Improving Type 2 Diabetes Phenotypic Classification by Combining Genetics and Conventional Risk Factors

Basma Abdulaimma
Department of Computer Science
Liverpool John Moores University
Liverpool, United Kingdom
b.t.abdulaimma@2015.ljmu.ac.uk

Abir Hussain
Department of Computer Science
Liverpool John Moores University
Liverpool, United Kingdom
A.Hussain@ljmu.ac.uk

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

Dhiya Al-Jumeily
Department of Computer Science
Liverpool John Moores University
Liverpool, United Kingdom
D.Aljumeily@ljmu.ac.uk

Paulo Lisboa
Department of of Applied Mathematics
Liverpool John Moores University
Liverpool, United Kingdom
P.J.Lisboa@ljmu.ac.uk

De-Shuang Huang
College of Electronics and Information Engineering
Tongji University
Shanghai 201804, China
dshuang@tongji.edu.cn

Naeem Radi
Al Khawarizmi International College
Dhabi, United Arab Emirates
n.radi@khawarizmi.com

***Abstract*—Type 2 Diabetes condition is a multifactorial disorder involves the convergence of genetics, environment, diet and lifestyle risk factors. This paper investigates genetic and conventional (clinical, sociodemographic) risk factors and their predictive power in classifying Type 2 Diabetes. Six statistically significant Single Nucleotide Polymorphisms (SNPs) associated with Type 2 Diabetes are derived by conducting logistic association analysis. The derived SNPs in addition to conventional risk factors are used to model supervised machine learning algorithms to classify cases and controls in genome wide association studies (GWAS). Models are trained using genetic variable analysis, genetic and conventional variable analysis, and conventional variable analysis. The results demonstrate of the three models, higher predictive capacity is evident when genetic and conventional predictors are combined. Using a Random Forest classifier, the Area Under the Curve=73.96%, Sensitivity=68.42%, and Specificity=78.67%.**

Keywords—Clinical data, Genetics, Machine Learning, Single Nucleotide Polymorphism, Type 2 Diabetes

I. INTRODUCTION

Currently, the prevalence of Type 2 Diabetes (T2D) throughout the world has reached epidemic proportions. In 2012, the World Health Organization (WHO) [1] estimated that 1.5 million deaths were directly attributed to diabetes, and that by 2030 diabetes will be the seventh leading cause of mortality worldwide [2]. T2D is the most predominant form of all types of diabetes [1]. T2D (also known as insulin resistance) is a chronic disease that occurs as a consequence of the ineffective use of insulin by body cells [1]. T2D remains the leading cause of serious long-term health complications [3]. It is responsible for most cases of blindness (Diabetic retinopathy), kidney failure and lower limb amputation [1]. Moreover, high glucose levels (raised blood sugar) or Hyperglycemia in the bloodstream can damage blood vessels which increase the likelihood of atherosclerosis (cardiovascular disease) and stroke and can cause nerve damage [3]. Until recently, T2D was recognized only in people who are over the age of 40, but currently children are also being diagnosed with T2D [4].

Researchers have indicated that T2D results from the convergence of genetics, environment, diet and lifestyle choices [5]. Various risk factors are involved in the development of T2D including obesity and overweight (with a body mass index (BMI) of 30 or more), family history, older age (people over the age of 40), ethnicity, and physical inactivity [6].

Since the completion of the Human Genome Project in 2003, researchers have confirmed that among the 3 billion base pairs of DNA, 99.9% are remarkably similar [7] with the remaining 0.1% making an individual unique. The 0.1% of variations are termed Single Nucleotide Polymorphisms (SNPs). A SNP is a single base-pair change in the genetic code (Deoxyribonucleic Acid (DNA Sequence)), and it is the main cause of human genetic variability [8]. Genotyping technology has facilitated rapid progress in genome-wide association studies (GWAS) typically used to study SNPs and their prevalence within and across different population groups [9]. More specifically, GWAS has seen widespread use in studies that investigate the genetic architecture of human disease in the entire genome [10]. Within these kinds of studies genetic markers that show evidence of increased predisposition to a complex disease, such as T2D and related traits, are identified as being important for furthermore in-depth analysis. Identifying high risk SNPs allows researchers to investigate the interactions between genes, the environment, and sociodemographic factors to provide a complete understanding of specific diseases, treatment options and prevention. In the case of T2D this might require a change in diet and lifestyle to prevent or delay the onset of the disease in high-risk individuals [11], [12].

