Journal of Electrical Engineering & Technology
https://doi.org/10.1007/s42835-021-00967-3

ORIGINAL ARTICLE

Prognosis of Alzheimer's Disease Progression from Mild Cognitive Impairment Using Apolipoprotein-E Genotype

M. Rohini[1] · D. Surendran[2] · S. Oswalt Manoj[1]

Received: 18 April 2021 / Revised: 2 November 2021 / Accepted: 13 November 2021

Abstract
Alzheimer's disease (AD), cerebrovascular disease, Lewy-body disease, and Frontal–temporal degeneration disease are the age-related cognitive impairments that cause dementia. However, AD is the primary cause of dementia that causes brain cell degeneration in the geriatric community. Brain cell degeneration is the crucial cause of AD, due to the abnormal accumulation of indissoluble clumps known as plaques and tangles in the human brain's neurons. Amyloid precursor protein levels and Apolipoprotein -E gene are the biomarkers of AD since it causes accumulations and hence blocks the neuron transport system throughout the body. The early onset of AD includes mild-cognitive impairment (MCI) that progresses to complete dementia. Many related works include AD prediction using clinical modality images and cognitive assessments scores of the individuals but have not addressed comparative genome study for significant subjects. However, there is a lack of affordable biomarkers for the effective early detection of high-risk individuals. In this study, we utilize one or more features of Magnetic Resonance Imaging (MRI) tests and Apolipoprotein-E genotype sequence that provides more significant biomarkers for the early prediction. The ML classifiers including Support vector classifier, Gaussian process, AdaBoost, Random Forest, Decision trees learns the subset of patterns that predicts the AD with gene descriptors from microRNA expression profile and the profiled gene pattern. These significant multiple gene descriptors provide a supportive prediction methodology that apply genotype strength with the ensemble classifiers. The final optimal model is given by validation evaluations. The support vector classifier and Random Forest classifiers had given consistent results for disease conversion and progression from MRI attributes and had given promising results with the validation that showed accuracy greater than 80% and F1 weighted score of 0.8 in disease classification and prognosis. The experimental results had proven 95% accuracy in the saliency values of APOE isoforms implemented in DragonNN framework that will vary AD pathogenic. Hence particular focus and clinical interventions can be given on Aβ genome dependent subjects that predicts the disease.

Keywords Magnetic resonance imaging (MRI) · Mild cognitive impairment (MCI) · Apolipoprotein-E genotype · Support vector classifier · Random forest classifiers · DragonNN

1 Introduction

Aging is the natural process every human being will acquire without any choice. Global Ageing community gets increasing due to improvement in medicines. The most imperial problem in the life of old aged people is the gradual impairment in their memory skills, which faints as the years' pass. Various studies show that the loss of memory and cognitive skills is due to the degeneration of their brain cells. The brain is composed of billions of these brain cells called neurons that communicate information to and from the brain and the entire body. Synapses connects these building block neurons. It transmits signals throughout the brain cells that aids memory, thinking, and decision making. Due to aging

✉ M. Rohini
rohinim@skcet.ac.in

D. Surendran
surendran.d@kpriet.ac.in

S. Oswalt Manoj
oswaltmanojibm@gmail.com

[1] Sri Krishna College of Engineering and Technology, Coimbatore, India

[2] KPR Institute of Engineering and Technology, Coimbatore, India

Published online: 29 November 2021

and several genetic reasons these neurons may degenerate. This degeneration occurs due to heredity, stress, lifestyle, and various other causes. When this degeneration is mild and expected, then the memory skills they lose seem tolerable. Once elderly individuals get abnormal memory loss, it is evident that Alzheimer's disease (AD) is already onset.AD is the high-risk factor that causes dementia [1]. Alzheimer's disease would exhibit numerous structural changes in the brain and behavioral changes in individuals before the actual onset. It causes vast depletion of memory cells. The two main factors that destroy brain cells are tangles and plaques. These are delicate specks formed by the abnormality of specific proteins in the brain, such as beta-amyloid and Tau proteins. Accumulation of plaques and tangles inside the brain and within the synapses blocks communication throughout the brain [2]. These chemical changes and electrical imbalance are because of Alzheimer's disease that occurs in the brain about ten to fifteen years before the actual onset of disease and memory loss. It is crucial to research why not all mild cognitively impaired individuals are subjected to AD. Their disease progression depends on various predictors such as Mini-Mental state examination (MMSE score), Hippocampus volume of the brain, and various other genetic factors. These genetic factors that cause late Alzheimer's disease are called biomarkers in the gene or cerebrospinal fluid [3]. These predictors are taken as input features into our classifier. We evaluated the predictors that classify these ensemble features as Mild-Cognitively impaired, Normal Cognitive due to aging, and Alzheimer's disease using Support vectors, Random Forest classifiers, and Ensemble Tree. Thus, clinicians perform initial treatments that may postpone disease onset by healing and rescuing the brain cells from degenerations. The present investigation classifies the predictor with average cognitive decline related to normal aging or the actual symptoms of the cognitive disorder for dementia in the early stage [4, 5]. The success of the research in this problem lies in the earlier diagnosis and classification of subjects into disease progressing groups or stable groups.

This study acquired Mini-Mental State Examination score, Clinical Dementia Rating, Estimated Total Intracranial Volume, Normalize Whole Brain Volume, and Atlas Scaling Factor for building random forest and SVM classifiers. The classifiers analyze the input feature predictors for early diagnosis of Alzheimer's disease. To improve the model's performance, we utilized blood sample data obtained from the public repository. We curated the gene expression data for predicting the deterministic gene variants that act as carriers of AD and increase the risk of the onset of AD in individuals. Thus, appropriate clinical interventions delay or cure AD's late-onset and treat AD's genetic carriers. Hence this optimal multitask classifier framework provides a clinically flexible strategy that utilizes real-time neuroimaging predictors to classify the individuals with Alzheimer's Disease and link the genome deep learning model that identifies the physiopathology cohorts in very early stages.

2 Related Works

Various scientific researches are ongoing that provide the solution to the AD disease sufferers in diagnosis. A set of research works include utilizing MRI, PET scan modality images and training the deep neural network layer that derives patterns of AD and normal brain. [6] Machine learning models have utilized Lewy Body disease and Micro RNA array structure for disease prediction. This work implemented ML classifiers and combined the total error rate that exceeds the threshold level the classifier is bagged with other datasets. [7] The work had constructed a class balanced and imbalanced risk prediction model and achieved an accuracy of 90% using all the clinical data sources available in single-dimensional space. It also calculated the matching gene pattern obtained by profiling the target. It further predicted candidate target genes from the miRNAs. Gene set enrichment analysis of the miRNA target genes revealed 6 functional genes included in the DHA signaling pathway associated with DLB pathology. Two of them are supported by gene-based association studies using many single nucleotide polymorphism markers (BCL2L1: $P = 0.012$, PIK3R2: $P = 0.021$). These gene-based associations are studies to predict the family history of dementia. Many studies proved good accuracy in the hippocampus volume of the human brain since degeneration is the highly vulnerable symptom once pathology begins. [8] proposed a logistic regression hypothesis in multiclass classification among the scattered data sources available from dementia cohorts, and accuracy is successfully derived [9]. Numerous investigations were performed on master framework answers for Alzheimer's illness and Mild Cognitive Impairment changes. The significance of interesting features in AD disease prediction involves selecting essential biomarkers that are applied with various AI algorithms that have shown good accuracy in the disease conversion of AD. Neuroimaging considers the fundamental and actual changes in the brain due to aging, and provides the disease conversion from mild cognitive impairment stages to AD with intellectual disability [10]. A few studies have implemented structural changes in the cerebrum from MRI imaging procedures and functional changes from PET imaging, which are biomarkers for detecting Alzheimer's disease in its early stages. MRI imaging highlights that structural feature extraction has been the subject of numerous trials. The information about the hippocampus volume is utilized to monitor fundamental changes in mind that occur due to neuron degeneration. It serves as an essential biomarker for disease visual proof caused by abnormal protein accumulation in neurons. Beta-amyloid and tau aggregations

are examples of such aggregates that block neurons and, as a result, induce cell death. [11] Because of advancements in the hippocampus and cortical mind dispersion, the cerebrum volume also changes with disease progression. Neuron degeneration, cell death, and genetic transformations all contribute to the structure, volume, and thickness changes that occur long before the onset of symptoms. Because of starting mind changes, cerebrum oversees without indicating any infection manifestations, and persistent synapse harm caused more social variations from the abnormalities [12].

