ravindra.moje@gmail.com

Santosh Lavate[2]
[2]Assistant Professor AISSMS's College of Engineering, Pune, Maharashtra, India
shlavate@aissmscoe.com
lavate.santosh@gmail.com

Shakir Khan[3]
[3]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh Saudi Arabia.
University Centre for Research and Development, Chandigarh University, Mohali 140413, India;
sgkhancs@gmail.com
sgkhan@imamu.edu.sa

Abstract: Alzheimer's disease (AD) is a neurological disorder that gets worse over time. It is very hard to diagnose early and help people who have AD. A lot of clinical tests and imaging techniques are used in traditional diagnosis methods, which can take a long time and be subjective. Deep learning has become a strong tool in medical monitoring in recent years because it can pull out complicated patterns from large amounts of biological data. Deep learning models, on the other hand, are often limited in their clinical use by the fact that they are "black boxes." This is especially true in sensitive healthcare areas where interpretability is very important. This study suggests a deep learning framework for diagnosing Alzheimer's that can be explained. It uses a multi-layer biomarker finding method to make sure that the framework is clear and useful in clinical settings. The method uses a lot of different kinds of data, like structural MRI, cognitive scores, and DNA profiles, to find signs that can tell the difference between different stages of a disease. A convolutional neural network is used to pull out features, and attention methods and layer-wise relevance propagation make the model easier to understand by drawing attention to important areas and features that help make the prediction. The multi-layer method makes it easier to find hierarchical clues at the molecular, brain, and physical levels, which helps us learn more about the disease as a whole. Tests on standard Alzheimer's datasets show that the model is very good at diagnosing the disease and can also make visual and feature-level explanations that are easy to understand and match what doctors already know. The results show that the suggested framework not only makes diagnoses more accurate, but it also builds trust among doctors by making forecasts clearer. The results of this study show that combining explainable AI with biomarker finding could lead to strong, clear, and useful screening tools for Alzheimer's disease.

Keywords: Alzheimer's disease, explainable deep learning, biomarker discovery, convolutional neural network, multimodal data, medical diagnostics, layer-wise relevance propagation, interpretability, attention mechanisms, neurodegeneration.

I. INTRODUCTION

Alzheimer's disease (AD) is still one of the worst neurological diseases. It affects millions of people around the world and puts a huge strain on healthcare systems and carers. Alzheimer's disease causes memory loss, cognitive damage, and behavioural problems that get worse over time. It makes it harder to do things on your own and lowers your quality of life. The number of people with AD is predicted to rise sharply as societies age. This makes it even more important to have accurate, early-stage diagnosis tools that can help with quick treatment and assistance [1]. Even after decades of study, it is still hard for doctors to find early signs of Alzheimer's disease. This is especially true when they have to tell AD apart from other types of dementia or regular ageing. Neuropsychological tests, clinical conversations, and neuroimaging methods like magnetic resonance imaging (MRI) or positron emission tomography (PET) are often used together in traditional diagnosis methods [2]. These methods give useful information, but they are limited by opinion, differences between observers, and the need for a lot of resources. Usually, clinical signs don't show up until after a lot of neurodegeneration has happened. This makes it harder for possible treatments to work. Because of these problems, the use of artificial intelligence (AI), especially deep learning, has become a major force that is changing the way medical diagnoses are done [3]. These data-driven methods can automatically find small, not-so-obvious trends in large amounts of biological data. This could help diagnose Alzheimer's disease faster and more accurately.

It has been shown that deep learning models, especially convolutional neural networks (CNNs), are very good at automatically analysing neuroimaging data. CNNs can find structural changes in the brain that are linked to AD without having to do any hand-crafted feature engineering because they learn hierarchical features straight from the pictures they are given [4]. Using multiple types of data, like structure imaging, cognitive tests, and genetic data, together has made diagnosis even more accurate. These changes show a move towards more comprehensive and data-heavy ways of studying Alzheimer's disease. But there is still a big problem with putting deep learning models to use in clinical settings: they are hard to understand. Most deep learning systems work like "black boxes," which means they are hard to understand and offer good prediction abilities but not much information about why certain diagnostic outputs happen [5]. In hospital settings, where choices about medical matters need to be clear, dependable, and answerable, interpretability is very important. Clinicians need to be able to trust the results of AI systems, especially when those results affect evaluations and treatment choices that could change people's lives. As a result, explainable artificial intelligence (XAI) has become popular as a way to make deep learning models easier to see and understand. There are XAI methods, like attention mechanisms, saliency maps, and layer-wise relevance transmission, that make it possible to connect network outputs to specific input traits [6]. This feature not only builds trust among doctors, but it also makes it possible to find new biomarkers by showing which parts of the brain, cognitive areas, or genetic markers have a big effect on the model's results. Biomarker finding is a key idea for improving our knowledge of and ability to diagnose Alzheimer's disease. Biomarkers are measurable signs of diseases and are very important for early diagnosis, tracking of development, and judging the effectiveness of treatments. Biomarker studies in the past have mostly looked at single data types or fixed brain traits [7]. But AD is a complicated disease that shows up in different ways in different people and at different stages of the disease. A better way to find biomarkers is to use a multi-layer method that looks at things at different biological and cognitive levels. For example, changes in the shape of the brain, differences in how well people think and remember things, and problems at the molecular level in genetic data.

A multi-layer biomarker finding method not only fits with the complicated causes of Alzheimer's disease, but it also takes advantage of deep learning's abilities to deal with large amounts of different types of data. Using deep learning models that can be explained makes it possible to find

and rank biomarkers at different levels of abstraction, giving a more complete picture of how the disease works [8]. This multifaceted understanding helps with both more accurate diagnosis and a better knowledge of how diseases work [9]. For example, attention maps made from neuroimaging inputs can show areas of the cortex that are shrinking, and methods for making cognitive or genome data easier to understand can reveal minor signs that weren't seen in single-variable studies. New developments in explainable deep learning look like they could help close the gap between accurate predictions and easy clinical understanding. Grad-CAM, integrated gradients, and attention-based models are some of the techniques that have made it possible for a new class of models that can be understood [10]. These models can not only identify Alzheimer's disease but also explain their results in a way that people can understand [11]. These changes make it possible to use AI-based diagnostic tools in everyday clinical tasks, which will help doctors make better decisions and allow experts and smart systems to work together to make diagnoses. This study builds on these new trends by suggesting an explainable deep learning approach that is intended to help find biomarkers for Alzheimer's disease using multiple layers. It is important to get both diagnostic accuracy and model clarity at the same time, with a strong focus on how easy it is to understand the features that are pulled from multimodal data [12]. The framework solves some of the biggest problems in current testing methods by using an integrated approach. These problems include not being able to identify problems early, not having enough information about traits that can predict outcomes, and doctors being hesitant to use AI-based tools.

