Deep Learning and Genome-Wide Association Studies for the Classification of Type 2 Diabetes

Basma Abdulaimma
Department of Computer Science
Liverpool John Moores University
Liverpool, United Kingdom
basmatalib77@yahoo.com

Paul Fergus
Department of Computer Science
Liverpool John Moores University
Liverpool, United Kingdom
P.Fergus@ljmu.ac.uk

Carl Chalmers
Department of Computer Science
Liverpool John Moores University
Liverpool, United Kingdom
C.Chalmers@ljmu.ac.uk

Casimiro Curbelo Montañez
Department of Computer Science
Liverpool John Moores University
Liverpool, United Kingdom
contact@acurbelo.com

***Abstract*—Genome-wide association studies (GWAS) have promised to significantly enhance our understanding of genetic based determinants of common complex diseases. A strong body of evidence suggested that genetic factors contribute significantly to the predisposition of Type 2 Diabetes (T2D). However, many studies have shown that single-locus analysis has demonstrated little effect in understanding the genetic architecture of complex human diseases, as is the case of GWAS. Traditional machine learning models, such as random forest and support vector machine have been widely used with genome-wide data as an alternative approach. However, there are still several challenges in modelling high-dimensional GWAS data. This paper addresses these issues using a deep learning framework to model the cumulative effects of Single Nucleotide Polymorphisms (SNP) for the classification of Type 2 Diabetes in the context of genome-wide data. The findings show that using 6609 SNPs it is possible to obtain (AUC=96.53%, Sens=93.91%, Spec=90.83%, Logloss= 32.33%, Gini=93.06%, MSE=9.50%). Using a deep learning approach, it is possible to capture the latent representation of genetic variants and the important interactions between them. Our approach holds great promise and warrants further study.**

Keywords— Classification, Deep Learning, Genome-Wide Association Studies, Machine Learning, Type 2 Diabetes

I. INTRODUCTION

According to the World Health Organization (WHO) [1], diabetes attributed to approximately 1.5 million deaths, which is set to be the seventh leading cause of mortality worldwide by 2030 [2]. As such, understanding the underlying cause of complex human diseases like T2D is high on government agendas. T2D is a multifactorial disorder. This means T2D is caused by the interactions between genetics, the environment, and a sedentary lifestyle [3]. However, genetic susceptibility has been established as a key risk factor. Twin studies have revealed that the concordance rate of T2D in monozygotic twins is approximately 70% compared with 20% to 30% in dizygotic twins [4].

With the evolution of less expensive high-throughput technologies [5], genome-wide association studies (GWAS) have become a vital approach within the field of genetics. In recent years, GWAS have succeeded in identifying genetic variants that show evidence of increased predisposition to a wide range of complex disorders, including Type 2 Diabetes (T2D), Schizophrenia, Epilepsy, Obesity, Cardiovascular Disease, and Hypertension [6]. GWAS utilise single-locus analysis to detect the main genetic effects associated with the phenotype (disease trait) in case-control studies by exploring each single nucleotide polymorphisms (SNP) individually [7]. Complex human diseases are polygenetic disorders that occur as a consequence of the non-linear interactions of multiple genetic loci [8]. This means that GWAS often fail to find the non-linear relationships between SNPs and the phenotype because of its reliance on multi-variable statistical approaches, such as logistic regression, which is more suitable for capturing linear interactions only.

Machine learning algorithms have been broadly utilised to model GWAS data, as shown in [9], [10], [11]. This is because machine learning algorithms, such as random forest [10], support vector machine [10], artificial neural network [12] can model complex relationships and interactions between features (SNPs) and their association with a phenotype of interest. The primary aim of these studies has focused on the detection of correlations between SNPs [9], [12], disease risk prediction [10], and feature selection [11]. Although they have produced some interesting results, many of them do not scale well when a much larger number of SNPs (almost one million SNPs and thousands of samples as in GWAS data) are introduced given the computing resources needed.

Deep learning (DL), however, is making major advances in solving big data problems where they have been used across many domains, which include image and speech recognition [13], [14], natural language processing [15], and pharmaceutical formulation analysis [16]. DL is a representation learning method that consumes raw data and

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

automatically discovers deep abstract representations to learn complex functions [17]. A key aspect of DL is its ability to automatically learn features from data and the interactions between data points using representation learning procedure [18]. Consequently, DL algorithms are useful for capturing the epistatic interactions between SNPs in GWAS. Therefore, in this paper, we consider the application of DL [17], [19], [20] for epistatic analysis and the classification of case (affected with the disease) and control (unaffected with the disease) observations in T2D.

The remainder of this paper is organised as follows. Section 2 provides details about the materials and methods used in this study. The results are presented in Section 3. The findings are discussed in Section 4. And lastly, the paper is concluded, and future work is presented in Section 5.