The present study employs different features from MRI reports such as MSME, ASF, NBW combined with Apolipoprotein E gene analysis and trains using deep learning classifiers (Fig. 1).

2.1 Materials and Methods

2.1.1 Participants

The data utilized in the present work are obtained from ADNI participants that describe the demented group studied from MRI and PET scans. The cross-sectional data of the first 600 controls from the cognitive complaint cohorts are analyzed. This data consists of 1800 participants over age 64, including 44% female and 66% male, whose clinical investigations include 6000 samples during various follow-ups. This study performs a novel methodology that differentiates each feature and compares it with the APO gene pattern that acts as a biomarker. [13].

ADNI launched in 2003 is the public repository providing MRI and PET scan data examined during various annual assessments and regular follow-ups of elderly individuals. The repository of high dimensional longitudinal collaborative data, including clinical, genetic, imaging, and biochemical values, the significant biomarkers for vast neurodegenerative cognitive diseases. provides more advanced genetic information, including APOE apolipoprotein E of the Homosapiens category, the protein encoded gene of the apoprotein chromosome. This chromosome carries the memory of pathology 'Y' gene across generations. With several types of genes present as C1, C2 clusters, the particular occurrence of the gene or mutations of the gene causes family dementia onset. This pattern compares other descriptors studied from MRI data that classify whether the pathology causes Type I or Type II imbalance in chromosome and genome levels. This implies whether the pathology affects neurodegeneration or cardiovascular illness in humans. ADNI consisted of gene expression profiling data from the samples collected from ADNI participants acquired during clinical investigations. Affymetrix Human Genome U219 is utilized by ADNI in gene expression profiling retaining the mRNA. The neurodegeneration scores of input feature classifiers for the present study which give MMSE (Mini-Mental State Examination), CDR (Clinical Dementia Rating), eTIV (Estimated Total Intracranial Volume), nWBV (Normalize WholeBrain Volume), ASF (Atlas scaling factor), APOLIPO(Apolipoprotein-E) are studied Table 1. Figure 2, 3) represents a correlation plot study that visualizes the correlation between gene variants acquired from NCBI samples in gene profiling stages.

2.1.2 Gaussian Process

The classification tasks are initially performed with the Gaussian process that chooses the first set of combined features given in Table 1. It is evaluated for kernel density distribution function. This distribution ensures a Gaussian process with the normal distribution of data and balanced classes. The kernel density distribution function is given by (1),

$$K\left(Z,Z'\right) = D\left[Y'||\alpha^2\right] \tag{1}$$

ruwhere the parameters, Ý, α gives the complexity and error rate of the model. These are the main and auxiliary features of different subjects with volumetric weights. This substitutes the cost of multi-classified data. The disease data of 500 subjects given in longitudinal dimension, the conversion ratio from MCI to Dementia calculates this process iterating the kernel distribution function. After prediction, validation is performed for a set of 172 filtered baseline studies that show a separating boundary between classes (Table 2).

Then within each of these two subgroups, the model considered the predictors that might split those subgroups. So, the following query arises at this diagnosis as to the survival rate greater than 78%. Then the disease was more likely to present symptoms for normal cognitive disability of older individuals whose survival rate is more. [17] proposed an algorithm that continues in a gaussian distribution manner until MMSE score and brain volume features are split out into smallest subsamples, the smallest subgroups that are grouped with kernel density function K(Z,Z']. Within each of these randomly generated trees, leaves of the tree and predictions were homogeneous. There are different

Table1 Neurodegeneration score for the studied sample

Groups	MCI	AD	NC
No of subjects	220	220	160
Age	60–80	> 70	> 60
MMSE	30	34	17
CDR	1	1.5	0.5
eTIV	1456	1558	1123
nwBV	0.88	0.98	0.72
ASF	1.56	1.76	0.8

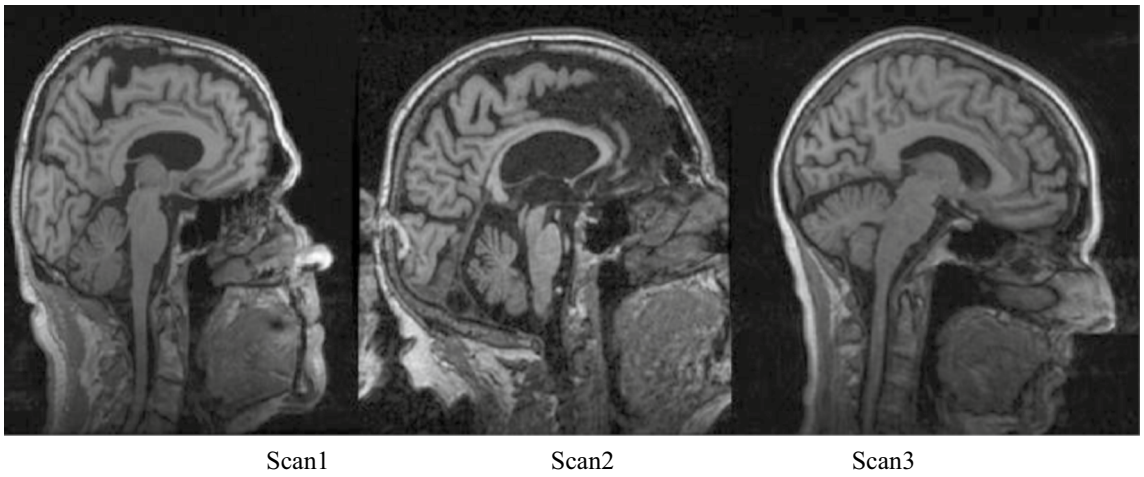

Fig. 1 MRI acquisition plane – Sagittal.Scan-1: cognitively normal aged individual, scan-2:AD individual and scan-3: Mild cognitively impaired individual (MCI)