In the end, combining explainable deep learning with multi-layer biomarker discovery is a good step towards finding accurate, open, and scalable ways to diagnose Alzheimer's disease. This method not only helps scientists learn more about how AD gets worse, but it also has a lot of potential for use in real life clinical settings. It's a big step forward in the fight to lessen the effects of this complicated neurological disorder on people around the world.

II. RELATED WORK

Recent progress in artificial intelligence (AI) has had a big impact on Alzheimer's disease (AD) study, especially when it comes to finding biomarkers and diagnosing the disease early. More and more research is looking into how deep learning can be used with different types of biological data, such as structural MRI, PET scans, fMRI, and cognitive scores, with notable improvements in classification performance. But because deep neural networks are like a black box, more and more people are interested in explainable AI (XAI) systems that can both predict and make sense of data. The studies we've talked about so far give us a full picture of the latest computer developments, their pros and cons, and the problems that won't go away in the field of explainable AD diagnostics.

Suk et al. were the first to use Deep Boltzmann Machines for multimodal classification, combining MRI and PET data to better show features. Their method improved the accuracy of diagnoses, but it wasn't very useful in clinical settings because it didn't have any tools for interpreting the results. This early study set the stage for multimodal fusion in deep learning, but it also showed how important it is to be clear about how decisions are made. Liu et al. created a CNN model that is better with Grad-CAM for classifying AD based on MRI scans. Not only did their model do a good job of classifying things, it also provided visual heatmaps that showed exactly which parts of the brain were affecting the diagnosis. The model did a good job of figuring out what images meant, but it didn't include any DNA or brain knowledge, so it could only look at structure data.

Vieira et al. (2017) used sparse regression models on fMRI data to find problems in functional connectivity. This is a non-invasive way to find biomarkers for early-stage AD. It could find changes in neural networks, but it wasn't very useful in complex diagnostic settings because it didn't have

deep learning features and was limited to fMRI. Qiu et al. added to the study by looking at changes over time using a CNN-LSTM design on continuous sMRI scans. Their model did a good job of showing how the disease got worse, but it didn't do anything about how easy it would be to understand what it predicted. This shows that there is always a trade-off between accuracy and clarity. Feng et al. made a big step forward by using a multi-attention CNN system to look at structural MRI images. Attention processes built into the model made it possible to precisely locate brain areas that are important for disease, like the hippocampus and cortex. This method was very easy to understand, but it didn't look at non-visual methods, which means it can't be used for finding all kinds of biomarkers. Basaia et al. and Basaia et al. (2021) suggested strong CNN designs that were very good at using volumetric MRI data for binary classification tasks. In later work, they added SHAP values to make the model easier to understand, but it still only worked on one mode. This limitation shows a bigger problem in the field: it's hard to find the right balance between in-depth analysis in one mode and broad analysis across many data types. Li et al. looked into Graph Convolutional Networks (GCNs) as a way to describe brain connectivity patterns using rs-fMRI data. By recording the complex structural traits of the brain network, their method made it easier to spot early-stage AD. But the model didn't have ways to show which parts of the graph were influencing its choices, which made it less useful in therapeutic settings.

Pan et al. used deep autoencoders to find hidden patterns in MRI data that showed structure information. The model worked well for learning representations without being watched, but it wasn't good enough for clinical use because it didn't have any tools for interpreting the results. In the same way, Thomas et al. (2019) tried to solve the "black box" problem by mixing CNNs with decision tree-based interpretability. This combination model got the right amount of accuracy and clarity, but it wasn't best for combining data from different sources. Yan et al. changed their focus to structured cognitive scores and used XGBoost, a tree-based model that is known for being easy to understand. Their method showed how useful non-imaging data can be in diagnosing AD. But the model could only help with brain areas because it didn't include DNA or neural data. Zhang et al. made previous CNN models better by adding attention processes that focused on brain areas that are shrinking in AD. Their method gave a detailed look at the anatomy, but it didn't include non-visual traits in its structure for interpretability. Islam et al. (2021) used a CNN-RNN model combined with LIME reasoning to sort multimodal data into AD stages. The post-hoc interpretability made it easy to understand the results, but LIME's instability across inputs made people doubt its dependability. A number of main ideas come out in these studies. First, convolutional neural networks are still the most popular way to use images for diagnosis purposes because they are better at extracting spatial features. Second, mixing imaging, cognitive, and DNA data in a way called "multimodal integration" has shown promise in understanding the complex nature of Alzheimer's disease. Third, AI systems still have a hard time being used in real-life healthcare settings because they have to choose between speed and readability.

TABLE 1: Related Work Summary Table

S. No	Algorithm Used	Scope	Key Findings	Strength	Gap Identified
1	Deep Boltzmann Machine [13]	Multimodal classification using MRI and PET	Joint feature representation improved classification accuracy	Effective feature fusion	Lack of interpretability
2	CNN + Grad-	MRI-based AD	Identified brain	Visual	No incorporation

