

Article

Deep Learning Framework for Complex Disease Risk Prediction Using Genomic Variations

Hadeel Alzoubi ^{1,*} , Raid Alzubi ¹  and Naeem Ramzan ² 

¹ Department of Computer Science, College of Computer Science and Information Technology, King Faisal University, Al-Ahsa 31982, Saudi Arabia

² School of Computing, Engineering and Physical Sciences, University of the West of Scotland, High Street, Paisley PA1 2BE, UK

* Correspondence: hmalzoubi@kfu.edu.sa

Abstract: Genome-wide association studies have proven their ability to improve human health outcomes by identifying genotypes associated with phenotypes. Various works have attempted to predict the risk of diseases for individuals based on genotype data. This prediction can either be considered as an analysis model that can lead to a better understanding of gene functions that underlie human disease or as a black box in order to be used in decision support systems and in early disease detection. Deep learning techniques have gained more popularity recently. In this work, we propose a deep-learning framework for disease risk prediction. The proposed framework employs a multilayer perceptron (MLP) in order to predict individuals' disease status. The proposed framework was applied to the Wellcome Trust Case-Control Consortium (WTCCC), the UK National Blood Service (NBS) Control Group, and the 1958 British Birth Cohort (58C) datasets. The performance comparison of the proposed framework showed that the proposed approach outperformed the other methods in predicting disease risk, achieving an area under the curve (AUC) up to 0.94.

Keywords: complex diseases risk prediction; feature selection; GWAS; machine learning; mutual information; SNP



Citation: Alzoubi, H.; Alzubi, R.; Ramzan, N. Deep Learning Framework for Complex Disease Risk Prediction Using Genomic Variations. *Sensors* **2023**, *23*, 4439. <https://doi.org/10.3390/s23094439>

Academic Editor: Bijan Najafi

Received: 13 February 2023

Revised: 5 April 2023

Accepted: 26 April 2023

Published: 1 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The human genome is the whole set of deoxyribonucleic acid (DNA) sequences in humans, which consists of approximately three billion base pairs [1,2]. Human genomes are almost identical; however, at least three million nucleotides per individual are different. The most common type of these genetic variations is Single Nucleotide Polymorphisms (SNPs). Studies have proven that SNPs are the most contributing markers in several complex and rare diseases [1]. Most SNPs are natural; however, certain SNPs are functional and affect the phenotype of interest, such as skin colour, height, infection, resistance, or responses to drugs.

Genome-wide association studies (GWASs) have proven their ability to unveil susceptibility variants in human diseases [3,4]. These studies provide a better understanding of diseases by enabling researchers to identify SNPs that significantly differ in frequencies between the affected and healthy individuals. GWASs have identified more than 4164 loci contributing to common complex diseases such as diabetes [5–7], cancer [8–10], and rheumatoid arthritis [11]. Moreover, GWASs allow researchers to develop models for complex disease risk prediction based on genetic information [12–16]. If the disease of interest can be identified at the early stage, specific therapy plans can be applied to delay or even prevent the outset of some diseases [17,18]. Building risk prediction models can contribute to personalised medicine becoming feasible by utilising an individual's genome to predict disease risk and the response to treatment. However, the critical issue is how to predict disease risk accurately from a huge number of SNPs.

Many machine learning algorithms have been applied to SNP data analysis in order to build a model that is able to predict disease risk for unseen samples. However, the process of developing a risk prediction system using genomic data is computationally expensive due to the extremely high dimensionality number of features. Consequently, it has become essential to provide the classifier with a manageable number of discriminative SNPs. Moreover, selecting the most convenient classification algorithm for building the prediction model is also a crucial step for achieving high model performance. Traditional machine learning techniques have been successfully employed in GWASs for risk prediction [19,20]. However, deep learning networks are one of the most popular and powerful machine learning techniques for pattern recognition in many fields such as image processing [21], natural language processing [22], and bioinformatics [23–25]. Deep learning has the potential to learn high-level complex and hierarchical data patterns more accurately than traditional machine learning techniques. Multi-Layer Perceptron (MLP) is one of the state-of-the-art neural networks that consists of three main layers: input, hidden, and output. The hidden layers consist of two or more layers of self-learning neurons, where the weights of fully connected neurons between adjacent layers are learned using the backpropagation algorithm. Once the weights are set, samples from unseen targets can be used as input, allowing the framework to perform prediction. Convolutional neural networks (CNN) have been used to extract the most informative features to classify Alzheimer’s disease patients [26]. There have been some attempts to apply deep neural networks in the SNP dataset with promising performance [27–29].

Machine learning techniques have been effectively used in a variety of disease risk prediction systems based on human genomic variations [30]. The authors in [31] highlighted the recent development of machine learning algorithms in the field of genomic variations. They illustrated the contribution of machine learning techniques in improving complex disease prediction models. In [32], the authors built a model for assessing the risk of type 1 diabetes (T1D). Their model employed Support Vector Machines (SVM) as the classification algorithm, which was fed with 100 SNPs selected using an ensemble feature selection technique to achieve an AUC of 0.84. The authors in [33] developed a breast cancer risk prediction model using a combination of the MeanDiff feature selection technique and KNN classifier. Their proposed model showed a 10% increment in terms of accuracy over the baseline classifier; however, the best accuracy achieved by their system was less than 60%.

The authors in [34] compared five different machine learning methods, Bayesian Networks, SVM, Random Forest, Radial Basis Function network, and Logistic Regression, to predict the risk factors of bipolar disorder. The best performance was achieved using Bayesian Networks with an AUC of 0.556. In [35], the authors developed an ensemble machine-learning technique for autism disease risk prediction. In [36], the authors assessed risk prediction for complex diseases by comparing the performance of four different prediction algorithms. The best performance was reported using sparse penalised approaches. The authors in [37] applied different machine-learning techniques to predict the risk of anorexia nervosa. The logistic regression with the lasso penalty technique performed slightly better than SVM and gradient-boosted trees. They applied their system to different sample sizes of the dataset and concluded that a larger sample size improves the machine learning risk prediction outcomes. A similar conclusion was drawn by [15], who employed logistic regression over the top 3000 associated SNPs on the WTCCC [38] for Crohn’s disease. In this paper [39], the authors compared the ability of allele counting, Logistic Regression (LR), and SVM to predict coronary artery disease risk and found the best AUC up to 0.60 to be achieved by LR. In order to assess the risk of T1D and rheumatoid arthritis (RA), the authors in [40] employed SVMs and random forest. Their proposed system successfully reached an AUC of 0.82 and 0.71 for T1D and RA, respectively. The aforementioned works applied different machine learning techniques and reached promising performance in some cases. In order to predict breast cancer risk, the authors of [41] proposed a system that adopts a gradient tree boosting method followed by an adaptive iterative SNP search. The authors aimed to capture the group of interacting SNPs over the

given disease. A combination of genomic data and demographic data has been used to predict the disease risk of breast cancer by [42]. Their system used a gradient tree-boosting method in both the selection and classification phases. However, many clinical and medical applications require more accurate prediction systems.

