Backpropagation Approach Supported by Image Compression Algorithm for the Classification of Chronic Condition Diseases

Abir J. Hussain, *Member, IEEE*, Dhiya Al-Jumeily, *Senior Member, IEEE, Ahmed J. Aljaaf* and Naeem Radi. *Senior Member, IEEE*

Abstract—Diabetes is one of the main public health chronic conditions that are potentially reaching epidemic proportions globally. Worldwide, the occurrence of these types of diseases are increasing sharply at a worrying degree, with death of around 18 million people every year from cardiovascular disease, for which diabetes and hypertension are major predisposing factors. Two major concerns are that much of this increase in Diabetes is predicated to be happened in developing countries, with a growing incidence of Type 2 Diabetes (T2D) at a younger age including some obese children even before puberty. However, in developed countries most people with diabetes are above the age of retirement. As such, understanding the aetiology of T2D is vital. It has been thought that T2D is resulting from the convergence of genetics, environment, diet and lifestyle risk factors; however, genetic susceptibility has been established as a key component of risk. Genome-wide association studies (GWAS) is a study design and analytic tool specifically developed for investigating the genetic architecture of human disease. The ultimate aim of GWAS is to identify the genetic risk factors for common complex diseases such as T2D. Traditional parametric statistical approaches such as linear modelling framework (e.g. logistic regression) have limited power for modelling the complexity of genotype-phenotype relationship that is characterized by non-linear interactions. These nonlinear interactions are necessary in discovering the aetiology of complex diseases. More specifically, the linear modelling model has some limitations such as examining each single nucleotide polymorphisms independently for the association to the phenotype ignoring the epistatic (gene-gene interactions) and non-genetics factors. This paper presents a novel approch based on the use of backpropogation technique inspired by image compression algorithm. The proposed classifier is fine-tuned for binary classification to predict those who could suffer from the disease among those who do not. Simulation results indicated that the proposed technique showed an area under the curve, true positive rate, true negative rate values of 0.92, 0.9 and 0.8 respectively when using 2500 hidden neurons.

Index Terms— Backpropagation, GWAS study, hierarchical neural networks, SNPs, Artificial Intelligence, genome, Type 2 Diabetes, logistic regression.

1 Introduction

Diabetes has emerged as one of the main alarms to human health in this century.

Type 2 Diabetes (T2D) is a multifactorial disorder and is the result of the complex interaction between genetic, environment and sedentary lifestyle [1]. T2D remains the leading cause of a serious long term health complications. It is responsible for most cases of blindness (Diabetic retinopathy), kidney failure and lower limb amputation. Moreover, high glucose levels (raised blood sugar) or Hyperglycaemia in the bloodstream can damage blood vessels which increases the likelihood of atherosclerosis (cardiovascular disease) and stroke cases and can cause nerve damage [2].

Until recently, T2D was recognised only in people who are over the age of 40 but currently children also are being diagnosed with this disease [3]. According to World Health Organisation (WHO)[1], Diabetes is one of a leading cause of death (2.7%) worldwide. In 2012, WHO[1] revealed that diabetes killed 1.5 million people in the world.

Beyond the human suffering, in the UK, the annual cost of T2D to the National Health Service is approximately £8.8 billion for direct cost which includes diagnosis, lifestyle interventions, management, complications, and ongoing treatment. In addition to that, there is an indirect cost that was estimated to be £13 billion, this includes mortality, sickness, reduced productivity among people who remain in work and informal care [4], [5]. Additionally, the International Diabetes Federation (IDF)[2] reported that approximately 12% of global heath expenditure is spent on diabetic people.

The identification of genetic markers that show evi-

- *A. J. Hussain, and D. Al-Jumeily, with the Department of Computer Science, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, UK. E-mail: {A.Hussain , D.AlJumeily}@ljmu.ac.uk.*
- *A. J. Aljaaf is with the University of Anbar, Ramadi, Iraq. E-mail: a.j.aljaaf@uoanbar.edu.iq*
- *N. Radi is with Al Khawarizmi International College, Dhabi, United Arab Emirates. E-mail: n.radi@khawarizmi.com.*

[1] http://www.who.int/

[2] http://www.idf.org/about-diabetes/facts-figures

dence of increase susceptibility to T2D and related traits will facilitate the translation of this genetic information to the clinical practice and may open up the opportunity to improve risk prediction of the disease and enable delay or prevention of disease onset and to reduce expenditures of cares.

It has been thought that complex diseases such as T2D involve multiple genetics with their interactions [6], [7]. Moreover, these genetics factors do not act independently but also interact with other factors such as environment, sociodemographic and clinical factors. In [8] suggested that the traditional parametric statistical approaches such as linear modelling (e.g. logistic regression) have limited power for modelling the complexity of genotype-phenotype relationship that is characterised by non-linear interactions. These nonlinear interactions are necessary in discovering the aetiology of complex diseases. More specifically, the linear modelling approach has some limitations such as examining each Single Nucleotide Polymorphisms (SNP) independently for the association to the phenotype ignoring the epistatic and non-genetics factors. The subset of SNPs that should be included in the analysis requires to be evaluated among a list of thousands or probably millions of candidates SNPs using advanced techniques such as filtering algorithms or wrapper algorithms [8], [9]. Consequently, these challenges of traditional approaches have led to search for alternative methods such as Artificial Intelligence (AI) methods and techniques. AI has already been successfully applied to a wide range of medical applications trying to contribute to the prediction of risk susceptibility to T2D. In [10], researchers considered generalised multifactor dimensionality reduction (GMDR) approach for detecting Gene-gene interaction. The study identified 24 core SNPs that appear to be important to T2D. Another study [7] also investigated gene-gene interaction using lasso-multiple regression approach. Researchers found that the SNPs from genes CDKN2BAS and KCNJ11 are significantly associated to T2D. Random Forest for GWAS has been implemented in [11] for exploiting SNP correlations. Moreover, support vector machine (SVM) has been proposed in [12] to investigate risk assessment related to specific traits. Furthermore, random forest [13], support vector machine [13], artificial neural network [14] were used to model complex relationships and interactions between features SNPs and their association to the phenotype.

This paper proposes the use of backpropagation and hierarchical neural networks for the detection of T2D Genome Wide Association Study (GWAS) data. The model is inspired by image compression hieratical neural network system for the utilisation of compressed SNPs to analyse nonlinear multilayered interactions of large numbers of genetic variants.

Examining the literture review, it is should be noted that, this paper presents a novel approach of using back-propagation algorithm inspired by image compression hierarchical neural network for the classification of type 2 diabetes GWAS analysis.