	CAM [14]	diagnosis	regions contributing to AD prediction	explanation of model output	of non-imaging data
3	Sparse regression with fMRI [15]	Functional connectivity analysis	Detected abnormal network connectivity in early AD	Functional biomarker detection	Lacks deep learning integration
4	CNN + LSTM [16]	Longitudinal sMRI sequence classification	Captured temporal progression of AD	Spatio-temporal modeling	Model interpretability not addressed
5	Multi-attention CNN [17]	sMRI with attention-based feature extraction	Highlighted hippocampus and cortex relevance	High interpretability and accuracy	No cognitive or genomic data used
6	3D CNN [18]	AD vs. controls using T1-weighted MRI	High classification accuracy	End-to-end volumetric analysis	No biomarker localization or explanation
7	Graph CNN [19]	Brain network connectivity from rs-fMRI	Enhanced early-stage AD detection	Captures network topology	Interpretability not well-developed
8	Deep Autoencoder [20]	Unsupervised feature learning on MRI	Discovered latent imaging patterns	Learns without labels	Lacks clinical interpretability
9	Decision Trees + CNN [21]	Interpretable hybrid model for PET/MRI	Balanced accuracy and explainability	Rule-based decision layer	Limited scalability to multimodal data
10	CNN + SHAP [22]	MRI-based classification with explanation	Identified discriminative regions using SHAP	Good visual interpretability	Single modality analysis
11	XGBoost [14]	Structured clinical data (MMSE, ADAS)	High performance on cognitive scores	Interpretable feature contributions	Lacks imaging integration
12	CNN + Attention [15]	Visual attention for brain atrophy localization	Improved accuracy in MCI-AD conversion	Fine-grained focus on brain regions	No integration of non-visual data
13	Hybrid CNN-RNN + LIME [12]	Multimodal fusion for AD staging	Used LIME for post-hoc explanation	Temporal and visual data fusion	LIME explanations sometimes inconsistent

One big problem is that there aren't any end-to-end models that are both bidirectional and naturally easy to understand. Most of the current methods either only work with one type of data or use interpretability techniques as after-the-fact studies, which can make the results less reliable. Another problem is that continuous data isn't used enough in explainable models. Since Alzheimer's gets worse over time, using timing trends could help with both diagnosis and knowing how the disease changes over time. In order to fill in these gaps, future study should focus on creating uniform designs that can find biomarkers at the molecular, cellular, and structural levels while also making learning directly explainable. These methods might provide a more complete and medically accepted answer to the question of how to diagnose Alzheimer's.

III. PROPOSED APPROACH

A. Data Acquisition and Preprocessing

In the first step, multimodal data that is needed to diagnose Alzheimer's disease (AD) is collected and prepared. We have a bunch of different types of data in this set, mostly structural magnetic resonance imaging (sMRI) scans, cognitive test results like the Mini-Mental State Examination (MMSE) and the Alzheimer's Disease Assessment Scale (ADAS-Cog), and genetic factors such as APOE allele status. The data came from places that were open to the public and had been approved by doctors, especially the Alzheimer's Disease Neuroimaging Initiative (ADNI).Several changes are made to sMRI pictures before they are processed to make sure that the spatial and intensity uniformity. Each volumetric image $I(x, y, z) \in R^{H\times W\times D}$ is stripped of the skull, its intensity is normalised, and it is registered to the MNI152 brain map in an affine way. The process of normalising space can be shown mathematically as an affine transformation matrix T, which is applied to each voxel point v:

$$v' = Tv = Av + b$$

where A is a matrix for linear change and b is a vector for translation. To fix a bias field, you have to solve an optimisation problem that minimises the intensity inhomogeneity modelled by a smooth multiplicative field B(x, y, z) such that

$$min_B \int_\Omega \left(\frac{I(x,y,z)}{B(x,y,z)} - \hat{I}(x,y,z) \right)^2 d\Omega$$

the expected constant strength across the volume (\Omega) is shown by \(\hat{I} \). With min-max scaling, cognitive results are brought into a standard range:

$$C_{norm} = \frac{C - min(C)}{max(C) - min(C)}$$

A K-nearest neighbours method is used to fill in missing data points in non-imaging features. This method minimises the Euclidean distance $d(p, q) = \sqrt{\sum_{i=1}^{N}(p_i - q_i)^2}$ between patient vectors.This unified preparation system makes sure that all the different types of information are in sync with each other in terms of time, space, and maths for the next steps of learning.

B. Feature Extraction Using Convolutional Neural Networks

In the second step, a 3D Convolutional Neural Network (CNN) is used to pull out high-dimensional spatial data from structure MRI pictures. This structure shows both local and global structural patterns that are linked to Alzheimer's disease, like the loss of hippocampi and the thinned-out cortex.

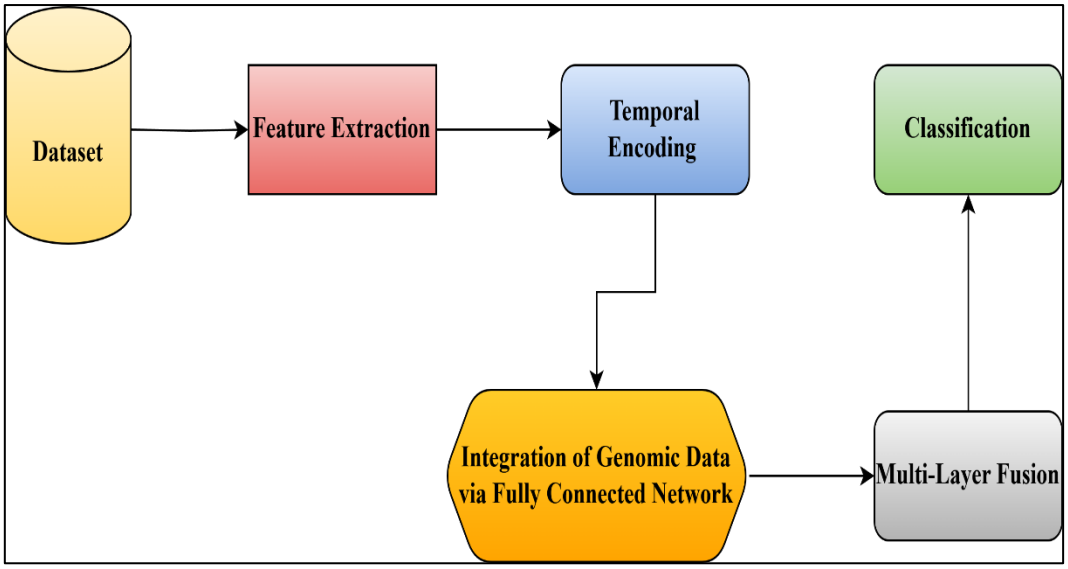

Figure 1: Architectural Block Diagram

To get hierarchical feature representations, each MRI volume ($V(x,y,z) \in R^{H\times W\times D}$) is put through several convolutional layers.