Risk prediction methods typically apply different techniques in order to select a manageable number of SNPs. Most studies rank SNPs based on the p -value of their association with the phenotype of interest to control the number of selected SNPs and use the top associated SNPs as input to a prediction algorithm [43]. However, the predictive power of these studies is relatively poor, and discarding SNPs with a low p -value could limit the opportunity to identify inter-SNP correlations [19,44]. Moreover, there are variations associated with many diseases that have not yet been identified; hereby, analysing an expanded list of SNPs may improve the prediction system performance. For example, in [19], it was suggested that considering uncommon and rare SNPs can improve risk prediction for some diseases such as Parkinson's disease using SVM. In addition, SNPs were selected for SVM by applying different p -value thresholds. Moreover, in [44], the BootRank technique was used in order to select robust informative SNPs to be used in a risk prediction model. The BootRank technique was combined with seven different classifiers to evaluate the performance of their proposed technique. Their model improved the ability to predict the disease risk of unseen individuals in the WTCCC data.

In this work, an accurate deep-learning framework for complex disease risk prediction has been proposed. An adequate subset of SNPs that are highly correlated and non-redundant has been selected using the Joint Mutual Information (JMI) method [45]. Then, the selected features were fed to an MLP that consists of an input layer, five hidden layers, and an output layer to train the prediction system. The proposed system was evaluated using datasets from WTCCC. The comparative experimental results demonstrate the ability of the proposed to accurately predict risk for different diseases as compared to the state-of-the-art approaches including [15,32,34,39,40,44,46–49], achieving an AUC of up to 0.94. The rest of this work is organised into three sections. Section 2 discusses materials and methods. Section 3 presents the experimental results and discussion. Finally, Section 4 concludes this work.

2. Materials and Methods

2.1. Genotype Datasets

Genotype data were obtained from the WTCCC [38] for seven different diseases. The diseases are Type 1 diabetes (T1D), Type 2 diabetes (T2D), inflammatory bowel disease (IBD), coronary artery disease (CAD), bipolar disorder (BD), rheumatoid arthritis (RA), and hypertension (HT), as presented in Table 1. Each disease dataset contains approximately 2000 cases. The control sets obtained from the UK National Blood Service Control Group (NBS) and 1958 British Birth Cohort (58C) contained 1500 individuals [38]. Each sample consists of 500,568 SNPs that were produced by an Affymetrix 500k chip sequencer. As recommended by the associated datasets, 809 samples and 30,956 SNPs were excluded due to deviation from Hardy–Weinberg equilibrium, bad quality, or bad clustering [38]. The dataset has been filtered to exclude SNPs based on the following threshold: a Minor Allele Frequency (MAF) of 1%, p -value $< 1 \times 10^{-3}$, and a missing rate of 5% [32,46,50–52]. As a result of the filtering, the final number of samples for each data set is presented in Table 1, with 469,606 SNPs for each one. To ensure that our results are not biased to cases or control, an equal number of samples for each class have been used. Where a group of healthy samples were randomly selected from UKBS and 58C to keep the case:control ratio at 50%:50% for each disease.

Mutation exists in the gene copy that is inherited from both parents. The allele frequencies are represented by A and B for the major allele frequency and minor allele frequency, respectively. Any given SNP could have the value of AA or BB to indicate that it is a homozygous SNP and the value of AB for a heterozygous SNP. In this proposed work, we used the additive model to encode SNPs. The encoding technique counts the

minor allele appearance. Consequently, the coding value 0 represents AA, 1 represents AB, and 2 represents BB. Finally, after implementing the aforementioned coding technique, the dataset is represented in numerical format.

Table 1. Case and control datasets used.

Dataset	No. of Samples	No. of Excluded Samples	No. of Samples after Filtration
Bipolar disorder (BD)	1998	129	1869
Coronary artery disease (CAD)	1998	62	1936
Inflammatory bowel disease (IBD)	2005	256	1749
Hypertension (HT)	2001	48	1953
Rheumatoid arthritis (RA)	1999	136	1863
Type 1 diabetes (T1D)	2000	37	1963
Type 2 diabetes (T2D)	1999	75	1924
UK National Blood Service (UKBS)	1500	42	1458
1958 British Birth Cohort (58C)	1504	24	1480

If a dataset consists of n samples and q SNPs, which can be represented by a $G = n \times q$ matrix, then G_{ij} is the number of the minor allele of SNP j for the sample i . Let Y_i be a binary indicator for the disease status of a given sample $i = 1, \dots, n$. The affected samples (case) are considered as having a positive class label ($Y_i = 1$) and the healthy ones (control) as having a negative class label ($Y_i = 0$).

2.2. Method

The proposed framework predicts the risk of an examined disease using SNP data. An MLP-based binary classifier has been developed to predict the disease risk status. A mutual information feature selection technique has been applied to decrease the feature space dimensionality and select the most discriminative SNPs. The dataset was split into (70%) training and (30%) testing sets, keeping the class ratio of each group similar to that of the whole dataset, and the testing data were only used for analysing the predictive power of the proposed system as illustrated in Figure 1. Five-fold cross-validation has been applied over the training data in order to perform feature selection. Finally, different performance metrics have been used to evaluate the predictive power of the proposed framework.

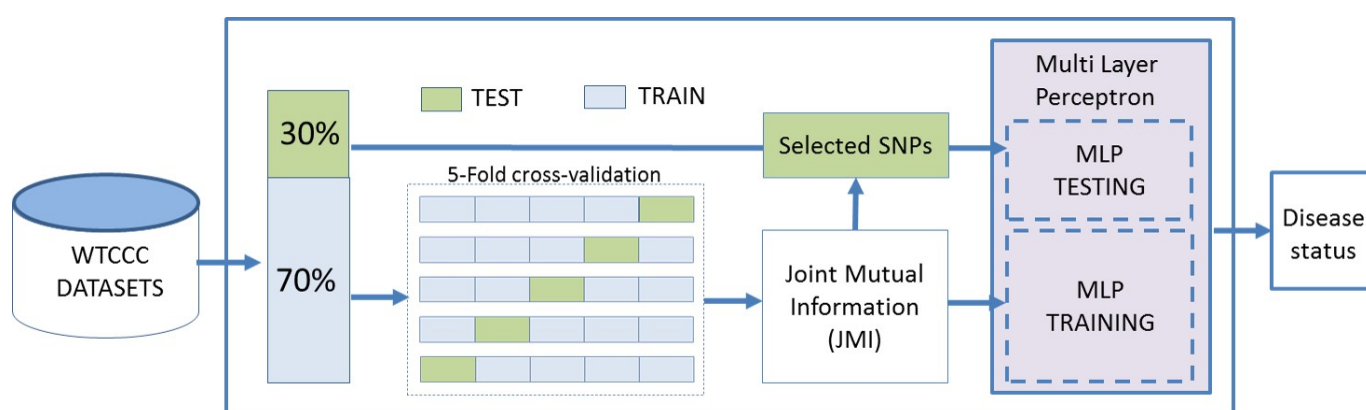


Figure 1. The proposed framework structure.