The remainder of this paper is organised as follows: The background study and related works are illustrated in Section 2. The network structure is described in Section 3, while the simulations results are presented in Section 4. Sections 5 and 5 demonstrate the discussion and conclusion of this paper, respectively.

2 BACKGROUND AND RELATED WORKS

2.1 Genome-Wide Association Studies

Genome-wide association studies (GWAS) is a study design and analytic tool specifically developed for investigating the genetic architecture of human disease [21]. The ultimate aim of GWAS is to identify the genetic risk factors for common complex and chronical diseases such as T2D, Schizophrenia, Epilepsy, Obesity, Cardiovascular Disease, and Hypertension [21], [15]. GWAS aims to find SNPs that occur frequently in individuals who are affected with a disease, than individuals that are unaffected with the disease.

The availability of genotyping technology has facilitated rapid progress in genome wide association studies. These genotyping technology specifically designed for assaying more than one million SNPs for example to sequence the entire human genome within a single day [22]. Recent DNA sequencing technology is Next Generation Sequencing (NGS), which is developed to provide tools to sequence DNA and RNA and to enable cost effective and rapid performance of sequencing genome in comparison to the previous one, Sanger sequencing [22], [23], [24]. There are two essential platforms consistently utilised for GWAS, including Illumina and Affymetrix platforms [21], [29]. Each of these techniques has offered different approach to measure and detect genomic variation (alleles).

Although, GWAS have significant impact on the area of human genetics, there are still challenges associated to computational and statistical methods causing issues when conducting such approach. These challenges include scalability, missing markers and complex traits [25]. GWAS datasets contains millions of SNPs with thousands of individuals therefore to perform GWAS, the algorithms should be extremely scalable to avoid consume huge amount of computational resources and to reduce the time that is used to conduct GWAS. In addition to that, to handle missing markers that is generated due to the absent of genetic variants availability as there are still many genetic markers that are not genotyped. One of the popular approach to handle missing markers is imputation method [26], which simply means to impute the unidentified markers by using the accessible SNPs databases such as the 1000 Genome Projects [27] and International HapMap Project [28]. One of the GWAS successful features is to detect a single gene related phenotype traits. However, this approach may not be successful in finding SNPs associated with complex traits/diseases such as T2D. As complex traits are more likely to be affected by multiple genes rather than a single gene, and each gene of these multiple genes separately may have a weak association with the disease as such it would be extremely difficult for a SNP with low marginal effects to be identified using single-locus methods. Hence, alternative approach such as multi-locus analysis needs to be conducted [21].

Complex diseases recognised to have a genetic component and they do not follow a simple pattern of inheritance and therefore they could not be explained or analysed based on the inheritance patterns of single gene diseases. As such, association analysis has been considered and it is applied on a case/control dataset that consist of a large number of unrelated samples, to detect genetic markers that are more frequently appears in cases (affected) rather than in controls(unaffected). This highlights the common disease - common variant hypothesis [29] indicating that common diseases are probably influenced by genetic markers that are relatively common in the population. Under this hypothesis, phenotype associated alleles are more likely established through using common genetic markers specifically SNPs that have been detected to compare between affected and unaffected samples. However, other researchers do not agree with this hypothesis as they suggested that common diseases would not be possible to be caused by common alleles and rather they would be influenced by rare variants [30], [33].

2.2 The Nurses' Health Study and the Health Professionals Follow-up Study data sets

The Nurses' Health Study (NHS) cohort and the Health Professionals Follow-up Study (HPFS) are used in this study, which are provided by the Genotypes and Phenotypes (dbGap) database [16].

The NHS was established in 1976. Participants were 121,700 female registered nurses between age 30 to 55 and residing in 11 U.S states. All nurses responded to mailed questionnaire requesting information related to their medical history and lifestyle characteristics. Since then, the Nurses have been requested twice a year to fill questionnaire and attain updated information (for instance information on newly diagnosed illness) [45]. All NHS members were requested to provide blood samples, in which 32,826 members responded. For T2D study, the cases and controls participants were selected from those who provided a blood sample. Cases participants were identified as those who have reported themselves to be affected by T2D and it was confirmed by a medical record validation questionnaire. Controls participants were defined as those without diabetes. The NHS participants consist of 1581 T2D cases and 1854 controls.

For the HPFS data set, a nested case-control study was completed in which data such as age and other clinical factors were collected through questionnaire before the blood draw in 1990. In this case, the diabetes status was self-reported and then confirmed with questionnaire to include 1338 control 1164 T2D cases.

Diabetes was diagnosed by the National Diabetes Data Group criteria before 1998 and the American Diabetes Association criteria in 1998-2002 [17], [18] [19].

PLINK v1.07 and v1.9 [20] for Windows are used to conduct data quality control (QC) and preliminary analysis. PLINK is a whole genome data analysis toolset which is developed for handling SNP data. PLINK data contains two files which include information associated to genetic data for each participant in the study as well as recording information related to participant's phenotype features such as affected (case sample) and unaffected (control sample).

The NHS and HPFS datasets are merged using PLINK (NHS and HPFS participants were genotyped using the Affymetrix Genome-Wide Human 6.0 array). Pre-established quality control protocols was performed [33].

Individuals QC: Discordant sex information (homozygosity rate between 0.2 and 0.8) were found in which 14 samples were removed from the dataset. Individuals with elevated missing data rates (genotype failure rate ≥ 0.05) and outlying heterozygosity rate (heterozygosity rate ± 3 standard deviations from the mean) were identified and 131 individuals are discarded from the analysis. Identity-by-descent (IBD) was estimated to remove duplicated or related individuals (IBD > 0.185). This resulted in eight individuals being excluded from the dataset. Individuals with divergent ancestry were identified using the 2nd principal component score < 0.061 resulting in 51 individuals being removed. 101 individuals were removed due to missing genotype data rate of 0.05.

Genetic Markers QC: it should be noted that genetic markers (SNPs) were removed from the analysis when SNPs with excessive missing data rates were identified resulting in this case with the exclusion of 29 SNPs. Another, 116863 variants with missing genotype rate of 0.01 and 178004 variants with minor allele frequency (MAF) $<$ 0.05 were eliminated. Furthermore, 2248 variants were removed due to Hardy-Weinberg Equilibrium (HWE) with p-value < 0.001 in control samples. Following the QC steps, there were 608342 markers with 0.961665 genotype rate in remaining samples.

