Classification of Autism Gene Expression Data using Deep Learning

Noura Samy[1], Radwa Fathalla [2], Nahla A.Belal[3], and Osama Badawy[4]
[1]Arab Academy for Science and Technology and Maritime Transport
[2,3,4]{radwa_fathalla, nahlabelal and obadawy}@aast.edu
[2]Collage of Computing and Information Technology,Alexandria,Egypt
[1]nou.moh91 @gmail.com

Abstract. Gene expression data is used in the prediction of many diseases. Autism spectrum disorder (ASD) is among those diseases, where gene expression data is analyzed for gene selection and classification. The difficulty of selection and identification of the ASD genes remains a major setback in the gene expression analysis of ASD. This paper aims to develop a model for classifying ASD subjects. The paper employed: Deep Belief Network (DBN) based on the Gaussian Restricted Boltzmann machine (GRBM). Restricted Boltzmann machine (RBM) is considered a popular graphical model that constructs a latent representation of raw data fed at its input nodes. The model is based on its learning algorithm, namely, contrastive divergence, and information gain (IG) is used as the criterion for gene selection. Our proposed model proves that it can deal with gene expression values efficiently and achieved improvements over classical classification methods. The results show that that the most discriminative genes can be selected and identified with its gene expression values. We report an increase of 8% over the highest achieving algorithm on a standard dataset in terms of accuracy.

Keywords: Restricted Boltzmann machine, information gain, feature analysis, gene expression, autism, deep learning.

1 Introduction

Autism spectrum disorders (ASD) (also known as Autism) among individuals are commonly found in societies. Autism refers to a group of developmental brain disorders. It includes a variety of symptoms and levels of impairment or disability. ASD are related to brain growth. It affects how people recognize and deal with others, causing problems in interaction and social communication. The disorder also includes specific patterns and frequent behavior. Autism has a prevalence ratio of 1:4 among males to females. Genetic composition and environment may play a role in the ASD. Researchers assert that autism has genetic causes, which are the main cause of the disease. Autism comes in different degrees. An individual can be mildly impaired by the symptoms or severely disabled [1]. A class of disorders distinguishes autism degree. These disorders are called Pervasive Developmental Disorders (PDDs). The PDDs includes: Autistic Disorder (classic autism: Fragile X syndrome), Asperser's Disorder (Asperser syndrome), Pervasive Developmental Disorder not Otherwise Specified (PDD-NOS), Rett's Disorder (Rett syndrome) and Childhood Disintegrative Disorder (CDD) [1]. It is difficult to explain the disease through mutations that appear, or through rare interactions that are multi-genes. Autism genetic variables or DNA does not change, but are inherited, affecting the so-called gene expression, such as Fragile X syndrome. There is no way to prevent autism spectrum dis-order. However, early diagnosis can improve behavior and skills [1]. Microarray technology allowed gene expression data to be highly available. Microarray allows the gathering of data, which determine the expression pattern of thousands of genes [2]. The profile pattern of mRNA might exhibit the genes or the environmental factors, which will cause the disorder. The larger number of features with small sample size is the cause of high dimensionality problems. This is the focal problem in classifying gene expression data [2]. Consequently, the gene selection can distinguish differentially expressed genes, or eliminate unrelated genes. There are two main types of gene selection: filter methods and wrapper methods. The first type of gene selection, filter methods, are efficient because it respects computational time and works without using classifiers [3]. Hence, filtering methods are selected in examining the high-dimensional data of microarray datasets. Information gain is used in our method to filter genes. The filter method calculates the degree of gene correlation of each gene to the target group through the internal characteristics of the data set [4]. There are two types of machine learning algorithms: supervised learning and unsupervised learning. Supervised machine learning involves algorithms that use input variables to forecast a target classification (i.e., the dependent variable). Supervised machine learning algorithms are either categorical or continuous. Among the supervised learning algorithms are: Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Bayesian Network (BN), Multilayer Perceptron's (MLP or NN), and Decision Trees (DT). The second type of machine learning algorithms, unsupervised learning, is also referred to as descriptive learning. Unsupervised learning is training through input data without any known output [5]. Examples of Unsupervised techniques include: Clustering and Self-Organizing Maps (SOM). Unsupervised techniques analyze the relationships among diverse genes. Lately, Deep learning methods have gained huge popularity due to their high performance. The hierarchical structure of deep neural network conduct the nonlinear transfer of the input data. During the analysis of gene expression, there are frequent challenges in the selection and identification of the most relevant genes to autism. This struggle is due to variations associated with the experiments. This struggle could also be due to the existence of alterations in the genes. In an autistic case, the variance may be associated to the existence of alterations in many genes. There are further difficulties found when there are small samples (in the range of hundreds) versus the large samples (in the range of tens of thousands) [6]. In this paper, a Deep Belief Network (DBN) based on Gaussian–Bernoulli Restricted Boltzmann Machine (GBRBM) is used as a classifier that employs deep learning for autism disease classification. The IG filter is used as a gene selector to remove irrelevant genes, and to select the most relevant genes. DBN is a stacking of RBMs. The conditional probabilities are considered as the input of the RBM. The conditional

