Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2016 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

Summer 6-27-2016

TOWARDS DEEP LEARNING IN GENOME-WIDE ASSOCIATION INTERACTION STUDIES

Suneetha Uppu
Curtin University, suneetha.uppu@postgrad.curtin.edu.au

Aneesh Krishna
Curtin University, a.krishna@curtin.edu.au

Raj P. Gopalan
Curtin University, r.gopalan@curtin.edu.au

TOWARDS DEEP LEARNING IN GENOME-WIDE ASSOCIATION INTERACTION STUDIES

Suneetha Uppu, Department of Computing, Curtin University, Perth, Australia, Suneetha.uppu@postgrad.curtin.edu.au

Aneesh Krishna, Department of Computing, Curtin University, Perth, Australia, a.krishna@curtin.edu.au

Raj P.Gopalan, Department of Computing, Curtin University, Perth, Australia, R.Gopalan@curtin.edu.au

Abstract

The complexity of phenotype-genotype mapping are characterised by non-linear interactions between gene-gene and gene-environmental factors. These interaction studies provide better understanding of underlying biological architecture of complex disease traits. A number of statistical and machine learning approaches have been proposed to identify multi-locus interactions between genetic variants and their association to a disease. However, the challenges hindering these approaches are missing heritability, curse of dimensionality, and computational limitations. Despite abundant computational methods and tools available to discover interactions, there have been no breakthrough methods that can demonstrate replicable results. In this paper, a deep feedforward neural network is trained to identify two-locus interacting genetic variants responsible for a disease risk. The method is evaluated on number of simulated datasets to predict the performance of the model. The results are encouraging with replicable results. Hence, the model is further evaluated to confirm its findings on a published genome-wide association dataset. The experimental results demonstrated significant improvements in the prediction accuracy over the previous approaches. The result ranks top 20 interactions among 35 polymorphisms associated with the disease.

Keywords: SNP interactions, epistasis, deep learning, high dimensional data, and gene-gene interactions.

1 INTRODUCTION

Genome-wide association studies (GWAS) have been successful in identifying the associated genetic variants for a number of complex diseases. A genetic variant occurs due to a change in single nucleotide adenine (A), guanine (G), thymine (T), or cytosine (C) in a certain stretch of DNA. These genetic variants occur throughout a person's DNA and are frequently referred to as single nucleotide polymorphisms (SNPs). GWAS predominantly focus on single locus approaches which scan for a single SNP at a time and their associations to a disease (McCarthy et al. 2008). However, a complex disease may not have a clear pattern of disease manifestation. It could be caused by non-linear interactions of genetic and environmental factors acting together or independently (Moore 2003). These interactions may either directly (by changing transcription or translation levels) or indirectly (by altering the protein product) alter the disease risk (Moore et al. 2005, Thornton-Wells et al. 2004). As a step forward in GWAS, a number of studies emerged to discover SNP interactions for a better understanding of underlying biological mechanisms of a complex disease (Cordell 2009, Gusareva et al. 2014).

The computational approaches are divided into three broad categories (Li et al. 2011) and are updated. The methods in the first category are based on exhaustive search. Multi dimensionality reduction (MDR) (Ritchie et al. 2001), multifactor dimensionality reduction based associative classification (MDRAC) (Uppu et al. 2016, Uppu et al. 2015), combinatorial partitioning methods (CPM) (Nelson et al. 2001), and logistic regression (LR) (Marchini et al. 2005) are some of the methods that exhaustively search all the SNP interactions. The second category of methods is based on stochastic search. BEAM (Bayesian epistasis association mapping) (Zhang et al. 2007), BOOST (Boolean operation based screening) (Wan et al. 2010), SNPHarvester (Yang et al. 2009) and epiMODE (Tang et al. 2009) are some of the pioneering works in stochastic searching methods. The third category relies on machine learning approaches such as tree-based and pattern recognition methods. Random forest (RF) (Bureau et al. 2005), Random Jungle (Schwarz et al. 2010) and SNPInterforest (Yoshida et al. 2011) are some of the popular tree based approaches. Some of the pattern recognition methods include support vector machine (SVM) (Chen et al. 2008, Wei et al. 2009) and neural networks (NNs) (Motsinger-Reif et al. 2008) (Upstill-Goddard et al. 2013). The success rate of genome-wide association interaction studies (GWAIS) are relatively high compared to GWAS. However, the increase of multiple testing burden due to small sample size, and no statistically significant findings for detecting interactions are the two major factors that affect these models (Gusareva et al. 2014).