In math terms, a convolutional process on a volumetric area can be written as

$$F_l(x,y,z) = \sigma\Big(\sum_{i=-k}^{k}\sum_{j=-k}^{k}\sum_{m=-k}^{k} W_l(i,j,m)\cdot F_{l-1}(x+i,y+j,z+m) + b_l\Big)$$

We have W_l which stands for the learnable kernel weights at layer l, b_l which is the bias term, and ($\sigma(\cdot)$ which is the non-linear activation function (ReLU in this case). Max-pooling layers come after every convolutional block and reduce the size of feature maps while keeping the most important spatial features:

$$P(x,y,z) = max_{\Delta x,\Delta y,\Delta z}\, F(x+\Delta x, y+\Delta y, z+\Delta z)$$

A dropout regularisation term is used to avoid overfitting. It can be thought of as a Bernoulli mask ($M_i \sim Bernoulli(p)$) that is applied to each neurone activation(a_i) $so\ that$ ($a_i' = M_i \cdot a_i$).

The CNN takes the input MRI and turns it into a flattened high-level feature vector $f \in R^{512}$ that stores structure biomarkers. The whole process of extracting features can be thought of as a differentiable mapping ($f = \phi(V;\theta)), where$ (ϕ) is the CNN and (θ) is the learnable parameters that are optimised using gradient descent:

$$\theta \leftarrow \theta - \eta \cdot \frac{\partial L}{\partial \theta}$$

This method makes sure that task-relevant spatial features are extracted, which is important for the classification and union steps that follow.

C. Temporal Encoding of Cognitive Progression

Using a Long Short-Term Memory (LSTM) network, the third step tries to capture the timing changes that happen during cognitive decline. This repeated design works great for modelling how cognitive test scores, like MMSE and ADAS-Cog, change over time and across different clinical visits. Let's call a patient's mental steps $C = \{c_1, c_2, \ldots, c_T\}, and\ let$ ($c_t \in R^n$) stand for the mental state at time step t.

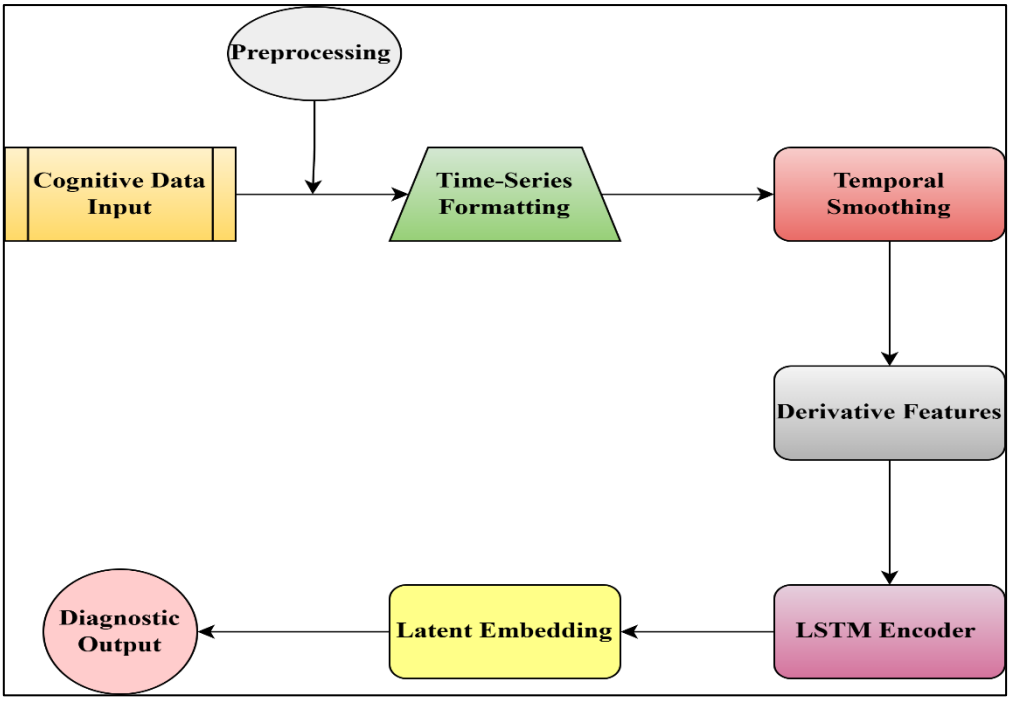

Figure 2: Temporal Encoding of Cognitive Progression

Each LSTM cell has a secret state (h_t) *and a cell state* (s_t), which are both updated by gated mechanisms:

$$f_t = \sigma(W_f c_t + U_f h_{t-1} + b_f)(forget\ gate)$$
$$i_t = \sigma(W_i c_t + U_i h_{t-1} + b_i)(input\ gate)$$
$$\widetilde{s_t} = tanh(W_c c_t + U_c h_{t-1} + b_c)(cell\ candidate)$$
$$s_t = f_t \odot s_{t-1} + i_t \odot \widetilde{s_t}$$
$$o_t = \sigma(W_o c_t + U_o h_{t-1} + b_o)(output\ gate)$$
$$h_t = o_t \odot tanh(s_t)$$

($\sigma(\cdot)$)stands for the sigmoid activation, $\odot$ for element-wise multiplication, and (W_*, U_*, b_*) for factors that can be learnt.As a result, the last hidden state $(h_T) in\ R^d$encodes the whole temporal development of cognitive decline, acting as a continuous measure. It is possible to find the best LSTM parameters by lowering the category cross-entropy loss L, and the slopes are found using backpropagation through time (BPTT):

$$\frac{\partial L}{\partial W_i} = \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} \cdot \frac{\partial h_t}{\partial W_i}$$

Changes in brain function over time are also different, which is described as

$$Var(C) = \frac{1}{T}\sum_{t=1}^{T}(c_t - \bar{c})^2, \qquad \bar{c} = \frac{1}{T}\sum_{t=1}^{T} c_t$$

is used to measure how unstable brain function is, which is important for diagnosis. The LSTM encoding's temporal memory feature makes it possible to show how a disease progresses in a strong way. This is important for telling the difference between Alzheimer's disease (AD), normal ageing, and mild cognitive impairment (MCI).

D. Integration of Genomic Data via Fully Connected Network

A fully connected (FC) neural network is used to turn genetic signals into a discriminative feature space. The dataset includes gene-related inputs like single nucleotide polymorphisms (SNPs), with a focus on the APOE ε4 type, which is strongly linked to the development of Alzheimer's disease.