II. Materials and Methods

A. Data Description

The Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS) in T2D are used in this study following the authorised access to the Database of Genotypes and Phenotypes (dbGap) [21]. The case and control participants were selected from those who provided a blood sample. Case participants were defined as those who reported themselves to be diagnosed by T2D, and a medical record assessment questionnaire confirmed it. Control participants were identified as those without diabetes. The Deoxyribonucleic acid **(**DNA) of case and control samples were genotyped at the Broad Centre for Genotyping and Analysis (CGA) via the Affymetrix Genome-Wide Human 6.0 array.

The dataset contains 6041 NHS and HPFS case-control samples with genotype information across 909622 SNPs. The NHS samples consist of 1581 T2D cases and 1854 controls, and the HPFS subjects comprise 1232 T2D cases and 1374 controls. Participants in the NHS dataset are defined as Hispanic or non-Hispanic, and each belongs to one of four racial categories (White, African-American, Asian or Other). Participants are predominantly White and non-Hispanic, representing 97.4% of the NHS subjects. The HPFS subjects belong to one of the four racial categories (White, Asian, African-American or Other). They are mainly White, representing 96% of the HPFS subjects.

B. Data Quality Control

PLINK v1.9 [22] is used on Windows 10 machine, with 12 GB of Memory and Intel(R) Core(TM) i7-6500U CPU @ 2.50 GHz, to conduct data quality control (QC) and preliminary analysis. PLINK is also utilised to merge the NHS and HPFS datasets (NHS and HPFS subjects were genotyped via the Affymetrix Genome-Wide Human 6.0 array). Before QC a series of steps were conducted to exclude useless information, the 0 Chromosome, non-T2D participants (65 NHS, and 68 HPFS), the HapMap controls (44 NHS, and 29 HPFS) were removed from the study. In addition, only those samples reported to belong to white ancestry were selected to reduce potential bias due to population stratification. QC assessments for individuals and genetic data are conducted separately, following pre-established quality control protocols and guidelines as recommended in [23]. In addition, QC parameters are tuned to meet the requirements of the analysis presented in this study.

Samples that met any of the criteria illustrated in Table I were discarded from the analysis.

TABLE I. Sample Quality Control

Samples Criteria	Number of Removed Samples
Discordant sex-homozygosity rate between 0.2 and 0.8.	14 Samples
Elevated missing data rates - genotype failure ≥ 0.05. Outlying heterozygosity rate ±3 standard deviations from the mean.	131 Samples
Duplicated or related individuals with Identity-by-Descent (IBD) > 0.185.	8 Samples
Divergent ancestry of the 2nd principal component score < 0.061.	51 Samples
Missing genotype data rate of 0.05.	101 Samples

Genetic variants (SNPs) that met any of the criteria in Table II were removed from the analysis.

TABLE II. Genetic Variants Quality Control

SNPs Criteria	Number of Removed SNPs
SNPs with excessive missing data rates.	29 SNPs
Missing genotype rate of 0.01.	116863 SNPs
Minor Allele Frequency (MAF) < 0.05.	178004 SNPs
Hardy-Weinberg Equilibrium (HWE) of p-value < 0.001 in control samples.	2248 SNPs

Following the QC analysis, there were 5393 samples (2481 cases, 2912 controls) and 608342 SNPs for each sample.

C. Logistic Regression Association Analysis

Statistical case-control association analysis is conducted in an unrelated, white racial subpopulation to compare the frequency of alleles or genotypes at genetic marker loci (SNP) between cases and controls of the merged version of T2D Geneva NHS and HPFS Datasets. Pearson's Chi-squared test (x^2) is used to test the null hypothesis (no association). Logistic regression under an additive genetic model was conducted to assess the association of all SNPs within the study with disease status of binary traits (0/1) for case and control subjects. Logistic regression association test is adjusted using Genomic Control (GC) to control population structure, the p-values are considered based on a GC inflation factor.

Let $Y \in \{0,1\}$ be a binary variable for disease status with 0 indicating control and 1 indicating case. Let $X \in \{0,1,2\}$ be a genotype at a particular SNP. Assuming that 0, 1, 2 represent homozygous major allele AA, heterozygous allele Aa and homozygous minor allele aa respectively. Logistic regression modelling is performed under an additive genetic model, and it is given as [24]:

The conditional probability of $Y = 1$ is

$$\theta(X) = P(Y = 1|X) \quad (1)$$

The logit function which is the inverse of the sigmoidal logistic function is represented as:

$$logit(X) = \ln \frac{\theta(X)}{1-\theta(X)} \quad (2)$$

The logit is given as a linear predictor function as follows:

$$logit(X) \sim \beta_0 + \beta_1 X \quad (3)$$

The logistic regression model is used to assess the association of all SNPs within the study with phenotype. Several p-value thresholds are considered including 5×10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} resulting in 7 SNPs, 13 SNPs, 23 SNPs, 103 SNPs, 766 SNPs, and 6609 SNPs respectively. These various subsets of SNPs are used to evaluate the predictive capacity of machine learning in discriminating between cases and controls in T2D GWAS data.

