Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Deep belief network-based approach for detecting Alzheimer's disease using the multi-omics data

Nivedhitha Mahendran, Durai Raj Vincent P M *

School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India

A R T I C L E I N F O

Article history:
Received 28 November 2022
Received in revised form 10 February 2023
Accepted 11 February 2023
Available online 13 February 2023

Keywords:
Alzheimer's disease
Gene Expression
DNA Methylation
Multi-omics
Feature Selection
Deep Belief Network

A B S T R A C T

Alzheimer's disease (AD) is the most uncertain form of Dementia in terms of finding out the mechanism. AD does not have a vital genetic factor to relate to. There were no reliable techniques and methods to identify the genetic risk factors associated with AD in the past. Most of the data available were from the brain images. However, recently, there have been drastic advancements in the high-throughput techniques in bioinformatics. It has led to focused researches in discovering the AD causing genetic risk factors. Recent analysis has resulted in considerable prefrontal cortex data with which classification and prediction models can be developed for AD. We have developed a Deep Belief Network-based prediction model using the DNA Methylation and Gene Expression Microarray Data, with High Dimension Low Sample Size (HDLSS) issues. To overcome the HDLSS challenge, we performed a two-layer feature selection considering the biological aspects of the features as well. In the two-layered feature selection approach, first the differentially expressed genes and differentially methylated positions are identified, then both the datasets are combined using Jaccard similarity measure. As the second step, an ensemble-based feature selection approach is implemented to further narrow down the gene selection. The results show that the proposed feature selection technique outperforms the existing commonly used feature selection techniques, such as Support Vector Machine Recursive Feature Elimination (SVM-RFE), and Correlation-based Feature Selection (CBS). Furthermore, the Deep Belief Network-based prediction model performs better than the widely used Machine Learning models. Also, the multi-omics dataset shows promising results compared to the single omics.

1. Introduction

The life expectancy of the people has improved with changes in standard of living; however, the aging population and age-related diseases are also equally growing. The major complaint among the elderly is Dementia, which occurs for two reasons, brain injury or neurodegenerative disease [1]. The former is always static, while the latter is progressive and fatal. One such deadly form of Dementia is Alzheimer's Disease (AD), often termed as the Neurodegenerative and proved to be a progressive disease, i.e., it develops gradually over a while [2]. It causes permanent and irreversible impairment to the neurons in the brain, majorly affecting the Cerebrum part of the brain [3]. The clinical symptoms of AD are continuous decline in cognitive activities, such as memory and thinking skills, which eventually hinder the patient's daily routine [4]. Several risk factors characterize AD; some clinically approved risk factors are age, high alcohol consumption, lifestyle, genetic factors, or depression [5]. Among these factors, genetics seems to contribute 70% in causing AD [6].

In neuropathological terms, AD is characterized as the deposition of Amyloid β peptides, neurofibrillary tangles, neural injury, and chronic neuroinflammation [6]. Also, the dysfunction caused in the brain parts such as the Hippocampus, Amygdala, and Cortical areas contributes significantly to causing AD.

According to the sources, it is one of the significant health care challenges in the current century, which affected around 5.5 million people (65 and older) worldwide [7]. The prediction says 1 in every 85 people will be down with AD by 2050 [8]. This cortical atrophy is 60–80% heritable to the first-degree relatives. AD is a progressive disease; it persists for many years, deteriorating the patient's health slowly and steadily, resulting in death [7]. The treatments for AD are mostly transient (cure for a short time), not long-term.

* Corresponding author.
E-mail addresses: nivedhitha.m2019@vitstudent.ac.in (N. Mahendran), pmvincent@vit.ac.in (D.R. Vincent P M).

https://doi.org/10.1016/j.csbj.2023.02.021

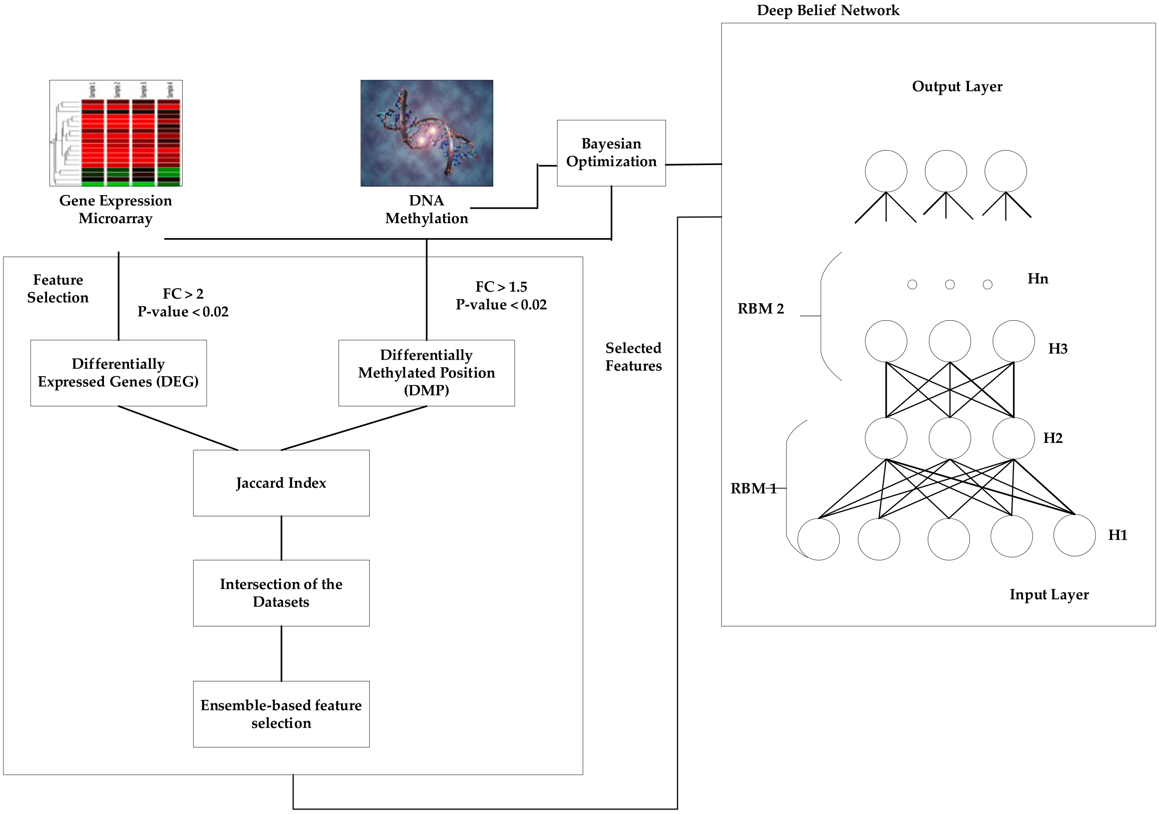

Fig. 1. The process involved in classification AD.