probabilities take into consideration the real values normalized to a range from [0, 1]. The output is the per class probability. The Contrastive Divergence (CD) method is used for training the model. The proposed model is tested using a gene expression dataset downloaded from NCBI for fragile X syndrome and it is proved effective in binary classification problems that contain gene expression values.

The current paper contains the following sections: Section 2 presents the related works regarding the research problem; Sections 3 illustrates the proposed model and the data used; Section 4 explains the experiments and the emerged results; and Section 5 concludes and explores future directions.

2 Related work

This section presents existing studies and work related to feature selection and classification of au-tism disorder. It reviews several works done in different machine learning methods that were applied on microarray gene expression data. Scientists can observe gene expression through microarray technology on a genomic scale. Microarray technology has increased the possibility of classification and diagnosis on the gene expression level.

Anibal Sólon Heinsfeld et al. [7], identify ASD with the application of deep learning algorithms, a brain-based disorder. They indicated in their study that ASD reflect social deficits and repetitive be-haviors. They objectively identified ASD participants, using data regarding functional brain imaging. They explained that the brain distinguished ASD from typically developing controls. The results achieved an accuracy of 70%.

Shilan S. Hameed, et al. [6] research used different statistical filters and a wrapper-based geometric binary particle swarm optimization-support vector machine (GBPSO-SVM) algorithm. In their study, they investigated the gene expression analysis of ASD. They explained that there is always difficulty in the process of seeking the genes relevant to autism.

Tomasz Latkowski and Stanislaw Osowski [8] study focused on an application of data mining. This application mainly recognized the significant genes and its sequences in a dataset of gene expression microarray. Their research used several analyses: the Fisher discriminant analysis, relief algorithm method, two sample t-test, Kolmogorov–Smirnov test, Kruskal Wallis test, stepwise regression meth-od, feature correlation with a class and SVM recursive feature elimination. These tests were used to assess the autism data. Accordingly, their research classification accuracy increased 78% after inte-gration.

Julian Tang, et al. [9] used the dimensionality reduction in assessing the autism data. This assess-ment is a popular technique to delete noise and redundant features. Results specified that the dimen-sionality reduction techniques categorized data into: feature extraction and feature selection.

In another study [10], Nur Amalina Rupawon and Zuraini Ali Shah applied several methods for data mining to select the informative genes of autism in gene microarray. Their research used a two-stage selection, genetic algorithm hybrid with three different classifiers, namely: “K-nearest Neighbor (KNN), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA)”. Their proposed model reported an 8% increase over the results obtained in [8].

Masayuki Tanaka et al. [11] suggested an inference of the RBM. Their research studied the input of the RBM (random binary variable). It also studied the straightforward derivation. They revealed the transition from the stacked RBM to the DBN-DNN. They indicated that the proposed inference improves the performance of the DNN. They also showed that the conventional inference was insignificant. Thus, the suggested inference is sensible than RBM conventional algorithm.

Jit Gupta et al. [12] projected the classification of gene expression data. In their study they used a Gaussian Restricted Boltzmann Machine (RBM). RBM is considered a machine learning model. This model focuses on the Neural Networks. In their research, RBM is applied on a binary classification problem. This aided in identifying if certain individuals are affected by lung adenocarcinoma. Thus, the analyses main focus was on the gene expression values and Random Forest used as a gene selector.

Kayleigh K. Hyde, et al [13] research reviewed various ASD literature, giving scholars a proposed method for identifying and describing supervised machine-learning trends. Their insights acted as empirical evidence used to fill in the academic gap found in literature. The emerged empirical evidence increased the body of mining ASD data, relating to clinically, computationally, and statistically sound approaches.

Lingyun Gao et al. [14] used several analyses for the data in their study. These analyses provided sufficient cancer classification

accuracy, using the entire set of genes as data. In the beginning of the research, the IG was initially employed to filter irrelevant and redundant genes. Next, the SVM was used to remove the noise in the datasets. Finally, IG-SVM was conducted to indicate and select the informative genes. Results of the IG-SVM showed a classification accuracy of 90.32% for colon cancer.