D. Deep Learning

Deep learning, based on a multi-layer feedforward artificial neural network trained with gradient descent using backpropagation is implemented in this analysis for classification tasks, based on the theoretical definitions in [20], [19]. Neurons represent the basic computational units of the network; each neuron takes n input values $x_1, x_2, \ldots x_n$, and a bias intercept term represented by $+1$ (not included in the input), which is a constant term used to overcome the problem related to the input pattern that is zero. The output is a hypothesis $h_{W,b}(x)$ where W and b are weight and bias parameters that can be learned from the input data, x. The neuron output is defined as:

$$h_{W,b}(x) = f(W^T x) = f(\sum_{i=1}^{n} W_i x_i + b) \quad (4)$$

where $f: \mathbb{R} \mapsto \mathbb{R}$ represents the non-linear activation function which is the rectifier linear unit (ReLU) used to compute a weighted sum of the inputs and is given according to:

$$f(x) = \max(0, x) \quad (5)$$

where x denotes the input to the neuron.

The neural network consists of input, hidden, output layers, and each contains n units (neuron). let n_l denote the number of layers in the network where l is a layer and L_l is a particular layer. Thus, L_1 is the input layer and L_{n_l} is the output layer in the network. First, the input vector is transmitted to the input neurons in the input layer, and then the outputs from the input neurons are passed to the hidden neurons in the hidden layer, which is the second layer. This process is continued until the last layer of the hidden layers is reached. Then, the outputs of this last hidden layer are sent to the output neurons in the output layer. In addition to the layers and neurons, the neural network consists of several parameters, including weight and bias. The parameter $(W, b) = \left(W^{(1)}, b^{(1)}, W^{(n)}, b^{(n)}\right)$ where $W_{ij}^{(l)}$ denotes the weight of the connection between unit j in layer l, and unit i in layer $l+1$. Additionally, the bias unit $b_i^{(l)}$, associated with unit i in layer $l+1$ is used with output value equal to $+1$. The number of units in layer l is represented by s_l, and bias unit $b_i^{(l)}$ is not counted with s_l. The output value of unit i in layer l is given by an activation vector $a_i^{(l)}$ which is equal to the total weighted sum of inputs (including the bias term), denoted by $z_i^{(l)}$. Thus, $a_i^{(l)} = f\left(z_i^{(l)}\right)$ where $z_i^{(l)}$ is given as:

$$z_i^{(l+1)} = \sum_{j=1}^{n} W_{ij}^{(l)} x_j + b_i^{(l)} \quad (6)$$

Given a fixed setting of parameters W, b the neural network hypothesis is defined as $h_{W,b}(x)$ which outputs the real number as given by:

$$h_{W,b}(x) = a_i^{(l)} = f\left(z_i^{(l)}\right) \quad (7)$$

The network is trained using training subjects $(x^{(i)}, y^{(i)})$ where $y^{(i)} \in \mathbb{R}^2$. The parameter x is a vector of input features of a sample and y is the outcome (in our case, a sample with T2D and a sample without T2D). With a fixed training set $\left\{\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(m)}, y^{(m)}\right)\right\}$ of m training examples, the neural network can be trained using gradient descent, and the overall cost function is defined as [19], [20]:

$$\begin{aligned} J(W, b) &= \left[-\frac{1}{m}\sum_{i=1}^{m} J\left(W, b; x^{(i)}, y^{(i)}\right)\right] \\ &+ \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2 \\ &= \left[-\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log h_{W,b}\left(x^{(i)}\right) + \right. \\ &\left. \left(1 - y^{(i)}\right) \log\left(1 - h_{W,b}\left(x^{(i)}\right)\right)\right] \\ &+ \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2 \end{aligned} \quad (8)$$

where the first term is a cross-entropy error function and the second term is a regularisation term which is also known as a weight decay term. The weight decay term helps to reduce the magnitude of the weights and prevents overfitting. Parameter λ is the weight decay parameter and, it controls the relative importance of the first and second terms.

Before training the neural network model, random initialisation of the parameter $W_{ji}^{(l)}$ and each $b_i^{(l)}$ are set to a value close to zero. This step stops the hidden layer units learning the same function of the input. The gradient descent updates parameters W, b as defined below:

$$W_{ij}^{(l)} := W_{ij}^{(l)} - \alpha \frac{\partial}{\partial W_{ij}^{(l)}} J(W, b) \quad (9)$$

$$b_i^{(l)} := b_i^{(l)} - \alpha \frac{\partial}{\partial b_i^{(l)}} J(W, b)$$

where α represents the learning rate.

The partial derivatives $\frac{\partial}{\partial W_{ij}^{(l)}} J(W, b; x, y)$ and $\frac{\partial}{\partial b_i^{(l)}} J(W, b; x, y)$ of the cost function $J(W, b; x, y)$ for a single example (x, y) are computed using the backpropagation learning process.