Even though, GWAIS can give new clues for understanding the underlying architecture of a disease, yet there are no breakthrough methods which could produce replicable results. Hence, this research has explored the application of deep learning techniques in GWAIS. Deep learning is a new area of machine learning that discover multiple levels of distributed representation with more abstraction (Bengio et al. 2015). It has emerged from advances in big data by using sophisticated algorithms and the power of parallel computation. A number of studies have produced promising results in image processing, natural language processing and speech recognition (LeCun et al. 2015). Many researchers believe that the deep learning will led to empirical success in many other domains such as bioinformatics (LeCun et al. 2015). Min reviewed the applications of deep learning to bioinformatics (Min et al. 2016). However, there are no deep learning studies in the literature for identifying SNP interactions responsible for complex diseases. Hence, in this paper, a deep feedforward neural network is trained to identify two-locus interactions between SNPs. The model is evaluated on case-control based simulated datasets. The experimental results demonstrated significant improvements in prediction accuracy over some of the previous approaches, such as classification based on predictive association rules (CPAR), gradient boosted machines (GBM), LR, MDR, MDRAC, naive Bayes, NNs, RF, and SVM. Further, the findings of the model are confirmed by validating it on a published dataset. The results ranked top 20 two-locus interactions responsible for the disease among 35 SNPs.

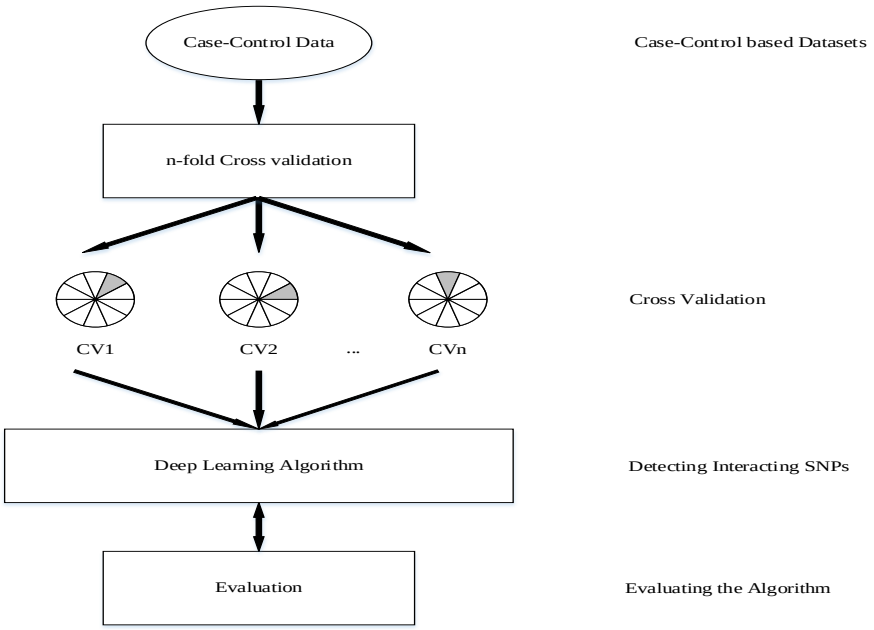

Figure 1. Overview of the deep learning model.

The structure of the paper is as follows: the overview of the deep learning method is presented and applied to the current problem in Section 2. Results are discussed and presented in Section 3. Finally, Section 4 includes the conclusion and future works.

2 METHODS

The trained model is based on feedforward neural networks (Bengio et al. 2015). The overall workflow of the model is represented in Figure 1, updated version of (Gola D 2015). The first stage is data input. Cross validation (CV) is performed in the second stage to assess the ability of the model to classify and to predict a disease status. The following stages encompass the deep learning algorithm and evaluation. The core algorithm classifies the two-locus genotype combinations, and identifies the high risk interacting SNPs associated to a disease. Finally, the classification results are evaluated by prediction accuracy, cross validation consistency (CVC), and classification error.