In [15], a novel approach was used for classifying and analyzing the cancer detection. The study mainly focused on the determination of cancer and its subtypes. The classification was performed on selected features, derived from both (1) Particle Swarm Optimization algorithm and (2) Ant colony Optimization algorithm. The study used breast cancer gene microarray datasets.

In another study [16], Andrey Bondarenko et al. compared between RBMs and DBNs classification performance against other accepted classifiers. Some of the classifiers that the study compared were: SVMs and Random Forest Trees. The study used several datasets: UCI datasets, and a proprietary document classification dataset, which was single mid-sized. In conclusion to their study, the existing approaches allowed RBMs and DBNs to cope with high dimensional data. RBMs allowed the performance of training on unlabeled data.

Johannes Smolander et al. [17] focused on the arrangement of patients with breast cancer and inflammatory bowel disease. Their analyses focused on high-dimensional gene expression data. They investigated classifiers that integrated deep belief networks and vector machines. By combining classifiers, it aided to solve high dimensionality problem in genomic data. The research studied a computational diagnostics task. The results of their study were able to introduce guidelines for the complex usage of DBN. Their study showed how DBN could be used to classify gene expression data from complex diseases.

James A. Koziol et al. [18] research study used Boltzmann machines for the classification problem regarding the diagnosis of hepatocellular carcinoma. Their study used two methods for the classification problem: "logistic regression and restricted Boltzmann machines (RBMs)". The analysis that was used in this study was the 10-fold cross-validation for the determination of operating characteristics of the classifiers. Results of the study indicated that: "RBMs typically had greater sensitivities, but smaller specificities, than logistic regression with the same input variables".

Xue Jiang et al. [19] study identified the vital genes that affect disease progression. The study sought to: (1) identify hierarchical structures, and (2) apprehend differential analysis of gene expression datasets. The research focused and used the restricted Boltzmann machine. The study conducted the investigation by using Huntington's disease mice at different time periods. In conclusion, the results showed that SRBM-II outperformed other traditional methods.

Shubhra Sankar Ray et al. [21] established a granular self-organizing map (GSOM). In the study, they combined a fuzzy rough set with the SOM. They explained that during the progress of the GSOM, weights of the neighborhood neurons and the winning neuron are updated. This update was caused by a modified learning procedure. Results of the study showed that GSOM have an effect on the clustering samples and the development fuzzy rough feature selection.

In the current study, the proposed model is inspired from [12]. This model used Gaussian restricted Boltzmann machine to solve the classification problem from cancer gene expression dataset. However, in the current study, we applied an RBM based model, which is devised by Masayuki Tanaka [11]. The inference for RBM will be used to classify autisms' gene expression. This study also incorporates information gain into the proposed model to be used as a gene selector.

3 Proposed Model

In this paper, a deep learning method for classification of gene expression data is presented. Prior studies focused on feature selection on both supervised and unsupervised learning. We decided to focus on the problem of supervised learning (classification) in autism, where the class labels are known beforehand. In order to exclude irrelevant genes from the given gene expression, we applied the information gain procedure to generate dependency weight vector. The weight vector denotes the correlation of the gene and ASD. The main outcome of this phase is choosing the most relevant genes to autism. In this work, we sort these genes' weight vector in a descending order. Based on a manually preset threshold, we select only the most relevant genes. The updated data records will be used during the classification steps in the next phase.

The main contribution of this study is the identification of the process of classifying gene expression. In the first phase, this study uses small sample data size with large number of genes (features) that manifests severe high dimensionality. High dimensionality is the main problem because its increase the computational complexity, increase the risk of over fitting and the sparsity of the data will grow.

According to studies, there is a huge amount of high-dimensional datasets. In addition, feature selec-tion has drawn great interest in the field of machine learning. Former studies addressed various ap-proaches, such as: clustering, regression and classification, in both supervised and unsupervised ways. Generally, the field is challenged by 3 difficulties: "Class ambiguity- Boundary complexity-Sample sparsity". Today, high-dimensional, data of small sample size are common in various fields; they include: "genetic microarrays previously presented, chemometrics, medical imaging, text recog-nition, face recognition, finance, and so on". The features of these problems hinder the execution of a reliable and efficient classification system. Thus, feature selection is significant to avoid this prob-lem [22]. The feature selection is used: (1) to identify different expressed genes, (2) to choose the appropriate features, and (3) to remove the irrelevant genes (not harming the remaining genes). The outcome of this phase is then passed to a classification module that uses a DBN based on GBRBM.