2.2.1. Feature Selection

The extremely large number of SNPs in the genome makes the application of machine learning techniques on SNP data computationally impossible. Consequently, the application of feature selection techniques is necessary for the selection of a significantly smaller subset of SNPs. Statistical and machine learning-based feature selection methods have demonstrated their ability to select an optimal SNP subset out of the whole genome [35,53].

In this work, JMI has been employed as a feature selection method in order to reduce computational complexity and improve risk prediction performance.

Mutual information feature selection methods have been widely applied in the biomedical field [54]. Mutual information is used to measure the features' relevancy and redundancy [55]. In the multivariate filter selection method, mutual information does not make any assumption of the data or change the original data representation [56]. In this work, JMI was used to measure the discriminative power of features and select a reduced set of SNPs to be injected into the prediction model.

JMI is a popular feature selection technique that selects a subset of features to maintain a high feature association and maximum correlation with the class of interest [45]. This method measures the information provided by the feature vector $s_1, s_2, \dots, s_q$ that decreases the uncertainty about the class label Y . JMI uses mutual Information to measure the amount of relevancy and redundancy between features. JMI calculates not only mutual information between features and the class label but also takes into consideration the correlation between the new feature and already selected features D , thus ensuring a good trade-off between relevancy and redundancy [45]. A higher JMI value for a feature s_i means that the feature s_i is relevant to the target Y and is highly complementary to the already picked features s_j, j in D . The JMI for a feature s_q is computed as shown in Equation (1) [57].

$$JMI(s_i) = \sum_{s_j \in D} I(s_i, s_j; Y) \quad (1)$$

where :

$$I(X, Y; Z) = I(X; Z|Y) + I(Y; Z) \quad (2)$$

$$I(Y; Z) = \sum_{y,z} p(y, z) \log \frac{p(y, z)}{p(y)p(z)} \quad (3)$$

$$I(X; Y|Z) = \sum_{y \in Y} p(y) \sum_{x \in X} \sum_{z \in Z} p(xz|y) \log \frac{p(xz|y)}{p(x|y)p(z|y)} \quad (4)$$

In this work, JMI was applied over the training set to select a subset of SNPs, D . F -fold cross-validation has been applied in order to create a matrix $q \times F$, with q being the number of SNPs of each sample and $F = 5$. At any given fold, the subset of selected SNPs is assigned to 1, and the value of 0 is assigned to the remaining unselected ones as presented in Equation (5).

$$d_j = \begin{cases} 0 & \text{if SNP } j \text{ is not selected} \\ 1 & \text{if SNP } j \text{ is selected} \end{cases} \quad (5)$$

At the end of the fifth fold, the accumulated weight W for a given SNP j is calculated as presented in Equation (6). The weight for each SNP represents how many times a given SNP was selected, as illustrated in Figure 2. For example, if an SNP weight is 1, that means it has been selected in all folds. An SNP weight will be 0.6 if it has been selected in three folds. An SNP weight will be 0 if it has not been selected in any fold. Only SNPs that have a weight exceeding a threshold value (a) will be propagated to the prediction model.

$$W_j = \frac{\sum_{i=1}^5 d_j}{5} \quad (6)$$

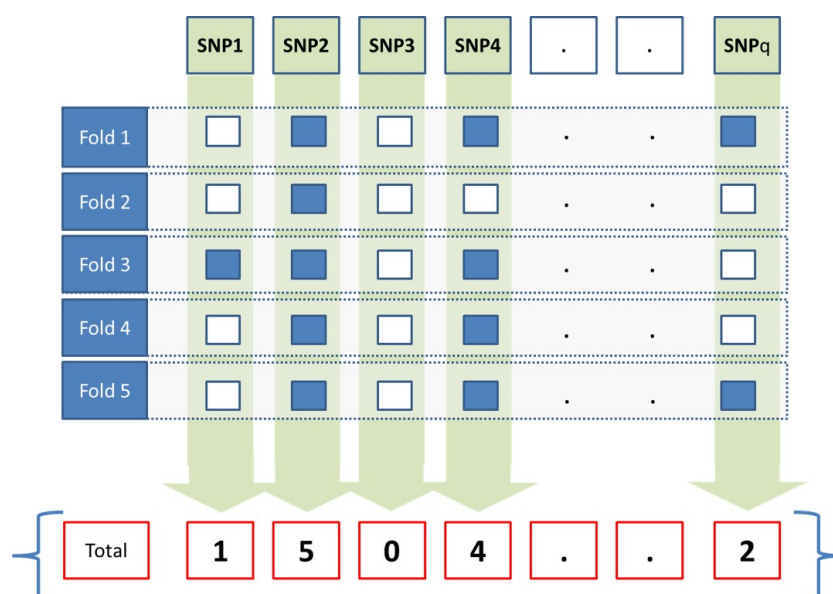


Figure 2. The feature selection method.

2.2.2. Deep Learning

Artificial neural networks (ANNs) are modelling tools inspired by the function of neurons in the human brain. These networks offer an alternative way to handle complex problems and are able to perform predictions for linear/nonlinear problems. Multi-Layer Perceptron (MLP) is one of the popular feed-forward neural networks that consists of an input, hidden, and output layer. In this work, the feedforward MLP consisting of one input layer and one output layer along with five hidden layers has been employed, as conceptualised in Figure 3. Each layer contains a number of neurons, which are interconnected in multiple layers by weighted connections.