2.1 Deep Learning Algorithm

The deep learning algorithm used in this paper focuses on feedforward deep networks (Bengio et al. 2015, Candel et al. 2015, LeCun et al. 2015). It is also known as multilayer perceptrons (MLPs) and comprise of multiple layers interconnected with neurons. The basic units of the network are neurons that are inspired from the human brain. The weighted combinations of the inputs are combined together to transmit the output parametric function by the connected neuron. These functions are nonlinear activation functions to compose affine transformations along with a bias, which represents the neuron's activation threshold. The generalised parametric activation output function of a neuron is: $y = b + wx$, where b represents bias, and w is the weight of the input x. The bias is included in all the layers in the architecture excluding the output layer. The number of neurons in each layer is represented as width of the model and the number of layers in the network is represented as depth of the model. Not only, weights and biases linking the neurons determine the output of the entire network, but it also depends on the width and depth of the network. Figure 2 is an illustrative example of a deep feedforward network with one input layer, three hidden layers (h_1, h_2 and h_3) and one output layer.

The activation function of j^{th} neuron in input layer is y_j and is represented as (LeCun et al. 2015):

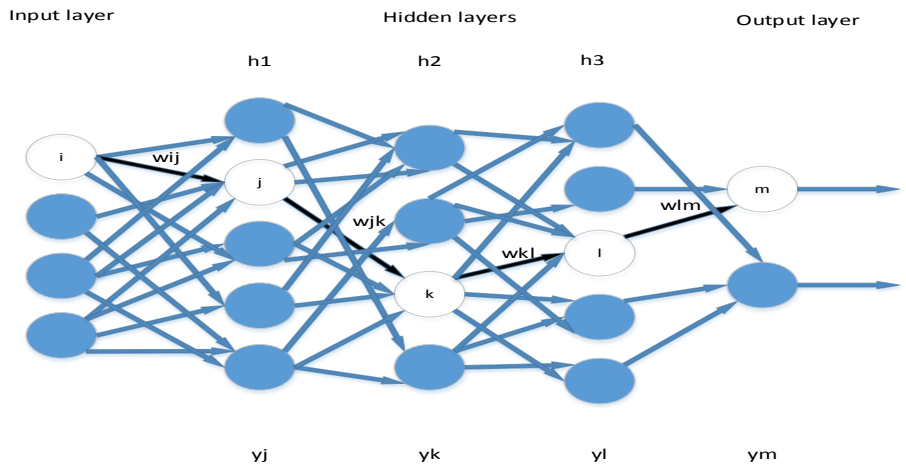

Figure 2. A four layer feedforward network (LeCun et al. 2015).

$$y_j = f\left(a_j\right) = f\left(\sum_{i\,\varepsilon\,input}^{n} w_{ij}x_i + b_i\right) \qquad (1)$$

In the above represented model, the deep learning algorithm is trained by three hidden layers which comprise of 50 neurons in each layer. The hyperbolic tangent activation function is used by the connected neuron to transmit the classification output. The hyperbolic tangent (tanh) function is a rescaled and shifted logistic function, whose symmetry is around 0 that allows the algorithm to converge faster, and is given by (Bengio et al. 2015):

$$f(a) = \tanh(a) = \frac{e^{2a} - 1}{e^{2a} + 1} \qquad (2)$$

The objective function of learning is to adapt the weights by minimising the loss. For training sample t, the cross entropy objective function is given by (Candel et al. 2015):

$$L(W, B|t) = -\sum_{y\in l}(\ln\left(o_y^t\right) * p_y^t + \ln\left(1 - o_y^t\right) * (1 - p_y^t)) \qquad (3)$$

where, W is the collection $\{w_i\}_{1:N-1}$ and B is the collection $\{b_i\}_{1:N-1}$. w_i and b_i are the weight matrix connecting layers i and i+1 for N layers and the vector columns of biases for layer i+1 respectively. Let y represent the output units and l represent the output layer. p denotes predicted output and o denotes actual output respectively. Multinomial distribution function is used along with cross entropy (log-loss) for the response variables in classification. In practise, most of the researchers use parallel versions of stochastic gradient descent (SGD) to minimise the log-loss by handling the memory efficiently. However, the execution time of the algorithm drops drastically. In this method, a lock-free approach is used to parallelise the SGD by sharing the memory with the possibility of overwriting (Candel et al. 2015, Recht et al. 2011).