A Gaussian RBM contains normalized real values on visibles between the ranges of [0:1]. In this study, we used a Deep belief network-Deep Neural Network (DBN-DNN) hybrid architecture. First, stand-alone RBMs are pre-trained, with seen data by the Contrastive Divergence (CD) training algo-rithm. Stacking the trained RBMs forms a DBN. The outputs of the trained RBM on the hidden nodes are supplied as the training data on the input nodes of the RBM of the next layer. Then, the DBN is unfolded into a DNN by adding a topmost layer with nodes corresponding to classes. The fi-nal move is fine-tuning the architecture parameters by Back-Propagation (BP) algorithm of the error on the topmost layer. The following block diagram explains our model (Fig. 1).

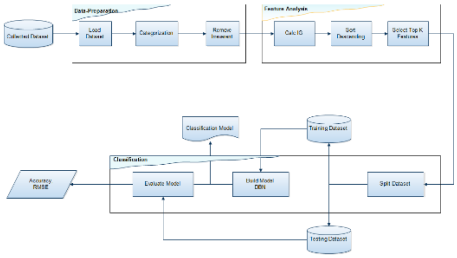

Fig.1. Block diagram of our proposed model.

3.1 Data preparation

The data records were subjected to a normalization process yielding the values in the ranged from 0 to 1.

3.2 Feature analysis

In this phase, the dataset is ready to be handled by feature selection. The study uses information gain (IG). Filter techniques (IG) contain several advantages. They are able to scale to very high-dimensional datasets. They are simple and fast to use. They are independent of the classification algorithm. The IG method assesses the applicability of features by observing the intrinsic properties of the data. Thus, the IG method is applied as a filter in gene selection. It helps rank the genes based on entropy. Entropy is an information theory measure [4]. It can be viewed as the expectation of how useful the information in a message, which is represented by IG value. Filter techniques (IG) relies on the degree of the entropy, which reflects the amount of information this attribute contributes to the data set. IG value is calculated for each feature. It aids to decide whether this feature is to be chosen, or not. The gene contribution with more information yields a high IG value [20]. Thus, IG values are sorted in descending order, and a cut-off point is applied. The study assumes that dataset have M instances. M= {1, 2, 3… m} with x classes. Calculate the entropy of the dataset using equation (1) [4]:

$$\text{Entropy}(M) = -\sum_{i=1}^{x} P(C_i, M) * \log P(C_i, M) \quad (1)$$

Where P (C_i, N) representsC_i,the ratio of and M where C_i,i= 1, 2,.., x are set of instances that belong to the i the class. The entropy of the dataset from gene y calculated using equation (2)[4]:

$$\text{Entropy}_y(M) = \sum_{j-1}^{n} \frac{|M_j|}{M} * \text{Entropy}(M_j) (2)$$

If y is gene that has distinct valued L = {L_1,L_2,…..L_n } and letting $M_j \in M \mid y=L_n$

Calculate IG value of gene u using equation (3)[4]:

$$IG(u) = \text{entropy}(M) - \text{entropy}_U(M) (3)$$

Because gene expression values are within different ranges for different, we need to normalize the calculated IG equation using the following equation (4).

$$X = (Y_i - \min(Y))/(\max(Y) - \min(Y)) (4)$$

Where Y= (Y_1 ... ... Y_n)and $X_i$is the ithe normalized data point. Sort descending according to normalized IG value calculated by Eq. (4). Select top of k mean (manually) that we only accept the most important genes. The result is transferred for use during classification phase in the next phase.

3.3 Classification

In this phase, we utilize the emerged filtered dataset originated from the previous phase of feature selection. The top k features are

the new data. This study builds a model. The model applied a novel inference for RBM. The inference of the RBM is a main key in the DBN-DNN. It is able to portray the proposed inference. It reflects the probabilistic properties of the RBM. Nevertheless, the exact calculation of the proposed inference is intractable. Consequently, the closed form approximation is also conducted. The CD training and the DBN-DNN is also applied with the pro-posed inference, which is devised by Masayuki Tanaka [11]. A RBM is depicted in Fig. 2. Accordingly, the visible layer is the input. It contains unlabeled data to the neural network. In addition, the hidden layer appeals to the features from the input data. Each neuron shows a different feature [11]. Works have defined RBM as a bipartite undirected graph. A RBM has m visible units $\vec{V} = (V_1, V_2, \dots, V_m)$, the input data, and n hidden units $\vec{H} = (H_1, H_2, \dots, H_n)$, the features [11]. A joint configuration $(\vec{v}, \vec{h})$ of the visible and hidden units has an energy given by Eq. (5) [11].