2.3 Analysis

Logistic regression is performed to look at the association of all SNPs within our dataset for the binary classification to distinguish between case (value 0) and control subjects (value 1). In this case, standard case-control association analysis is performed to extract information from unrelated white racial subpopulation which has allowed us to provide information about the frequency of alleles or genotypes at genetic marker loci (SNP) to compare between cases and controls of the PLINKV.19 merged data of NHS and HPFS Datasets. Pearson's Chi-squared test (x^2) is used to test the null hypothesis (no association).

Logistic regression is utilised (as illustrated in Algorithm 1) to statistically analysis genetic model. A p-value threshold of 10^{-2} is considered resulting in 6609 SNPs.

Allelic association test has been performed to explore the association between single allele of the SNP and the disease trait specifically for T2D among our dataset. Manhattan plot has been used to visualize the results of the association as represented in Figure 1. In this plot each dot represents a single SNP and the x-axis corresponds to a chromosome location or number, while the y-axis is the negative of the log of observed p-value of the SNPs. The smallest p-value has the strongest associations and it will appeared at the uppermost in Manhattan plot.

Algorithm 1 Logistic Regression

```
Let $Y \in \{0,1\}$ a binary variable, 0 for control and 1 for status and Let $X \in \{0,1,2\}$ be a genotype at a particular SNP.
    Let 0, 1, 2 represent homozygous major allele $AA$, heterozygous allele $Aa$ and homozygous minor allele $aa$ respectively.
Y = 1, IFF $(X) = P(Y = 1|X)$
$logit(X) = \ln \frac{\theta(X)}{1-\theta(X)}$
$logit(X) \sim \beta_0 + \beta_1 X$
```

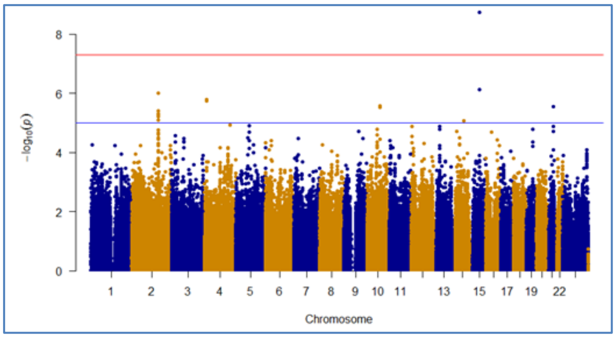


Fig.1. Manhattan Plot for Logistic Regression Analysis. Showing the SNPs that reached Bonferroni Level of Significant, Red Line.

3 Network Structures

3.1 Backpropagation Algorithm (BP)

In this work, backpropagation learning algorithm for neural networks is used for the classification of cases and control T2D genetic data. The algorithm is also referred to as error-backpropagation. This algorithm can be used to train multilayer feedforward neural network using gradient descent. The learning algorithm was developed by different researchers independently. It was investigated by Werbos in 1974 [31], then by Parker in 1982 [32] and rediscovered independently by Rumelhart, Hinton and Williams in 1986 [34]. An algorithm that is closely related was proposed by Le Cun [35].

Consider a neural network consisting of an output layer and a hidden layer. We refer to the input units by the index k, the output units by the index i, and the hidden neurons by the index j.

Let the number of inputs to be M, the number of outputs to be N, and the number of hidden units is S. Let y represents the N-tuple outputs of the output layer, and x is the M-tuple inputs to the network. There are two sets of matrices representing the weights of the network. The weights matrix that connects the input neurons to the hidden layer is represented by W_{jk}^1 and has $S \times M$ elements. While, the weights matrix that connects the hidden layer to the output layer is referred to by W_{ij}^2 and contains $N \times S$ elements. The biases of the network can either be represented separately in the neural network or by adding an extra input line of value one for each layer of the network.

Now, if an input pattern p is given to the network, the hidden neuron j receives a net input nj determined as follows:

$$n_j^p = \sum_{k=1}^{M} w_{jk}^1 x_k^p \tag{1}$$

and the output of this unit is:

$$V_j^p = f(n_j^p) = f(\sum_{k=1}^{M} w_{jk}^1 x_k^p) \tag{2}$$

where f is usually a nonlinear transfer function and has to be differentiable.

The output of the hidden layer is the input to the next layer which is the output layer in this case. Therefore, the net input to the output unit i is:

$$n_i^p = \sum_{j=1}^{S} w_{ij}^2 V_j^p = \sum_{j=1}^{S} w_{ij}^2 f(\sum_{k=1}^{M} w_{jk}^1 x_k^p) \tag{3}$$

and the output unit i produces the following output value:

$$y_i^p = f(n_i^p) = f(\sum_{j=1}^{S} w_{ij}^2 V_j^p) = f(\sum_{j=1}^{S} w_{ij}^2 f(\sum_{k=1}^{M} w_{jk}^1 x_k^p)) \tag{4}$$

Notice that, the transfer function at the output layer can be different to the transfer function in the hidden layer. However, for simplicity we will assume that they are the same in all the layers of the network.

Let d_i^p represent the target value of the output unit i when the input pattern p is represented to the input of the network. Then, the error produced at this unit is:

$$e_i^p = d_i^p - y_i^p \tag{5}$$

The overall network error or the cost function per pattern is described by the following equation:

$$J = \frac{1}{2}\sum_{i=1}^{N}[e_i^p]^2 = \frac{1}{2}\sum_{i=1}^{N}[d_i^p - y_i^p]^2 \tag{6}$$

By substituting the output y, the above equation becomes:

$$J = \frac{1}{2}\sum_{i=1}^{N}[d_i^p - f(\sum_{j=1}^{S} w_{ij}^2 f(\sum_{k=1}^{M} w_{jk}^1 x_k^p))]^2 \tag{7}$$

The change in the weight that connect a hidden unit to an output unit is given by the following equation:

$$\Delta w_{ij}^{2} = -\eta \frac{\partial J}{\partial w_{ij}^{2}} \tag{8}$$

where η is a positive real value representing the learning rate, and

$$\frac{\partial J}{\partial w_{ij}^{2}} = \frac{\partial}{\partial w_{ij}^{2}}\left(\frac{1}{2}\sum_{i=1}^{N}[d_i^p - y_i^p]^2\right) = \frac{\partial}{\partial w_{ij}^{2}}\left(\frac{1}{2}\sum_{i=1}^{N}[d_i^p - f(\sum_{j=1}^{S} w_{ij}^{2} f(\sum_{k=1}^{M} w_{jk}^{1} x_k^p))]^2\right)$$
$$= -(\mathrm{d_i^p} - y_i^p) f'(n_i^p) V_j^p$$