2.2 Case-control based Datasets

Simulated datasets based on case-control data are generated in two different scenarios. In the first scenario, six two-locus models with different penetrance values are simulated for 20 SNPs with two functional SNPs (P1 and P2) and 18 non-functional SNPs. 400 samples are generated by varying case-control ratios of 1:1, 1:2, 1:4, and 1:6. 100 datasets are simulated for each model. Hence, 2400 datasets are generated in total for first scenario (Uppu et al. 2015). In the second scenario, datasets are replicated as in the study performed by Velez (Velez et al. 2007). The datasets are generated for single locus and two-locus models with 20 SNPs using GAMETES (Urbanowicz et al. 2012). It is a fast and flexible tool used to generate complex n-locus simulated models with random architecture. Each genetic model is distributed across seven heritability (0.01, 0.025, 0.05, 0.1, 0.2, 0.3 and 0.4), and two

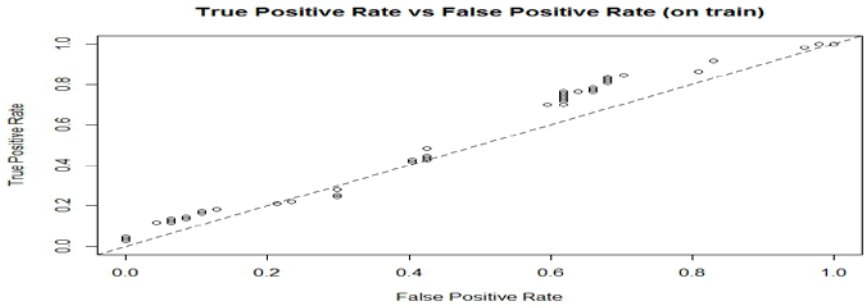

Figure 3. Performance of the deep learning model while training the data.

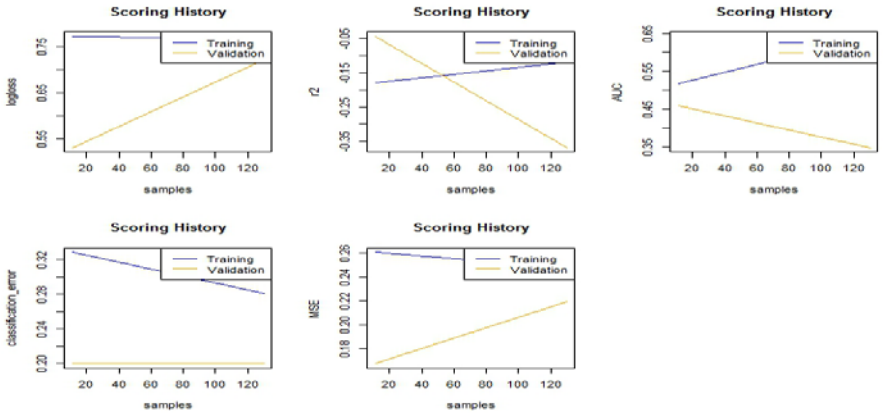

Figure 4. Scoring history of samples vs metrics of deep learning model.

different minor allele frequencies (0.2 and 0.4). In total, 14 models are generated in accordance to Hardy-Weinberg proportions. 100 datasets are generated for each model by varying case-control ratios (1:1, 1:2 and 1:4) and sample size (400,800 and 1600). Hence, for each locus, 12,600 datasets are generated. In total 25,200 datasets are generated in the second scenario (Uppu et al. 2015). The model is further analysed using the data obtained from a whole genome association study (González et al. 2007, Juan R González) to confirm the findings (Uppu et al. 2016). The data comprise of 110 cases and 47 controls, which makes 157 samples in total with 134 missing values. The attributes for each sample consist of an identifier, case or control status, sex, arterial blood pressure, protein levels, and observations of 35 SNPs. SNPs are bi-allelic, and their genotype combinations are numerically represented by zero for common homozygous (AA), one for heterozygous (Aa), and two for variant homozygous (aa). The case or control status represents zero for controls and one for cases.