Recent researches and advancements in bioinformatics have led to high-throughput omics data, which contributes significantly to gaining insights about the disease at a detailed molecular level and assists in developing disease-modifying treatments [9]. Along with the bioinformatics methods, the development in Artificial Intelligence, Machine Learning has also led to numerous research opportunities in understanding the disease. Recently, the research is concentrated on integrating various omics data [9], for example, combining Gene Expression with Data Methylation or integrating Copy Number Validation (CNV) with DNA methylation, and so on. Integration of omics data assists in overcoming the difficulties in understanding the disease and developing advanced accurate models that incorporate the nature of biology [10].

As the advancements in acquiring omics data have been tremendous, the interest has grown towards computer-aided methods such as Machine Learning. The omics data is widely utilized in the Machine Learning domain for taking the diagnosis and treatment a step forward [11]. Though the idea seems interesting, there are critical challenges in executing it. One such challenge is the biological process involved in two different omics datasets, which are interactive and sometimes interdependent. A countable number of studies related to omics data integration in Cancer were carried out over the years [12–16], though only fewer models are developed in diagnosing and predicting other diseases. One of the "other diseases" is AD; most of the studies carried out in AD are based on Phenotypes, i.e., Brain images (MRI, CT, PET, etc.). Genotypic analyses are rare and upcoming, as it is difficult to obtain data from the brain tissues [17]. However, recent technologies and advancements have made it possible. A study based on the Single Polynucleotide Polymorphism (SNP) with APOE from Genome-wide Association Study (GWACS) has divulged that an increase and decrease in the APOE gene increases the risk of AD [18]. However, many other risk factors are associated with AD, which cannot be described with SNP alone. Thus, integrating omics data will reveal various other biomarkers, with which an accurate predictive model can be built [19,20].

One other critical challenge in integrating the omics data is that they are High-Dimensional-Low-Sample-Size (HDLSS) types, which requires high-ended computational methods to break down the features to find the essential biomarkers and build a predictive model. Thus, Artificial Intelligence, Machine Learning, and Deep Learning come in handy.

Artificial intelligence approaches, such as machine learning and deep learning are increasingly popular in the genetic research. Deep learning techniques are more advanced than the machine learning approaches, especially when there is HDLSS issue, as they can handle feature selection as part of the classification process. AD is a progressive neurological disorder found mostly in the older age groups [21–25]. The diagnosis is mostly done through brain imaging. Thus, the research is more focused on the imaging datasets. For instance, in this work, Le net architecture and the CNN techniques are applied on the MRI data in classification of AD and non-AD [26]. There are numerous works that are carried out in this area [26–31]. The main challenge is working with the imaging dataset is sometimes the results may be inaccurate. Thus, there must be more research works carried out in the gene expression based datasets. However, the problem with the gene expression dataset is the HDLSS. Studies are done to handle the HDLSS issue with the help of the machine learning and deep learning approaches.

Although Machine Learning is effective on other types of data, it is not suitable for HDLSS datatypes, it requires multiple stages of optimizations, and the steps must be predefined. Machine Learning learns and gains knowledge from past data and makes knowledgeable decisions based on the acquired information. Although it is a black box, Deep Learning is on the fly and accurate on HDLSS datasets. The Deep Learning algorithms create an artificial neural network as layers, which can learn and make intelligent "human-like" decisions on its own. (Fig. 1).

2. Materials and methods

2.1. Data Source

In this study we have used two types of datasets, Microarray gene expression, and DNA Methylation. DNA Microarray gene expression data are collected using the microscopic slide that consists of massive amount of gene sequences. The mRNAs are captured in these slides, which is a template for building the proteins. These microarray chips are highly helpful in analysing the gene expression. DNA Methylation is the process of making chemical modifications in the DNA. It is participates in gene expression and cell differentiation. During the methylation process, the methyl groups are transferred to cytosine on the C5 position to form methyl cytosine. DNA methylation influences the gene activities that is important for performing cognitive functions.

Three heterogeneous large-scale datasets have been used in this study accessed from the Gene Expression Omnibus (GEO) data repository under the National Centre for Biotechnology Information (NCBI). Among the three datasets, two are gene expression, and the other one is the DNA Methylation profile. The datasets with accession numbers GSE33000 [32], GSE44770 [33], and GSE80970 [34] associated with Alzheimer's are extracted from the Human Prefrontal Cortex.

The AD gene expression datasets are integrated to expand the size of the sample. The final integrated set has 696 records with 203 features. Among the 696 records, 439 are cases (AD), and 257 are controls (non-demented). The DNA Methylation dataset, which was extracted with the help of Illumina Human Methylation, has 142 records with 503 features. Among the 142 samples, 68 are controls (AD), and 74 are cases (non-demented).

We used the R Studio to implement the proposed framework. For pre-processing the datasets, we used the limma package in R, which is powerful for analysing the gene expression datasets. The GEOquery package is used to download the datasets from the GeoOmnibus database. For the SVM we used the e1071 package, and deepnet for the deep belief network.

2.2. Data Preparation

For using the data for further evaluation, the data needs to be pre-processed. The raw integrated gene expression data is normalized using the Z-Score normalization after applying the $\log_{10}$ transformation [35]. The Z-Score normalization is done to make the data comparable across all the experiments.

If $g_1, g_2, \ldots. g_n$ are the respective genes and I the intensity, Z-Score is given by [35],

$$ZScore = \frac{I_g - mean(I_{g_1,g_2,\ldots.,g_n})}{SD_{g_1,g_2,\ldots.,g_n}}$$

The raw DNA Methylation data with Beta-Values, we have calculated the M-Values. M-Values is a metric widely used in Methylation studies to measure the methylation levels (highly methylated and unmethylated). With 'I' being the intensity, M-Value is given by [36],

$$MValue = \log\frac{\max(I_{n,methy}, 0) + \alpha}{\max(I_{i,unmethy}, 0) + \alpha}$$

We have used quantile normalization to eliminate the background signals and systematic errors along with the M-Values. The data distribution before and after normalization of the combined dataset can be seen in Figs. 2 and 3.

2.3. Feature selection

In this study, we have used multi-omics data, which has its difficulties. It is challenging to combine two datasets with different biological meanings and retrieve them for various purposes. Thereby, their distributions vary. Gene expression is a sequence of processes that eventually result in the formation of proteins through translation and transcription [37]. DNA Methylation is part of Epigenetics (heritable changes in the gene expression), which modifies the gene by adding the Methyl group (CH_3) to the DNA, affecting the gene expression [38].