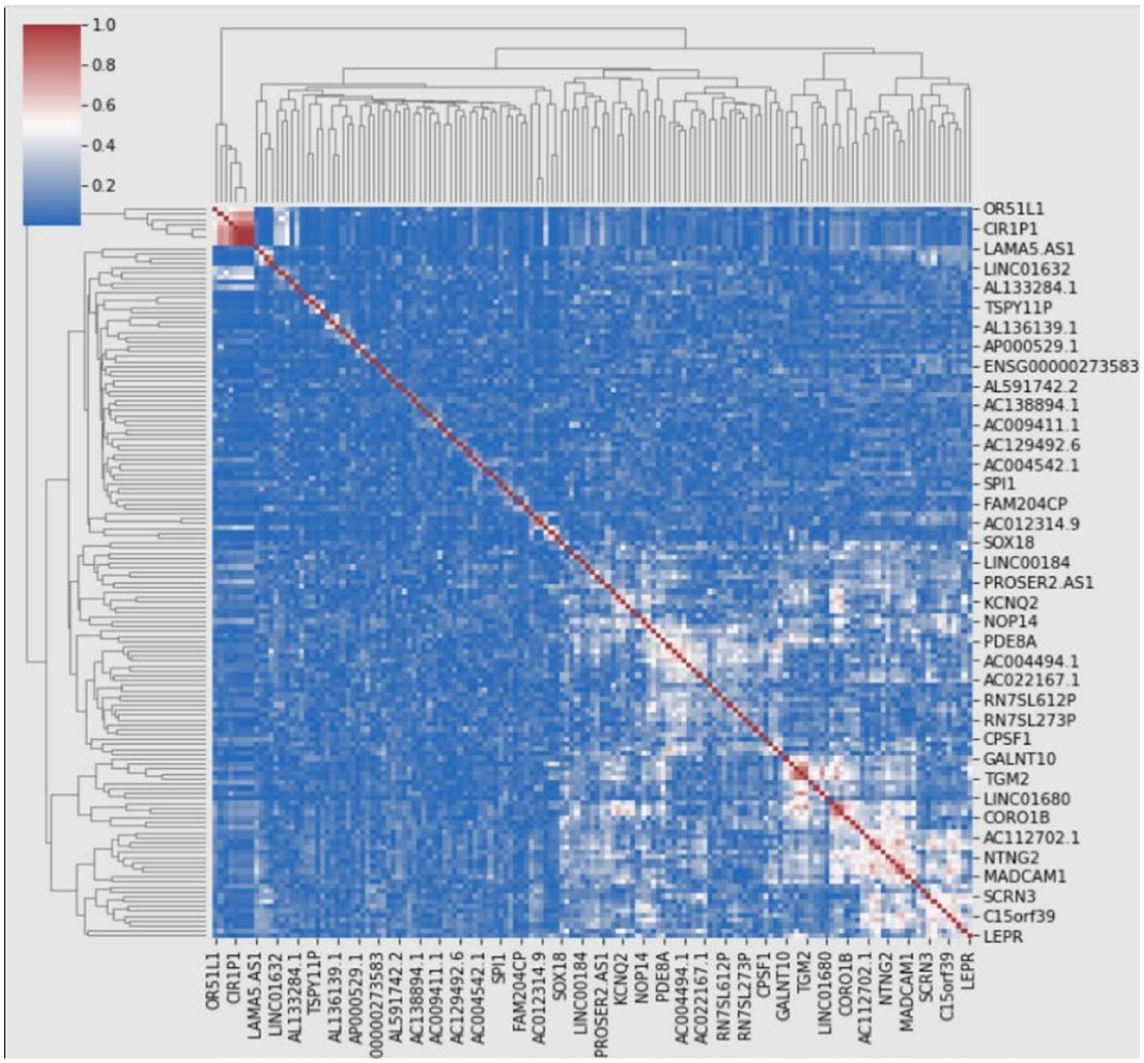

Fig. 2 Gene Co-relation Analysis of Transcriptome data

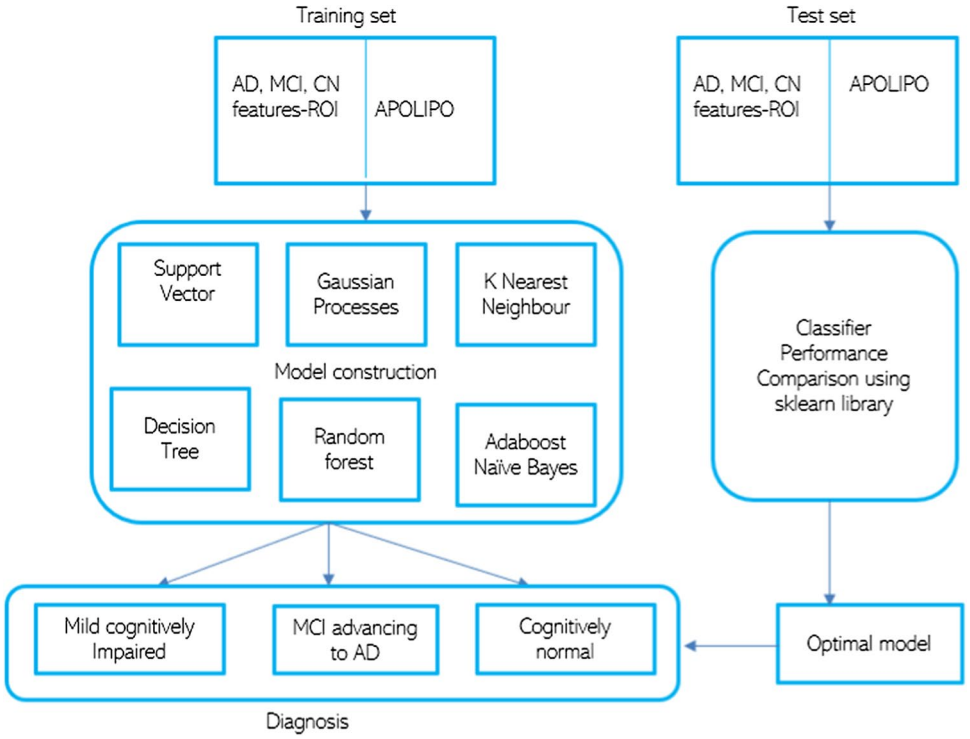

Fig. 3 System Architecture of Multi-classifier Machine learning model in AD diagnosis

Table 2 Comparison of various diagnosis models and modalities for disease classification

Modality	Diagnostic models	Accuracy	F1-Score
MRI	Randomn Forest	0.83	0.80
ADNI Test		0.82	0.59
AIBL		0.76	0.69
FHS			
COGNITIVE ASSESSMENT			
ADNI Test	Randomn Forest		
AIBL		0.95	0.95
FHS		0.91	0.77
MRI			
ADNI Test	SVM	0.83	0.80
AIBL		0.82	0.59
FHS		0.76	0.69
COGNITIVE ASSESSMENT			
ADNI Test	SVM		
AIBL		0.95	0.95
FHS		0.91	0.77
		0.76	0.51
	Gaussian	0.82	0.81
Apolipoprotein-E genotype	SVM	0.82	0.82
Learning model	Random Forest	0.81	0.80
Genetic Risk Prediction	0.98	0.9	

measures of impurities measures based on probabilities of classes that resulted from decision trees. Let us consider the classification of disease using random forest, [18] where the homogenous classes and purity considers the calculated prediction probability. Class (C) probability for the (N) random subsamples gives the total classes that trees resulted in through each iteration. For the predictor variables Xi…, Xn, the misclassification error is given as the difference between Gini index (P_{CN}), calculated in (2), for the probabilities of features assigned to class C that includes the subjects with progressive or stable cognitive impairment for variables N. Prediction probability and misclassification prediction given in (3), (4)

$$P_{CN} = 1 \Big/ N_n \sum_{xi\ in\ ci}^{xn} (1) Y_i = C_i \tag{2}$$

$$1 - P_{CN} = 1 - 1 \Big/ N_n \sum_{xi\ in\ ci}^{xn} (1) Y_i = C_i \tag{3}$$

$$\sum PCnNn = 1 - \sum P^2 C_i N_j \tag{4}$$

2.1.3 Ada Boost Algorithm

Considered subject n=i_1, i_2, i_3, …, i_{n-1}, the classification subjects are given as Demented and non-demented. This score is a bootstrap for iteration that starts from previous variables. The conversion of the predicted class from

weaker to stronger attribute that has the best discriminatory characteristic is given to the classifier. Their process involves specifying the base classifier and its input data. Ensemble of the classifier is considered and the error rate from the Ada Boost model is calculated as the difference between continuous and multi-class categorical measures. The probability P_{ada} given in (5) for the number of training samples predicting the class with N instances is given as,

$$P_{ada} = 1/2 \log \left[1 - total_{error}\right] \tag{5}$$

Suppose the training set is implemented and learned with the ensemble of classifier and has given a low error rate, the result is passed to the next index value in performance iteration. This error rate is compared and the minimum error rate derived from classifier performance is considered for prediction [19]. This works by considering the voting results given by each iteration of the ADA boost classifier.

2.1.4 Support Vector Classifier

This classifier generates a hyperplane for linearly separable two-dimension data. The hyperplane passed through mild dementated cognitively normal, normal, and AD control groups. When concatenated with feature descriptors, the genetic information uses another function for mapping the two-dimensional space to higher-dimensional space.

The equation of hyperplane is assumed as Y'=mx+c, where two descriptors are substituted and the average given

$$Y' = m\left[x1 + x2/2\right] + c \tag{6}$$

$$F\left(Y'\right) = Z\left[Y' - Y'/2\right] \tag{7}$$

where for given output Y', the hyperplane conversion parameter is given as feature vector Z. The decision boundary is derived as a minimum of several iterative training points. The function of matching APOLIPO protein is given as a hyperparameter for linear separable plane conversion of disease data in spatial feature sets. (7) Function F(Y) predicts classification results came out when kernel Y' (6), is chosen. In multi-dimensional space, each pixel is separated from linear non-separable data.