The backpropagation algorithm first performs a feedforward pass to compute all the activations $a_i^{(l)}$ and the output value of $h_{W,b}(x)$ in the network. An error term is calculated for each node i in layer l then this error term is propagated backward to the previous layers through the network to adjust the weights for each node i in layer l. Finally, the gradient descent is applied to minimise the overall cost function $J(W, b)$.

Momentum training and learning rate annealing are advanced optimisation tuning parameters that are used to modify backpropagation to allow previous iterations to influence the current version. The velocity vector is defined as follows:

$$v_{t+1} = \mu v_t - \alpha \nabla L(\theta_t)$$
$$\theta_{t+1} = \theta_t + v_{t+1} \quad (10)$$

where θ denotes the parameters W and b. The momentum coefficient is represented by μ, and the learning rate is α. The nesterov accelerated gradient method is used with momentum updates, and it modifies the updates as follows:

$$v_{t+1} = \mu v_t - \alpha \nabla L(\theta_t + \mu v_t)$$
$$W_{t+1} = W_t + v_{t+1} \quad (11)$$

E. Performance Measures

The performance of the proposed DL algorithm is measured using the Area Under the Curve (AUC), Sensitivity, Specificity, Logarithmic Loss, Gini, and MSE values. The dataset is split randomly into training (80%) to train the models, validation (10%) and testing (10%) to evaluate model performance on unseen data.

Sensitivity and specificity are utilised to measure the positive and negative predictive capabilities of classifiers in binary classification. Sensitivity refers to the true positive rate, which describes the ability of the test to classify people correctly with T2D. While Specificity describes the true negative rate, which is the ability of the test to classify people correctly without T2D.

Furthermore, in this analysis the area under the curve (AUC) and the receiver operating characteristic curve (ROC curve) is used to assess and compare classifiers performance, these are both widely used evaluation techniques for binary classification studies [25]. The AUC value represents the probability of correct classification for positive and negative instances; the positive class will be ranked higher thus a higher AUC means a better classification [25]. While the ROC curve is a graphical plot to display the performance of a binary classification model. It is created by plotting the true positive rate (also known as sensitivity) against the false positive rate which can be represented as (1-specificity) [25].

The Gini coefficient can be derived from Area Under the ROC curve (AUC) $Gini = 2 * AUC - 1$. It represents the area between the ROC curve and the diagonal. The Gini coefficient is usually used in binary classification problems, Gini value above 60% is considered a good model.

Logarithmic Loss (Logloss) is a classification loss function often used to measure the performance of a classification model where the prediction input is a probability value between 0 and 1. Logloss increases as the predicted probability (accuracy) decrease. A Logloss value of 0 is an indication of a perfect model where the model correctly classifies all class instances.

The Mean Squared Error (MSE) performance metric is used to measure the average of the squares of the errors which is the difference between the actual values and the predicted values. An MSE value close to 0 means that the model correctly classifies all class instances.

F. Machine Learning Model Parameters

A deep learning classifier is used for the binary classification of T2D using a various subset of features and is benchmarked with random forest (RF) classifier model using the same set of features. The number of trees is set to 400 with a maximum tree depth of 40 to train RF models. For DL models, a RectifierWithDropout activation function is used with input dropout ratio set to 0.1 and hidden dropout ratios for each layer set to 0.5. Epochs are set to 100 iterations for models using features extracted based on 10^{-2}, 10^{-3}, 10^{-4} thresholds, while 10 epochs are specified for models based on 10^{-5}, 10^{-6}, 5×10^{-8}. Four hidden layers with 10 neurons are used for models using extracted features based on 10^{-2}, 10^{-3}. While the remaining models (10^{-4}, 10^{-5}, 10^{-6}, 5×10^{-8}) utilise two hidden layers with 10 neurons. Early stopping is adopted using a stopping metric set to logloss and a stopping tolerance and stopping rounds coefficient set to 1×10^{-2} and 5 respectively. The learning rate and momentum are experimentally detected. The learning rate is configured to 0.005 with rate annealing, and rate decay set to 1×10^{-6} and 1 respectively. Momentum start is set to 0.5 with momentum stable sets to 0 and momentum ramp to 1×10^{6}. The max w2 coefficient is set to 10. The implementation, evaluation, and visualization of the predictive classification models were performed using H2O package in R software.

III. RESULTS

This section presents the classification results for T2D obtained using the DL and RF classifiers. RF is used to benchmark the performance. Several association analysis thresholds are considered including 5×10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} resulting in 7 SNPs, 13 SNPs, 23 SNPs, 103 SNPs, 766 SNPs, 6609 SNPs respectively.

Table III illustrates the performance metrics for the DL classifier for the validation set. Metric values for 5×10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} were obtained using an optimized F1 threshold with values 0.4601, 0.4728, 0.4452, 0.3586, 0.5749, 0.3959 respectively.