The only common characteristic among the Microarray data and the DNA Methylation data is that they are both HDLSS datasets, one vital issue most researchers have with genetic data. It is dangerous to use the data, resulting in high variance and an acute overfitting problem. To overcome the HDLSS challenge, an appropriate algorithm for selecting only the necessary features is necessary. Several works in the literature have used the conventional dimension reduction algorithms; however, these algorithms don't consider the biological meaning of the data and fail to identify the relationship between the omics data. Moreover, machine learning methods are more suitable for datasets with large sample sizes. Thus, in this study, we have designed a framework for feature selection by considering the characteristics of Differentially Methylated Positions (DMP) and Differentially Expressed Genes (DEG).

The proposed feature selection is carried out in two steps. Gene Expression researchers seek to identify the Differentially Expressed (DE). Thus, the first step is to determine the Differentially Expressed Genes (DEG) and Differentially Methylated Positions (DMP). In the Microarray data analysis, the DEGs are identified through the ratio between the gene expression level in the target and the control sample. The ratio is then scaled with the help of 2 logarithms, and the resultant is termed the Log2 Ratio. Further, the absolute value of the log2 ratio is the Fold Change (FC). FC is the intuitive method widely used in finding out the DEGs. Also, Significance Analysis of Microarrays (SAM) is used to find out the statistical significance of the genes, and each gene is assigned with a P-Value. The common threshold used for FC are $|FC= > 1.5$ or $|FC= > 2$ and for P-Value its P-Value < 0.01. In this study, we have maintained a threshold of $|FC= > 2$ and P-Value < 0.01 for the Microarray Gene Expression Data. For finding out the DMPs, the same criteria as the Microarray can be followed. The threshold we maintained for DMPs is $|FC= > 1.5$ and P-Value < 0.01. We have selected the DMPs with CpG sites found under 1500 base pairs from the start site of the transcription. The methylation near the transcription sites has a high probability of regulating gene expression. Once the DEGs and DMPs are identified, we used Jaccard Similarity to find the similarity between the two datasets. Jaccard Similarity Coefficient is a standard statistical method, in recent days used commonly in gene expression data to find the similarity between two sets. It is primarily composed of sets, objects, unions, and intersections. The cut-off for Jaccard index varies from 0 to 1. Closer to one, more are the features similar to each other.

To find out the similarity using the Jaccard index, we first created a randomized version control using the two omics datasets. For each gene set, we generated 'm′ sized gene set with respect to the gene set size distribution of the original gene set. Further, we have random sampled 'm′ genes from the original gene set, where the probability of sampling for each gene is proportional to their occurrence in the original dataset. For all pairs of genes (from two different datasets), we calculated the set similarity using the Jaccard coefficient of edges (m, n),

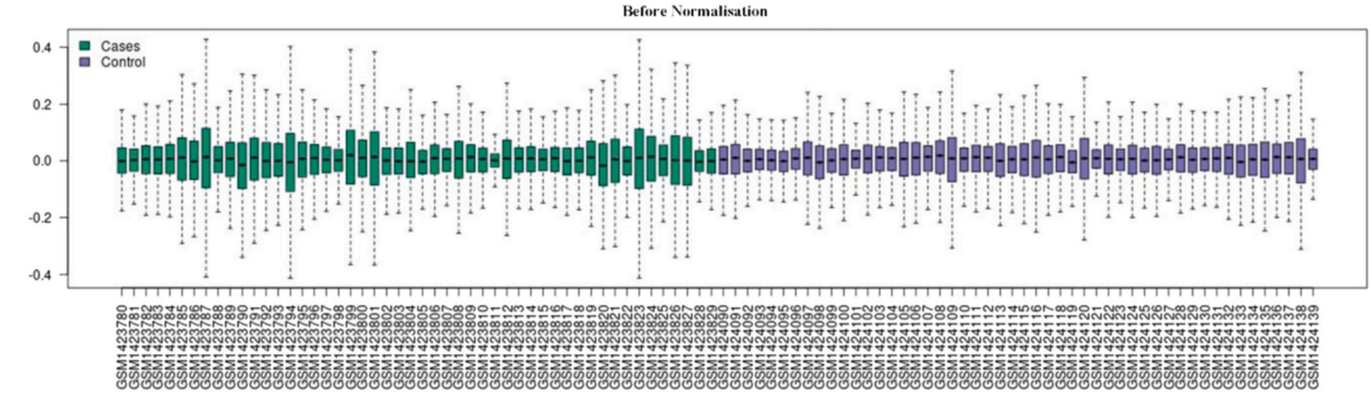

Fig. 2. Sample dataset before normalisation.

$$Jaccar(m, n) = \frac{|Set1_m \cap Set2_n|}{|Set1_m \cup Set2_n|}$$

Where, S_m and S_n represent the boolean values of the gene expression data of gene m and gene n.

The cut-off for Jaccard similarity coefficient used in this study is 0.8. In the second phase, we applied intersection to merge the two datasets. The hypothesis here is that if the gene is down or upregulated in Microarray and its respective CpG site is hyper or hyporegulated, it has significant involvement in causing the disease. The DMPs have their respective genes. The sample size after combining both the datasets is 838 with 513 cases and 325 controls. After the intersection of both datasets, we need to ensure that the effect from the datasets is minimum as the sample size has increased. Thus, we have performed the variance partitioning to assess the contribution of each feature to variation that happens to the response variable. We used the variation partitioning method in R to find out the variance that are attributable to the multiple features in the dataset. The variation partitioning method requires a dataset that is normalized already. As we have normalized the datasets using the standard normalization techniques mentioned above, we have applied the variance partitioning on the integrated dataset. The significant features from both the datasets are already ranked based on the FC and p-values. However, after combining the dataset, we need to find out whether the features in the dataset still contributes to the response variable and does not deviate much.

We designed a setup using R to find out the differences, using the F-test to find out the variance among the groups is higher than the variance within the group. We identified the statistically significant features using the p-value. The threshold for p-value is kept as less than 0.01. The features that do not satisfy the threshold are eliminated and further processed using the ensemble feature selection. Further, the features age and gender have very low variance, 0.08 and 0.05 respectively, which indicates they do not contribute much to the target variable and the classification models implemented. Thus, we have eliminated the age and gender features from the dataset.

2.4. Ensemble feature selection

There are four feature selection techniques, filter, wrapper, embedded, and ensemble. This study has built ensemble-based feature selection techniques, which will be applied as the second layer of feature selection. Ensemble-based models combine different leaners, which are then put together using an aggregation technique. The first step in building an ensemble is to choose the different feature selection approaches, and the second step is to combine the outputs and offer a single decision. In this study, we have used five filter techniques (ReliefF, Information gain, Signal-to-noise ratio, Mutual Information, and Chi-squared) into an ensemble and combined their outputs using an aggregation technique.