During kernel initialization and conversion from 2D space to n-dimensional kernel mapping function given as polynomial radial density sigmoid function. This is the optimization function with its kernel distribution function for spatial features such as MMSE score and Clinical Dementia Rating. The performance evaluation and parameter tuning are implemented that handles the distance of measure between Y' and y' of distributed features.

2.1.5 Ensemble Tree

The present ensemble classifier is used with random forest, where the trees are constructed to utilize the clinical data and cognitive descriptors. Each tree-based gradient descent follows learning and training from each row vector and column vectors available in the data source iteratively. We have calculated the learning and error rate by constructing the N depth of decision trees. Where each tree is compared with N-depth of column descriptors, N-tree minimum split, minimum samples, and learning rate which reached around (0.1 to 0.2). This process is repeated for N slices of feature vectors.

Gaussian process, decision tree, random forest, support vector classifier, and Ada boost were used for model construction. Using K-fold validation each dataset is mapped with Apolipoprotein E genetic information and RNDmin value is given to each layer in the learning model as implemented in [20]. The model learning is performed step by step and a highly voted prediction algorithm is mapped for genomic validation with the comparison graph. The model is fine-tuned using completed in-depth real-time data that helped us derive error rates. We analyzed to implement a feature selection algorithm that minimizes error rate with value classifier and equivalent descriptors in the selection process.

The final ensemble model was constructed by iterating and wrapping up 4 clinical features which showed improved performance. MRI features acquired by ADNI are thus learned with these ensemble models and are classified based on volumetric brain changes and results are shown in Fig. 4. which gives classification and conversion probability applying the functions (1) to (7) in the high dimensional feature vectors. Thus, the ML classifier is the linear training model where distributed nonlinear features are trained in non-linear space with kernel transformation function given in (8), (9):

$$\theta = \sum_{n=1}^{i} |\lambda i| \tag{8}$$

$$K\left(m, m^{'}\right) = e \log\left(\theta\right) \tag{9}$$

These measures give hyperparameter optimization during the training process that eliminates random subsets of irrelevant categorical classes. Thus, the pre-processing of sparse input features $\tilde{\lambda}_{1,\,2\ldots i}$ and subset of feature visualization obtained from machine learning classifier is intuitive in obtaining disease-related datasets, combining these disease predictions that are matched with gene sequences and familial causes of disease onset is that are evaluated in next stage classifier. The kernel transformation function applied

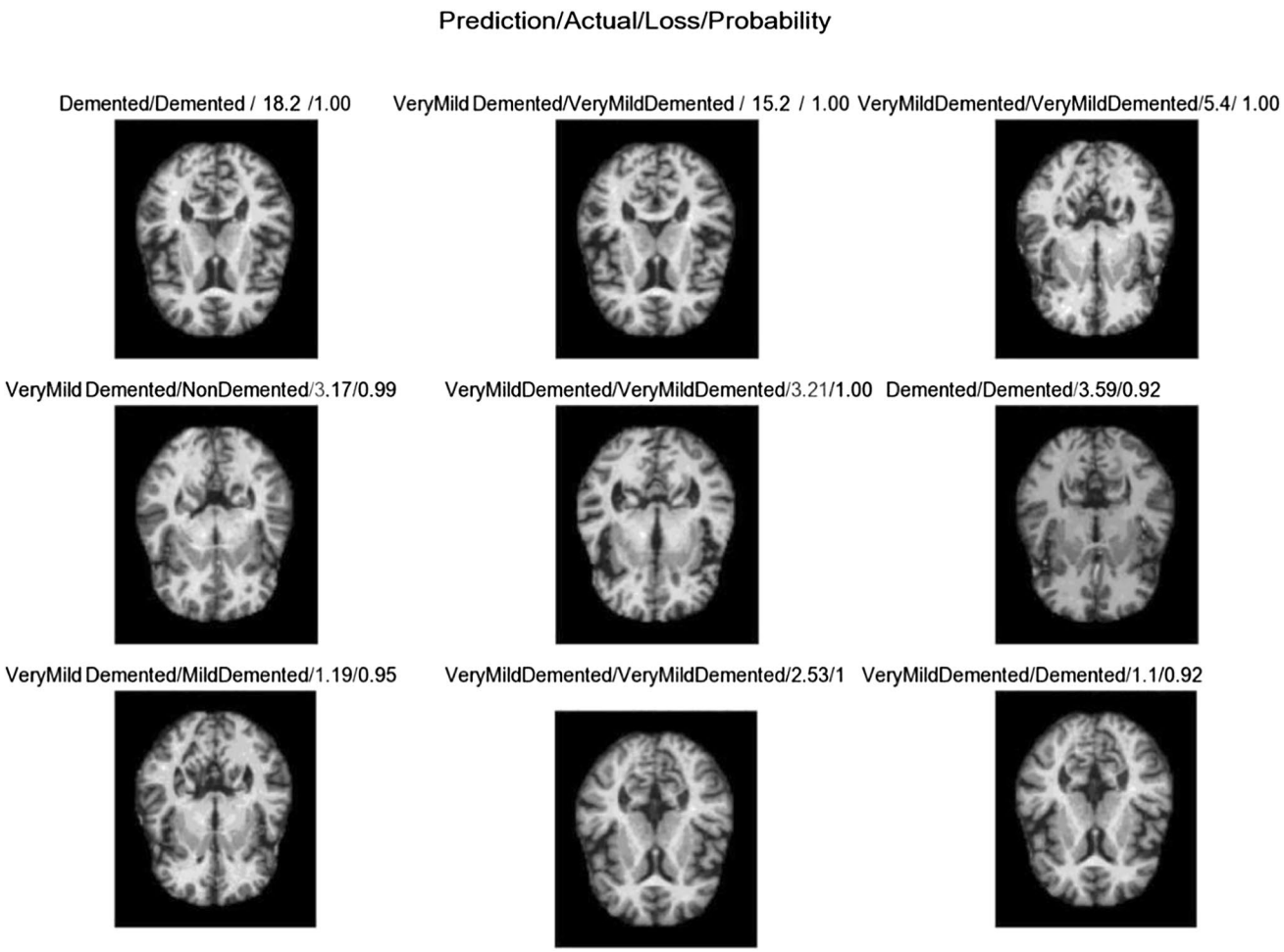

Fig. 4 Classifier results of AD prediction and MCI to AD conversion probability

between input non-linear space features 'm' mapped to the linear space 'm`' for Naive classification,

2.1.6 Gene Visualization

ADNI consisted of Gene expression profiling data from 800 ADNI participants from ADNI genetic study cohorts. Affymetrix Human Genome U219 is utilized by ADNI in gene expression profiling withholding the mRNA [21]. There are various frameworks to extract transcriptome patterns from the ADNI gene expression dataset. The disease-causing gene has resulted from RNA produced by DNA in individual brain cells of humans. The long non-coding RNA sequence and splicing code details are studied by deep networks. DNA methylation from each cell identifies the pathogenic regions of the genome that are significant carriers of AD, known as promoters. The protein transcript genome available in the chromosome is the carrier of disease phenotype in several familial hierarchies transmitted pathologies. Early prognosis is the key factor for the present study. Thus, utilizing gene sequences present in chromosomes available in cerebrospinal fluid tissues of the cohorts provides various interactions phenotypes with the proteins. We use the CGACCGAACTCC allele [adenine (A), cytosine (C), guanine (G), thymine (T)] in the study of the Apolipo genotype that is available in Amyloid Precursor Protein. Where C and T are the common and variant allele gene pattern that causes gene mutation from 1–23 copies to 1500 copies associated with the datasets available from several chromosome studies. Each genotype contributes to a weak and robust association of potential pathology causing abnormal protein deposition and neurodegeneration. [22] proven that APO-€2, -€3, -€4 are the Apolipoprotein variants the combination of which causes the onset of Alzheimer's disease in the late seventies. This study considered a microarray of transcriptomes from aging community cohorts [23]. Chromosome 9 groups study showed rs429358 and common allele variant genotype rs7412 are the transcriptome variants co-related to all the gene causing pathological carriers.