TABLE III. PERFORMANCE METRICS FOR THE DL VALIDATION SET

p-value	AUC	Sens	Spec	Logloss	Gini	MSE
10^{-2}	0.9591	0.9636	0.8589	0.3153	0.9183	0.0931
10^{-3}	0.8640	0.8581	0.7307	0.5411	0.7281	0.1781
10^{-4}	0.6691	0.9127	0.2863	0.6751	0.3382	0.2412
10^{-5}	0.6073	0.9818	0.0940	0.6766	0.2146	0.2418
10^{-6}	0.6037	0.9927	0.0982	0.6779	0.2074	0.2424
5×10^{-8}	0.5794	0.9818	0.0427	0.6825	0.1588	0.2447

Table IV provides the performance metrics for the DL for the test set. Metric values for 5×10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} were gained using an optimized F1 threshold with values 0.4512, 0.4665, 0.4476, 0.3378, 0.5293, 0.5102 respectively. Comparatively the predictive results are lower than those obtained using the validation set.

The classification accuracy of the DL classifier model shows significant improvement with values ranging between 57.94% for 5×10^{-8} and 95.91% for 10^{-2} in the validation set. This is also the case for the test set with values of 52.58% and 96.53% for 5×10^{-8} and 10^{-2} p-value thresholds, respectively. Sensitivity and specificity metrics for the DL validation and test sets are imbalanced for lower p-value thresholds. However, this is not the case with higher thresholds 10^{-2} where the number of SNPs increase.

TABLE IV. PERFORMANCE METRICS FOR THE DL TEST SET

p-value	AUC	Sens	Spec	Logloss	Gini	MSE
10^{-2}	0.9653	0.9391	0.9083	0.3233	0.9306	0.0950
10^{-3}	0.8233	0.875	0.6030	0.5658	0.6466	0.1897
10^{-4}	0.6391	0.9425	0.1908	0.6826	0.2782	0.2448
10^{-5}	0.6025	0.9527	0.1259	0.6791	0.2050	0.2430
10^{-6}	0.5744	0.9831	0.0572	0.6841	0.1489	0.2455
5×10^{-8}	0.5258	1.0000	0.0038	0.6908	0.0517	0.2488

Table V shows the performance metrics of RF for the validation set. Metric values for 5×10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} were obtained using an optimized F1 threshold with values 0.2863, 0.1686, 0.1417, 0.46, 0.48, 0.5 respectively.

TABLE V. PERFORMANCE METRICS FOR THE RF VALIDATION SET

p-value	AUC	Sens	Spec	Logloss	Gini	MSE
10^{-2}	0.7438	0.9163	0.2863	0.6519	0.4872	0.2297
10^{-3}	0.7034	0.8909	0.2863	0.6527	0.4068	0.2301
10^{-4}	0.6506	0.8727	0.3290	0.6567	0.3012	0.2322
10^{-5}	0.5532	0.9890	0.0427	0.7318	0.1064	0.2639
10^{-6}	0.5586	0.9963	0.0128	0.7149	0.1172	0.2557
5×10^{-8}	0.5713	0.9963	0.0085	0.6879	0.1426	0.2468

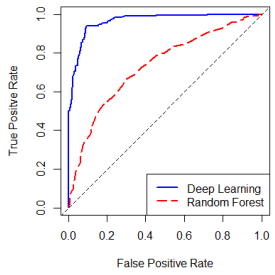
Table VI presents the performance metrics for the RF test set. Metric values for 5×10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} were obtained using an optimized F1 threshold with values 0.2687, 0.045, 0.056, 0.395, 0.455, 0.515 respectively. The classification accuracy and the corresponding performance metrics partially show in some cases, less values than those obtained by validation set.

The classification prediction accuracy for the RF classifier model shows 17.25% improvement for the validation test and 19.67% improvement in the test set when increasing the threshold values from 5×10^{-8} to 10^{-2}. Sensitivity and specificity metrics for the RF validation and test sets are imbalanced for all p-value thresholds.

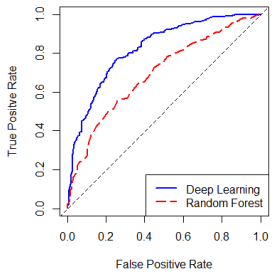
TABLE VI. PERFORMANCE METRICS FOR THE RF TEST SET

p-value	AUC	Sens	Spec	Logloss	Gini	MSE
10^{-2}	0.7324	0.8310	0.4656	0.6553	0.4649	0.2313
10^{-3}	0.6942	0.9560	0.1488	0.6549	0.3884	0.2313
10^{-4}	0.6334	0.9594	0.1526	0.6625	0.2669	0.2351
10^{-5}	0.5715	1.0000	0.0038	0.7349	0.1430	0.2631
10^{-6}	0.5507	1.0000	0.0000	0.7248	0.1015	0.2601
5×10^{-8}	0.5357	1.0000	0.0076	0.6947	0.0714	0.2506

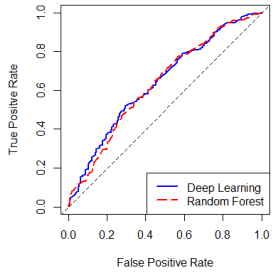
Fig.1 presents the ROC curves for DL and RF classifiers. The performance for the DL and RF is relatively similar using 5×10^{-8}, 10^{-6}, 10^{-5}, and 10^{-4}. However, the DL classifier model outperforms the RF model using 10^{-3} and 10^{-2} thresholds.