- ReliefF – This filter approach calculated the proxy statistics of the features, using the quality and relevance to the target feature. The proxy statistics are the feature scores or weights that ranges from − 1 to + 1.
- Information gain – It calculates the reduction in entropy (disorder) from the transformation dataset. It can be used as a filter based feature selection approach by estimating the information gain of each feature with respect to the target feature.
- Signal-to-noise ratio – SNR filter approach is used to estimate the minimal variation within each group and mean expression between the groups. Ranks are given to all the features based on the expression levels identified with the help of SNR test statistics.

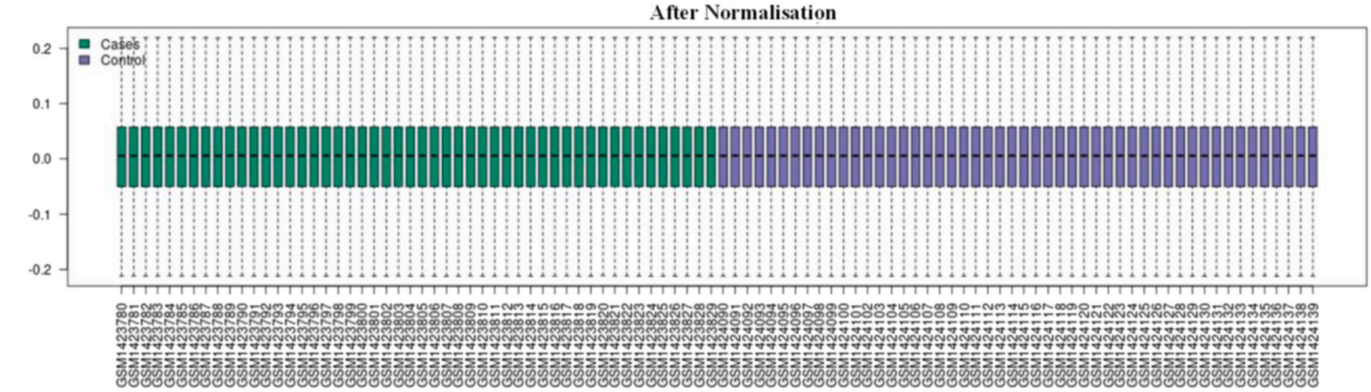

Fig. 3. Sample dataset after normalisation.

- Mutual Information – It is a measure to calculate the average uncertainty reduction of one feature that results from the learning value of another feature. It helps in identifying the mutual dependence between the features.
- Chi-squared test: Chi-squared test is generally used to test the independence between the two events in statistics. As a feature selection technique, the chi-square value for two features are calculated and the target feature, then choose the features based on the chi-square ranks.

3. Deep belief neural network

This paper implements the Deep Belief Network with the RBM. DBN is composed of stochastic and latent variables, binary and called the hidden layers. DBN is a probabilistic generative model constructed using the Restricted Boltzmann Machines (RBMs), representing the layers of the architecture [39].

RBMs are commonly defined as the generative model that is based on energy. It consists of two layers, the first layer being the visible layer and the second layer being the hidden layer, having nodes connected across the layers [40]. The RBMs involve weight, bias, and activation function. The nodes in the visible layer take a low-level feature from the dataset under learning and pass it to the first hidden layer. Then, it is multiplied with a weight, and a bias is added to the result. The resultant will be produced after processing through an activation function, which produces the output for the hidden node [41].

4. Training RBM

Provided the observed state, the energy joint configuration for the RBMs visible unit and hidden units (vu, hu) is written as,

$$E_{JC}(vu, hu) = -\sum_{a=1}^{A} y_a vu_a - \sum_{b=1}^{B} z_b hu_b - \sum_{b=1}^{B}\sum_{a=1}^{A} W_{ba} vu_a hu_b \tag{1}$$

where y_a is the bias of the ath visible unit and z_b is the bias of the bth hidden unit, W_{ba} is the weight initialized to the connection between the ath visible unit and bth hidden unit. RBM allots a probability for each configuration of (vu, hu). The relationship between the probability distribution and the energy function is given by,

$$p(vu, hu) = \frac{1}{N_c} \times e^{-E_{JC}(vu,hu)} \tag{2}$$

where, N_c is the normalization constant or partition function and it is written as the summation of all possible connections between the visible and the hidden layers.

$$N_c = \sum_{vu,hu} e^{-E_{JC}(vu,hu)} \tag{3}$$

If there are 'm' visible units and 'n' hidden units, the conditional probability for the visible units vu configuration, given the hidden units hu configuration is written as,

$$p(vu|hu) = \prod_{a=1}^{m} p(vu_a \mid hu) \tag{4}$$

Similarly, the conditional probability of the hidden units (hu) configuration, given the visible units (vu) configuration is written as,

$$p(hu|vu) = \prod_{b=1}^{n} p(hu_b|vu) \tag{5}$$

When the RBM is of binary units with vu_a and $hu_b \in \{0, 1\}$, the probability is given as,

$$p(hu_b = 1|vu) = \sigma\left[z_b + \sum_{b=1}^{B} vu_a W_{ba}\right] \tag{6}$$

and,

$$p(vu_b = 1|hu) = \sigma\left[y_c + \sum_{a=1}^{A} hu_b W_{ba}\right] \tag{7}$$

The visible unit vector's probability is written as,

$$p(vu) = \sum_{hu} p(vu, hu) = \sum_{hu} p(vu|hu)p(hu) \tag{8}$$

The probability can be increased by adjusting the biases and the weights to lower the energy of a specific vector and raise the energy of other vectors. The RBM's learning algorithms basically implement the log-likelihood with gradient ascent. The log probability with respective weights after computing the derivative is given by,

$$\frac{\partial logp(vu)}{\partial W_{ba}} = \langle vu_a hu_b\rangle_{data} - \langle vu_a hu_b\rangle_{model} \tag{9}$$

$\langle . \rangle_{data}$ is the expectation of the data distribution and $\langle . \rangle_{model}$ is the expectation of the model distribution. $\langle vu_a hu_b\rangle_{model}$ requires huge number of computation steps and Gibbs sampling, which, unfortunately, takes long time. Therefore, Hinton proposed CD-k (Contrastive Divergence), which speeds up the learning process. In CD-k, the $\langle vu_a hu_b\rangle_{model}$ is replaced with $\langle vu_a hu_b\rangle_k$, k having the small values. The following rules are implemented in order to correct the bias and weight in the network,

$$\Delta W_{ab} = \varepsilon(\langle vu_a hu_b\rangle_{data} - \langle vu_a hu_b\rangle_{k)}$$

$$\Delta z_b = \varepsilon(\langle hu_b\rangle_{data} - \langle hu_b\rangle_k$$

$$\Delta y_a = \varepsilon(\langle vu_a\rangle_{data} - \langle vu_a\rangle_k$$

Where, ε is the learning rate.