$$E(\vec{v}, \vec{h}) = \sum_{i=1}^{m} a_i v_i - \sum_{j=1}^{n} b_j h_j - \sum_{i=1}^{m}\sum_{j=1}^{n} v_i h_j w_{ij} \quad (5)$$

Where v_i and h_j are the binary states of the visible and hidden units, respectively; a_i, b_j are the biases, and w_{ij} is a real valued weight associated with each edge in the network. The building block of a RBM is a binary stochastic neuron [11]. Fig.3 shows how to obtain the state of a hidden neuron given a visible layer (data). A RBM can be seen as a stochastic neural network. First, weights w_{ij} are randomly initialized. Then, the data to be learned is set at the visible layer. Now, the state of the neurons at the hidden layer is obtained by Eq. (6) [1].

$$P(h_j = 1|\vec{v}) = \text{sig}\,(b_j + \sum_i v_i w_{ij}) \quad (6)$$

So the conditional probability of h_j being 1 is the firing rate of a stochastic neuron with a sigmoid activation function by Eq. (7) [11].

$$\text{sig}(x) = \frac{1}{1 - e^{-x}} (7)$$

Learning in a RBM is achieved by the Contrastive Divergence (CD) algorithm Fig.2. The first step of the CD algorithm is $< v_i h_j >_0$.The next step is the "reconstruction" of the visible layer by Eq. (8) [11].

$$P(v_i = 1|\vec{h}) = \text{sig}\,(a_j + \sum_j h_j w_{ij}) (8)$$

This step is denoted as $< h_j v_i >_0$. The new state of the hidden layer is obtained using the result of the reconstruction as the input data, and this step is denoted as $< v_i h_j >_1$. Finally, the weights and biases are adjusted in the following Eq. (9, 10, 11) [11].

$$\Delta w_{ij} = \varepsilon < v_i h_j >_0 - < v_i h_j >_1 (9)$$

$$\Delta a_i = \varepsilon(v_i^0 - v_i^1) \quad (10)$$

$$\Delta b_j = \varepsilon(h_j^0 - h_j^1) \quad (11)$$

Where ε is the learning rate. We applied the novel inference of the RBM. RBM is a main player in the DBN-DNN [11]. The proposed inference considers the probabilistic properties of the RBM. DBN is built by stacking RBMs. Thus, the more levels the DBN has, the deeper the DBN is. The hidden neurons in a RBM1 capture the features from the visible neurons. Then, those features become the input to RBM2, and so on until the RBM is reached.

A DBN extracts features from features in an unsupervised manner (deep learning). The DBN is a directed generative model is calculating by Eq. (12) [11].

$$P(x, h^1, h^2, \dots h^l) = P(x|h^1)P(h^1|h^2) \dots P(h^{l-2}|h^{l-1})P(h^l, h^{l-1}) \quad (12)$$

Where all the conditional layers $P(h^i|h^{i-1})$ are factorized conditional distributions for which the computation of probability. After the neural network parameters for each layer, the weights W and the biases b have been initialized by RBMs. The DBN learning is called fine-tuning. Fine-tuning uses the class-label information of the training data set that was omitted in pre-training. The main goal of this research is to able to generalize to new unseen samples. So, a back-propagate was needed in the final layer of the derivatives. This is calculated by Eq. (13, 14, and 15) [11].

$$\frac{\theta L}{\theta W_{ij}} = \delta_j \frac{\theta \mu_j}{\theta W_{ij}} + T_j \frac{\theta p_j^2}{\theta W_{ij}} \quad (13)$$

$$\delta_j = \frac{\theta L}{\theta \mu_j} = \frac{\theta L}{\theta O_j}\frac{\theta o_j}{\theta n_j}\frac{\theta n_j}{\theta \mu_j} \quad (14)$$

$$T_j = \frac{\theta L}{\theta P_j^2} = \frac{\theta L}{\theta O_j}\frac{\theta O_j}{\theta n_j}\frac{\theta n_j}{\theta p_j^2} \quad (15)$$

Where j is the output nodes, i is the input node L -loss function, O_j - the output of the j-the node and n_j-the input of the activation function of j-the node. The input of the activation function for the proposed algorithm is expressed by Eq. (16) [11].