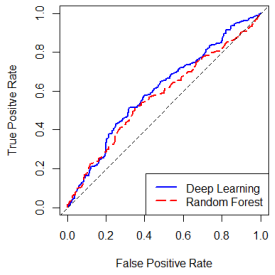
(a) ROC for 10^{-2}



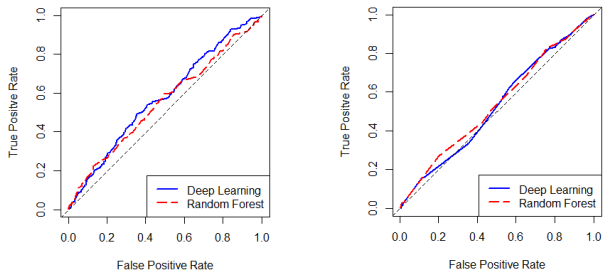
(b) ROC for 10^{-3}



(c) ROC for 10^{-4}



(d) ROC for 10^{-5}



(e) ROC for 10^{-6} (f) ROC for 5×10^{-8}

Fig. 1. From (a) to (f) performance ROC curves for DL and RF test sets using p-value threshold between 5×10^{-8} and 10^{-2}

IV. Discussion

The genetic constructs for complex human diseases are complicated [26]. Instead of a single allele or single gene, such complex disorders stem from the interactions and contributions of multiple genes. Although GWAS have identified genetic variants that show increased susceptibility to many complex disorders, single-SNP omits such interactions among genetic variants. In the area of bioinformatics where large scale and complex biological data structures exist, researches have focused on the use of traditional machine learning algorithms such as random forest [11], support vector machine [27], [28] to perform multi-SNPs analysis. Using Statistically significant or suggestive SNPs, the models fail to classify phenotype, and this is often caused by the fact that most of these SNPs are false positives. The prediction capacity of disease risk based on highly ranked SNPs demonstrates little predictive power [29]. Therefore, it may be beneficial to raise the number of SNPs utilised in the classification analysis, as shown in [30] and [31] to enhance performance. In this paper, we presented a robust methodology for high-dimensional GWAS data through the application of DL. DL is capable of discovering sophisticated and complex structures in high-dimensional data through its characteristics of feature representation-learning. This allows latent representations in data to be extracted to discover interactions between SNPs.

Using a DL classifier model, we investigated and evaluated classification capacity in distinguishing between cases and controls in T2D genomic data using different features size configurations. The best result obtained with a p-value threshold of 10^{-2} (6609 SNPs) (AUC=96.53%, Sens=93.91%, Spec=90.83%, Logloss= 32.33%, Gini=93.06%, MSE=9.50%). However, a clear deterioration in performance is evident when the p-value threshold is decreased. Using Bonferroni genome-wide significance threshold of 5×10^{-8} (7 SNPs) attained the worst results (AUC=52.58%, Sens=100%, Spec=0.38%, Logloss= 69.08%, Gini=5.17%, MSE=24.88%). It is clear a much higher predictive accuracy is obtained by increasing the number of SNPs. Sensitivities and specificities for DL models are imbalanced for all p-value thresholds excluding 10^{-2} with 93.91% for sensitivity and 90.83% for specificity. Although the results for p-value = 10^{-3} (766) are encouraging given that a considerably smaller number of SNPs are required. The reason for this is because the algorithm has the ability to transform big data into abstract representations using the concept of hierarchical explanatory for automatic feature learning.

The RF algorithm is used as a baseline model to compare the classification performance obtained by the DL model. RF is a method that has been successfully used in genetic studies [9], [10]. In this analysis, the results show that the best classification accuracy (73.24%) was achieved with a p-value threshold of 10^{-2} (6609 SNPs). In general sensitivities and specificities were instable for all thresholds used in the analysis. This indicates that the RF classifier has low discriminatory capacity for this given dataset when separating case/control observations. More importantly, the RF classifier showed significantly lower results than the DL using the 10^{-2} threshold. This is likely caused by the fact RF models find it difficult to process high dimensionality data as is the case in this study (5393 samples, 6609 SNPs) [17].

The results in this paper outperform most previous studies in the area of T2D GWAS data classification. Table VII lists previous works. The closest being Kim *et al.* [32] who in their study used the same dataset (NHS-HPFS) and classifier model (DL) as used in this study. They utilised different genetic association mappings and a different set of features. They achieved (AUC=93.1%) while our study obtained (AUC=96.53%, Sens=93.91%, Spec=90.83%, Logloss= 32.33%, Gini=93.06%, MSE=9.50%).