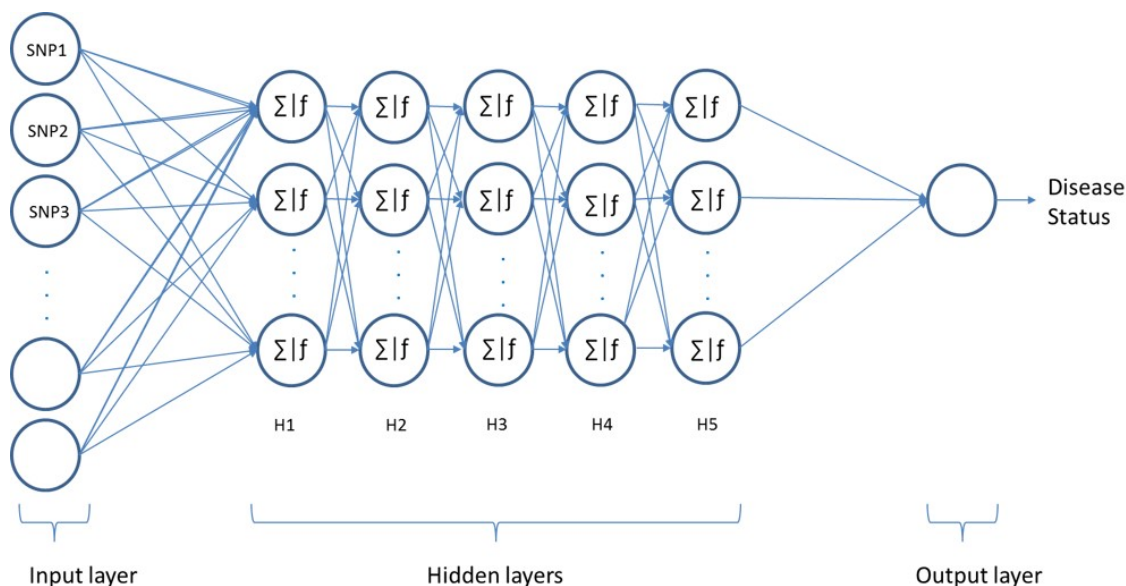


Figure 3. The multilayer perceptron framework.

The input feature vectors are passed through the multiple hidden layers downstream to the output layer [58]. The feature vectors are combined with weights to identify the informativeness of the inputs to the next layer. For any given neuron in layer L , the input is the sum of the weights for each neuron with a bias after applying an activation function in layer $L - 1$.

Given data input $x_i (i = 1, 2, 3, \dots, N)$, the neural model output y can be gained by Equation (7)

$$y = f\left(\sum_{i=1}^N W_i x_i + b_i\right) \quad (7)$$

where W is the model weight, b is the bias vector, and f is the activation function.

In this work, an MLP has been employed in order to identify patients with a certain disease, as conceptualised in Figure 3. The input layer consists of N nodes and considers SNPs as features. The output layer consists of one neuron (affected or healthy). The proposed models' hyperparameters have been optimised using a grid search of the k -fold cross-validation technique with $k = 3$. This technique can ensure how accurately the model would perform in practice and avoid overfitting. Since we are implementing our model on different diseases, a modular model consisting of multiple modules is used. While similarities between models are possible, training the models separately means that all the architectures are optimised using various hyperparameters. After implementing a grid-based search, all possible hyperparameter value combinations have been examined. The best performances have been achieved with five hidden layers and 512 neurons in each hidden layer for all dataset models. Different activation functions for the hidden layers were evaluated: the tanh function performed the best out of the examined activation functions in three datasets, namely RA, T1D, and T2D, while the relu function performed better in CAD and IBD. The best performance of the HT dataset has been achieved using the sigmoid function. The softmax function was used in the output layer for all models. The softmax activation function used in our model is presented in Equation (8).

$$\text{softmax}(x) = \frac{e^{(x_i)}}{\sum_{j=1}^k e^{(x_j)}} \quad (8)$$

where x is the input vector to the softmax function, x_i is the i th element of the input vector, and k is the number of classes.

Different optimizers were used as a learning algorithm. The Adam optimizer outperforms the compared optimisation functions in most models. For BD and HT datasets, the best optimizers were NADAM and rmsprop activation functions, respectively. However, for the other parameters, all models achieved their best performance using the same values. In order to avoid model overfitting, a dropout technique that drops neurons randomly along with their connections has been used with a probability of 0.6. The best performance was achieved using 200 epochs for all models. Finally, the best learning rate was achieved using a 0.001 learning rate. The proposed models can be validated using test data in order to demonstrate their high-performance ability. Possible hyperparameter values are given in Table 2.

Table 2. Hyperparameters and their range for MLP models.

Hyperparameter	Description	Range
Activation function	Neuron's activation function	Relu, Sigmoid, tanh
Optimizer	The optimisation algorithm that performs the learning process in a neural network	rmsprop, NADAM, ADAM, SGD
Epochs	Number of learning iterations	50, 100, 200, 300
Learning Rate	Weight change updated during learning	0.001, 0.0001, 0.00001
No. of hidden nodes	No. of neurons in the hidden layer	64, 128, 256, 512
Dropout	Dropping out nodes during training	0.2, 0.4, 0.6
Mini batch size	Group size submitted to model during training	16, 32, 64, 100

After building the networks and optimising the parameters on the seven aforementioned complex disease datasets, we came up with more than one model: the first one uses the same hyperparameters for three datasets (RA, T1D, and T2D), and the model uses the Adam optimizer and tanh activation function. A slight difference was implemented by using the relu activation function in the CAD and T1D datasets. On the other hand, two more models were implemented using NADAM, tanh and rmsprop, sigmoid as optimiser and activation functions, respectively. Finally, the proposed models can be validated using test data in order to demonstrate their high-performance ability.

The cross-entropy cost function, which is explained in Equation (9), has been used to estimate the output error.

$$Loss = -\frac{1}{N} \sum_{i=1}^N [t_i \log(p_i) + (1 - t_i) \log(1 - p_i)] \quad (9)$$

for N data points, where t_i is the truth value taking a value 0 or 1, and p_i is the Softmax probability for the i th data point.

2.2.3. Evaluation

Different experiments were conducted to evaluate the performance of the examined deep learning prediction architecture in terms of accuracy (Equation (10)), sensitivity (Equation (11)), precision (Equation (12)), F1-score (Equation (13)), AUC, and Matthews correlation coefficient (MCC) (Equation (14)). In order to compute the metrics, different values were calculated: 1. True positive (TP): the number of samples that were correctly identified to be corresponding to the targeted disease. 2. False positive (FP): the number of samples that were wrongly identified to correspond to the targeted disease. 3. False negative (FN): the number of samples that were wrongly classified as healthy. 4. True negative (TN): the number of samples that were correctly classified as healthy.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (10)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (11)$$

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$F1 \text{ score} = 2 \cdot \frac{Precision \cdot sensitivity}{Precision + sensitivity} \quad (13)$$

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (14)$$

Each dataset was split into two main parts: 70% for training the system and 30% for testing the system. Furthermore, in order to identify the optimum features subset, a five-fold cross-validation technique was implemented to train data during the feature selection process. To this end, the training samples were shuffled and split into five groups. The splitting kept the ratio of the classes similar in each group to that of the original dataset. Then, the experiment of selecting features was repeated five times, and at each fold we used one group for testing the model and the remaining four groups as a training set. The final classification performance results were computed using the 30% of the original dataset that was unseen by the feature selection and training phases.

At the end of the fifth fold of the feature selection process, each SNP had a weight value, depending on how many times the SNP has been selected. SNPs with weights larger than the threshold value α were selected for the final feature vector.