$$n = \frac{\mu}{\sqrt{1+\rho^2\pi/8}} \quad (16)$$

The derivatives of the hidden layer for the proposed algorithm is calculated by Eq. (17, 18, 19, 20) [11].

$$\frac{\theta L}{\theta w_{ij}} = \delta_j \frac{\theta \mu_j}{\theta w_{ij}} + T_j \frac{\theta \rho^2}{\theta w_{ij}}, \quad (17)$$

$$\delta_j = \alpha_j \frac{\theta n_j}{\theta \mu_j}, \quad (18)$$

$$T_j = \alpha_j \frac{\theta n_j}{\theta \rho_j^2}, \quad (19)$$

$$\alpha_j = \left[\sum_k \left\{\delta_k \frac{\theta \mu k}{\theta o_j} + T_k \frac{\theta \rho_k^2}{\theta o_j}\right\}\right] \frac{\theta o_j}{\theta n_j}, \quad (20)$$

Where the node of the previous layer identify with k, $\delta_k$and T_k are the propagated derivatives from the previous layer.

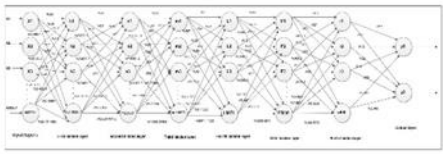

Fig.2. Proposed RBM model.

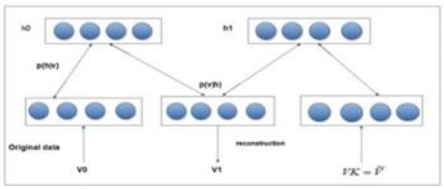

Fig.3.Data and reconstruction in the CD training [11].

4 Experiment Results and Discussion

This section explains the data used for experimentation and the results obtained, along with analysis of the achieved results. This study uses MATLAB for implementation on a machine with Intel Processor, 2.8 GHz and 64-bit architecture and on Windows 10 Professional.

4.1 Experiment Dataset

The datasets used in this study are available on NCBI [23]. In this experiment, we used two datasets. The first dataset refers to the fragile X syndrome gene expression dataset (GEO: GSE7329). The first dataset contains 30 samples (observations) with 43,931 genes (features) in addition to the ground truth class. Classes divided to 15 autistics and 15 non autistic.
The second dataset refers to the peripheral blood lymphocytes (GEO: GSE25507). The second dataset contains 146 samples and 54,613 genes and is used in previous studies [6], [8]. The ground truth class is also given. Classes divided to 69 non autistic and 77 autistics.

4.2 Performance evaluation

The proposed model evaluated through Accuracy and RMSE .We evaluated RMSE of the training and test data using equation (21,22)[11].

$$\text{RMSE} = \sqrt{\frac{1}{MN}\sum_{K=1}^{N} ||y_k - f(X_k)\ ||_2^2} \quad (21)$$

Where M is number of output classes, N is number of data, k the input,y_k the ground truth output and represent the inference of the trained DBN.

$$\text{Accuracy} = \frac{Tp}{N} \quad (22)$$

Where True Positive (TP): These refer the positive tuples that were correctly labeled by the classifier.

4.3 Experiment

The experiments focused on an inference for RBM, which is applied to build a DBN. And, it uses IG for dimensionality reduction. Masayuki Tanaka inspired this approach [11]. The data are not binary data. Succeeding, before dimensionality reduction, different architectures are built, used to obtain the highest accuracy with the best architecture (table. 1). Here, the architectures are given in order.

- A_1: Five-hidden-layers with [43931-4096-1024-512-128-32-2] nodes.
- A_2: Six-hidden layers with [43931-4096-2048-1024-512-128-32-2] nodes (Fig.2).
- A_3: Four – hidden- layers with [43931-1024-512-128-32-2] nodes.

The training hyper parameters used in DBN are as follows: total number of iterations = 100, mini-batch data = 10, Learning Step Size = 0.01. Then, it uses IG for dimensionality reduction. 10000 genes (features) are selected from the dataset. The second architecture (A_2) achieves higher accuracy than the others before using feature selection.
As for the second dataset, we apply the 6-layered architecture [54,613-4096-2048-1024-512-128-32-2], similar to A_2, shown (Fig.2). After dimensionality reduction, the dataset reduces from 54,613 to 10000 with the same weight and bias parameters. Classical classifiers are compared against this architecture in table. 2

4.4 Result and Discussion

Table.1. Comparison of different classifiers on dataset-1.