Machine learning and predictive modeling have become important tools in a variety of medical domains [13], [14] particularly biomedical research [15], [16], [17], [18]. Investigators have successfully applied machine learning to model the relationships between combinations of SNPs, the environmental, clinical factors and human disease. This paper builds on early positive results in this area and investigates genetic and clinical factors and their relation to T2D. Utilizing three different machine models, the predictive capacity of SNPs, clinical data and both of these combined, are evaluated to determine their predictive capacity in distinguishing between cases and controls in GWAS.

The remainder of this paper organized as follows. Section II provides details about the database used followed by the steps conducted in quality control and association analysis. Additionally, details about the classification models adopted and the evaluation techniques considered are demonstrated. The results are presented in section III, while the findings are discussed in section IV, before the paper is concluded in section V.

II. Materials and Methods

A. Data Description

The data, for this study, was obtained following authorized access to the Database of Genotypes and Phenotypes (dbGaP) [19]. The Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS) in T2D (Study Accession: phs000091.v2.p1) are used in this paper. The NHS and HPFS cohorts are 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 NHS was started in 1976; participants include 121,700 female registered nurses aged between 30 and 55 years of age that reside in 11 U.S states. The HPFS study was established in 1986; participants include 51,529 male health professionals aged between 40 and 75 years that reside in 50 U.S states. All participants responded to a mailed questionnaire requesting information related to their medical history and lifestyle characteristics. Since then, on a 2 to 4-year cycle, cohort members have been asked to provide dietary information using validated semi-quantitative food frequency questionnaires. Participants were also requested to provide blood samples, in which 32,826 members of the NHS and 18,225 members of the HPFS responded. The case and control participants were selected from those who provided a blood sample. Cases were identified as those who reported themselves to be affected by T2D, and it was confirmed by a medical record validation questionnaire. Controls were defined as those without diabetes. The DNA of case and control participants were genotyped at the Broad Centre for Genotyping and Analysis (CGA) using the Affymetrix Genome-Wide Human 6.0 array.

A total of 6041 NHS and HPFS case-control subjects with genotype information across 909622 SNPs successfully passed the initial quality control at the Board CGA and were used as a final version of the dataset. The NHS subjects consist of 1581 T2D cases and 1854 controls, and the HPFS subjects comprise 1232 T2D cases and 1374 controls. The NHS and HPFS participants belong to one of four racial categories (White, African-American, Asia or Other). Participants are predominantly white representing 97.4% and 96% of the NHS and HPFS subjects, respectively.

Clinical and dietary data information is also collected for NHS and HPFS participants including age, gender, Body Mass Index (BMI), alcohol intake, smoking status, physical activity, height, weight, family history of diabetes among first degree relatives, high blood pressure, high blood cholesterol, polyunsaturated fat intake, magnesium intake, cereal fibre intake, and glycaemic load. A comprehensive description of GENEVA NHS and HPFS dataset can be found in the quality control report for the GENEVA NHS and HPFS T2D project [20], [21].