3. Results and Discussion

The threshold value was selected experimentally, by evaluating different values and selecting the optimal value, as shown in Figure 4 and Table 3. In most datasets, the best prediction accuracy was achieved using a threshold of 0.6. However, in predicting the risk of CAD and HT, the best performance was achieved using a threshold value of 0.8 and 1, respectively.

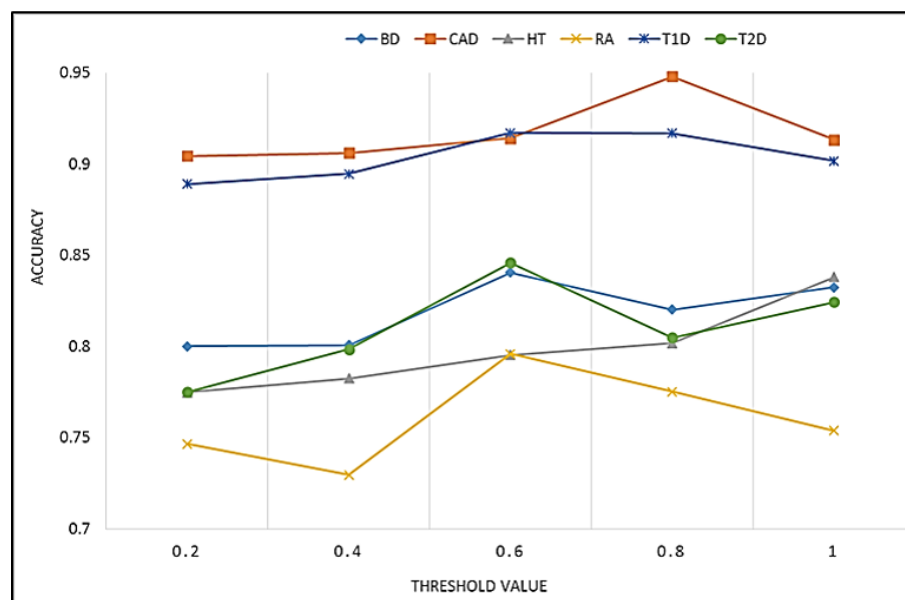


Figure 4. The performance evaluation using different thresholds over seven datasets.

Table 3. The number of selected features at each threshold value for all datasets.

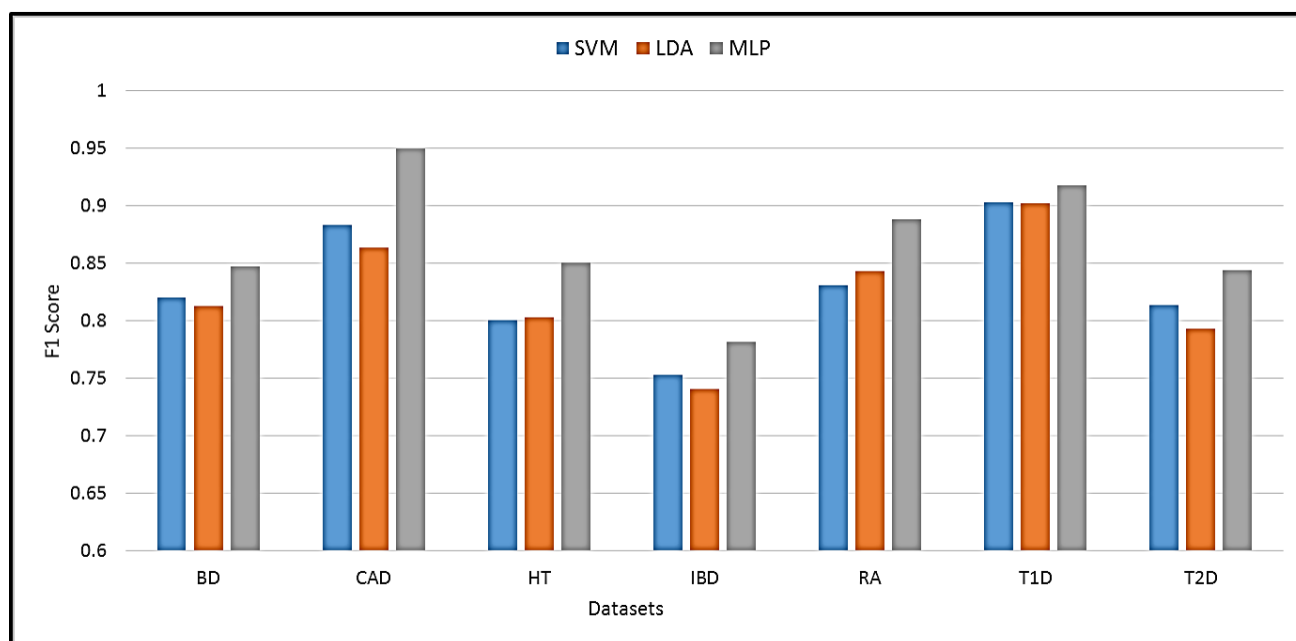
Fold	BD	CAD	HT	IBD	RA	T1D	T2D
0.2	1991	2053	1767	1988	1758	1603	2224
0.4	1167	1099	1147	1183	1128	1121	1121
0.6	830	794	878	832	882	907	750
0.8	602	607	695	597	705	764	555
1	410	447	513	400	527	605	350

The performance of the examined prediction deep neural network approach for each dataset in terms of accuracy, sensitivity, precision, F1 score, and MCC is presented in Table 4. The achieved results demonstrate the ability of the proposed disease risk prediction system to perform accurately. The affected samples were identified with an accuracy range between 0.796 to 0.948 for seven different complex disease datasets. Regarding the sensitivity and precision values, the proposed system was able to detect most patients in the datasets with high sensitivity values in most cases, ranging from 0.798 for HT disease to 0.934 for T1D disease. Moreover, very few healthy samples were identified as a case with precision values ranging from 0.83 to 0.966 for all diseases, apart from IBD disease, with a precision of 0.726. The proposed system performed the best in identifying the risk of CAD with an F1 score reaching 0.95. In predicting the risk of T1D, T2D, BD, HT, and RA, the F1 score of the proposed system ranged between 0.84 and 0.92. However, predicting the risk of IBD disease was the most challenging, achieving an F1 score of up to 0.782. For the MCC, the performance of our proposed system ranged between 0.606 and 0.901.

Table 4. Prediction performance for complex diseases.