5. Deep belief network training

The RBMs are stacked together in a greedy approach to form a Deep Belief Network. We have proposed a Deep Belief Network with simple stopping criteria in this study. The hyperparameters must be fine-tuned in any deep nets for better results. It is tricky to choose appropriate hyperparameters in deep learning networks. The performance of the nets with certain configurations might not be the same as the performance when the configuration is changed. Bayesian optimization records previous results and checks the probabilistic model to select the next set of parameters. Thus, it is effective in choosing optimal hyperparameters. Therefore, we have used Bayesian Optimization to optimize DBN hyperparameters. It aims to construct a probabilistic model considering the objective function, such as Accuracy, Root Mean Squared Error. Bayesian optimization can choose better settings within fewer iterations. It is promising because it takes informative decisions in choosing the next hyperparameter. The objective value function in our model is the test dataset accuracy. We aim to find the best test accuracy within determining bounds for the parameters chosen, 8–12 hidden layers, learning rate between 0.01 and 0.2, 200–400 no. of nodes per layer, and dropout rate between 0.6 and 0.9.

For choosing the features from the omics data, and the hyperparameters for the DBN, the proposed feature selection method, along with the Bayesian Optimization techniques, is performed with 5-fold cross-validation. The learning is performed for 300 epochs, and the parameter combination with the highest test accuracy is selected. An average of all the 5-fold learnings hyperparameter is applied for the final proposed model. The input layer in the proposed model includes the Gene expression and DNA Methylation data. Our problem is a binary classification; therefore, the output layer has two nodes with integer encoding. We used the Rectified Linear Unit (ReLU) as the activation function and the softmax regression in the output layer to normalize the value between 0 and 1. We have used

Table 1
The Bayesian Optimization Results.

5-Fold Cross Validation	Learning Rate	Dropout Rate	Hidden Layers	Number of Nodes per layer	Test Accuracy
1	0.025	0.685	12	256	0.992
2	0.017	0.887	7	380	0.997
3	0.015	0.785	6	280	0.994
4	0.012	0.897	8	311	1.000
5	0.010	0.910	9	352	0.995
Average	**0.02**	**0.83**	**8**	**315**	**0.9956**

the cross-entropy cost function to evaluate how wrong the model is. The cross-entropy cost function, in simple terms, is the measure of the distance between the probability distribution of the actual and the predicted vector.

The critical issue in any machine learning or deep learning model is overfitting. Therefore, to avoid overfitting, we have formulated a simple stopping criterion. After every 50 epochs, the test accuracy of the last 15 epochs will be compared with the average test accuracy to check whether the accuracy is converging or decreasing. Also, the training accuracies will also be compared in the same way. If both the conditions are satisfied, the learning will be stopped.

6. Results

6.1. Hypotheses

We have carried out this study by framing the following hypothesis,

1. Combining the gene expression microarray data and the DNA Methylation data will produce a better prediction of Alzheimer's.
2. The proposed feature selection approach considers the data's biological characteristics and improves the learning model's performance by choosing the critical features.
3. The Deep Belief Network with an added stopping criteria will improve the learning accuracy compared to the machine learning model.

We have compared the proposed method and the existing methods to verify the hypothesis mentioned above. We have done a two-phased analysis to evaluate the assumptions. In the first phase, we implemented the existing feature selection techniques to enhance the performance of the conventional machine learning models and the Deep Belief Network with the added stopping criteria using the single omics dataset (gene expression microarray and the DNA Methylation). In the second phase, we have implemented the models with the multi-omics data, combined data with the help of the proposed feature selection method.

We have split the data into a training set and the testing set and applied the five-fold cross-validation. In the first phase, we trained SVM, Naïve Bayes, and proposed a Deep Belief Network with features selected by SVM-RFE and Correlation-based Feature Selection (CBS) using the single omics data. The parameters for the SVM kernel are Radial Basis Function (RBF), complexity set to 1 and Gamma as auto; the parameters for DBN are chosen with the Bayesian Optimization's help. The resultant parameters are shown in Table 1. After applying Bayesian optimization, the final parameters used in the DBN are the Learning rate = 0.02, Dropout = 0.83, hidden layers = 8, and the number of nodes per layer is 315.

The SVM-RFE and CBS are applied to the single omics data, and the AD classification is done with the help of SVM, NB, and the proposed DBN. The results are tabulated in Table 2. The table, we can notice that the average number of genes selected by SVM-RFE is 49, and CBS is 21.6. The number of CpG sites selected by SVM-RFE and CBS is 107 and 81.6, respectively. The selected features are inputted into the classification methods. The average accuracy of SVM-RFE is 0.692, and CBS is 0.725. Also, we can notice that the DBN has better accuracy with both SVM-RFE and CBS than SVM and NB.

The DEG and DMP are combined to form a multi-omics dataset in the second phase. The SVM-RFE and CBS feature selection techniques are applied to the combined dataset. The results are tabulated in Table 3. The table shows that the average number of genes selected by all three feature selection methods is more than the features selected in the single omics data. However, the average accuracy is better with the multi-omics data than the single omics data. The average genes selected by SVM-RFE is 120.6, CBS is 94, and the average of the proposed method is 39. The proposed method has less average than the other two methods. Still, the accuracy of the proposed method is considerably better than the other two methods. Also, we can notice that the accuracy of DBN is slightly better than the other classification methods. The average accuracy of SVM-RFE with SVM, NB, and DBN with multi-omics data is 78%, 76%, and 82%, respectively. The average accuracy of the proposed feature selection and DBN produces better accuracies in all five folds. (Table 4).

The framework is further tested using two other datasets (GSE109887 and GSE118553) both extracted from the temporal gyrus region of the brain. The GSE109887 has 78 samples with 29 control and 54 cases, and GSE118553 has 83 samples with 29 control and 54 cases. The classification accuracy of the proposed model with the testing multi-omics dataset is 82%, which is better than the accuracy obtained in the cross validation. The implemented framework offers better results than the standard feature selection and classification algorithms, the results are not promising when compared to the models implemented with the imaging datasets. For instance, the accuracy of CNN on AD neuroimaging initiative (ADNI) dataset is 97.8%, whereas, the framework implemented using the multi-omics dataset is 82%. However, there is a lot of scope for research and improvement in the Alzheimer's omics data.

7. Discussion and conclusion

AD, in recent days, is common among the elderly and is considered deadly. It ends fatally if not taken care of with appropriate treatment. The development is slow and progressive. Several kinds of researches are going on to find the gene representatives and early prediction of AD. The maximum of the researches is focused on the single omics dataset, which has the problem of HDLSS. Recently, few works of literature have been focused on the multi-omics data, integrating the single omics data. Integrating single omics data into multi-omics data will solve the issue of HDLSS moderately. Thus, we have designed a framework to select the important features considering their biological characteristics from the multi-omics data. We implemented the techniques using the single omics (Gene Expression Microarray and DNA Methylation) and the multi-omics dataset. The proposed feature selection is based on the Jaccard Index. We have applied the Fold Change and Z-Score to normalize and find out the DMPs and DEGs from the dataset. The multi-omics dataset is formed by combining both the datasets (Gene Expression Microarray and DNA Methylation). Then the Jaccard Index is applied to find out the similarity between the genes from the Microarray and DNA Methylation. The DBN with the added stopping criteria is also

Table 2
Feature Selection using the Single Omics AD dataset.