This protein information causing frontal–temporal neuron degeneration consisting of RNA-binding proteins that imply cellular expressions and numerous mutations of genotypes causes future pathology. K-means clustering is performed and the genomic dataset that resulted in the predictor gene acts as classifier threshold for the optimal classifiers, we derive in the previous algorithm [24]. This is an unsupervised methodology that improves classification accuracy since we implement selecting Chromosome 9 and genes from trisomy. The gene pattern and variant combination are checked before each cluster formed for rs429358 and rs7412 among different alleles of gene sets participated in our study. At each stage in training the optimal classifier, we update the clustered and classified genotype expressions that are the main carriers of pathology in familial genes. The classification of each gene variant in C9 chromosomal rs429358 gene sequence and its variance for the number of SNP (Single nucleotide polymorphisms) molecules that cause further mutations are acquired in the form of one-hot encoding [25]. This encodes the base in DNA sequence as four-dimensional vectors given in (10).

```
APOE genome sequences
----------------------
DNA Sequence #1:
 CCGAGGGCTA ... CGCGGACACC
One hot encoding of Sequence #1:
 [[0. 0. 0. ... 1. 0. 0.]
 [1. 1. 0. ... 0. 1. 1.]
 [0. 0. 1. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
```

(10)

We accomplished three steps in genetic data processing. In the first step, we extracted ADNI datasets consisted of gene expression variants for more than 100 samples, with a relative mean of 3.5 and above. In the second step, the K means clustering model implemented utilizing genome dataset trains the model with function P(g|y) =¥ (E (Log (K carrier)), where K denotes the clustered phenotype resulted from the clustering algorithm. This algorithm is estimated for various hyper-parameter tuning that are the carriers of other genetic disorders. As result, the bio-marker investigation of the genotype obtained from cluster results has proven as the accurate predictor for disease onset [26–28]. Consider if λ denotes a binding factor of DNA molecule in genome variant, a beta ensemble learning model to tune the parameter inputs for input vectors where gene allele is considered as (aλ′s, bλ′s, cλ′s, dλ′s …) € P, where P is the actual predictor in the given gene expression. We can use the clustering algorithm for a set of input genomic data, thus maximizing the input feature of C9FM genomic dataset E (Log (K carrier)). In the third step, training the model is performed with initial values of K, the selection of biomarker gene-phenotype for the optimal classifier is achieved. The feature selection in the clustering algorithm is performed with each associated allele, the sparse subset and correlation feature is obtained by Lasso Regression Co-efficient in each iteration.

Fig. 5 visualizes the DragonNN framework environment that utilizes the acquired genome data in training the neural network, that discriminates the DNS sequences binding to a transcription factor that is significant for AD with the common sequence.

In the third approach, we used the Lasso regression function to evaluate the minority allele gene combinations present in the variance of genes across all groups. This function groups the set of genes with null hypothesis mean calculation across all DNA binding motifs of the Apo-€ gene. As result is the prediction P representing the importance of allele combination in group variable Y, N representing the significant and non-significant carriers. This ensures the Mild cognitive impairment to Alzheimer's disease

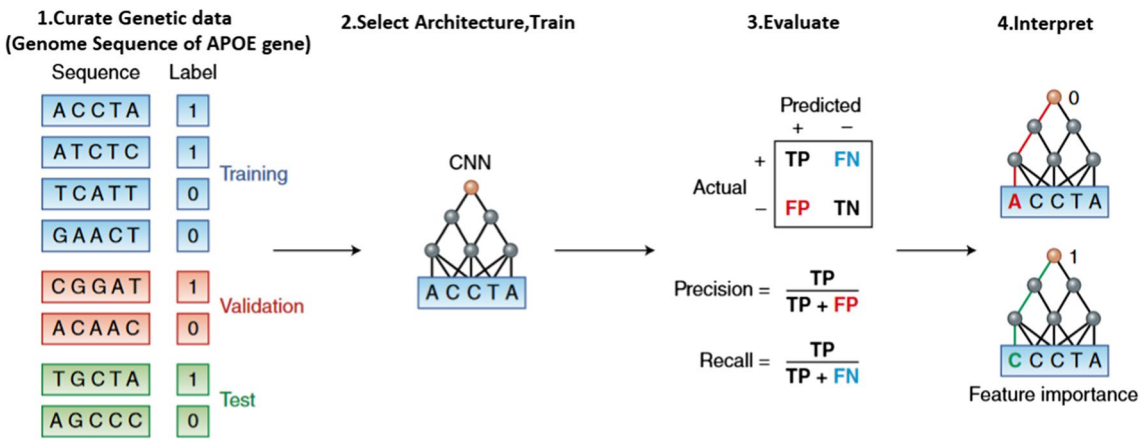

Fig. 5 Training CNN model to recognize genetic variant patterns across space

conversion based on regression factor using SNP-miRNA allele within chromosome studied as in Fig. 6

3 Results and Discussion

The software library required for the present study is scikit-learn: machine learning in Python — scikit-learn 0.24.1 library implemented in Jupyter Python notebook. Performance evaluations are calculated with JDK 1.8/Net Beans 8.2 as the front end utilizing the MRI attributes obtained from ADNI. The dragon toolkit provided the model interpretation and DNA sequence simulations that ensure a benchmark in the AD prediction. The prediction results obtained from deep learning python modules are integrated with Java packages as native libraries and results are obtained. The bagging of data samples outperformed in tree structures of random forest, which are minimized by the final ensemble model that had balanced each dataset without affecting performance. The novelty of the present study is the inclusion of various assessment scores and MRI features in the ensemble model to train and discriminate between AD, MCI to AD, and NC classes. Combining each feature with matching Apolipoprotein gene allele serves as a benchmark standard for dementia diagnosis and hence a vital discriminator feature (Fig. 7). Biomarker interaction in diseases such as cardio-vascular disease emphasizes the success of machine learning prediction models [29]. These biomarkers are common among cognitive disabilities since cardiovascular-affected individuals possess a lack of nutrition and oxygen supplied to brain cells. Hence this attracts several novel biomarkers that help in dementia prediction.

Similarly, we have plotted the accuracy and mean squared error to monitor the performance for positive response variables that signifies the polygenic risk factors from samples. The effective way to measure the success of a deep network is by the classification of sequence with novel test data set consisting of data that are not observed at all during training. Here, we evaluate the genome learning model on the test set and plot the result as a confusion matrix. Almost every test data variant should be correctly classified.

The correlation measure between various structural and demographic AD features are given by the probability of feature vectors that classifies the disease given in (11),

$$C = \sum_{i,j=0}^{N} \frac{[1-\mu x][1-\mu y]}{\sqrt{\mu x.\mu y}} \tag{11}$$

We found that support vector classifiers achieved an accuracy of above 80% for N fold cross-validation and 0.78 F1 weighted scores with low mean accurate percentage error (MAPE) (Fig. 9). The most significant features are correlated for each classifier and voting of performance is derived. After analyzing the performance of the study data presented below, the entire dataset is evaluated for an accurate prediction model. This is implemented by applying tenfold cross-validation and high interesting features such as (age, MMSE, ETIV, Apolipo-E) are used and specific gene pattern is given for testing. From,6000 samples of miRNA genome sequence obtained from data source 'APOE ε4 genotype' are checked with the following miRNA pattern GSM735100, GSM735101, GSM735102, GSM735103, GSM735104, GSM735105, GSM735106, GSM735107, and so on [30],which are unseen samples obtained from ADNI for validation of the model.