2.3 Evaluation

The predicted model identifies two-locus SNP interactions and their associations with a disease risk. The models identified by the core algorithm are evaluated by n-fold cross validation and by splitting the data for validation. In the first approach, the data is equally split into n subsets without losing any data. One split is considered for testing and n-1 splits are used for training. The algorithm runs on training data for each split by excluding different splits for testing data. The model is trained for each subset of the data. This process is repeated for each fold. Hence, for n-fold CV, the core algorithm runs n times. Finally, the overall best model is selected with the highest CVC and lowest classification error. Ten-fold cross validation is performed on the model. That is, a model is built 10 times excluding one 10th of the data each time. The model is assessed by the remaining one tenth of the data. The overall best model is evaluated by selecting the model with high prediction accuracy, low classification error, and high cross validation consistency. CVC is used to determine the statistical significance of the predicted model by 1000 fold permutation strategy (Uppu et al. 2015). The p-

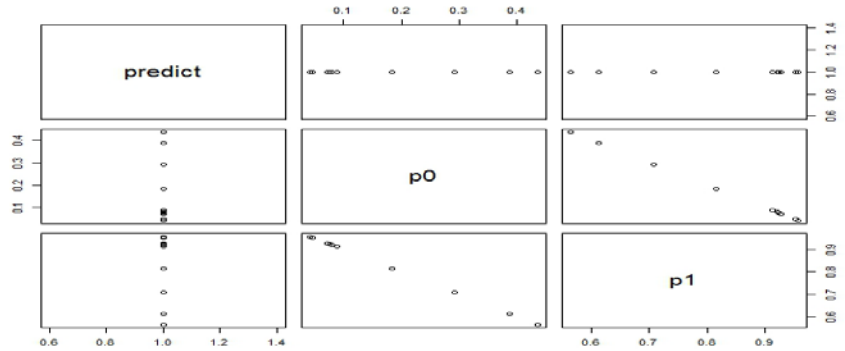

Figure 5. Predicting test data on the deep learning model.

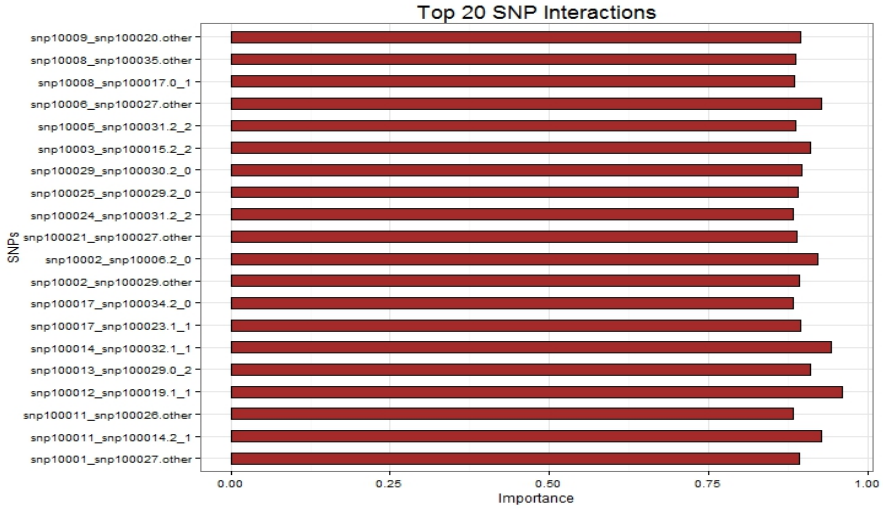

Figure 6. Top 20 interacting SNPs identified by the deep learning model.

values are compared with 0.05 in determining the significance of the findings. In the second approach, the data is randomly split into three parts with 80% of data for training, 10% of data for validation, and remaining 10% data for testing. The performance of the training, validation and testing data are evaluated by determining the metrics of the models. Further, training time and speed of the predicted model is evaluated. Finally, the hypothesis test has been performed (under the null hypothesis of no interaction derived by random permutations of the phenotypes) over the best model chosen to evaluate its statistical significance.

3 RESULTS AND DISCUSSION

Several experiments are performed over simulated datasets and the published dataset to evaluate the accuracy of the trained deep feedforward neural network. The goal of this study is to determine whether the model is a better approach for identifying SNP interactions in genome-wide interaction studies. It identifies statistically significant genotype combinatorial associations based on cases and controls. Despite improving the accuracy of the model, the approach will still reduce false positive errors by permutation testing under the null hypothesis. The approach is developed and analysed in R using H2O package (Aiello et al. 2015).