The Findings in this paper are encouraging. Using a DL model to classify T2D GWAS data could provide a starting point for researchers and professionals investigating the aetiology of disease. It could lead to improved diagnostic testing, the prevention of the disease onset, and advances in personalised medicine. Furthermore, it may be possible to mitigate the progression of the disease and its complications.

TABLE VII. Previous Works Predictive Metrics of T2D

Paper	Year	Classifier	AUC	Sens	Spec
Ban *et al.* [27]	2010	SVM	0.653	0.567	0.739
López *et al.* [10]	2018	RF LR* SVM	0.853 0.835 0.825		
Botta *et al.* [9]	2014	RF TT**	0.758 0.834		
Malovini *et al.* [33]	2012	HNB***	0.92	0.89	0.93
Kim *et al.* [32]	2018	DL	0.931		
Gul *et al.* [31]	2014	LR	0.965		
Proposed Method	2020	DL	0.9653	0.9391	0.9083

*Logistic Regression
**T-Trees
***Hierarchical Naïve Bayes

V. Conclusion

In this paper, we presented a robust framework for the classification of T2D genetic data. We investigated the potential use of DL for the classification of GWAS. This study utilised existing datasets obtained from the Genotypes and Phenotypes (dbGap) database. Several stringent QC assessment steps followed by logistic regression association analysis adjusted GC was performed for single-SNP analysis. Using 5393 samples of T2D case-control and 6609 SNPs for our classification analysis we achieved (AUC=96.53%,

Sens=93.91%, Spec=90.83%, Logloss= 32.33%, Gini=93.06%, MSE=9.50%). Reducing the number of SNPs, noticeable deterioration in performance was observed with classification accuracies results of 82.33%, 63.91%, 60.25% 57.44%, and 52.58% for 766 SNPs, 103 SNPs, 23 SNPs, 13 SNPs, and 7 SNPs respectively.

Although this paper has presented encouraging results, a more in-depth investigation is still required to further analyse the results. Furthermore, unsupervised deep learning stacked autoencoders are widely utilised to learn the compressed representation of the data input in many domains, and this would be worth considering further [34], [35]. For example, stacked autoencoders have been applied successfully to learn the abstract representation of SNP data and to study epistatic interactions between SNPs for the classification of preterm birth in African-American woman [36].

In future work, we will consider deep learning stacked autoencoder [37] to learn epistatic interactions between SNPs in large-scale GWAS data. This compressed representation of SNP data will be used to evaluate the predictive capacity of SNPs by classifying samples as either case or control in the T2D dataset.

Overall, the proposed methodology is robust and contributes to the bioinformatics research field and computational biology, and provides new insights into the potential use of DL algorithms when analysing high-dimensional GWAS data that we believe warrants further investigation.

ACKNOWLEDGEMENT

The dataset(s) used for the analyses described in this manuscript were obtained from the database of Genotype and Phenotype (dbGaP) found at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000091.v2.p1. The NHS and HPFS is part of the Gene Environment Association Studies initiative (GENEVA, http://www.genevastudy.org) funded by the trans-NIH Genes, Environment, and Health Initiative (GEI).

REFERENCES

[1] World Health Organization, "Global Report on Diabetes," 2016.

[2] C. D. Mathers and D. Loncar, "Projections of Global Mortality and Burden of Disease from 2002 to 2030," PLoS Med., vol. 3, no. 11, p. e442, Nov. 2006.

[3] V. Lyssenko et al., "Clinical Risk Factors, DNA Variants, and the Development of Type 2 Diabetes," N. Engl. J. Med., vol. 359, no. 21, pp. 2220–2232, Nov. 2008.

[4] F. Medici, M. Hawa, A. Ianari, D. A. Pyke, and R. D. G. Leslie, "Concordance rate for Type II diabetes mellitus in monozygotic twins: actuarial analysis," Diabetologia, vol. 42, no. 2, pp. 146–150, Jan. 1999.

[5] S. Behjati and P. S. Tarpey, "What is next generation sequencing?," Arch. Dis. Child. Educ. Pract. Ed., vol. 98, no. 6, pp. 236–238, 2013.

[6] X. Guo, N. Yu, F. Gu, X. Ding, J. Wang, and Y. Pan, "Genome-Wide Interaction-Based Association of human diseases - A survey," Tsinghua Sci. Technol., vol. 19, no. 6, pp. 596–616, Dec. 2014.

[7] G. M. Clarke, C. A. Anderson, F. H. Pettersson, L. R. Cardon, A. P. Morris, and K. T. Zondervan, "Basic statistical analysis in genetic case-control studies," Nat. Protoc., vol. 6, no. 2, pp. 121–133, Feb. 2011.

[8] M. R. Robinson, N. R. Wray, and P. M. Visscher, "Explaining additional genetic variation in complex traits," Trends Genet., vol. 30, no. 4, pp. 124–132, Apr. 2014.

[9] V. Botta, G. Louppe, P. Geurts, and L. Wehenkel, "Exploiting SNP Correlations within Random Forest for Genome-Wide Association Studies," PLoS One, vol. 9, no. 4, p. e93379, Apr. 2014.