The final dementia diagnosis is calculated by applying this genotype strength with the ensemble classifiers and is analyzed for model error and accuracy as given in Fig. 8.The saliency values for the bases CGACCGAAC TCC appearing in the DNA sequence resembles the motif that are carriers of pathogenic variants and is interpreted as in Fig. 8. During validation, SVM and deep neural network classifiers showed the best performance. During the validation phase, the training outperformed the biased

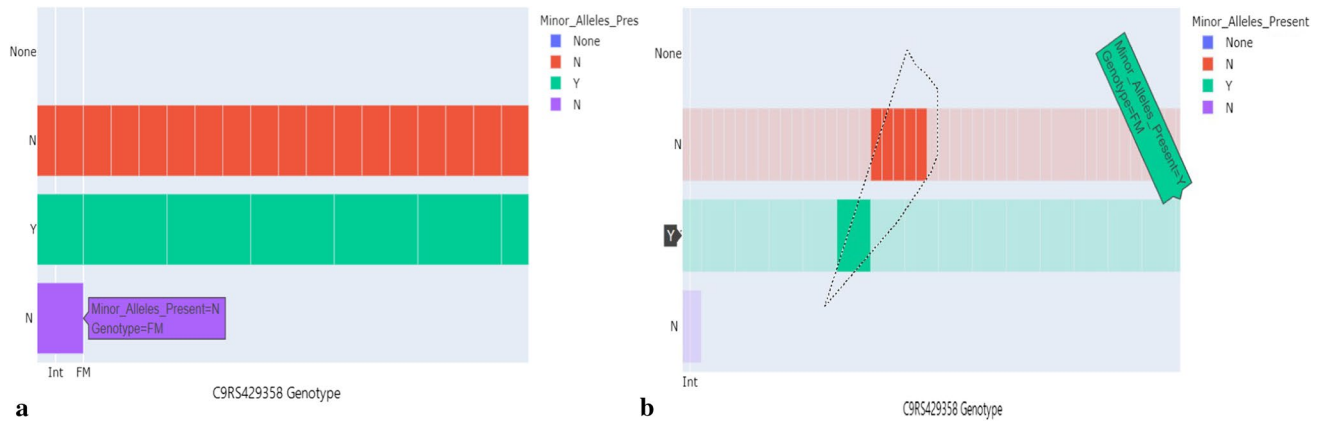

Fig. 6 **a** Frequency of various gene allele SNPs in Chromosome9 **b**Lasso Regression Co-efficient

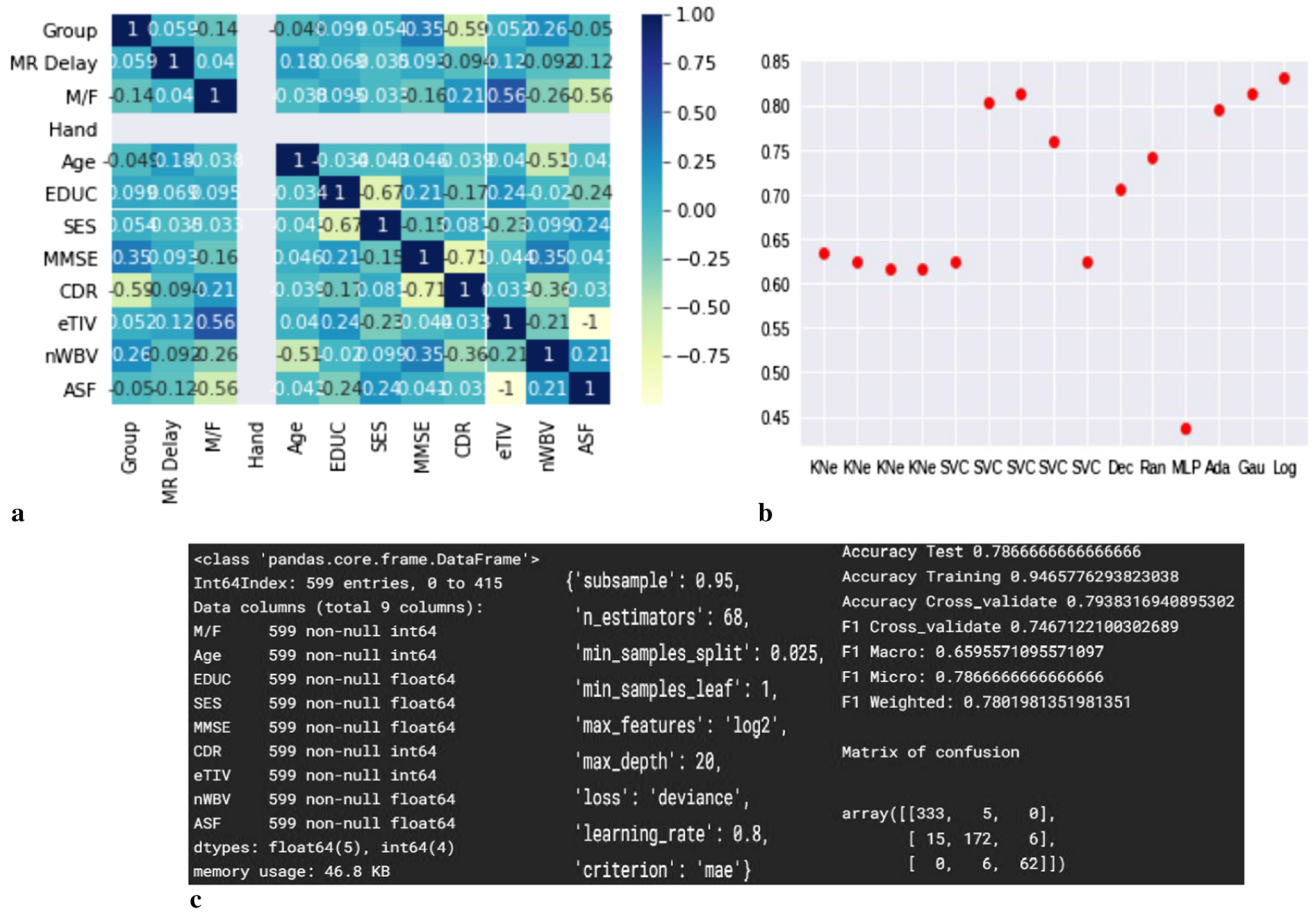

Fig. 7 **a** Co-relation between multi-classifiers **b** Classifier performance **c** Optimal model accuracy

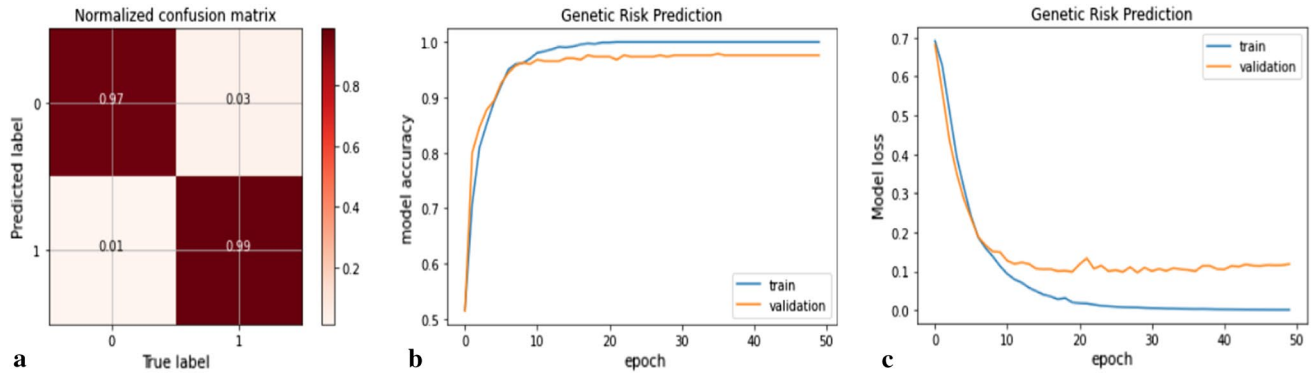

Fig. 8 Genetic Risk Prediction Model Performance

genome datasets and high dimensional information such as the number of samples and high pathogenic risk distribution, which are insufficient to train the perceptions of deep neural networks. In further experiments, we iterated several network parameters to improve performance such as feedback weights, mean square error, and early stopping that had shown increased performance which is more varied from the default initial setting.