	Accuracy	Sensitivity	Precision	F1-Score	MCC
BD	0.839	0.812	0.882	0.846	0.697
CAD	0.948	0.934	0.966	0.950	0.891
HT	0.838	0.798	0.904	0.848	0.685
IBD	0.796	0.847	0.726	0.782	0.606
RA	0.885	0.884	0.886	0.885	0.764
T1D	0.917	0.901	0.936	0.918	0.901
T2D	0.846	0.857	0.831	0.844	0.696

The proposed MLP was compared with other state-of-the-art machine learning techniques. The best performance with the compared techniques was achieved using SVM and linear discriminant analysis (LDA). A comparison of SVM, LDA, and the proposed MLP performance in terms of F1 score is shown in Figure 5. It is evident that MLP achieves the highest prediction performance with an improvement of 1.2% to 7.9% over SVM and LDA. The best improvement was achieved in the HT dataset, while the lowest was obtained in the T1D dataset.

**Figure 5.** Performance comparison of the proposed method against SVM and LDA classifiers.

Finally, we compared our proposed system prediction performance against other studies conducted on the WTCCC datasets as presented in Table 5. Comparing with these studies, we can guarantee that their systems dealt with datasets that have the same properties, the same number of controls and cases, and the same genotyping density. The AUC of the compared methods varied between 0.56 and 0.90 depending on the dataset and the algorithm. The proposed system outperformed the other frameworks for all datasets. The improvement of the proposed system in terms of AUC was less than 4% in identifying affected samples of T1D, T2D, and BD datasets. A better improvement of approximately 9% was achieved in predicting the risk of IBD. The risk prediction for RA, CAD, and HT was the best with an improvement of more than 15% over the best competitors.

Applying deep learning techniques to complex disease genomic datasets is not a trivial task, and the pre-processing of the data can be highly affected by many factors leading to a

severe impact on the final conclusions. The proposed framework was able to select a subset of high discriminative SNPs that contributed to improving the prediction ability. A number of SNPs that have been identified with high discriminative values in our proposed system have been previously identified to be associated with diseases in other published works. Out of the selected SNPs, 75 SNPs were identified to be highly correlated with different diseases in the original dataset [38], 23 SNPs were identified in a study conducted over the same dataset [59], and 9 SNPs were identified in a study conducted only on the HT dataset [60].

Table 5. Performance comparison of the proposed prediction system and studies conducted on the WTCCC dataset in terms of AUC.

Disease/Method	T1D	T2D	BD	IBD	CAD	RA	HT
Proposed Model	0.92	0.85	0.84	0.79	0.94	0.89	0.84
BootRank [44]	0.90	0.82	0.83	0.70	0.72	0.74	0.68
GWASRank [44]	0.88	0.69	0.68	0.67	0.72	0.75	0.65
LO, AC [46]	0.75	0.60	0.67	0.63	0.60	0.67	0.61
DeepCOMBI [49]	0.65	0.65	0.65	0.65	0.65	0.65	0.65
SVM [40]	0.82	-	-	-	-	0.71	-
GWASselect [47]	0.79	-	-	-	-	-	-
SVM, LR [32]	0.89	-	-	-	-	-	-
Forward ROC [48]	-	-	-	-	-	0.71	-
LR, SVM, RF, BN [34]	-	-	0.56	-	-	-	-
Elastic-net [15]	-	-	-	0.64	-	-	-
LR, AC, SVM [39]	-	-	-	-	0.60	-	-

4. Conclusions

In this work, a deep learning approach using MLP has been proposed to predict the risk of complex diseases based on genomic variations. The proposed approach exploits the JMI filter feature selection method in order to select a subset of SNPs with high discriminative power. The selected features are then fed to an MLP-based prediction algorithm to distinguish between healthy and affected samples. The proposed model has been evaluated on seven state-of-the-art datasets from WTCCC, UKNBS, and 58C. The experiment results demonstrate the superiority of the proposed model as compared to the traditional machine learning techniques, achieving an F1-score of 0.94. Moreover, the obtained results have been compared with state-of-the-art methods that were applied on the same datasets. An improvement in terms of an AUC of up to 22% compared to previous methods was achieved using the proposed approach. The proposed framework was also able to identify a number of SNPs that have high discriminative value and were previously identified to be linked with diseases in other published work. Taking into consideration the obtained prediction performance, as well as the performance of other methods proposed in the literature, it is evident that the proposed approach is applicable and efficient for complex disease risk prediction from SNP data.

Author Contributions: Conceptualization, H.A., R.A. and N.R.; methodology, H.A., R.A. and N.R.; software, H.A. and R.A.; validation, H.A., R.A. and N.R.; formal analysis, H.A., R.A. and N.R.; investigation, H.A., R.A. and N.R.; resources, N.R.; data curation, H.A. and R.A.; writing—original draft preparation, H.A., R.A. and N.R.; writing—review and editing, H.A. and R.A.; visualization, H.A. and R.A.; supervision, N.R.; project administration, N.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported through the Annual Funding track by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Project No. GRANT 3306).