Therefore, the change in the weight of the output unit i is:

$$\Delta w_{ij}^{2} = \eta e_i^p\, f'(n_i^p) V_j^p \tag{9}$$

Let

$$\delta_i^p = e_i^p\, f'(n_i^p) \tag{10}$$

then

$$\Delta w_{ij}^{2} = \eta \delta_i^p V_j^p \tag{11}$$

In order to update a weight that connects the input layer to the hidden layer, the chain rule can be used as follows:

$$\Delta w_{jk}^{1} = -\eta \frac{\partial J}{\partial w_{jk}^{1}} = -\eta \frac{\partial J}{\partial V_j^p} \frac{\partial V_j^p}{\partial w_{jk}^{1}} \tag{12}$$

where we have:

$$\frac{\partial J}{\partial V_j^p} = -\sum_{i=1}^{N} e_i^p\, f'(n_i^p) w_{ij}^{2} \tag{13}$$

and

$$\frac{\partial V_j^p}{\partial w_{jk}^{1}} = f'(n_j^p) x_k^p \tag{14}$$

This means that, we have:

$$\begin{aligned}\Delta w_{jk}^{1} &= \eta f'(n_j^p) x_k^p \sum_{i=1}^{N} \delta_i^p w_{ij}^{2} \\ &= \eta \delta_j^p x_k^p\end{aligned} \tag{15}$$

where δ_j^p is given as follows:

$$\delta_j^p = f'(n_j^p) \sum_{i=1}^{N} w_{ij}^{2} \delta_i^p \tag{16}$$

The algorithm starts by initialising the weights and biases of the network to small random values. Then, the input patterns are presented to the network. In this case, the network calculates the output values produced at each layer. These outputs are used as inputs to the next layer. Once the external outputs are calculated, the error signals are determined. The error signals are propagated backwards to update the values of the weights matrix. This is the online training algorithm which can minimise the cost function for small values of the learning rate by following the local gradient. On the other hand, batch training, where the weights are only updated when all input patterns are presented to the network, requires additional memory at each unit. Therefore, the online training is faster especially for a very large training set. Algorithm 2 summaries these steps involved in the backpropagation learning.

Algorithm 2 Backpropagation Algorithm

```
1: Initialise the network by selecting small random values for the weights and biases of the network.
2: Select an input pattern to be presented to the network.
3: Calculate the output of each layer until the external output is determined.
4: Compute the delta values for the output layer as demonstrated before.
5: Compute the deltas for the proceeding layers by propagating the errors backwards.
6: Determine the change in the weights and update the weights of the network for all layers of the network as follows:
       $\mathrm{w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t)}$
```

3.2 Hierarchical Neural Network

Namphol and Chin [38] proposed compressor and decompressor (CODEC), which is called Hierarchical Neural Networks (HNN) as illustrated in Fig. 2 for image compression. The system consists of three sections, which are the input-layer, the hidden-layer, and the output-layer sections. At the input layer section, there are M blocks corresponding to the number of nonoverlapped blocks of the image. Each block is constructed from P2 nodes, corresponding to the size of each block of the image. The hidden-layer section is constructed from the combiner, the compressor, and the decombiner. The first structure acts as a multiplexer to the input subblocks and contains less nodes than the number of neurons of the input layer. The second structure has less neurons than the combiner and its outputs represent the compressed data, which are stored for latter processing. The reconstruction of the compressed image involves decombining the stored data using the decombiner which acts as the demultiplexer. The output of the decombiner is forwarded to the output-layer section in which the compressed image data is reconstructed.

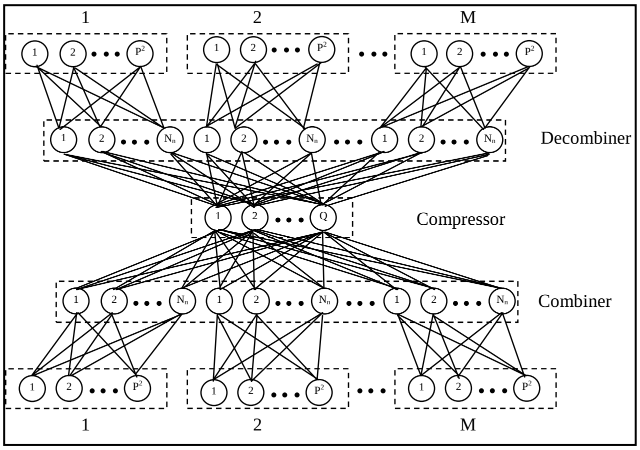


Fig. 2. Hierarchical CODEC system [38].

The network has symmetric structure which allows it to be trained outside-in (i.e., the outer layers are trained first, then the inner layers are trained). The training procedure used Nested Training Algorithm (NTA) which is an extended version of backpropagation.

In this case, the input training set to the network is defined as

$$X_t^{[m]} = \left\{x_{t,1}^{[m]} + x_{t,2}^{[m]}, \dots . x_{t,p}^{[m]}\right\} \quad (17)$$

The output of the hidden layer is defined as

$$h_{t,r}^{[m]} = f\left(\sum_{k=1}^{P^2} w_{r,k}^{[m]} x_{t,k}^{[m]} + b_r^{[m]}\right) \quad (18)$$

where $w_{r,k}^{[m]}$ is the weight from the k$^{\text{th}}$ node of the input layer to the r$^{\text{th}}$ node of the hidden layer.

The output for the output layer is calculated as

$$y_{t,k}^{[m]} = f\left(\sum_{r=1}^{N} w_{k,r}^{[m]} h_{t,r}^{[m]} + b_k^{[m]}\right) \quad (19)$$

In our works, Hierarchical Neural Network is used as an unsupervised learning method [38], which was adapted to identify SNP interactions

3.3 Evaluation and Validation

Performance of the proposed BP inspired by the hierarchical neural network classifier is measured using a receiver operating characteristic curve (ROC) and measuring the area under the curve (AUC), along with various other performance metrics including; true positive rate (TPR), true negative rate (TNR), Gini, Logarithmic Loss, and Mean Squared Error (MSE). The data is randomly divided into three smaller sets, 80% for training the models, 10% for tuning, and 10% to test the models performance.

TPR and TNR are used to measure the positive and negative predictive capabilities of classifiers in binary classification settings.

Furthermore, in this analysis the ROC curve is widely used to assess and compare classifiers performance. 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). AUC value represents the probability of a correct classification as for the positive and negative instances; the positive class will be ranked higher thus a higher AUC means a better classification. An ideal model would have a point in the upper North West corner of the ROC curve, which means that the model has accurately classified all people with T2D. In contrast, a model with random prediction performance will fall along the diagonal line of the ROC curve, in which TPR and FPR are equal over all different decision thresholds. The Gini coefficient value can be derived from Area Under the ROC curve (AUC), where $Gini = 2 * AUC - 1$. It represents the area between the ROC curve and the diagonal line, i.e. random prediction. The Gini coefficient is usually used in binary classification settings, Gini closer to 1 indicates higher learning rate.