B. Data Preprocessing

Data quality control (QC) and preliminary analysis are performed using PLINK v1.07 and v1.9 [22] for Windows. PLINK is also used to merge the NHS and HPFS datasets (NHS and HPFS participants were genotyped using the Affymetrix Genome-Wide Human 6.0 array) and filtering procedures. Before QC, the 0 Chromosome was removed, and non-T2D participants, i.e. other types of diabetes (65 NHS, 68 HPFS), the HapMap controls (44 NHS, 29 HPFS) and those belonging to ethnicity other than white (61 NHS, 103 HPFS) were excluded from the study. This study is restricted to white ancestry to reduce potential bias due to population stratification. The dataset was subjected to pre-established quality control protocols as recommended in [23]. In addition, quality control parameters are tuned to meet the requirements of the analysis presented in this study. Quality control assessments for individuals and genetic data are conducted separately.

Individual QC: Samples with discordant sex information (homozygosity rate between 0.2 and 0.8) were identified resulting in 14 samples being 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 resulting in 131 individuals being 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 Marker QC: Genetic Markers (SNPs) that met any of the following criteria were removed from the analysis. SNPs with excessive missing data rates were identified resulting in 29 SNPs being excluded. 116863 variants with missing genotype rate of 0.01 and 178004 variants with minor allele frequency (MAF) < 0.05 were removed. 2248 variants removed due to Hardy-Weinberg Equilibrium (HWE) with p-value < 0.001 in control samples. Following the QC steps, there were 5393 individuals (2481 cases, 2912 controls) and 608342 markers with a 0.961665 genotype rate in the remaining samples.

C. Association Analysis

For association analysis, the case-control study design is

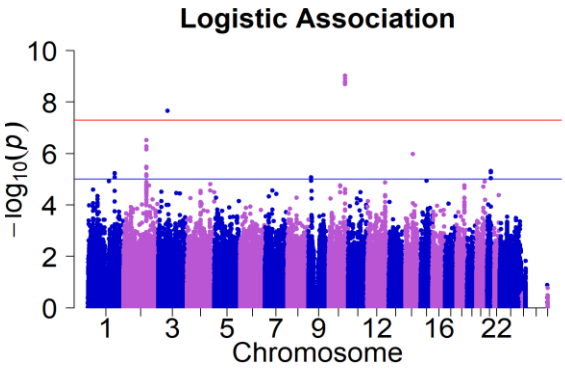

Fig. 1. Manhattan Plot for Logistic Regression Analysis. Showing the SNPs that reached Bonferroni Level of Significant, Red Line.

TABLE I. SNPS FROM LOGISTIC ASSOCIATION

Chr	Gene	SNP	P-Value
10	*TCF7L2*	rs4132670	9.253×10^{-10}
10	*TCF7L2*	rs12243326	1.164×10^{-9}
10	*TCF7L2*	rs12255372	1.591×10^{-9}
10	*TCF7L2*	rs7901695	1.701×10^{-9}
10	*TCF7L2*	rs4506565	1.992×10^{-9}
3	*ADAMTS9*	rs2371765	2.206×10^{-8}

used to obtain statistically significant SNPs associated to T2D. Allelic and Logistic association analyses are conducted for preliminary and confirmatory explorations, respectively. Six SNPs from logistic association analysis reached Bonferroni corrected genome-wide significance threshold of 5×10^{-8} including (rs4132670, rs12243326, rs12255372, rs7901695, rs4506565, rs2371765) as demonstrated in Table I, these are located in chromosome (Chr) 10 and 3.

Fig. 1 illustrates the logistic regression model for association analysis showing the level of statistical significance as measured by the negative log of the corresponding p-value, for each SNP. The red line corresponds to the Bonferroni level of significance and the SNPs that reached this threshold were considered to be statistically significant. These six SNPs were extracted and reformatted to construct a new dataset that is used for T2D classification and risk prediction using several machine learning algorithms.

D. Classification Models

Seven supervised machine learning algorithms have been selected for binary classification of T2D (control = 1, case = 2). The performance of each model is measured using the Area Under the Curve (AUC), Sensitivity and Specificity values. The dataset is split randomly into training (80%) to train the models and testing (20%) to evaluate model performance on unseen data. Several evaluations are considered which includes modelling using genetic features only, genetics and clinical features, and clinical features only.