[10] B. López, F. Torrent-Fontbona, R. Viñas, and J. M. Fernández-Real, "Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction," Artif. Intell. Med., vol. 85, pp. 43–49, Apr. 2018.

[11] T.-T. Nguyen, J. Huang, Q. Wu, T. Nguyen, and M. Li, "Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests," BMC Genomics, vol. 16, no. Suppl 2, 2015.

[12] C. L. Koo, M. J. Liew, M. S. Mohamad, and A. H. Mohamed Salleh, "A Review for Detecting Gene-Gene Interactions Using Machine Learning Methods in Genetic Epidemiology," Biomed Res. Int., vol. 2013, no. 432375, 2013.

[13] A. Krizhevsky, I. Sutskever, and H. Geoffrey E., "ImageNet Classification with Deep Convolutional Neural Networks," in NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, 2012, pp. 1097–1105.

[14] G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, Nov. 2012.

[15] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (Almost) from Scratch," J. Mach. Learn. Res., vol. 12, pp. 2493–2537, 2011.

[16] S. Ekins, "The Next Era: Deep Learning in Pharmaceutical Research," Pharm. Res., vol. 33, no. 11, pp. 2594–2603, Nov. 2016.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, May 2015.

[18] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," Brief. Bioinform., vol. 18, no. 5, pp. 851–869, Jul. 2017.

[19] A. Ng, "Sparse autoencoder," CS294A Lect. notes, pp. 1–19, 2011.

[20] A. Candel, V. Parmar, E. LeDell, and A. Arora, "Deep Learning With H2O," 2018.

[21] K. A. Tryka et al., "NCBI's Database of Genotypes and Phenotypes: dbGaP," Nucleic Acids Res., vol. 42, pp. D975–D979, Jan. 2014.

[22] S. Purcell et al., "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," Am. J. Hum. Genet., vol. 81, no. 3, pp. 559–575, Sep. 2007.

[23] C. a Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan, "Data quality control in genetic case-control association studies," Nat. Protoc., vol. 5, no. 9, pp. 1564–1573, Sep. 2010.

[24] X. Wang, C. Baumgartner, D. C.Shields, H.-W. Deng, and J. S.Beckmann, Application of Clinical Bioinformatics, vol. 11. Dordrecht: Springer Netherlands, 2016.

[25] D. J. Hand, "Measuring classifier performance: a coherent alternative to the area under the ROC curve," Mach. Learn., vol. 77, no. 1, pp. 103–123, Oct. 2009.

[26] K. J. Mitchell, "What is complex about complex disorders?," Genome Biol., vol. 13, no. 237, 2012.

[27] H.-J. Ban, J. Y. Heo, K.-S. Oh, and K.-J. Park, "Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine," BMC Genet., vol. 11, no. 26, 2010.

[28] C. J. C. Tello, D. Hernández-Ramírez, and C. A. García-Sepúlveda, "Support vector machine algorithms in the search of KIR gene associations with disease," Comput. Biol. Med., vol. 43, no. 12, pp. 2053–2062, 2013.

[29] F. Mittag et al., "Use of support vector machines for disease risk prediction in genome-wide association studies: Concerns and opportunities," Hum. Mutat., vol. 33, no. 12, pp. 1708–1718, Dec. 2012.

[30] Z. Wei et al., "From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes," PLoS Genet., vol. 5, no. 10, p. e1000678, Oct. 2009.

[31] H. GÜL, Y. AYDIN SON, and C. AÇIKEL, "Discovering missing heritability and early risk prediction for type 2 diabetes: a new perspective for genome-wide association study analysis with the Nurses' Health Study

and the Health Professionals' Follow-Up Study," Turkish J. Med. Sci., vol. 44, no. 6, pp. 946–954, 2014.

[32] J. Kim, J. Kim, M. J. Kwak, and M. Bajaj, "Genetic prediction of type 2 diabetes using deep neural network," Clin. Genet., vol. 93, no. 4, pp. 822–829, Apr. 2018.

[33] A. Malovini, N. Barbarini, R. Bellazzi, and F. De Michelis, "Hierarchical Naive Bayes for genetic association studies," BMC Bioinformatics, vol. 13, no. S14, Sep. 2012.

[34] L. Vařeka and P. Mautner, "Stacked Autoencoders for the P300 Component Detection," Front. Neurosci., vol. 11, no. 302, May 2017.

[35] L. Deng, C. Fan, and Z. Zeng, "A sparse autoencoder-based deep neural network for protein solvent accessibility and contact number prediction," BMC Bioinformatics, vol. 18, no. 569, Dec. 2017.

[36] P. Fergus, A. Montanez, B. Abdulaimma, P. Lisboa, C. Chalmers, and B. Pineles, "Utilising Deep Learning and Genome Wide Association Studies for Epistatic-Driven Preterm Birth Classification in African-American Women," IEEE/ACM Trans. Comput. Biol. Bioinforma., 2018.

[37] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, Aug. 2013.