Logarithmic Loss (Logloss) is a classification loss function which measures the performance of a classification model where the prediction input is a probability value between 0 and 1. An ideal Logloss value would be 0, which is an indication of a perfect model that correctly classifies T2D from healthy individuals.

The Mean Squared Error (MSE) is another performance metric measures the difference between actual values and the predicted values. MSE value closer to 0 denotes that the classification model is correctly classifies T2D from healthy individuals.

BP inspired by HNN is used to discover the epistatic interactions between SNPs before it is fine-tuned for the classification of T2D and is benchmarked with the supervised Deep Learning (DL) classifier and Random Forest (RF) classifier model. This study specifies 400 trees to train RF models with a maximum tree depth of 40. For DL a RectifierWithDropout activation function is employed with a number of epochs is set to 100 iterations and four hidden layers with 10 neurons. Input dropout ratio set to 0.1 and hidden dropout ratios for each layer set to 0.5. Early stopping is adopted using stopping metric set to Logloss and stopping tolerance and stopping rounds to 1×10^{-2} and 5, respectively. The learning rate and momentum are experimentally revealed. The learning rate is configured to 0.005 with rate annealing, and rate decay is set to 1×10^{-6} and 1, respectively. Momentum is set to 0.5 with momentum stable to 0 and momentum ramp to 1×10^{6}.

For first iteration of HNN models (2500 hidden neurons), a RectifierWithDropout activation function is used, a number of epochs are set to 10 iterations and two hidden layers with 10 neurons. For the second (2500, 1500 hidden neurons) and third (2500, 1500, 700 hidden neurons) classifiers, a RectifierWithDropout activation function is used, a number of epochs are set to 10 iterations and two hidden layers with 20 neurons. For all HNN classifiers an adaptive learning rate is adopted with parameters ρ and ε set to 0.99 and 1×10^{-8} respectively. Simulation, visualisation and evaluation were carried out using H2O machine learning platform in R software environment. It should be noted that due to the large number of SNPs and big genetic data used for our machine learning algorithms, all other types of machine learning algorithms failed to run. Subsections below describes the tuning parameters applied with BP classifier model.

Input dropout ratio parameter randomly specifies selected neurons to be ignored during each of the training iterations to improve generalization process. While the Hidden dropout ratios parameter can randomly select neurons from each hidden layer to be ignored during each of the training iterations to improve generalisation process. The stopping metric parameter determines the early stop of training process when the model's misclassification rate does not improve, while the stopping tolerance parameter allocating a threshold to stop training iterations when tolerance is crossed. The stopping rounds parameter quits the training in case the selection for stopping metric does not enhance for a particular value of training iterations. Max w2 parameter is a maximum on the sum of the squared incoming weights into any one neuron. It is useful when activation function is set to Rectifier. This help stability for Rectifier. The learning rate is a hyper-parameter measured as the difference between the forecasted and actual values to adjust training of BP. While rate annealing parameter is the inverse of the number of training samples that requires to trim the learning rate by 50%. Rate annealing cuts down the learning rate to freeze into local minima in the optimization concept. While the rate decay parameter controls the change of learning rate across layers.

Momentum is a method to improve both learning speed and accuracy. There are three momentum attributes including; momentum start, momentum stable, and momentum ramp. The momentum start is employed to control the amount of momentum at the start of learning process. Momentum stable is utilized to regulate the amount of learning in which momentum increases. Finally, momentum ramp will control the final momentum value reached after momentum ramp training samples.

4 Simulation and Discussion

The results of examining the proposed model for the T2D data set is acquired using HNN and BP. In comparison, Deep Learning (DL) and Random Forest (RF) classifiers are used to benchmark the performance of combined HNN and BP (HNN+BP). This evaluation considers SNPs generated with a p-value threshold of 10^{-2} resulting in 6609 SNPs. HNN uses these SNPs to extract the abstract representation of the features and to capture the non-linear epistatic interactions between SNPs. The results based on several HNN iterations. The first HNN consists of 2500 hidden neurons while the second and third HNN use (2500, 1500), and (2500, 1500, 700) hidden neurons respectively.

Table 1 illustrates the performance metrics of HNN+BP, DL, and RF for the validation set. Metric values for the first HNN+BP (2500 hidden units), second HNN+BP (2500,1500), and third HNN+BP (2500,1500,700) were obtained using optimized F1 threshold with values 0.3412, 0.3192, and 0.3474 respectively. The performance metrics for DL and RF were acquired using optimised F1 threshold with values 0.37, and 0.5, respectively.

TABLE 1

PERFORMANCE METRICS FOR HNN+BP, DL, RF FOR VALIDATION SET

Metric	HNN+BP1 (%)	HNN+BP2 (%)	HNN+BP3 (%)	DL (%)	RF (%)
AUC	95.46	92.06	86.89	97.81	74.36
TPR	93.09	90.90	86.18	97.09	91.63
TNR	88.46	78.63	72.64	88.88	28.63
Logloss	32.41	39.32	47.10	28.78	65.19
Gini	90.93	84.13	73.79	95.62	48.72
MSE	07.78	11.29	15.01	08.12	22.97

HNN+BP1=2500 hidden units.
HNN+BP2=2500,1500 hidden units.
HNN+BP3=2500,1500,700 hidden units.

Table 2 presents the performance metrics obtained using the test set for HNN+BP, DL, and RF. Metric values for the first HNN+BP (2500 hidden units), second HNN+BP (2500,1500), and third HNN+BP (2500,1500,700) were obtained using optimised F1 threshold with values 0.2770, 0.4340, and 0.3424, respectively. The performance metrics for DL and RF were gained using optimized F1 threshold with values 0.3683, and 0.515 respectively. The results are lower than those produced using validation set except for HNN (2500, 1500, 700) hidden units as 2.49% improvement is observed.