Feature Selection Algorithm	5 Fold Cross Validation	Learning Algorithm	No of Gene Selected	Accuracy	No of CpGs selected	Accuracy
SVM-RFE	K = 1	SVM	25	0.564	164	0.628
		Naïve Bayes		0.625		0.615
		Proposed DBN		0.792		0.657
	K = 2	SVM	18	0.763	78	0.634
		Naïve Bayes		0.785		0.627
		Proposed DBN		0.824		0.668
	K = 3	SVM	4	0.786	226	0.751
		Naïve Bayes		0.792		0.775
		Proposed DBN		0.847		0.794
	K = 4	SVM	2	0.693	56	0.683
		Naïve Bayes		0.735		0.705
		Proposed DBN		0.786		0.751
	K = 5	SVM	30	0.663	11	0.675
		Naïve Bayes		0.718		0.725
		Proposed DBN		0.805		0.797
Average			**49**	**0.692**	**107**	**0.699**
CBS	K = 1	SVM	26	0.654	122	0.685
		Naïve Bayes		0.625		0.677
		Proposed DBN		0.688		0.704
	K = 2	SVM	34	0.712	54	0.724
		Naïve Bayes		0.749		0.733
		Proposed DBN		0.781		0.790
	K = 3	SVM	7	0.759	11	0.769
		Naïve Bayes		0.787		0.745
		Proposed DBN		0.812		0.826
	K = 4	SVM	30	0.742	153	0.780
		Naïve Bayes		0.738		0.792
		Proposed DBN		0.768		0.814
	K = 5	SVM	11	0.631	68	0.653
		Naïve Bayes		0.682		0.691
		Proposed DBN		0.751		0.781
Average			**21.6**	**0.725**	**81.6**	**0.744**
Proposed Feature Selection	K = 1	SVM	22	0.718	102	0.697
		Naïve Bayes		0.768		0.708
		Proposed DBN		0.795		0.735
	K = 2	SVM	32	0.754	46	0.680
		Naïve Bayes		0.749		0.678
		Proposed DBN		0.796		0.720
	K = 3	SVM	4	0.746	12	0.756
		Naïve Bayes		0.780		0.754
		Proposed DBN		0.806		0.797
	K = 4	SVM	2	0.815	17	0.808
		Naïve Bayes		0.802		0.756
		Proposed DBN		0.863		0.844
	K = 5	SVM	18	0.785	9	0.799
		Naïve Bayes		0.782		0.780
		Proposed DBN		0.814		0.843
Average			**15.6**	**0.784**	**37.2**	**0.757**

proposed. The parameters of the DBN are optimized using the Bayesian Optimization technique, and the average of 5 folds is applied to the final model.

We have framed three hypotheses, and to evaluate the hypothesis, the single omics, and multi-omics dataset are used with existing feature selection techniques and the proposed feature selection technique. Furthermore, we have implemented the conventional machine learning models and the proposed DBN to validate the feature selection technique. The results are tabulated; from the results, we can notice that all three hypotheses framed are validated and satisfied.

The existing feature selection techniques PCA and CBS are implemented with both single omics and multi-omics datasets. In the multi-omics data experiment, we have noticed that the iteration K= 3 has the highest accuracy of any other iterations. Among the genes selected in the fold three, MS4A4A is reported in the AlzGene Database, and BEX2 seems important in inhibiting neuronal differentiation. As a result of the study, it can be found that the results are better in the case of the multi-omics data. Fig. 4 shows the accuracy of the prediction models with the widely used feature selection methods and the proposed feature selection technique with 5-Fold Cross-Validation. The dark brown bars show the accuracy of DEGs, and light brown bars show the DMPs from the single omics dataset. The pink trend line on top of the bars shows the accuracy of the models implemented with the multi-omics dataset. The plot clearly shows the upper hand of the models implemented with the multi-omics dataset.

Furthermore, the proposed feature selection and the DBN prediction model perform considerably better than the other two Machine Learning models. Thus, the hypothesis proposed in this study is validated. Although the DBN model produces better results than the Machine Learning methods, further enhancements must improve accuracy. It is critical to find the risk factors and the genes responsible for AD, as it is the common form of Dementia yet uncertain of the cause. The future work will include analyzing the multi-omics dataset and finding out the genes responsible for causing the AD using computational methods, and validating the results with the help of bioinformatics tools.

Table 3
Feature Selection using the Multi-omics AD dataset.

Feature Selection Algorithm	5 Fold Cross Validation	Learning Algorithm	No of Gene Selected(Gene Expression + DNA Methylation)	Accuracy
SVM-RFE	K = 1	SVM	252	0.692
		Naïve Bayes		0.649
		Proposed DBN		0.817
	K = 2	SVM	178	0.785
		Naïve Bayes		0.812
		Proposed DBN		0.847
	K = 3	SVM	10	0.829
		Naïve Bayes		0.80
		Proposed DBN		0.861
	K = 4	SVM	138	0.768
		Naïve Bayes		0.774
		Proposed DBN		0.827
	K = 5	SVM	25	0.681
		Naïve Bayes		0.741
		Proposed DBN		0.818
Average			**120.6**	**0.78**
CBS	K = 1	SVM	115	0.682
		Naïve Bayes		0.701
		Proposed DBN		0.719
	K = 2	SVM	210	0.737
		Naïve Bayes		0.761
		Proposed DBN		0.818
	K = 3	SVM	25	0.791
		Naïve Bayes		0.798
		Proposed DBN		0.841
	K = 4	SVM	102	0.788
		Naïve Bayes		0.804
		Proposed DBN		0.814
	K = 5	SVM	18	0.689
		Naïve Bayes		0.712
		Proposed DBN		0.787
Average			**94**	**0.762**
Proposed Feature Selection	K = 1	SVM	35	0.751
		Naïve Bayes		0.797
		Proposed DBN		0.817
	K = 2	SVM	39	0.781
		Naïve Bayes		0.824
		Proposed DBN		0.818
	K = 3	SVM	36	0.775
		Naïve Bayes		0.787
		Proposed DBN		0.89
	K = 4	SVM	44	0.842
		Naïve Bayes		0.824
		Proposed DBN		0.868
	K = 5	SVM	41	0.812
		Naïve Bayes		0.811
		Proposed DBN		0.837
Average			**39**	**0.815**

Table 4
Selected genes for the proposed method (Multi-omics dataset).