The deep feedforward model is trained with an input layer, three hidden layers (each layer with 50 neurons), and an output layer. The distribution function of response variable is set to multinomial along with cross-entropy loss function. As a preliminary analysis, a series of simulated datasets in two scenarios (2400 datasets in scenario 1 and 25,200 in scenario 2) are analysed on the model. Results

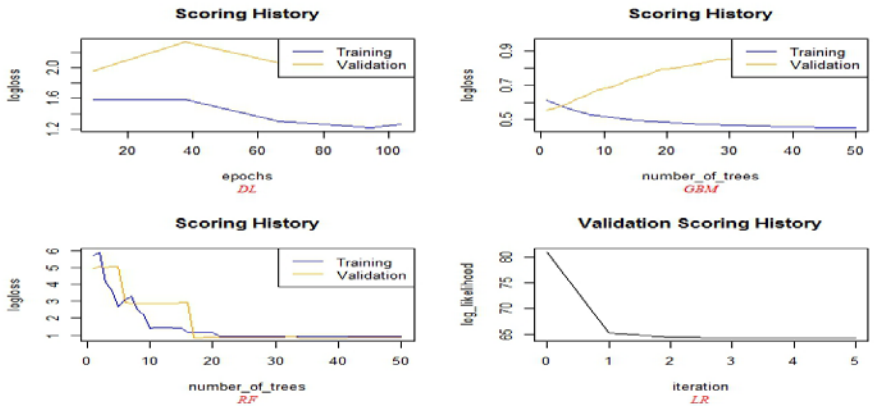

Figure 7. *Comparing the deep learning model with Gradient Boosted Machines (GBM), Random Forest (RF), and Logistic Regression (LR).*

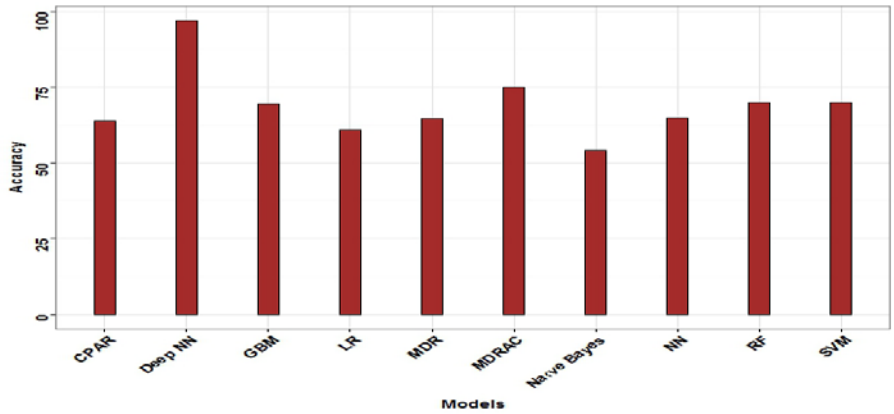

Figure 8. *Prediction accuracy of the deep learning model compared with the previous approaches.*

demonstrated improved prediction accuracy for almost all the datasets. The results are quite interesting and encouraging. Hence, the model is further analysed and validated on the published genome-wide data (González et al. 2007). Figure 3 shows the performance of the model on the entire data during training. False positive rate increases as true positive rate increases. Figure 4 represents the scoring history of the model. Timestep (a unit of measurement for the x-axis) and metrics (a unit of measurement for the y-axis) are the arguments available in the scoring history of the model. Timestep for the model must be either epochs or samples. Metrics of the model are log-loss, r2 (a measure of goodness-of-fit for linear regression), area under the curve (AUC), mean square error value (MSE), and classification error. It is observed that as the number of training samples increased, synchronization and model convergence decreased. When the number of training samples is too low, network communication dominated the runtime by affecting the computational performance. Figure 5 shows the test data predicted for the classification on the trained deep neural network. The model predicted single-locus and two-locus SNPs interactions associated to the disease. Top 20 highly ranked SNPs acting independently or due to two-way interactions are plotted as a bar chart in Figure 6.

The deep learning model is further analysed and compared with some of the previous approaches, such as MDR, MDRAC, RF, LR, GBM, naïve Bayes, CPAR, SVM, and NN. The accuracy of the trained model has the highest prediction accuracy of 97.01% when compared with other previous approaches. The best two-way SNP interaction identified by the model is snp100012 (presence of Aa/aA) and snp100019 (presence of Aa/aA). The java implementation of MDR (version-3.0.2) (Ritchie et al. 2001) is used to analyse the published data. The best two-locus model identified by MDR is snp10001 and snp10005, providing a training accuracy of 64.72 % and cross-validation consistency of 10 out of 10.