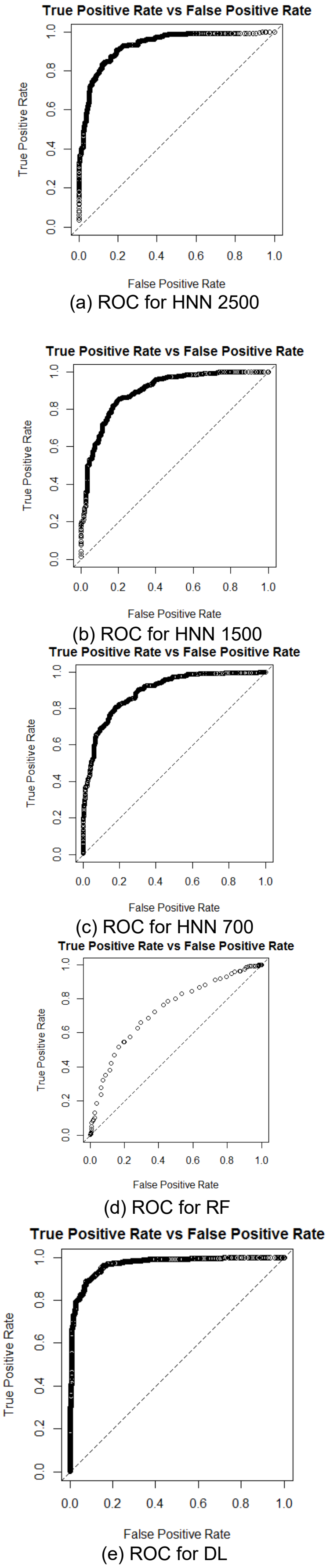
TABLE 2

PERFORMANCE MEASURE FOR HNN+BP, DL, RF FOR THE TEST SET

Models	AUC (%)	TPR (%)	TNR (%)	Logloss (%)	Gini (%)	MSE (%)
HNN+BP1	92.89	90.87	80.53	45.82	85.78	11.62
HNN+BP2	89.50	85.47	79.77	47.61	79.01	13.91
HNN+BP3	89.38	90.20	69.84	41.55	78.77	13.34
DL	96.74	96.28	84.35	31.17	93.49	8.93
RF	73.24	83.10	46.56	65.53	46.49	23.13

The classification accuracy of HNN+BP shows a progressive deterioration as the input raw features are steadily compressed down to 700 hidden neurons using validation and test sets. DL classifier achieved comparable results to those produced using HNN (2500 compressed units) in the validation set. The results evidently show that HNN+BP outperforms RF classifier.

Fig.3 presents the ROC curves for HNN, DL and RF models classifiers. Despite the gradual deterioration in the performance of HNN, the hierarchal neural network achieved accuracy values above 89% with 700 hidden units. This is significant in comparison to the RF which achieves accuracy value slightly higher than 73%.



(a) ROC for HNN 2500

(b) ROC for HNN 1500

(c) ROC for HNN 700

(d) ROC for RF

(e) ROC for DL

Fig. 3. Performance ROC curves for the test set. (a) to (c) for HNN. (d) for RF. (e) for DL.

2 Rational Discussion

Genetic association studies have significantly expanded the understanding of the genetic variants that cause predisposing to complex human diseases.

The last decade has seen a significant expansion in our knowledge of genetic variants associated with common diseases. Critical to this has been the ability to measure genetic variation at hundreds of thousands of markers across the human genome, in large numbers of individuals. Genome-wide association studies exploited these technological developments in large case-control studies, with unprecedented success to find the non-linear relationships exist in genotype-phenotype interactions as standard multi-variable statistical approach such as logistic regression is more suitable for capturing linear interactions but not the epistatic interactions present among SNPs. This work focuses on detecting epistatic interactions in high-dimensional T2D GWAS data. The selected set of SNPs is used for the binary classification of phenotypes outcome, with the support of back prorogation algorithm and hierarchical neural network.

Using HNN, the results in the simulation process was tested to demonstrate a gradual deterioration as the number of features compressed down to 700 hidden units. However, the classification accuracy value of the 700 compressed neurons remains reasonable with 89.38% in the test set. The best result obtained using 2500 compressed units (AUC=92.89%, TPR=90.87%, TNR=80.53%, Logloss= 45.82%, Gini=85.78%, MSE=11.62%).

RF and DL algorithms are used to benchmark HNN classification performance. RF is a prevalent method that is increasingly and successfully used in genetic studies [16], [23], [42], [43]. In this analysis, the result shows that using 6609 SNPs it was possible to achieve 73.24% classification accuracy. TPR and TNR are instable indicating that RF classifier has the low discriminatory capacity for this given dataset to separate cases and controls phenotypes. In comparison, even though HNN showing trivial deterioration ranging from AUC=92.89% to AUC=89.38% these results remain significantly higher than what RF is achieved. This is because, in HNN CODEC system, the multiple hidden layers compress the input features into abstract representations with modelling the complexity of non-linearity of genotype-phenotype interactions generally observed in genetic data. This algorithm outperforms the traditional supervised classification models and offers a powerful way to enhance GWAS data analysis.

Using 6609 SNPs to train DL it was possible to obtain high performance result of (AUC=96.74%, TPR=96.28%, TNR=84.35%, Logloss= 31.17%, Gini=93.49%, MSE=8.93%). HNN with 2500 compressed units (initially 6609 SNPs) achieved less predictive accuracy than DL using (6609 SNPs), the results still comparable and significant for both models with AUC=92.89% for HNN and AUC=96.74% for DL. Obtaining high results despite the fact that the original data compressed gradually from 6609 SNPs to 700 SNPs with a lower predictive accuracy of 89.38% is encouraging and demonstrate the potential of applying BP with HNN for the classification of type 2 diabetes GWAS data.

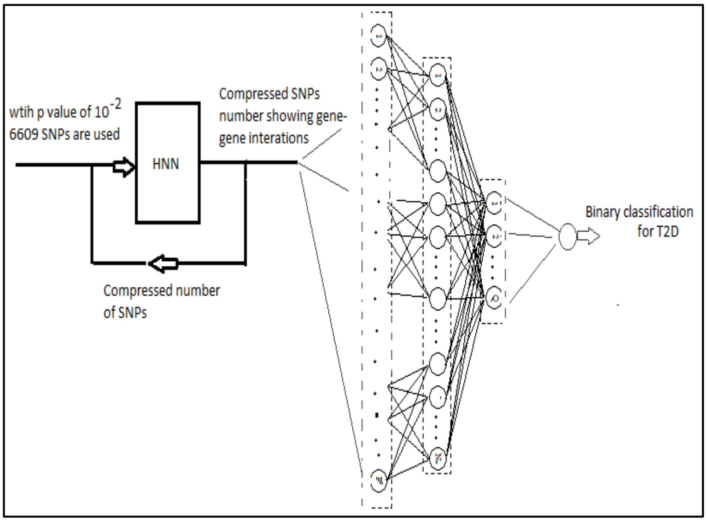

wtih p value of 10-2
6609 SNPs are used
HNN
Compressed SNPs number showing gene-gene interations
Compressed number of SNPs
Binary classification for T2D

Fig. 4. Proposed BP with the HNN binary genetic data classification for T2D.