	Dataset- 1 before dimensionality reduction				Dataset -1 after dimensionality reduction			
	Decision Tree	K-NN	Naïve Bayes	DBN	Decision Tree	K-NN	Naïve Bayes	DBN/IG hybrid
Accuracy	90%	70%	80%	98.77%	53.3%	83.3%	86.67%	98.64%
RMSE	-	-	-	0.667	-	-	-	0.667

Table.2. Comparison of different classifiers on dataset-2.

	Dataset- 2 before dimensionality reduction			*Dataset -2 after dimensionality reduction*
	(GBPSO-SVM) algorithm [6]	Relief algorithm SVM [8]	DBN	DBN/IG hybrid
Accuracy	92.1%	78%	98.66%	98.62%
RMSE	-	-	0.667	0.499

Table.3. Comparison time on dataset1&2.

	Dataset- 1 dimensionality reduction		*Dataset -2 dimensionality reduction*	
	Before	After	Before	After
	Time	Time	Time	Time
Accuracy	2:30h	50 min	3:10h	1h
RMSE	2:30h	50 min	3:10h	1 h

The experiment is implemented on matlab2018A and on a machine that having CPU 2.8 GHz and 64 bit. The research evaluated the algorithm before using feature selection. This allows the research to obtain the best accuracy with the best architecture. Table.1 illustrates the comparison of dataset-1 before and after reduction in classical classifiers, included in our experiments. Also, the table compares between proposed model before using feature selection (represented in the column called DBN) and after feature selection (represented in the column DBN/IG). The research obtains that the highest accuracies using DBNs on both datasets. When the IG is used, the result decreases contrary to expectations. Further, the time decreases approximately to its third. This outcome saves computational power and reduces memory consumption. The shortcut visible nodes (from 43,931 to 10000) lead to decrease equations and calculations, while preserving the performance measures. When the proposed model is compared to other classical classifiers (such as Decision Tree, K-Nearest Neighbor (K-NN) and Naïve Bayes), or related research [6], [8] the results shows that our proposed model surpasses all the other algorithms by a large margin (as shown in Table 1 and 2).

5 Conclusion

In this study, the newly proposed (IG/DBN) hybrid model is applied to the classification of autistic gene expressions based on RBMs. This research supports the early diagnosis for ASDs. It explains that ASDs is significant. It is considered as a clinical best practice. When ASDs is spotted early, it leads to early intervention. Nevertheless, diagnosis before the age of 3 years remains a challenge. For example, there are many cases in which some symptoms of ASDs overlap with normal developmental variance [24]. RBMs have been shown to be a highly performing classification tool in areas of image analysis, video sequence, and motion capture [12]. The utilized datasets consist of real-valued genes. The competitive performance, which is reported in these applications, inspired us to investigate its use in the analysis of gene expression values. We developed a novel framework for RBMs based on the Tanaka's approach [11]. This study also uses IG as a filter to reduce dataset dimensionality. In this study, the research ranked features, and selected the top k, reducing the size of the feature vector. This greatly reduced the model complexity in terms of the

density of parameters to learn. Thus, it elevated computation efficiency of the proposed framework. In all cases, with/without dimensionality reduction, the presented classification tool achieved unprecedented accuracy figures, beating the state-of-the-art on both datasets. This proved that stacked RBMs forming a DBN learn the data model in a more accurate and efficient manner, overcoming the bottleneck of the disproportion between the size of samples and features.

In future work, modern GPUs should be used instead of CPUs for increasing the speed of the computation. According to research, modern GPUs can be 2500 times faster than CPUs. This will allow stacking more RBM layers, in hope of achieving even higher accuracies and ultimately supporting real time processing.