The best two-locus SNP interactions identified by MDRAC (Uppu et al. 2015) is SNP 100033 and SNP 10005 with the prediction accuracy of 75.15 %. Even though, the accuracy is better than previous approaches, the accuracy of the model is low compared to the trained deep neural network model. Further, the performance of the model is poor in the presence of genetic heterogeneity and phenocopy (Uppu et al. 2016). The dataset is also used to analyse LR using LogicFS (Schwender et al. 2008) available for R. The accuracy of best two-way interaction model identified by LR is 60.91%. When the number of SNPs increased, it is observed that searching the interacting SNPs among all the possible logic trees became computationally hard. LR uses the simulated annealing as a searching algorithm by improving the variable selection. However, measuring the importance of interacting variables is restricted only to binary variables. GBM, RF, and Naïve Bayes are analysed using H2O interface (Aiello et al. 2015) developed for the R environment. GBM built gradient boosted classification trees on the dataset. The prediction accuracy of the model is 69.47%. RF analysis determines the importance of variables that allows for possible interactions. The prediction accuracy of the model is 70.06%. The power of RF is reduced as it requires a marginal effect in at least one of the SNP interacting pair. However, it is observed that RF outperforms the prediction when the trees do not exhibit a correlation with each other. Figure 7, demonstrates the scoring history of GBM, RF, and LR along with the deep learning method. Naïve Bayes classifier is analysed, whose prediction accuracy is 54.14% with high classification error compared with other methods. Further, the dataset is evaluated on CPAR, SVM, and NN using weka tool whose accuracies to detect two-locus interacting SNPs are 63.69%, 70.07% and 64.97% respectively. The detailed evaluation of some these approaches were studied in the previous research using simulated datasets on various epistatic models (Uppu et al. 2014). The accuracy of all these models is represented as a bar chart in Figure 8.

4 CONCLUSION

In this paper, a deep neural network is trained to detect SNP interactions in genetic and epidemiologic studies of complex diseases. The approach is evaluated for two-locus SNP interactions using simulated case-control datasets and a published dataset. The experimental results of simulated datasets in both scenarios demonstrated significant improvements in the prediction accuracy over the previous machine learning methods. The results are quite interesting and encouraging. Hence, the approach is further analysed and validated on a genome-wide association study. The experimental results for the published data confirmed the improved prediction accuracy of the model over previous methods. Further, the results showed top 20 highly ranked single-locus and two-locus SNPs responsible for the disease manifestation. Currently, the approach is validated only for two-locus case-control based interaction studies. It will be extended over higher order interaction studies, and family-based association studies. Furthermore, studies will be conducted to investigate the performance of the proposed approach in the presence of missing data, genotypic error, phenocopy, and genetic heterogeneity.

References

Aiello, S., Kraljevic, T. and Maj, P. (2015). Package ‘h2o’.

Bengio, Y., Goodfellow, I. J. and Courville, A. (2015). Deep learning. An MIT Press book in preparation. Draft chapters available at http://www. iro. umontreal. ca/~ bengioy/dlbook.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P. and Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. Genetic epidemiology, 28 (2), 171-182.

Candel, A., Parmar, V., LeDell, E. and Arora, A. (2015). Deep Learning with H2O.

Chen, S. H., Sun, J., Dimitrov, L., Turner, A. R., Adams, T. S., Meyers, D. A., Chang, B. L., Zheng, S. L., Grönberg, H. and Xu, J. (2008). A support vector machine approach for detecting gene-gene interaction. Genetic epidemiology, 32 (2), 152-167.

Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. Nature Reviews Genetics, 10 (6), 392-404.

Gola D, M. J. J., van Steen K, König IR. (2015). A roadmap to multifactor dimensionality reduction methods. Briefings in Bioinformatics doi: 10.1093/bib/bbv038

González, J. R., Armengol, L., Solé, X., Guinó, E., Mercader, J. M., Estivill, X. and Moreno, V. (2007). SNPassoc: an R package to perform whole genome association studies. Bioinformatics, 23 (5), 654-655.