Folds	Genes selected in each fold
K = 1 (35)	TMSL3, SAT1, SYTL4, SLC25A5, STAG2, SLC9A6, SRPX, TA2, ARMC5, ELF4, BEX2, DDX3Y, BEX5, CNKSR2, ELK1, MORF4L2, FLNA, MORC4, GPRASP1, MID1IP1, MAGED1, MCTS1, MAP7D2, RRAGB, NGFRAP1, RPGR, PAK3, RNF128, PIGA, RBMX, PIR, PJA1, MS4A6A, DTNA, TCEAL4
K = 2 (39)	TMSL3, SAT1, SYTL4, SLC25A5, STAG2, SLC9A6, SRPX, TA2, ARMC5, ELF4, BEX2, DDX3Y, BEX5, CNKSR2, ELK1, MORF4L2, FLNA, MORC4, GPRASP1, MID1IP1, MAGED1, MCTS1, MAP7D2, RRAGB, NGFRAP1, RPGR, PAK3, RNF128, PIGA, RBMX, PIR, PJA1, MS4A6A, DTNA, TCEAL4, CRH, ZNF438, CD59, BDNF
K = 3 (38)	TMSL3, SAT1, SYTL4, SLC25A5, STAG2, SLC9A6, SRPX, TA2, ARMC5, ELF4, BEX2, DDX3Y, BEX5, CNKSR2, ELK1, MORF4L2, FLNA, MORC4, GPRASP1, MID1IP1, MAGED1, MCTS1, MAP7D2, RRAGB, NGFRAP1, RPGR, PAK3, RNF128, PIGA, RBMX, PIR, PJA1, MS4A6A, DTNA, TCEAL4, CRH, ZNF438, HAP-1
K = 4 (42)	TMSL3, SAT1, SYTL4, SLC25A5, STAG2, SLC9A6, SRPX, TA2, ARMC5, ELF4, BEX2, DDX3Y, BEX5, CNKSR2, ELK1, MORF4L2, FLNA, MORC4, GPRASP1, MID1IP1, MAGED1, MCTS1, MAP7D2, RRAGB, NGFRAP1, RPGR, PAK3, RNF128, PIGA, RBMX, PIR, PJA1, MS4A6A, DTNA, TCEAL4, CRH, ZNF438, CD59, BDNF, NRN1, WSB2, HAP-1
K = 5 (41)	TMSL3, SAT1, SYTL4, SLC25A5, STAG2, SLC9A6, SRPX, TA2, ARMC5, ELF4, BEX2, DDX3Y, BEX5, CNKSR2, ELK1, MORF4L2, FLNA, MORC4, GPRASP1, MID1IP1, MAGED1, MCTS1, MAP7D2, RRAGB, NGFRAP1, RPGR, PAK3, RNF128, PIGA, RBMX, PIR, PJA1, MS4A6A, DTNA, TCEAL4, CRH, ZNF438, CD59, BDNF, NRN1, WSB2
Intersection of genes selected in all five folds	TMSL3, SAT1, SYTL4, SLC25A5, STAG2, SLC9A6, SRPX, TA2, ARMC5, ELF4, BEX2, DDX3Y, BEX5, CNKSR2, ELK1, MORF4L2, FLNA, MORC4, GPRASP1, MID1IP1, MAGED1, MCTS1, MAP7D2, RRAGB, NGFRAP1, RPGR, PAK3, RNF128, PIGA, RBMX, PIR, PJA1, MS4A6A, DTNA, TCEAL4

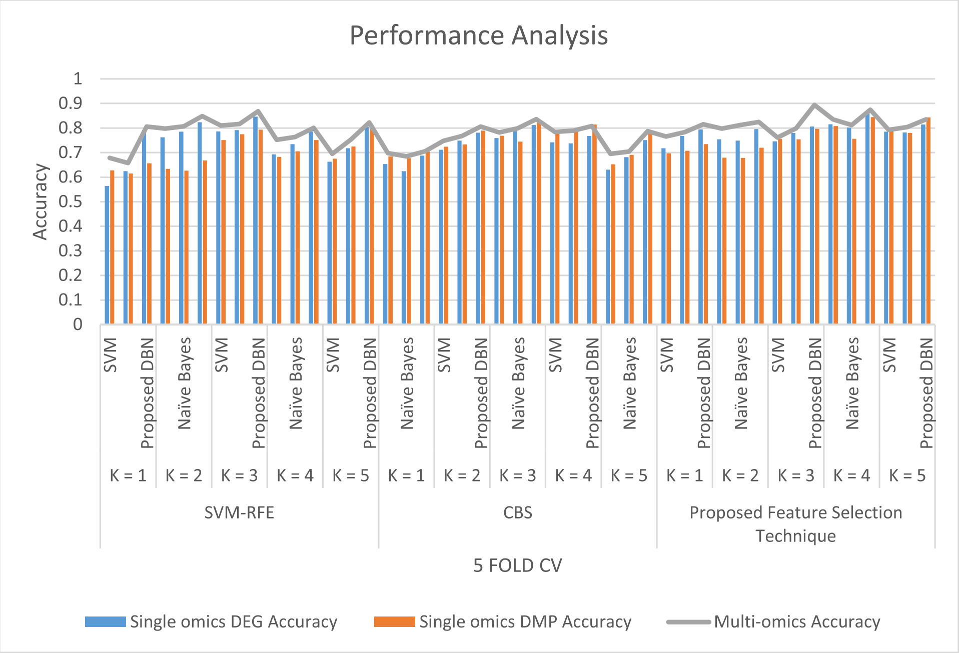

Fig. 4. The Accuracy of Feature Selection and the Prediction Models in 5-Fold CV.