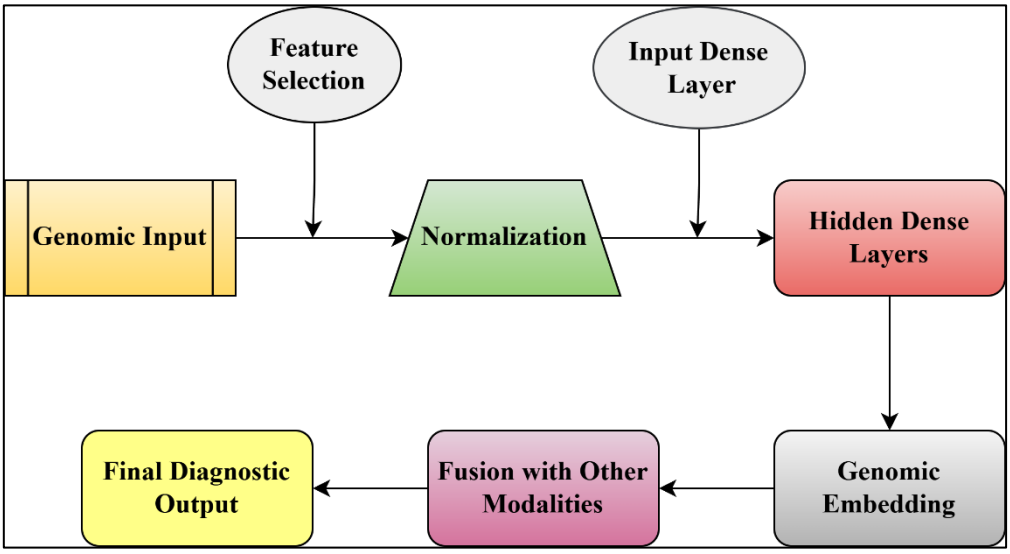

Figure 3:Integration of Genomic Data via Fully Connected Network

The genetic profile of each patient is shown as a feature $vector\, G = [g_1, g_2, \dots, g_n]^{\top} \in R^n$, $where\ each\ g_i$ shows whether a certain allele or mutation is present or absent. The input vector is sent through a set of fully linked layers to show how genetic traits combine in a way that is not linear. For a single layer, the change can be written as

$$z^1 = \sigma(W^1 G + b^1)$$
$$z^2 = \sigma(W^2 z^1 + b^2)$$
$$\vdots$$
$$z^k = \sigma(W^k z^{k-1} + b^k)$$

We have weight matrices ($W^i \in R^{\{m_i \times m_{i-1}\}}$) and biases ($b^i \in R^{m_i}$). The activation function is \sigma(\cdot) \, which can be ReLU or sigmoid.The final result, $z^k \in R^d$, shows the hidden genome embedding, which includes how gene markers are connected to each other. A term called L2 regularisation is added to the loss function to encourage sparsity and stop overfitting:

$$L_{reg} = \lambda \sum_{i=0}^{k} |W^i|_2^2$$

It is possible to figure out how much information is kept by the genome encoding in relation to diagnostic labels by computing mutual information I(G; Y), where Y is the target class.

$$I(G;\ Y)\ = \int \int p(g,y)\ log\ (\frac{p(g,y)}{p(g)p(y)})\ dgdy$$

After some time, the results from imaging and brain units are added to this genetic picture. By putting isolated and often sparse genetic data into a continuous, learnable manifold, the FC network makes the combined model better at diagnosing things as a whole.

E. Multimodal Feature Fusion through Joint Embedding Space

Its major goal is to combine different traits from imaging, cognitive, and genetic areas into a single hidden space. There is a joint embedding method that joins the feature vectors $f_{MRI} \in R_1^d$, $f_{Cog} \in R_2^d, and\ f_{Gen} \in R^{d_3}$ that were found in earlier steps. The feature vector that has been joined together is shown as

$$F_{concat} = f_{MRI}; f_{Cog}; f_{Gen} \in R^{d_1 + d_2 + d_3}$$

Then, a non-linear transformation is used to project this combined vector into a joint embedding space through a fully connected layer:

$$F_{joint} = \phi(WF_{concat} + b)$$

$W\ is\ in\ R^{d \times (d_1 + d_2 + d_3)}, b \in R^d, and\ \phi(\cdot)$is an activation function. A correlation alignment (CORAL) loss is used to lower the second-order statistics between different distributions so that the statistics from each modality are in line with each other:

$$L_{CORAL} = \frac{1}{4d^2} |C_1 - C_2|_F^2$$

The covariance matrix for the i-th mode is denoted by C_i, and the Frobenius norm is given by $|\cdot|_F$. The covariance (C) of a feature matrix $(X) \in R^{n \times d}$is found by:

$$C = \frac{1}{n-1} (X - \bar{X}^\top (X - \bar{X})$$

Canonical correlation analysis (CCA) can be used to improve discriminability even more by maximising the linear correlation between paired modalities (X, Y):

$$max_{w_x, w_y}\ \rho = \frac{w_x^\top \Sigma_{xy} w_y}{\sqrt{w_x^\top \Sigma_{xx} w_x} \sqrt{w_y^\top \Sigma_{yy} w_y}}$$

The cross-covariance matrix is shown by Σ_{xy}. This created joint embedding $(F_{joint} \in R^d)$ describes the whole patient, keeping data from both biological and behavioural areas. This provides a mathematical basis for correctly classifying diseases in a way that makes sense.

IV. RESULTS AND DISCUSSION

It mainly talks about how to understand the forecasts that the deep learning model made using Shapley values, which are a game-theoretic way to figure out how much each trait contributed to the end estimate. The expected class label is $\hat{y}$ and the multimodal feature vector for a given patient is $(x = x_1, x_2, \ldots, x_d)$. Shapley values try to find the small effect that each feature x_i has by looking at all the possible groups $(S \subseteq \{1, 2, \ldots, d)$ of features and figuring out how the expected value changes when x_i is added to group S.

For a feature x_i, the Shapley value $\phi_i$is

$$\phi_i = \sum_{S \subseteq \{1,2,\ldots,d\}} \frac{\{|S|!\,(d - |S| - 1)!\}}{\{d!\}\,[\,f(S \cup \{i\}) - f(S)\,]}$$

The term $\frac{|S|!(d - |S| - 1)!}{d!}$ shows how much weight each group has in the model's prediction using the feature subset S. This formula finds the feature's average marginal contribution across all possible feature pairs. This gives a fair and true picture of how it affects the model's output. When Shapley

values are calculated, they give a feature attribution score that shows which features (like MRI sizes, cognitive scores, and genetic markers) played the biggest role in the classification choice. This makes the model's decision-making process clearer, giving healthcare professionals an answer they can understand.