10-fold cross-validation with 3 repetitions is employed in this analysis to repeatedly split the training data into 10-fold repeated 3 times.

Sensitivity and specificity are used to represent the number of correctly identify case and control participants. Sensitivity refers to the true positive rate which describes the ability of the test to correctly classify people with T2D. While Specificity describes the true negative rate which is the ability of the test to correctly classify people without T2D [24].

Furthermore, in this analysis the area under the curve (AUC) and the receiver operating characteristic curve (ROC curve) are used to assess and compare classifiers performance, both quality measures are widely used to assess binary classifiers [25].

Seven supervised machine learning algorithms that are specific for modelling dichotomous data are investigated in this paper. The selected machine learning algorithms fall into two categories either developed to model the non-linear or the linear effects. The former includes Stochastic Gradient Boosting (GBM), Support Vector Machines with Radial Basis Function Kernel (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), Classification and Regression Trees (CART), Monotone Multi-Layer Perceptron Neural Network (MONMLP). While the later includes Lasso and Elastic-Net Regularized Generalized Linear Models (GLMNET). The implementation, comparison, and the evaluation of the predictive classification models were performed using R software specifically caret package [26].

Each of the machine learning model mentioned above is automatically tuned during models' training to best adapt for a given dataset through an automatic grid search. Table II demonstrates tuning parameters for three analyses: genetic analysis, genetic and clinical analysis, clinical analysis. These tuning parameters values were selected among different options since they optimized the ROC values for the models.

III. RESULTS

Several analyses are considered to investigate the risk prediction of T2D including genetic features, genetic with clinical features, and clinical features only.

The first analysis was conducted using genomic features only; these include rs4132670, rs12243326, rs12255372, rs7901695, rs4506565, and rs2371765. The results presented in Table III show that sensitivities and specificities are imbalanced for all the models, sensitivities are lower than specificities. This indicates that the selected features for these models are inadequate at distinguishing between cases and controls. This analysis also reveals that the performance using AUC for linear and nonlinear classifiers are almost the same ranging between 57.09% for the RF and SVM classifiers and 57.84% for the KNN. Fig. 2 illustrates the ROC curve for the chosen models.

A separate analysis is conducted using clinical variables only these include Body Mass Index (BMI), alcohol intake (Alcohol), smoking status (SMK), physical activity (ACT), family history of diabetes (Famdb), high blood pressure (Hbp), high blood cholesterol (Chol), AGE and SEX. The results in Table IV show that the RF classifier yields the best accuracy measure of 72.41%. Although RF produced the best AUC performance, the model can classify unaffected (control) better than affected (case) classes with 68.42% and 75.81% for sensitivity and specificity, respectively. The AUC values for KNN and RPART

TABLE II. TUNING PARAMETER FOR MODELS

Classifier	Parameters	Best Tuning GWAS	Best Tuning Clinical	Best Tuning GWAS&Clinical
GLMNET	Alpha lambda	alpha=0.1 lambda= 0.008532955	alpha=0.55 lambda= 0.003759095	alpha=1 lambda= 0.003759095
GBM	n.trees interaction.depth shrinkage n.minobsinnode	n.trees = 50 interaction.depth = 1 shrinkage = 0.1 n.minobsinnode = 10	n.trees = 150 interaction.depth = 1 shrinkage = 0.1 n.minobsinnode = 10	n.trees = 150 interaction.depth = 1 shrinkage = 0.1 n.minobsinnode = 10
SVM	Sigma C	sigma= 0.2731565 C = 0.25	sigma= 0.08706031 C = 0.25	sigma= 0.04792686 C = 1
KNN	k	k = 9	k = 9	k = 9
RF	mtry	mtry = 2	mtry = 2	mtry = 2
RPART	cp	cp= 0.001526718	cp= 0.01156677	cp = 0.01156677
MONMLP	hidden1 n.ensemble	hidden1 = 1 n.ensemble = 1	hidden1 = 1 n.ensemble = 1	hidden1 = 1 n.ensemble = 1

are lower than other classifiers. However, RPART is the only classifier with sensitivity higher than specificity (70.11%, 68.44%) which means that RPART model can better separate cases than controls. Fig. 3 shows the ROC curve for the selected models.