Figure 4 illustates the proposed system which consists from a HNN that can reduce and compact the number of SNPS and the BP neural network structure which can perform the binary classification. While Table 3 provides summary about the parameter selections for our experiemnts.

TABLE 3

SUMMARY OF THE PARAMETERS SELECTIONS FOR THE PROPOSED MODELS

Model	No of Neurons for HNN	Size DL	Epochs	*f (x)*
HNN+BP1	2500	2 hidden layers with 10 neurons	10	*max(0, x)*
HNN+BP2	2500 then1500	2 hidden layers with 20 neurons	10	*max(0, x)*
HNN+BP3	2500 then1500 then 700	2 hidden layers with 20 neurons	10	*max(0, x)*
Momentum term and learning rates are experimentally determined.				

Table 4 shows various litreture review for the prediction of T2D using genetic data. Ban et al. [47] and Bae et al [37] looked at the used of SVM for the early prediction of T2D. Their results has indicated that SVM can generate reasnable acucracy using the korean cohort study and the T2D-gene consolutium data sets, respectively. Kim et al [39] looked at the use of the NHS and the HPFS data sets for the early prediction of T2D HPFS and the NHS data sets repectively, while our reesarch has shown AUC of 0.96 for the use of DL for a combined HPFS and NHS data set.

TABLE 4

THE PREDICTION OF T2D USING GENETIC DATA AND MACHINE LEARNING ALGORITHMS

Study	Model	P-value	AUC	Sens	Spec	Data Set
Ban et. al. [47]	SVM	<0.6	0.65	56%	73%	Korean cohort studies
Bae et. al. [37]	SVM	$<10^{-6}$	0.89	-	-	T2D-GENE
Kim et al [39]	Deep NN	$<10^{-2}$	0.93	-	-	HPFS
Kim et al [39]	Deep NN	$<10^{-2}$	0.92			NHS
Proposed model	HNN1+ DL	$<10^{-2}$	0.92	0.90	0.80	HPFS+NHS
This study	DL	$<10^{-2}$	0.96	0.96	0.84	HPFS+NHS

3 CONCLUSION

This paper reports a comprehensive expermental study to explore a novel approach for classifying Type 2 diabe-tes genetics data. Backpropagation algorithm inspired by hierarchical neural network has been proposed and used for identifying epistatic interactions and classification of high-dimensional GWAS data in T2D. The datasets which was used for this study is provided by the Genotypes and Phenotypes (dbGap) database. Various stringent quality control assessment steps followed by logistic regression association analysis adjusted GC are performed for single-SNP analysis. Using 5393 T2D case-control samples, we achieved (AUC=92.89%, TPR=90.87%, TNR=80.53%) using 2500 compressed neurons.

The results provided by this study is very encroaching, however, more improvement is required, by furthering the work of parameters tuning and optimisation of proposed model to improve the classification results.

The fact that we use p-value threshold to extract a subset of SNPs this is a standard and common way widely used in the literature, however, using this method there is a significant possibility to include redundant variables. Using Linkage Disequilibrium (LD) pruning method can reduce the number of redundant predictors.

ACKNOWLEDGMENT

The authors would like to acknowledge Khawarizmi International College (Abu Dhabi, UAE) for the financial support for this research. 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 Nurses' Health

Study (NHS) and Health Professionals' Follow-up Study (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).

The author will also like to give their appreciation for The University of Anbar, Iraq, for some experiments that was carried out by their staff.

References

[1] J. Gulcher and K. Stefansson, "Clinical risk factors, DNA variants, and the development of type 2 diabetes.," N. Engl. J. Med., vol. 360, no. 13, p. 1360; author reply 1361, 2009.

[2] S. E. Inzucchi, R. M. Bergenstal, J. B. Buse, M. Diamant, E. Ferrannini, M. Nauck, A. L. Peters, A. Tsapas, R. Wender, D. R. Matthews, American Diabetes Association (ADA), and European Association for the Study of Diabetes (EASD), "Management of hyperglycemia in type 2 diabetes: a patient-centered approach: position statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD).," Diabetes Care, vol. 35, no. 6, pp. 1364–1379, 2012.

[3] S. Fazeli Farsani, M. P. Van Der Aa, M. M. J. Van Der Vorst, C. A. J. Knibbe, and A. De Boer, "Global trends in the incidence and prevalence of type 2 diabetes in children and adolescents: A systematic review and evaluation of methodological approaches," Diabetologia, vol. 56, no. 7, pp. 1471–1488, 2013.

[4] N. Hex, C. Bartlett, D. Wright, M. Taylor, and D. Varley, "Estimating the current and future costs of Type1 and Type2 diabetes in the UK, including direct health costs and indirect societal and productivity costs," Diabet. Med., vol. 29, no. 7, pp. 855–862, 2012.

[5] S. P. Deng and D. S. Huang, "SFAPS: An R package for structure/function analysis of protein sequences based on informational spectrum method," Methods, vol. 69, no. 3, pp. 207–212, 2014.

[6] J. F. Xia, X. M. Zhao, J. Song, and D. S. Huang, "APIS: Accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility," BMC Bioinformatics, vol. 11, 2010.

[7] M. H. Wanga, J. Li, V. S. Y. Yeung, B. C. Y. Zee, R. H. Y. Yu, S. Ho, and M. M. Y. Waye, "Four pairs of gene-gene interactions associated with increased risk for type 2 diabetes (CDKN2BAS-KCNJ11), obesity (SLC2A9-IGF2BP2, FTO-APOA5), and hypertension (MC4R-IGF2BP2) in Chinese women," Meta Gene, vol. 2, no. 1, pp. 384–391, 2014.

[8] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies," Bioinformatics, vol. 26, no. 4, pp. 445–455, 2010.

[9] T. W. T. C. C. Consortium, "Genome-wide association study of 14 000 cases of seven common diseases and 3 000 shared controls," Nature, vol. 447, no. 7145, pp. 661–678, 2007.

[10] Z. Zhu, X. Tong, Z. Zhu, M. Liang, W. Cui, K. Su, M. D. Li, and J. Zhu, "Development of GMDR-GPU for Gene-Gene Interaction Analysis and Its Application to WTCCC GWAS Data for Type 2 Diabetes," PLoS One, vol. 8, no. 4, 2013.

[11] 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, 2014.