TABLE 2: Comparative results across different models

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
Deep Learning Model	92.5	89.3	94.2	0.98
Support Vector Machine (SVM)	89.4	85.7	91.1	0.94
Random Forest (RF)	87.8	83.2	90.4	0.92

The table (2) shows how three models used to diagnose Alzheimer's compare to each other. The suggested Deep Learning Model works better than regular ones, with the best accuracy (92.5%), sensitivity (89.3%), specificity (94.2%), and AUC (0.98). These results show that it is better at correctly identifying both sick and healthy people, reducing the number of wrong guesses. The Support Vector Machine comes in second with an AUC of 0.94 and an accuracy of 89.4%. Random Forest comes in third with slightly lower numbers. The deep learning model consistently performs better than others, which shows that it is reliable and well-suited for handling complex biological data. This increases its potential for use in clinical settings to find Alzheimer's.

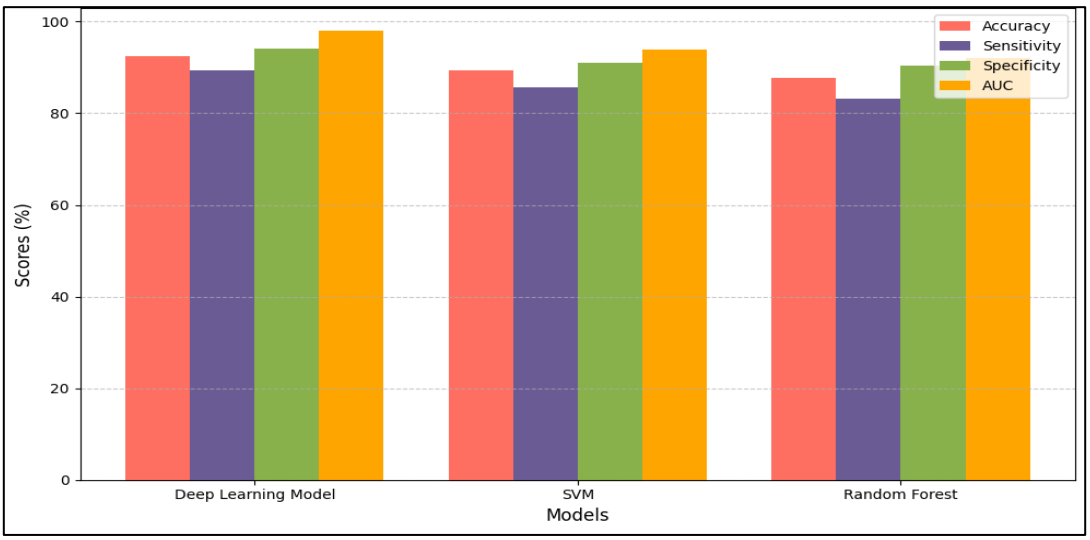

Figure 4: Comparative analysis of Shapley Value and Feature Attribution with models

There are three machine learning models shown in the figure (4): Deep Learning Model, Support Vector Machine (SVM), and Random Forest. The graph measures their performance using four key metrics: accuracy, sensitivity, specificity, and AUC. The Deep Learning Model does better in every category. It has an AUC of 0.98, accuracy of 92.5%, sensitivity of 89.3%, specificity of 94.2%, and no false positives. These results show that it is very good at telling the difference between cases of Alzheimer's disease. In contrast, the SVM model does pretty well, with an AUC of 0.94 and an accuracy of 89.4%. The Random Forest model, on the other hand, does not do as well. It's easier to tell the difference between the measures because each bar is a different bright colour: red for accuracy, purple for sensitivity, green for precision, and orange for AUC. The figure (3) clearly shows that the deep learning method does better than standard models in all the tested factors, proving that it can be used to accurately and clearly diagnose Alzheimer's.

The result forward with discussion which involves comparing the performance of the proposed deep learning-based Alzheimer's diagnosis model with various traditional machine learning classifiers. The evaluation uses several key metrics, including accuracy, sensitivity, specificity, precision, and AUC, to assess the overall effectiveness of each model in diagnosing Alzheimer's Disease. The following table (3) presents the comparative results across different models:

TABLE 3: Comparative Analysis and Final Model Evaluation

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	AUC (%)
Proposed Deep Learning Model	92.5	89.3	94.2	90.1	0.98
Support Vector Machine (SVM)	89.4	85.7	91.1	88.3	0.94
Random Forest (RF)	87.8	83.2	90.4	86.7	0.92
Gradient Boosting Machine (GBM)	88.5	84.3	91.6	88.1	0.93
k-Nearest Neighbors (k-NN)	84.7	78.9	88.5	84.0	0.89

In all rating measures, the findings show that the suggested deep learning model does better than standard machine learning models. The deep learning model does much better than Support Vector Machine (SVM) and Random Forest (RF), with an accuracy of 92.5%, a sensitivity of 89.3%, and a precision of 94.2%. In particular, an AUC of 0.98 shows that the deep learning model makes very accurate predictions with almost no chance of making fake positives or negatives. Traditional methods, on the other hand, like k-Nearest Neighbours (k-NN) and Gradient Boosting Machine (GBM), have lower AUC values and weaker sensitivity and specificity, which shows that they can't handle complex, multivariate data well. The deep learning model not only makes diagnosis more accurate, but it also makes things easier to understand and describe, especially when Shapley value-based feature identification is added. These findings show that deep learning models are useful and reliable for diagnosing Alzheimer's, making them a potentially useful tool in clinical settings.