A combination of six genetic variables with nine clinical variables is used as input features for the third analysis. The results in Table V show that the best classification accuracy of 73.96% was obtained by the RF algorithm. The AUC values for GLMNET, GBM, SVM, KNN, RF, RPART, MONMLP with this analysis yielded better results than using clinical or genomic data separately. As illustrated in Fig. 5, the predictive values of the machine learning models used in this investigation are due to clinical data, with slight evidence arising from genetic data. Body Mass Index (BMI) was significantly important for all models apart from GLMNET. Moreover, the importance of other clinical variables including Famdb, Hbp, Chol, SMK, Sex, Alcohol, ACT, and AGE appeared varied among these seven models. For the RPART model, the rank features for ACT, SMK, AGE, and SEX seemed completely trivial. For the GLMNET model the Alcohol, ACT, and AGE were considered not relevant.

The importance of genetic variables, in relation to the predictive values for these seven algorithms, is varied, but they always proved to be less relevant in comparison to clinical variables. Although all six genetic variables are used by SVM, RF, and MONMLP, their rank measurement is low. For GLMNET, GBM and RPART not all genetic variables were considered, and they show minor to no influence on the predictive results. Fig. 4 presents the ROC curve for the selected models.

TABLE III. PREDICTIVE RESULTS FOR GENETIC ANALYSIS

Classifier	Sensitivity	Specificity	Accuracy
GLMNET	0.2587	0.8417	0.5746
GBM	0.2546	0.8399	0.5718
SVM	0.2424	0.8485	0.5709
KNN	0.2668	0.8417	0.5784
RF	0.2505	0.8417	0.5709
RPART	0.2607	0.8382	0.5737
MONMLP	0.2668	0.8313	0.5728

TABLE IV. PREDICTIVE RESULTS FOR CLINICAL ANALYSIS

Classifier	Sensitivity	Specificity	Accuracy
GLMNET	0.6189	0.8065	0.7202
GBM	0.6484	0.7706	0.7144
SVM	0.6526	0.7778	0.7202
KNN	0.5642	0.7151	0.6457
RF	0.6842	0.7581	0.7241
RPART	0.7011	0.6703	0.6844
MONMLP	0.6716	0.7634	0.7212

TABLE V. PREDICTIVE RESULTS FOR GENETIC AND CLINICAL

Classifier	Sensitivity	Specificity	Accuracy
GLMNET	0.6484	0.8029	0.7318
GBM	0.6632	0.7742	0.7231
SVM	0.6737	0.7760	0.7289
KNN	0.5684	0.7133	0.6467
RF	**0.6842**	**0.7867**	**0.7396**
RPART	0.7011	0.6703	0.6844
MONMLP	0.6821	0.7706	0.7299

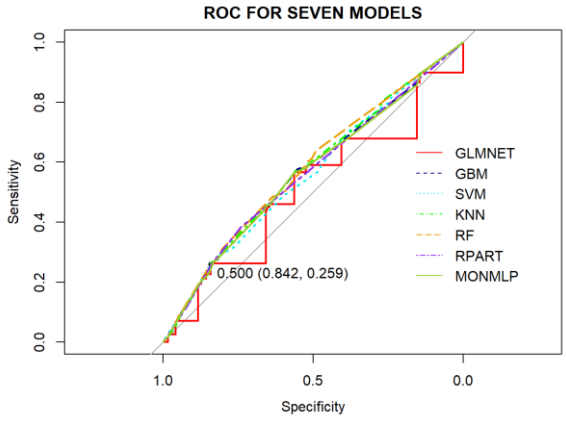

Fig. 2. ROC Curve for Seven Models using Genetic Features.