Gusareva, E. S. and Van Steen, K. (2014). Practical aspects of genome-wide association interaction analysis. Human genetics, 133 (11), 1343-1358.

Juan R González, L. A., Elisabet Guinó, Xavier Solé, and Víctor Moreno. SNPs-based whole genome association studies.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. Nature, 521 (7553), 436-444.

Li, J., Horstman, B. and Chen, Y. (2011). Detecting epistatic effects in association studies at a genomic level based on an ensemble approach. Bioinformatics, 27 (13), i222-i229.

Marchini, J., Donnelly, P. and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature genetics, 37 (4), 413-417.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics, 9 (5), 356-369.

Min, S., Lee, B. and Yoon, S. (2016). Deep Learning in Bioinformatics. arXiv preprint arXiv:1603.06430,

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human heredity, 56 (1-3), 73-82.

Moore, J. H. and Williams, S. M. (2005). Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. Bioessays, 27 (6), 637-646.

Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W. and Ritchie, M. D. (2008). Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genetic epidemiology, 32 (4), 325-340.

Nelson, M., Kardia, S., Ferrell, R. and Sing, C. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Research, 11 (3), 458-470.

Recht, B., Re, C., Wright, S. and Niu, F. (2011). Hogwild: A lock-free approach to parallelizing stochastic gradient descent. 693-701.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F. and Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. The American Journal of Human Genetics, 69 (1), 138-147.

Schwarz, D. F., König, I. R. and Ziegler, A. (2010). On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics, 26 (14), 1752-1758.

Schwender, H. and Ickstadt, K. (2008). Identification of SNP interactions using logic regression. Biostatistics, 9 (1), 187-198.

Tang, W., Wu, X., Jiang, R. and Li, Y. (2009). Epistatic module detection for case-control studies: a Bayesian model with a Gibbs sampling strategy. PLoS Genet, 5 (5), e1000464.

Thornton-Wells, T. A., Moore, J. H. and Haines, J. L. (2004). Genetics, statistics and human disease: analytical retooling for complexity. TRENDS in Genetics, 20 (12), 640-647.

Uppu, S. and Krishna, A. (2016). Evaluation of associative classification-based multifactor dimensionality reduction in the presence of noise. Network Modeling Analysis in Health Informatics and Bioinformatics, 5 (1), 1-9.

Uppu, S., Krishna, A. and Gopalan, R. P. (2014). Detecting SNP Interactions in Balanced and Imbalanced Datasets using Associative Classification. Australian Journal of Intelligent Information Processing Systems, 14 (1),

Uppu, S., Krishna, A. and Gopalan, R. P. (2015). A Multifactor Dimensionality Reduction Based Associative Classification for Detecting SNP Interactions. Neural Information Processing, Springer, 328-336.

Uppu, S., Krishna, A. and Gopalan, R. P. (2015). Rule-based analysis for detecting epistasis using associative classification mining. Network Modeling Analysis in Health Informatics and Bioinformatics, 4 (1), 1-19.

Upstill-Goddard, R., Eccles, D., Fliege, J. and Collins, A. (2013). Machine learning approaches for the discovery of gene–gene interactions in disease data. Briefings in bioinformatics, 14 (2), 251-260.

Urbanowicz, R. J., Kiralis, J., Sinnott-Armstrong, N. A., Heberling, T., Fisher, J. M. and Moore, J. H. (2012). GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. BioData mining, 5 (1), 1-14.

Velez, D. R., White, B. C., Motsinger, A. A., Bush, W. S., Ritchie, M. D., Williams, S. M. and Moore, J. H. (2007). A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genetic epidemiology, 31 (4), 306-315.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L. and Yu, W. (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. The American Journal of Human Genetics, 87 (3), 325-340.

Wei, Z., Wang, K., Qu, H.-Q., Zhang, H., Bradfield, J., Kim, C., Frackleton, E., Hou, C., Glessner, J. T. and Chiavacci, R. (2009). From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet, 5 (10), e1000678.

Yang, C., He, Z., Wan, X., Yang, Q., Xue, H. and Yu, W. (2009). SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. Bioinformatics, 25 (4), 504-511.

Yoshida, M. and Koike, A. (2011). SNPInterForest: a new method for detecting epistatic interactions. BMC bioinformatics, 12 (1), 469.

Zhang, Y. and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. Nature genetics, 39 (9), 1167-1173.