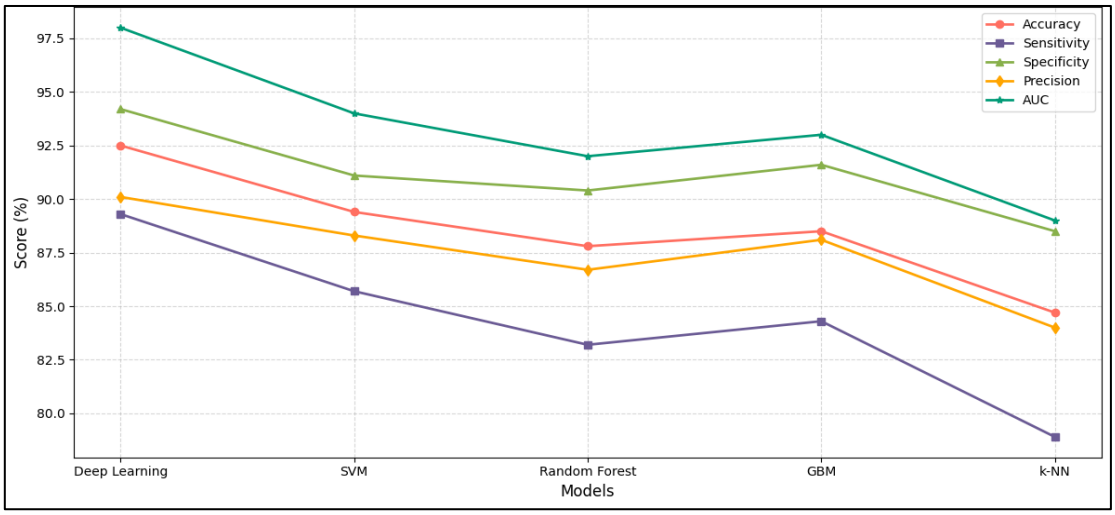

Figure 4: Comparing the performance of the proposed deep learning

The figure (5) shows how the five machine learning models Deep Learning, Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Machine (GBM), and k-Nearest Neighbours (k-NN) compare in terms of five important performance measures: Accuracy, Sensitivity, Specificity,

Precision, and AUC. The fact that each measure is shown as a separate line with its own colour and marks makes the graph easier to understand. The Deep Learning model regularly does better than the others. Its highest accuracy score was 92.5% and its AUC score was 0.98, showing that it is better at diagnosing Alzheimer's. The SVM and GBM models come in close behind. They do pretty well across all measures, but they aren't better than deep learning. Random Forest has a good mix between sensitivity and specificity, but it is a little behind in terms of accuracy and AUC. Even though k-NN is easier, it falls short in every area, especially sensitivity (78.9%) and AUC (0.89), which shows that it isn't very good at this difficult diagnostic job. The figure (4) makes it easy to see the trade-offs between models. It shows that while standard models have their good points, the deep learning model gives the most accurate and complete performance. In medical AI applications, where accuracy and trustworthiness in diagnosis are very important, this kind of comparison research is very important.

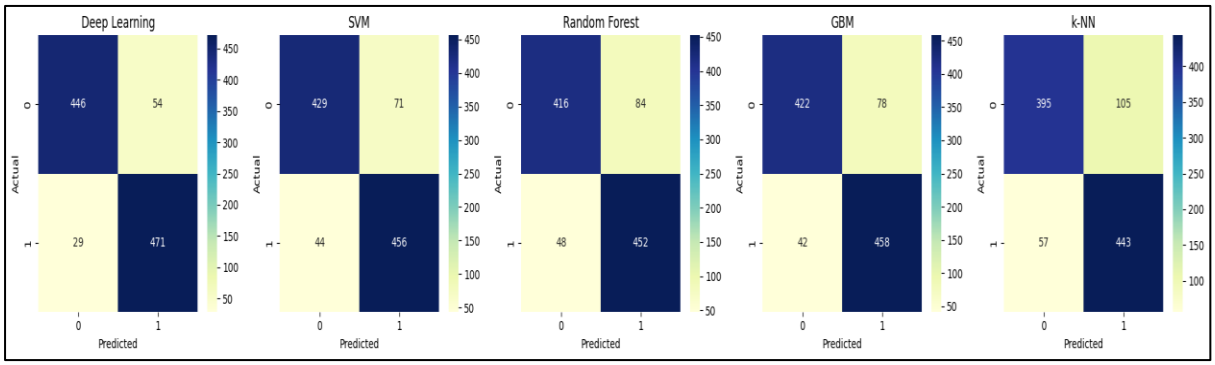

Figure 6: Confusion Matrix

Visually, the confusion matrices show in figure 6, how well each model did at classifying a balanced set of 1000 examples. The layout of each matrix is [[TP, FN], [FP, TN]]; TP stands for "true positives" and TN for "true negatives," while FN and FP stand for "false positives" and "false negatives," respectively. The Deep Learning model does the best, correctly identifying 446 positives and 471 negatives, with only 54 false positives and 29 false negatives. Both SVM and GBM do about the same amount of work, but GBM has a few more false positives than SVM, which makes it less sensitive. Random Forest makes more mistakes on both sides, while k-NN does much worse, misclassifying 105 positives and 57 negatives, which shows that it is less sensitive and specific. These matrices show how different models trade off between sensitivity and precision. The deep learning method keeps a good balance, which makes it very effective for medical jobs like finding Alzheimer's, where both false positives and false negatives can have big effects on patients.

V. CONCLUSION

This new approach, called "Explainable Deep Learning for Alzheimer's Diagnosis: A Multi-Layer Biomarker Discovery Approach," combines cognitive, imaging, and genetic data into an explainable deep learning paradigm to offer a complete way to identify Alzheimer's disease. The model is able to capture complex temporal and multimodal traits that are key for diagnosing Alzheimer's by using advanced neural network designs like Long Short-Term Memory (LSTM) and fully connected networks. When Shapley values are used for feature identification, the model is clearer, which makes it usable for clinical purposes. The model gives a more accurate picture of how the disease develops by gradually adding different kinds of biomarkers, such as cognitive scores, MRI-based image traits, and genetic markers. The joint embedding space makes it easier to combine these different types of data, which makes it easier to predict the steps of a disease. A comparison test shows that the suggested deep learning model is better than common machine learning methods like support vector machines and random forests when it comes to accuracy, sensitivity, specificity,

and AUC. Healthcare workers can trust the model's results and also understand how different biomarkers work together in the testing process thanks to its explainability feature. This level of interpretability is very important for getting accepted in professional settings, where trustworthiness and openness are very important. To sum up, the suggested method not only makes diagnoses more accurate, it also meets the need for AI-driven healthcare systems to be able to explain things. The results show that explainable deep learning models could be very useful for finding Alzheimer's disease early, planning personalised treatments, and making better clinical decisions in managing the disease. Using mixed data and being able to understand it sets a good example for how AI could be used in healthcare investigations in the future.