References

1. Joshua A. Gordon.: A Parent's Guide to Autism Spectrum Disorder: National institute of Mental Health, USA, pp. 1-27(2018)
2. Pushpa. M and Swarnamageshwari M.: Review on Feature Selection of Gene Expression Data for Autism Classification: International Journal of Innovative Research in Science. Engineering and Technology, Vol. 5, pp. 3166-3170(2016)
3. Patharawut Saengsiri1, Sageemas Na Wichian2, PhayungMeesad and Herwig Unger.: Integrating Feature Selection Methods for Gene Selection: Semantic Scholar, pp.1-10(2015)
4. Chyh-Ming Lai1, Wei-Chang Yeh and Chung-Yi Chang.: Gene selection using information gain and improved simplified swarm optimization: Neurocomputing, pp1-32(2016)
5. Jiawei Han, MichelineKamber and Jian Pei.: Data Mining Concepts and Techniques, 3rd ed. Elsevier, pp. 1-740(2012)
6. Shilan S. Hameed, Rohayanti Hassan, Fahmi F. Muhammad.: Selection and classification of gene expression in autism disorder: Use of a combination of statistical filters and a GBPSOSVM algorithm. PLoS ONE 12(11): e0187371, pp.1-25(2017)
7. Anibal Sólon Heinsfelda, Alexandre Rosa Francob, c, d, R. Cameron Craddockf, g, AugustoBuchweitzb, d, e, Felipe Meneguzzia.: Identification of autism spectrum disorder using deep learning and the ABIDE dataset, Elsevier. pp. 16-23(2017)
8. Tomasz Latkowski and Stanislaw Osowski.: Data mining for feature selection in gene expression autism data, Elsevier. pp. 864–872(2015)
9. Tang, J., Alelyani, S., & Liu, H.: Feature selection for classification: A review. In Data Classification: Algorithms and Applications.pp. 37-64(2014)
10. Nur Amalina Rupawon1 and Zuraini Ali Shah.: Selection of Informative Gene on Autism Using Statistical and Machine Learning Methods: UTM Computing Proceedings Innovations in Computing Technology and Applications, VOI. V1, pp. 1-8(2016)
11. Masayuki Tanaka and Masatoshi Okutomi.: A Novel Inference of a Restricted Boltzmann Machine. In: IEEE. 22nd International Conference on Pattern Recognition; Tokyo, pp. 1-6(2014)
12. Jit Gupta, Indranil Pradhan and Anupam Ghosh.: Classification of Gene Expression Data using Gaussian Restricted Boltzmann Machine (GRBM): International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 5 Issue: 6, IJRITCC (2017)
13. Kayleigh K. Hyde, Marlena N. Novack, Nicholas LaHaye, Chelsea Parlett-Pelleriti , Raymond Anden, Dennis R. Dixon and Erik Linstead .: Applications of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review : Review Journal of Autism and Developmental Disorders. Springer, USA, pp. 1-19(2019)
14. LingyunGao , Mingquan Ye , Xiaojie Lu , Daobin Huang.: Hybrid Method Based on Information Gain and Support Vector Machine for Gene Selection in Cancer Classification :Genomics, Proteomics & Bioinformatics, Volume: 15 Issue:6, ELSEVIER (2017)
15. R KajaNisha and A Sheik Abdullah.: Classification of Cancer Microarray Data with Feature Selection using Swarm Intelligence Techniques: Acta Scientific Medical Sciences, Volume 3 Issue 7(2019)
16. Andrey Bondarenko, Arkady Borisov, Riga Technical University .: Research on the Classification Ability of Deep Belief Networks on Small and Medium Datasets: Information Technology and Management Science, pp. 60-65(2013)
17. Johannes Smolander, Matthias Dehmer and Frank Emmert Sterib .: Comparing deep belief networks with support vector machines for classifying gene expression data from complex disorder: Open Bio, pp. 1-26(2017)
18. James A. Kozio, EngM. Tan, LipingDai, PengfeiRen, and Jian-Ying Zhang.: Restricted Boltzmann Machines for Classification of Hepatocellular Carcinoma: Computational Biology Journal, Volume 2014, pp. 1-5(2014)
19. XueJiang, Han Zhang, Feng Duan, and XiongwenQuan.: Identify Huntington's disease associated genes based on restricted Boltzmann machine with RNA-seq data: BMC Bioinformatics, pp. 1-13(2017)
20. Nermeen A. Shaltout, Mahmoud El-Hefnawi, Ahmed Rafea, and Ahmed Moustafa.: Information Gain as a Feature Selection Method for the Efficient Classification of Influenza Based on Viral Hosts: Proceedings of the World, London, Vo1 I, pp. 1-7(2014).
21. Shubhra Sankar Ray, Avatharam Ganivada, and Sankar K. Pal.: A Granular Self-Organizing Map for Clustering and Gene Selection in Microarray Data. IEEE, pp. 1-17(2015)
22. Bolón-Canedo, Verónica, Sánchez Maroño, Noelia, Alonso-Betanzos, Amparo.: Feature Selection for High-Dimensional Data, Artificial Intelligence: Foundations, Theory, and Algorithms, Springer,pp.1-163(2015)
23. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov
24. Jariya Chuthapisith and Nichara Ruangdaraganon.: Early Detection of Autism Spectrum Disorders, Autism Spectrum Disorders: The Role of Genetics in Diagnosis and Treatment, Stephen Deutsch, IntechOpen, DOI: 10.5772/17482(August 1st 2011).