An explicit comparison of a deep convolutional neural network is performed with the existing studies, which utilized MRI modality features and Mini-Mental State Examination scores obtained from subjects of ADNI, AIBL

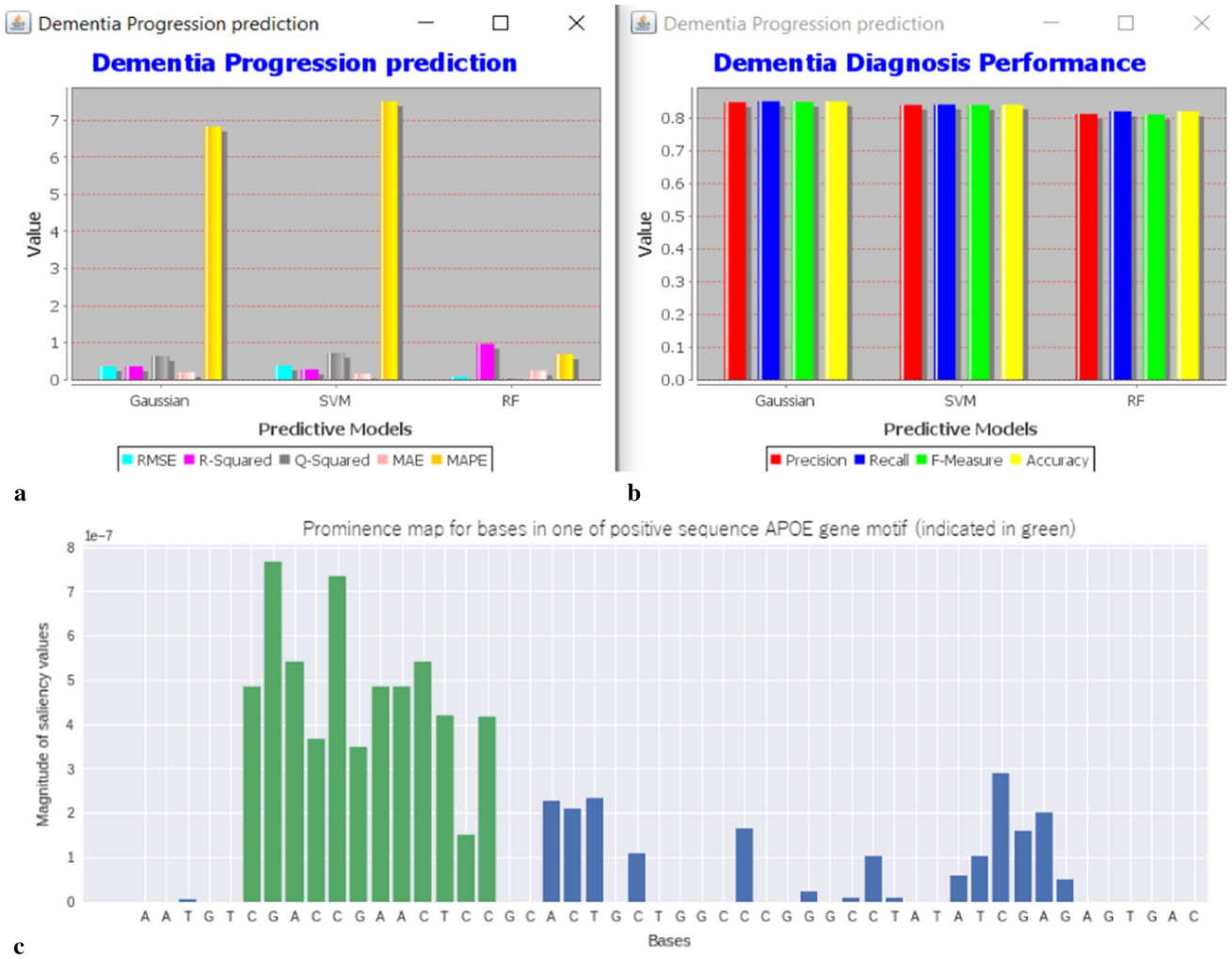

Fig. 9 Performance of Different Predictive models.**a** Dementia Progression Prediction Error. **b** Dementia Diagnosis Model Performance, **c** DragonNN model Saliency Feature Map for Genome Variants

(Australian Imaging Biomarkers and Lifestyle Study of Ageing), FHS (Framingham Heart study). The selected MRI features and non-imaging genetic patterns from DNA sequence predicted Alzheimer's disease status, and performance metrics are evaluated. We noted cognitive assessment scores displayed higher accuracy and F1 score.

Several state-of-art works had achieved genetic variation and pathogenesis involving small gene samples that are independent of structural AD features. However, the significance of genome variants relevant to disease is not well-addressed across the related researches. The present study showed various protein levels, such as tau, a beta-amyloid precursor that changes the MMSE score below 25. All the control groups from mild cognitive impairment with Tau, PTau, αβ levels exhibit positive conversion from MCI to AD. The ensemble machine learning model combines image features such as intracranial volume of the brain, assessment scores, and gene variants to implement co-relation between each model. The learning system finds the different slices of input images to calculate the disease conversion and assess performance. The model achieved 86% accuracy in disease prediction parallelly the regression task carries the sampled data output given for each AD cohorts.

Future research needs animal models implementing the proposed study to identify dynamic variations between pathologies for disease detection. Yet, the animal studies and current research lack all confining views for pathological variations in cognitive diseases like AD. Only a concordant result from the structural and genetic features will result in suitable conversion from pre-clinical experiments to practical clinical therapy. The limitations of this study include the absence of an external validation mechanism that will ensure the stability of current research. The investigation resulted in a notable correlation between structural brain changes with biomarkers and gene patterns. Although subjects in this study are from the ADNI database, we can

demonstrate fundamental pathological differences between disease groups. Independent validation sets are essential for in-depth exploration and micro perspective experiments. Transgenic models would be implemented and validated in future studies to support accurate results obtained with Aβ gene patterns. A sizeable potential investigation of real-time follow-up patients with MCI is necessary to address the prognostic challenges for practical implementation in a clinical discipline.

4 Conclusion and Future Work

Among various cognitive disabilities, dementia is the most life-threatening disease that affects the elderly community, caused by brain cell degeneration. We choose a selected subset of features from the sparse ADNI dataset and visualized the first-stage disease diagnosis with high dimensional learning model. Then early predictors such as MMSE score, brain volume, and Apolipo gene pattern that act as biomarkers for AD are combined and are trained with the machine learning classifiers for disease diagnosis. We came out with the best classification performance and chosen the optimal model that had given 85% accuracy in classifying the predictors that cause dementia. The implementation in DragonNN model had shown performance of different predictive models, dementia progression prediction error, model performance and model saliency feature map for Genome Variants that supports in early disease diagnosis and thus appropriate clinical interventions would prevent the disease progression. Feasibly, future work could accommodate the optimization process in hyperparameters and may explore higher dimensional data that are segmented to fit the models. The present work could act as a pre-trained model for preliminary classification and can be extended with transfer learning models and can be validated with a large set of experimentations with results that achieve high classification accuracy. Furthermore, we may analyze the neural network architecture with hyperparameter variables for significant gene sequence classification and identifying the DNA binding motif. By including the investigation of Omics analysis of gene sequences such as Single nucleotide polymorphisms (SNPs) will improve this model's accuracy before it is implemented in the healthcare industry.

Acknowledgements Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The Gene expression dataset used in the study is obtained from genetic studies of ADNI repository.

Funding No Financial and material support is obtained by any of the authors for this research.

Declarations

Conflict of interest All the authors in the paper have no conflict of interest.