[12] Z. Wei, K. Wang, H.-Q. Qu, H. Zhang, J. Bradfield, C. Kim, E. Frackleton, C. Hou, J. T. Glessner, R. Chiavacci, C. Stanley, D. Monos, S. F. A. Grant, C. Polychronakos, and H. Hakonarson, "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, 2009.

[13] López, B., Torrent-Fontbona, F., Viñas, R., Fernández-Real, J.M.: Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction. Artif. Intell. Med. 85, 43–49 (2018).

[14] Koo, C.L.C., Liew, M.M.J., Mohamad, M.S., Salleh, A.H.M.: A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. Biomed Res. Int. 2013, 13 (2013).

[15] D.-S. Huang and X. Huang, "Improved performance in protein secondary structure prediction by combining multiple predictions," Protein Pept. Lett., vol. 13, no. 10, 2006.

[16] K. A. Tryka, L. Hao, A. Sturcke, Y. Jin, Z. Y. Wang, L. Ziyabari, M. Lee, N. Popova, N. Sharopova, M. Kimura, and M. Feolo, "NCBI's database of genotypes and phenotypes: DbGaP," Nucleic Acids Res., vol. 42, no. D1, pp. 975–979, 2014.

[17] M. C. Cornelis, L. Qi, C. Zhang, et al. Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. Ann Intern Med. 2009;150(8):541–550.

[18] L. Qi, M. C. Cornelis, P. Kraft, et al. Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Hum Mol Genet. 2010; 19(13):2706–2715.

[19] C. C. Laurie, K. F. Doheny, D. B. Mirel DB, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol. 2010; 34(6):591–602.

[20] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham,

"PLINK: A tool set for whole-genome association and population-based linkage analyses," Am. J. Hum. Genet., vol. 81, no. 3, pp. 559–575, 2007.

[21] W. S. Bush and J. H. Moore, "Chapter 11: Genome-Wide Association Studies," PLoS Comput. Biol., vol. 8, no. 12, 2012.

[22] 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.

[23] C. S. Pareek, R. Smoczynski, and A. Tretyn, "Sequencing technologies and genome sequencing," J. Appl. Genet., vol. 52, no. 4, pp. 413–435, 2011.

[24] J. Xuan, Y. Yu, T. Qing, L. Guo, and L. Shi, "Next-generation sequencing in the clinic : Promises and challenges," Cancer Lett., vol. 340, no. 2, pp. 284–295, 2013.

[25] X. Zhang, S. Huang, Z. Zhang, and W. Wang, "Chapter 10: Mining Genome-Wide Genetic Markers," PLoS Comput. Biol., vol. 8, no. 12, 2012.

[26] P. Fomby and A. J. Cherlin, "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing," vol. 72, no. 2, pp. 181–204, 2011.

[27] A. Auton, et. al., "A global reference for human genetic variation," Nature, vol. 526, no. 7571, pp. 68–74, 2015.

[28] T. International and H. Consortium, "The International HapMap Project.," Nature, vol. 426, no. 6968, pp. 789–796, 2003.

[29] R. Shields, "Common disease: Are causative alleles common or rare?," PLoS Biol., vol. 9, no. 1, pp. 9–10, 2011.

[30] E. T. Cirulli and D. B. Goldstein, "Uncovering the roles of rare variants in common disease through whole-genome sequencing," Nat. Reniews Genet., vol. 11, pp. 415–425, 2010.

[31] P. Werbos, "Beyond Regression: New tools for prediction and analysis in the behaviour sciences". PhD. Thesis, Harvard university, 1974.

[32] D. B. Parker,. "Learning logic", Technical report TR-47, centre for computational research in economics and management science, Massachusetts institute of technology, Cambridge, 1985.

[33] D.-S. Huang, L. Zhang, K. Han, S. Deng, K. Yang, and H. Zhang, "Prediction of Protein-Protein Interactions Based on Protein-Protein Correlation Using Least Squares Regression," Curr. Protein Pept. Sci., vol. 15, no. 6, pp. 553–560, 2014.

[34] D.E., Rumelhart, G. E. Hinton, and R. J. Williams. "Learning presentation by back-propagating errors," Nature, 323, pp. 533-536, 1986.

[35] Y. Le Cun, Une Procédure d' apprentissage pour réseau à seuil assymétrique. In cognitiva 85: A la frontiére de l'intelligence artificielle des sciences de la connaissance des neurosciences (Paris 1985), 599-604. Paris: CESTA.

[36] D. S. Huang and H. J. Yu, "Normalized feature vectors: A novel alignment-free sequence comparison method based on the numbers of adjacent amino acids," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 10, no. 2, pp. 457–467, 2013.

[37] S. Bae and T. Park, "Risk prediction using common and rare genetic variants: Application to Type 2 diabetes", IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1757-1760. 2017

[38] A. S. H, . Namphol, M. Chin, Arozullah, "Image compression with a hierarchical neural network," IEEE Transactions on aerospace and electronic systems, 32(1), pp. 326-337, 1996.

[39] J. Kim, M.J. Kwak and M. Bajaj, "Genetic prediction of type 2 diabetes using deep neural network", Clinical Genetics, pp. 822-829, 2018

[40] D. S. Huang, X. M. Zhao, G. Bin Huang, and Y. M. Cheung, "Classifying protein sequences using hydropathy blocks," Pattern Recognit., vol. 39, no. 12, pp. 2293–2300, 2006.

[41] D.-S. Huang and C.-H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data.," Bioinformatics, vol. 22, no. 15, pp. 1855–62, 2006.

[42] S. P. Deng, L. Zhu, and D. S. Huang, "Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks," BMC Genomics, vol. 16, no. 3, 2015.

[43] M. B. Kursa, "Robustness of Random Forest-based gene selection methods," BMC Bioinformatics, vol. 15, no. 1, p. 8, 2014.

[44] P. Donaldson, A. Daly, L. Ermini, and D. Bevitt, Genetics of Complex Disease. New York: Garland Science, Taylor & Francis Group, 2016.

[45] J. Graffelman and B. S. Weir, "Testing for Hardy – Weinberg equilibrium at biallelic genetic markers on the X chromosome," vol. 116, no. 6, pp. 558–568, 2016.

[46] Y. Zhang, Y. Liu, Y. Liu, Y. Zhang, and Z. Su, "Genetic Variants of Retinoic Acid Receptor-Related Orphan Receptor Alpha Determine Susceptibility to Type 2 Diabetes Mellitus in Han Chinese," Genes., 2016

[47] H.J. Ban, J.Y. Heo, K.S. Oh, K.J. Park, " Identification of type 2 diabetes-associated combination of SNPs using support vector machine", BMC genetics, 2010.