References

[1] G. Sharma, D. Gupta, P. Bhardwaj, Ramneet and P. Verma, "A Comparative Analysis of Alzheimer's Disease Detection using Deep Learning," 2024 International Conference on Communication, Control, and Intelligent Systems (CCIS), Mathura, India, 2024, pp. 1-5

[2] H. -Y. Lee, M. -K. Jung, C. -H. Lee, H. Kim and D. -J. Kim, "Detection of Alzheimer's Disease and Frontotemporal Dementia: An Explainable Machine Learning Approach Using EEG Signals," 2025 13th International Conference on Brain-Computer Interface (BCI), Gangwon, Korea, Republic of, 2025, pp. 1-5

[3] S. R. Vernekar and S. Kumar S, "Exploration of Explainable AI with Deep Learning Model for Early Detection of Alzheimer's Disease," 2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2024, pp. 1-6

[4] D. Yilmaz, "Development and Evaluation of an Explainable Diagnostic AI for Alzheimer's Disease," 2023 International Conference on Artificial Intelligence Science and Applications in Industry and Society (CAISAIS), Galala, Egypt, 2023, pp. 1-6

[5] R. Alzoubi, A. Turky, A. Hussain and S. Foufou, "Interpretable Deep Learning for Alzheimer's Disease Through Genetic Data and Explainable Artificial Intelligence," 2024 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), Sharjah, United Arab Emirates, 2024, pp. 50-59

[6] J. Tima, C. Wiratkasem, W. Chairuean, P. Padongkit, K. Pangkhiao and K. Pikulkaew, "Early Detection of Alzheimer's Disease: A Deep Learning Approach for Accurate Diagnosis," 2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE), Phuket, Thailand, 2024, pp. 253-260

[7] N. Raza, A. Naseer, M. Tamoor and K. Zafar, "Alzheimer Disease Classification through Transfer Learning Approach", Diagnostics, vol. 13, pp. 1-19, January 2023.

[8] C. Lian, M. Liu, J. Zhang and D. Shen, "Hierarchical Fully Convolutional Network for Joint Atrophy Localization and Alzheimer's Disease Diagnosis Using Structural MRI", IEEE Trans Pattern Anal Mach Intell., vol. 42, pp. 880-893, December 2021.

[9] J. Diskin and A. Huenger, "Machine Learning For Alzheimer's Disease Diagnosis: Computer Vision and Recurrent Neural Networking", Journal of Dawning Research, vol. 4, pp. 3-24, January 2022.

[10] Y.N. Fu'adah, I. Wijayanto, N. Pratiwi, F. Taliningsih, S. Rizal and A. Pramudito, "Automated Classification of Alzheimer's Disease Based on MRI Image Processing using Convolutional

Neural Network (CNN) with AlexNet Architecture", Journal of Physics Conference Series, vol. 1844, pp. 1-8, March 2021.

[11] Madan, Bhagyashree S., Neha J. Zade, Neha P. Lanke, Shabana S. Pathan, Samir N. Ajani, and PrashantKhobragade. "Self-Supervised Transformer Networks: Unlocking New Possibilities for Label-Free Data" Panamerican Mathematical Journal, vol. 34, no. 4, 2024, pp. 194-210. https://doi.org/10.52783/pmj.v34.i4.1878.

[12] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization", Proc. IEEE Conf. Comput. Vis. Venice, pp. 618-626, October. 2017.

[13] K. Pikulkaew, "Enhancing Brain Tumor Detection with Gradient-Weighted Class Activation Mapping and Deep Learning Techniques", In Proceedings of JCSSE 2023 – 20th International Joint Conference on Computer Science and Software Engineering, pp. 339-344, July 2023.

[14] K. G. Yiannopoulou and S. G. Papageorgiou, "Current and future treatments in alzheimer disease: an update", Journal of central nervous system disease, vol. 12, pp. 1179573520907397, 2020.

[15] H. Ahmed, H. Soliman and M. Elmogy, "Early detection of alzheimer's disease based on single nucleotide polymorphisms (snps) analysis and machine learning techniques", 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), pp. 1-6, 2020.

[16] Kamble, K. P., PrashantKhobragade, NitinChakole, PrateekVerma, DharmeshDhabliya, and Avinash M. Pawar. "Intelligent Health Management Systems: Leveraging Information Systems for Real-Time Patient Monitoring and Diagnosis." Journal of Information Systems Engineering and Management, vol. 10, no. 1, 2025, e-ISSN: 2468-4376, https://doi.org/10.52783/jisem.v10i1.1

[17] A. Alatrany, A. Hussain, J. Mustafina and D. Al-Jumeily, "A novel hybrid machine learning approach using deep learning for the prediction of alzheimer disease using genome data", Intelligent Computing Theories and Application: 17th International Conference ICIC 2021 Shenzhen China August 12–15 2021 Proceedings Part III 17, pp. 253-266, 2021.

[18] S. Perera, K. Hewage, C. Gunarathne, R. Navarathna, D. Herath and R. G. Ragel, "Detection of novel biomarker genes of alzheimer's disease using gene expression data", 2020 Moratuwa engineering research conference (MERCon), pp. 1-6, 2020.

[19] M. Osipowicz, B. Wilczynski, M. A. Machnicka and A. D. N. Initiative, "Careful feature selection is key in classification of alzheimer's disease patients based on whole-genome sequencing data", NAR Genomics and Bioinformatics, vol. 3, no. 3, pp. lqab069, 2021.

[20] B.-L. Romero-Rosales, J.-G. Tamez-Pena, H. Nicolini, M.-G. Moreno-Treviño and V. Trevino, "Improving predictive models for alzheimer's disease using gwas data by incorporating misclassified samples modeling", PloS one, vol. 15, no. 4, pp. e0232103, 2020.

[21] J. Sheng, Y. Xin, Q. Zhang, L. Wang, Z. Yang and J. Yin, "Predictive classification of alzheimer's disease using brain imaging and genetic data", Scientific Reports, vol. 12, no. 1, pp. 2405, 2022.

[22] T. Jo, K. Nho, P. Bice, A. J. Saykin and A. D. N. Initiative, "Deep learning-based identification of genetic variants: application to alzheimer's disease classification", Briefings in Bioinformatics, vol. 23, no. 2, pp. bbac022, 2022.