CRediT authorship contribution statement

Nivedhitha Mahendran: Writing – original draft, Visualization. **Durai Raj Vincent:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Jo T, Nho K, Saykin AJ. Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. Front Aging Neurosci 2019;11:220.
[2] Reitz C, Brayne C, Mayeux R. Epidemiology of Alzheimer disease. Nat Rev Neurol 2011;7(3):137–52.
[3] Salat DH, Kaye JA, Janowsky JS. Selective preservation and degeneration within the prefrontal cortex in aging and Alzheimer disease. Arch Neurol 2001;58(9):1403–8.
[4] Lawrence AD, Sahakian BJ. Alzheimer disease, attention, and the cholinergic system. Alzheimer Dis Assoc Disord 1995.
[5] Reitz C, Mayeux R. Alzheimer disease: epidemiology, diagnostic criteria, risk factors and biomarkers. Biochem Pharmacol 2014;88(4):640–51.
[6] Chen H, He Y, Ji J, Shi Y. A machine learning method for identifying critical interactions between gene pairs in Alzheimer's disease prediction. Front Neurol 2019;10.
[7] Ren J, Zhang B, Wei D, Zhang Z. Identification of methylated gene biomarkers in patients with Alzheimer's disease based on machine learning. BioMed Res Int 2020;2020.
[8] Wang L, Liu ZP. Detecting diagnostic biomarkers of Alzheimer's disease by integrating gene expression data in six brain regions. Front Genet 2019;10:157.
[9] Badhwar A, McFall GP, Sapkota S, Black SE, Chertkow H, Duchesne S, et al. A multiomics approach to heterogeneity in Alzheimer's disease: focused review and roadmap. Brain 2020;143(5):1315–31.
[10] Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. Brief Bioinforma 2018;19(6):1370–81.
[11] Singhal P, Verma SS, Dudek SM, Ritchie MD. Neural network-based multiomics data integration in Alzheimer's disease (July). *Proc Genet Evolut Comput Conf Companion*2019:403–4.
[12] Ma S, Ren J, Fenyö D. Breast cancer prognostics using multi-omics data. AMIA Summits Transl Sci Proc 2016;2016:52.
[13] Lin E, Lane HY. Machine learning and systems genomics approaches for multi-omics data. Biomark Res 2017;5(1):2.
[14] Wang D, Gu J. Integrative clustering methods of multi-omics data for molecule-based cancer classifications. Quant Biol 2016;4(1):58–67.
[15] Peng C, Li A, Wang M. Discovery of bladder Cancer-related genes using integrative heterogeneous network modeling of multi-omics data. Sci Rep 2017;7(1):1–11.
[16] Yuan L, Guo LH, Yuan CA, Zhang Y, Han K, Nandi AK, et al. Integration of multi-omics data for gene regulatory network inference and application to breast cancer. IEEE/ACM Trans Comput Biol Bioinforma 2018;16(3):782–91.
[17] Park C, Ha J, Park S. Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. Expert Syst Appl 2020;140:112873.
[18] Kamboh, M.I., Demirci, F.Y., Wang, X., Minster, R.L., Carrasquillo, M.M., Pankratz, V.S.,. & Logue, M.W. (2012). Genome-wide association study of Alzheimer's disease. Translational psychiatry, 2(5), e117-e117.
[19] Ljubic B, Roychoudhury S, Cao XH, Pavlovski M, Obradovic S, Nair R, et al. Influence of medical domain knowledge on deep learning for Alzheimer's disease prediction. Comput Methods Prog Biomed 2020:105765.
[20] Moreira LB, Namen AA. A hybrid data mining model for diagnosis of patients with clinical suspicion of Dementia. Comput Methods Prog Biomed 2018;165:139–49.
[21] Srinivasan K, Ankur A, Sharma A. Super-resolution of magnetic resonance images using deep convolutional neural networks (June). 2017 IEEE Int Conf Consum Electron-Taiwan (ICCE-TW)2017:41–2.
[22] Agarwal P, Wang HC, Srinivasan K. Epileptic seizure prediction over EEG data using hybrid CNN-SVM model with edge computing services. MATEC Web Conf 2018;Vol. 210:03016. (EDP Sciences).
[23] Chakriswaran P, Vincent DR, Srinivasan K, Sharma V, Chang CY, Reina DG. Emotion AI-driven sentiment analysis: A survey, future research directions, and open issues. Appl Sci 2019;9(24):5462.
[24] Khan MA, Muhammad K, Sharif M, Akram T, de Albuquerque VHC. Multi-class skin lesion detection and classification via teledermatology. IEEE J Biomed Health Inform 2021;25(12):4267–75.
[25] Khan MA, Zhang YD, Sharif M, Akram T. Pixels to classes: intelligent learning framework for multiclass skin lesion localization and classification. Comput Electr Eng 2021;90:106956.
[26] Sarraf S, Tofighi G. Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks. arXiv Prepr arXiv 2016;1603:*08631*.
[27] Sarraf S, Tofighi G. Deep learning-based pipeline to recognize Alzheimer's disease using fMRI data (December). *2016 Future Technol Conf (FTC)*2016:816–20.
[28] Farooq A, Anwar S, Awais M, Rehman S. A deep CNN based multi-class classification of Alzheimer's disease using MRI (October). *2017 IEEE Int Conf Imaging Syst Tech (IST)*2017:1–6.
[29] Ji H, Liu Z, Yan WQ, Klette R. Early diagnosis of Alzheimer's disease using deep learning (June). *Proc 2nd Int Conf Control Comput Vis*2019:87–91.
[30] Ramzan F, Khan MUG, Rehmat A, Iqbal S, Saba T, Rehman A, et al. A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and residual neural networks. J Med Syst 2020;44(2):1–16.
[31] Tufail AB, Ma YK, Zhang QN. Binary classification of Alzheimer's disease using sMRI imaging modality and deep learning. J Digit Imaging 2020;33(5):1073–90.

[32] Narayanan M, Huynh JL, Wang K, Yang X, et al. Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases. Mol Syst Biol 2014;10:743. Jul 30.

[33] Zhang B, Gaiteri C, Bodea LG, Wang Z, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell 2013;153(3):707–20. Apr 25.

[34] Smith RG, Hannon E, De Jager PL, Chibnik L, et al. Elevated DNA methylation across a 48-kb region spanning the HOXA gene cluster is associated with Alzheimer's disease neuropathology. Alzheimers Dement 2018;14(12):1580–8. (Dec).

[35] Cheadle C, Vawter MP, Freed WJ, Becker KG. Analysis of microarray data using Z score transformation. (May). J Mol Diagn 2003;5(2):73–81. https://doi.org/10.1016/S1525-1578(10)60455-2. PMID: 12707371; PMCID: PMC1907322.

[36] Du P, Zhang X, Huang C, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinforma 2010;11:587.

[37] Molla M, Waddell M, Page D, Shavlik J. Using machine learning to design and interpret gene-expression microarrays. AI Mag 2004;25(1):23. https://doi.org/10.1609/aimag.v25i1.1745

[38] Rauschert S, Raubenheimer K, Melton PE, et al. Machine learning and clinical epigenetics: a review of challenges for diagnosis and classification. Clin Epigenet 2020;12:51. https://doi.org/10.1186/s13148-020-00842-4

[39] An, Ning & Jin, Liuqi & Ding, Huitong & Jiaoyun, Yang & Yuan, Jing (2020). A deep belief network-based method to identify proteomic risk markers for Alzheimer disease.

[40] (a) Lopes N, Ribeiro B, Gonçalves J. Restricted Boltzmann Machines and Deep Belief Networks on multi-core processors. 2012 Int Jt Conf Neural Netw (IJCNN), Brisb, QLD 2012:1–7. https://doi.org/10.1109/IJCNN.2012.6252431;
(b) Le Roux N, Bengio Y. Representational power of restricted boltzmann machines and deep belief networks. (Jun). Neural Comput 2008;20(6):1631–49. https://doi.org/10.1162/neco.2008.04-07-510