Ethical approval This article does not contain any studies with animals or human performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Bergeron MF, Landset S, Zhou X, Ding T, Khoshgoftaar TM, Zhao F, Du B, Chen X, Wang X, Zhong L, Liu X. (2020) Utility of MemTrax and machine learning modeling in classification of mild cognitive impairment. J Alzheimer's Disease. 1–4.
2. Choi H, Jin KH (2018) Alzheimer's disease neuroimaging initiative. Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. Behav Brain Res 15(344):103–109
3. Gill S, Mouches P, Hu S, Rajashekar D, MacMaster FP, Smith EE, Forkert ND, Ismail Z, (2020) Alzheimer's disease neuroimaging initiative. Using machine learning to predict dementia from neuropsychiatric symptom and neuroimaging data. J Alzheimer's Disease: 1–2.
4. Zhang Q, Sidorenko J, Couvy-Duchesne B et al (2020) Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture. Nat Commun 11:4799. https://doi.org/10.1038/s41467-020-18534-1
5. Shigemizu D, Akiyama S, Asanomi Y, Boroevich KA, Sharma A, Tsunoda T, Sakurai T, Ozaki K, Ochiya T, Niida S (2019 Dec) A comparison of machine learning classifiers for Dementia with Lewy bodies using miRNA expression data. BMC Med Genom 12(1):1
6. Al-Khuzaie FEK, Bayat O, Duru AD (2021) Diagnosis of Alzheimer disease using 2D MRI slices by convolutional neural network. Appl Bionics Biomech. https://doi.org/10.1155/2021/6690539
7. Kim JP, Kim J, Park YH, Park SB, San Lee J, Yoo S, Kim EJ, Kim HJ, Na DL, Brown JA, Lockhart SN (2019) Machine learning based hierarchical classification of frontotemporal Dementia and Alzheimer's disease. NeuroImage Clin 23:101811
8. Katabathula S, Wang Q, Xu R (2021) Predict Alzheimer's disease using hippocampus MRI data: a lightweight 3D deep convolutional network model with visual and global shape representations. Alz Res Therapy 13:104. https://doi.org/10.1186/s13195-021-00837-0
9. Loewenstein DA, Curiel RE, Duara R, Buschke H (2017) Novel cognitive paradigms for the detection of memory impairment in preclinical Alzheimer's disease. Assessment, 1073191117691608.
10. Morris JC, Storandt M, Miller JP et al (2011) Mild cognitive impairment represents early-stage Alzheimer's disease. Arch Neurol 58(3):397–405. https://doi.org/10.1001/archneur.58.3.397
11. Van Rossum IA, Vos S, Handels R, Visser PJ (2010) Biomarkers as predictors for conversion from mild cognitive impairment to

Alzheimer-type Dementia: Implications for trial design. J Alzheimers Dis 20:881–891
12. Petersen RC, Parisi JE, Dickson DW, Johnson KA, Knop- man DS, Boeve BF, Jicha GA, Ivnik RJ, Smith GE, Tangalos EG, Braak H, Kokmen E (2006) Neuropathologic features of amnestic mild cognitive impairment.
13. Plant C, Teipel SJ, Oswald A, Bohm C, Meindl T, Mourao-Miranda J, Bokde AW, Hampel H, Ewers M (2010) Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. Neuroimage 50:162–174
14. Rohini.M, Surendran.D (2020), Toward Alzheimer's disease classification through machine learning, Soft Computing:
15. A Fusion of Foundations, Methodologies and Applications, https://doi.org/10.1007/s00500-020-05292-x
16. Rohini.M, Surendran.D, (2019) Classification of neurodegenerative disease stages using ensemble machine learning classifiers. Procedia Comput Sci 165(219):66–73
17. Van Cauwenberghe C, Van Broeckhoven C, Sleegers K (2016) The genetic landscape of Alzheimer disease: clinical implications and perspectives. Genet Med 18:421–430
18. Li Y, Yao Z, Yang Y, Zhao F, Fu Y, Zou Y, Hu B, (2020) Alzheimer's Disease Neuroimaging Initiative. A Study on PHF-Tau Network Effected by Apolipoprotein E4. Am J Alzheimer's Dis Other Dementias®. 17; 35:1533317520971414
19. Zhu F, Li X, Tang H, He Z, Zhang C, Hung GU, Chiu PY, Zhou W (2020) Machine learning for the preliminary diagnosis of dementia. Sci Program 7:2020
20. Sirkis DW, Geier EG, Bonham LW, Karch CM, Yokoyama JS (2019) Recent advances in the genetics of frontotemporal dementia. Current Genetic Med Rep 7(1):41–52
21. Zhang F, Li Z, Zhang B, Du H, Wang B, Zhang X (2019) Multimodal deep learning model for auxiliary diagnosis of Alzheimer's disease. Neurocomputing 7(361):185–195
22. Bettens K, Sleegers K, Van Broeckhoven C (2010) Current status on Alzheimer disease molecular genetics: from past, to present, to future. Hum Mol Genet 19(R1):R4-11
23. Ciani M, Benussi L, Bonvicini C, Ghidoni R (2019) Genome wide association study and next generation sequencing: a glimmer of light toward new possible horizons in frontotemporal dementia research. Front Neurosci 16(13):506
24. Filippi M, Agosta F, Ferraro PM (2016) Charting frontotemporal dementia: from genes to networks. J Neuroimaging 26(1):16–27
25. Qiu S, Joshi PS, Miller MI, Xue C, Zhou X, Karjadi C, Chang GH, Joshi AS, Dwyer B, Zhu S, Kaku M (2020) Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain 143(6):1920–1933
26. Zhou F, Chen D, Chen G, Liao P, Li R, Nong Q, Meng Y, Zou D, Li X (2021) Gene set index based on different modules may help differentiate the mechanisms of Alzheimer's disease and vascular dementia. Clin Interv Aging 16:451
27. Lee G, Nho K, Kang B, Sohn KA, Kim D (2019) Predicting Alzheimer's disease progression using multi-modal deep learning approach. Sci Rep 9(1):1–2
28. Reus LM, Pasaniuc B, Posthuma D, Boltz T, Ferrari R, Hernandez DG, Nalls MA, Rohrer JD, Ramasamy A, Kwok JB, Dobson-Stone C (2021) Gene expression imputation across multiple tissue types provides insight into the genetic architecture of frontotemporal dementia and its clinical subtypes. Biol Psychiat 89(8):825–835
29. Rasmussen IJ, Tybjærg-Hansen A, Rasmussen KL, Nordestgaard BG, Frikke-Schmidt R (2019) Blood–brain barrier transcytosis genes, risk of dementia and stroke: a prospective cohort study of 74,754 individuals. Eur J Epidemiol 34(6):579–590
30. Altinkaya E, Polat K, Barakli B (2020) Detection of Alzheimer's disease and dementia states based on deep learning from mri images: a comprehensive review. J Inst Electron Comput 1(1):39–53
31. Wang G, Zhang DF, Jiang HY, Fan Y, Ma L, Shen Z, Bi R, Xu M, Tan L, Shan B, Yao YG (2019) Mutation and association analyses of dementia-causal genes in Han Chinese patients with early-onset and familial Alzheimer's disease. J Psychiatr Res 1(113):141–147

Ms. M. Rohini has 8 years of experience in Teaching UG and PG Computer Science and Engineering and currently working as Assistant Professor in the Department of CSE, Sri Krishna College of Engineering and Technology, Coimbatore, Tamilnadu. Pursuing her Ph.D in Information and Communication Engineering in Anna University, Tamilnadu. Her Research interests includes Machine learning and Deep learning.

Dr. D. Surendran has 20 years of experience in Teaching UG and PG Computer Science and Engineering and currently working as Professor in the Department of CSE, KPR Institute of Engineering and Technology, Coimbatore, Tamilnadu. He Obtained PG in Computer Science and Engineering in 2004 and Ph.D. in Information and Communication Engineering from Anna University in 2011.His research interests include Cloud Computing, Semantic Technologies, IoT. He has guided 5 Research scholars for their Ph.D in CSE.

Dr. S. Oswalt Manoj has 11 years of experience in Teaching UG and PG Computer Science and Engineering courses and currently working as Assistant Professor in the Department of Computer Science and Business System, Sri Krishna College of Engineering and Technology, Coimbatore, Tamilnadu. He obtained PG in Computer Science and Engineering in 2010 and Ph.D. in Information and Communication Engineering from Anna University in 2021.His research interests include Deep Learning, Machine Learning, Soft Computing. He had published research articles in more than 20 International Journals.