Institutional Review Board Statement: Ethical review and approval were waived for this study due to all the data used being publicly available.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source datasets are publicly available from their respective sources.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Collins, F.S.; Brooks, L.D.; Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **1998**, *8*, 1229–1231. [\[CrossRef\]](#) [\[PubMed\]](#)
- Davis, S.; Pettengill, J.B.; Luo, Y.; Payne, J.; Shpuntoff, A.; Rand, H.; Strain, E. CFSAN SNP Pipeline: An automated method for constructing SNP matrices from next-generation sequence data. *PeerJ Comput. Sci.* **2015**, *1*, e20. [\[CrossRef\]](#)
- Visscher, P.M.; Brown, M.A.; McCarthy, M.I.; Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **2012**, *90*, 7–24. [\[CrossRef\]](#)
- International Parkinson Disease Genomics Consortium. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: A meta-analysis of genome-wide association studies. *Lancet* **2011**, *377*, 641–649. [\[CrossRef\]](#)
- Sladek, R.; Rocheleau, G.; Rung, J.; Dina, C.; Shen, L.; Serre, D.; Boutin, P.; Vincent, D.; Belisle, A.; Hadjadj, S.; et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **2007**, *445*, 881. [\[CrossRef\]](#) [\[PubMed\]](#)
- Tsai, F.J.; Yang, C.F.; Chen, C.C.; Chuang, L.M.; Lu, C.H.; Chang, C.T.; Wang, T.Y.; Chen, R.H.; Shiu, C.F.; Liu, Y.M.; et al. A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. *PLoS Genet.* **2010**, *6*, e1000847. [\[CrossRef\]](#)
- Li, H.; Gan, W.; Lu, L.; Dong, X.; Han, X.; Hu, C.; Yang, Z.; Sun, L.; Bao, W.; Li, P.; et al. A genome-wide association study identifies GRK5 and RASGRP1 as type 2 diabetes loci in Chinese Hans. *Diabetes* **2013**, *62*, 291–298. [\[CrossRef\]](#)
- Shiraishi, K.; Kunitoh, H.; Daigo, Y.; Takahashi, A.; Goto, K.; Sakamoto, H.; Ohnami, S.; Shimada, Y.; Ashikawa, K.; Saito, A.; et al. A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the Japanese population. *Nat. Genet.* **2012**, *44*, 900. [\[CrossRef\]](#)
- Hu, Z.; Wu, C.; Shi, Y.; Guo, H.; Zhao, X.; Yin, Z.; Yang, L.; Dai, J.; Hu, L.; Tan, W.; et al. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12. 12 and 22q12. 2 in Han Chinese. *Nat. Genet.* **2011**, *43*, 792. [\[CrossRef\]](#)
- Xu, J.; Mo, Z.; Ye, D.; Wang, M.; Liu, F.; Jin, G.; Xu, C.; Wang, X.; Shao, Q.; Chen, Z.; et al. Genome-wide association study in Chinese men identifies two new prostate cancer risk loci at 9q31. 2 and 19q13. 4. *Nat. Genet.* **2012**, *44*, 1231. [\[CrossRef\]](#)
- Eyre, S.; Bowes, J.; Diogo, D.; Lee, A.; Barton, A.; Martin, P.; Zhermakova, A.; Stahl, E.; Viatte, S.; McAllister, K.; et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **2012**, *44*, 1336. [\[CrossRef\]](#) [\[PubMed\]](#)
- Janssens, A.C.J.; van Duijn, C.M. Genome-based prediction of common diseases: Advances and prospects. *Hum. Mol. Genet.* **2008**, *17*, R166–R173. [\[CrossRef\]](#) [\[PubMed\]](#)
- Jostins, L.; Barrett, J.C. Genetic risk prediction in complex disease. *Hum. Mol. Genet.* **2011**, *20*, R182–R188. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kruppa, J.; Ziegler, A.; König, I.R. Risk estimation and risk prediction using machine-learning methods. *Hum. Genet.* **2012**, *131*, 1639–1654. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kooperberg, C.; LeBlanc, M.; Obenchain, V. Risk prediction using genome-wide association studies. *Genet. Epidemiol.* **2010**, *34*, 643–652. [\[CrossRef\]](#) [\[PubMed\]](#)
- Evans, D.T. A SNP Microarray Analysis Pipeline Using Machine Learning Techniques. Ph.D. Thesis, Ohio University, Athens, OH, USA, 2010.
- Qi, Q.; Liang, L.; Doria, A.; Hu, F.B.; Qi, L. Genetic predisposition to dyslipidemia and type 2 diabetes risk in two prospective cohorts. *Diabetes* **2012**, *61*, 745–752. [\[CrossRef\]](#) [\[PubMed\]](#)
- Goh, C.; Schumacher, F.; Easton, D.; Muir, K.; Henderson, B.; Kote-Jarai, Z.; Eeles, R. Genetic variants associated with predisposition to prostate cancer and potential clinical implications. *J. Intern. Med.* **2012**, *271*, 353–365. [\[CrossRef\]](#)
- Mittag, F.; Büchel, F.; Saad, M.; Jahn, A.; Schulte, C.; Bochdanovits, Z.; Simón-Sánchez, J.; Nalls, M.A.; Keller, M.; Hernandez, D.G.; et al. Use of support vector machines for disease risk prediction in genome-wide association studies: Concerns and opportunities. *Hum. Mutat.* **2012**, *33*, 1708–1718. [\[CrossRef\]](#)
- Botta, V.; Louppe, G.; Geurts, P.; Wehenkel, L. Exploiting SNP correlations within random forest for genome-wide association studies. *PLoS ONE* **2014**, *9*, e93379. [\[CrossRef\]](#)
- Maier, A.; Syben, C.; Lasser, T.; Riess, C. A gentle introduction to deep learning in medical image processing. *Z. Med. Phys.* **2019**, *29*, 86–101. [\[CrossRef\]](#)
- Kim, Y. Convolutional neural networks for sentence classification. *arXiv* **2014**, arXiv:1408.5882.
- Elgart, M.; Lyons, G.; Romero-Brufau, S.; Kurniansyah, N.; Brody, J.A.; Guo, X.; Lin, H.J.; Raffield, L.; Gao, Y.; Chen, H.; et al. Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Commun. Biol.* **2022**, *5*, 856. [\[CrossRef\]](#) [\[PubMed\]](#)
- Li, Y.; Huang, C.; Ding, L.; Li, Z.; Pan, Y.; Gao, X. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods* **2019**, *166*, 4–21. [\[CrossRef\]](#)

25. Tabares-Soto, R.; Orozco-Arias, S.; Romero-Cano, V.; Bucheli, V.S.; Rodríguez-Sotelo, J.L.; Jiménez-Varón, C.F. A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. *PeerJ Comput. Sci.* **2020**, *6*, e270. [[CrossRef](#)] [[PubMed](#)]
26. Alatrany, A.S.; Khan, W.; Hussain, A.J.; Mustafina, J.; Al-Jumeily, D. Transfer Learning for Classification of Alzheimer's Disease Based on Genome Wide Data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**. [[CrossRef](#)]
27. Liu, L.; Meng, Q.; Weng, C.; Lu, Q.; Wang, T.; Wen, Y. Explainable deep transfer learning model for disease risk prediction using high-dimensional genomic data. *PLoS Comput. Biol.* **2022**, *18*, e1010328. [[CrossRef](#)] [[PubMed](#)]
28. Abd El Hamid, M.M.; Omar, Y.M.; Shaheen, M.; Mabrouk, M.S. Discovering epistasis interactions in Alzheimer's disease using deep learning model. *Gene Rep.* **2022**, *29*, 101673. [[CrossRef](#)]
29. Uppu, S.; Krishna, A.; Gopalan, R.P. A Deep Learning Approach to Detect SNP Interactions. *JSW* **2016**, *11*, 965–975. [[CrossRef](#)]
30. Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A.W.; O'Sullivan, J.M. A review of feature selection methods for machine learning-based disease risk prediction. *Front. Bioinform.* **2022**, *2*, 927312. [[CrossRef](#)]
31. Ho, D.S.W.; Schierding, W.; Wake, M.; Saffery, R.; O'Sullivan, J. Machine learning SNP based prediction for precision medicine. *Front. Genet.* **2019**, *10*, 267. [[CrossRef](#)]
32. Wei, Z.; Wang, K.; Qu, H.Q.; Zhang, H.; Bradfield, J.; Kim, C.; Frackleton, E.; Hou, C.; Glessner, J.T.; Chiavacci, R.; et al. From disease association to risk assessment: An optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet.* **2009**, *5*, e1000678. [[CrossRef](#)] [[PubMed](#)]
33. Hajiloo, M.; Damavandi, B.; HooshSadat, M.; Sangi, F.; Mackey, J.R.; Cass, C.E.; Greiner, R.; Damaraju, S. Breast cancer prediction using genome wide single nucleotide polymorphism data. *BMC Bioinform.* **2013**, *14*, S3. [[CrossRef](#)]
34. Pirooznia, M.; Seifuddin, F.; Judy, J.; Mahon, P.B.; Potash, J.B.; Zandi, P.P.; Bipolar Genome Study (BiGS) Consortium. Data mining approaches for genome-wide association of mood disorders. *Psychiatr. Genet.* **2012**, *22*, 55. [[CrossRef](#)] [[PubMed](#)]
35. Alzubi, R.; Ramzan, N.; Alzoubi, H.; Amira, A. A hybrid feature selection method for complex diseases SNPs. *IEEE Access* **2017**, *6*, 1292–1301. [[CrossRef](#)]
36. Abraham, G.; Kowalczyk, A.; Zobel, J.; Inouye, M. Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genet. Epidemiol.* **2013**, *37*, 184–195. [[CrossRef](#)] [[PubMed](#)]
37. Guo, Y.; Wei, Z.; Keating, B.J.; Hakonarson, H. Machine learning derived risk prediction of anorexia nervosa. *BMC Med. Genom.* **2015**, *9*, 4. [[CrossRef](#)] [[PubMed](#)]
38. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **2007**, *447*, 661. [[CrossRef](#)]
39. Davies, R.W.; Dandona, S.; Stewart, A.F.; Chen, L.; Ellis, S.G.; Tang, W.W.; Hazen, S.L.; Roberts, R.; McPherson, R.; Wells, G.A. Improved prediction of cardiovascular disease based on a panel of single nucleotide polymorphisms identified through genome-wide association studies. *Circ. Cardiovasc. Genet.* **2010**, *3*, 468. [[CrossRef](#)]
40. Roshan, U.; Chikagoudar, S.; Wei, Z.; Wang, K.; Hakonarson, H. Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. *Nucleic Acids Res.* **2011**, *39*, e62. [[CrossRef](#)]
41. Behravan, H.; Hartikainen, J.M.; Tengström, M.; Pylkäs, K.; Winqvist, R.; Kosma, V.M.; Mannermaa, A. Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls. *Sci. Rep.* **2018**, *8*, 13149. [[CrossRef](#)]
42. Behravan, H.; Hartikainen, J.M.; Tengström, M.; Kosma, V.M.; Mannermaa, A. Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. *Sci. Rep.* **2020**, *10*, 11044. [[CrossRef](#)] [[PubMed](#)]
43. Mittag, F.; Römer, M.; Zell, A. Influence of Feature Encoding and Choice of Classifier on Disease Risk Prediction in Genome-Wide Association Studies. *PLoS ONE* **2015**, *10*, e0135832. [[CrossRef](#)] [[PubMed](#)]
44. Manor, O.; Segal, E. Predicting disease risk using bootstrap ranking and classification algorithms. *PLoS Comput. Biol.* **2013**, *9*, e1003200. [[CrossRef](#)] [[PubMed](#)]
45. Bennasar, M.; Hicks, Y.; Setchi, R. Feature selection using joint mutual information maximisation. *Expert Syst. Appl.* **2015**, *42*, 8520–8532. [[CrossRef](#)]
46. Evans, D.M.; Visscher, P.M.; Wray, N.R. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum. Mol. Genet.* **2009**, *18*, 3525–3531. [[CrossRef](#)]
47. He, Q.; Lin, D.Y. A variable selection method for genome-wide association studies. *Bioinformatics* **2010**, *27*, 1–8. [[CrossRef](#)]
48. Ye, C.; Cui, Y.; Wei, C.; Elston, R.C.; Zhu, J.; Lu, Q. A non-parametric method for building predictive genetic tests on high-dimensional data. *Hum. Hered.* **2011**, *71*, 161–170. [[CrossRef](#)]
49. Mieth, B.; Rozier, A.; Rodriguez, J.A.; Höhne, M.M.; Gornitz, N.; Müller, K.R. DeepCOMBI: Explainable artificial intelligence for the analysis and discovery in genome-wide association studies. *NAR Genom. Bioinform.* **2021**, *3*, lqab065. [[CrossRef](#)]
50. Rich, S.; Goodarzi, M.; Palmer, N.; Langefeld, C.; Ziegler, J.; Haffner, S.; Bryer-Ash, M.; Norris, J.; Taylor, K.; Haritunians, T.; et al. A genome-wide association scan for acute insulin response to glucose in Hispanic-Americans: The Insulin Resistance Atherosclerosis Family Study (IRAS FS). *Diabetologia* **2009**, *52*, 1326–1333. [[CrossRef](#)]
51. Michel, S.; Liang, L.; Depner, M.; Klopp, N.; Ruether, A.; Kumar, A.; Schedel, M.; Vogelberg, C.; von Mutius, E.; von Berg, A.; et al. Unifying candidate gene and GWAS Approaches in Asthma. *PLoS ONE* **2010**, *5*, e13894. [[CrossRef](#)]
52. Kang, G.; Childers, D.K.; Liu, N.; Zhang, K.; Gao, G. Genome-wide association studies of rheumatoid arthritis data via multiple hypothesis testing methods for correlated tests. *BMC Proc.* **2009**, *3*, S38. [[CrossRef](#)] [[PubMed](#)]

53. Uppu, S.; Krishna, A.; Gopalan, R. A review on methods for detecting SNP interactions in high-dimensional genomic data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2016**, *15*, 599–612. [[CrossRef](#)] [[PubMed](#)]
54. Miller, D.J.; Zhang, Y.; Yu, G.; Liu, Y.; Chen, L.; Langefeld, C.D.; Herrington, D.; Wang, Y. An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions. *Bioinformatics* **2009**, *25*, 2478–2485. [[CrossRef](#)] [[PubMed](#)]
55. Battiti, R. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* **1994**, *5*, 537–550. [[CrossRef](#)]
56. Meyer, P.E.; Schretter, C.; Bontempi, G. Information-theoretic feature selection in microarray data using variable complementarity. *IEEE J. Sel. Top. Signal Process.* **2008**, *2*, 261–274. [[CrossRef](#)]
57. Brown, G.; Pocock, A.; Zhao, M.J.; Luján, M. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *J. Mach. Learn. Res.* **2012**, *13*, 27–66.
58. Riedmiller, M.; Braun, H. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In Proceedings of the IEEE International Conference on Neural Networks, San Francisco, CA, USA, 28 March–1 April 1993; pp. 586–591.
59. Mieth, B.; Kloft, M.; Rodríguez, J.A.; Sonnenburg, S.; Vobruha, R.; Morcillo-Suárez, C.; Farré, X.; Marigorta, U.M.; Fehr, E.; Dickhaus, T.; et al. Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Sci. Rep.* **2016**, *6*, 36671. [[CrossRef](#)]
60. Pahikkala, T.; Okser, S.; Airola, A.; Salakoski, T.; Aittokallio, T. Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations. *Algorithms Mol. Biol.* **2012**, *7*, 11. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.