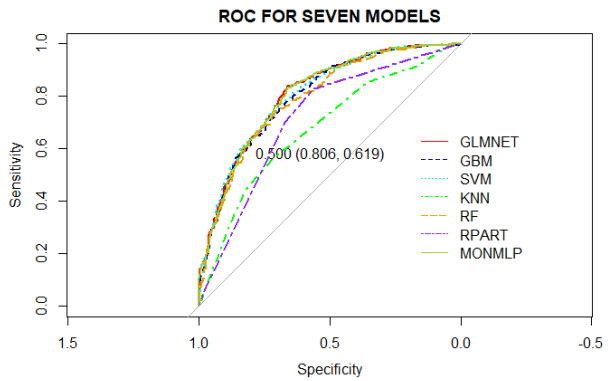

Fig. 3. ROC Curve for Seven Models using Clinical Features.

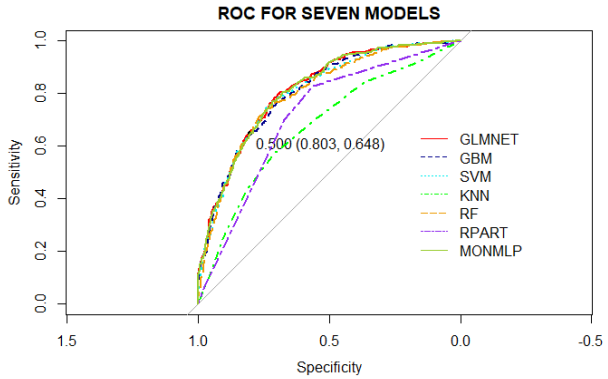

Fig. 4. ROC Curve for Seven Models using a Combination of Genetic and Clinical Features.

IV. Discussion

In this paper, our interpretation of the results is predominantly based on investigation conducted to determine the most relevant features for the classification of T2D in a case-control study. Genetic variables obtained from logistic regression association analysis, mainly SNPs variables, and clinical/sociodemographic variables are investigated to effectively understand and identify predisposition to T2D using advanced machine learning algorithms.

In the first analysis, genomic data variables extracted from logistic association analysis consisting of the six most significant SNPs are utilized as input features for the machine learning models. In general, as shown in Table III the classification accuracy of all seven machine learning models are low and almost show similar accuracy values ranging from 57.09% for RF to 57.84% for KNN. The low values of the predictive accuracy for the selected models is an indication that genomic data particularly SNPs passed GWAS failed to classify case and control observations, and these are often due to the fact that these SNPs are false positives. So far, the prediction of disease risk based on highly significant association SNPs demonstrated little predictive power [27]. This can be explained due to the limited heritability [28], which means how much of the phenotypic variance (combines the genotype variance with the environmental variance) is due to genetic variance [29].

A much higher predictive accuracy is obtained using clinical variables solely. Among the selected models, the RF achieved the best accuracy measure at 72.41% with 68.42% for sensitivity and 75.81% for specificity. Moreover, the predictive accuracy when employing both genomic and clinical data as input features showed satisfactory results as the RF classifier again achieved the best results at 73.96%. Comparatively, GLMNET, GBM, SVM, KNN, RF, RPART, MONMLP yielded better results than using clinical or genomic data separately. The interpretation of the results suggested that the improvement of classification prediction accuracy for all classifiers is entirely due to clinical variables, with no predictive value emerging from genotype variables alone. This is confirmed through the use of variable importance as illusterated in Fig. 5. Although, the variables for each model showed the disparity in relation to their rank measurement. Variable importance of the tested models shows that clinical data specifically BMI is the most associated variable in comparison to other features including clinical and genetic data. Although the predictive power is mainly due to the clinical variables, however, we could claim that combining genetic and clinical information might have more significant utility for T2D prediction than employing genetic or clinical data separately. For instance, RF classification accuracy values improved dramatically from 57.09% for genetic variables to 72.41% and 73.96% for clinical variables and the joint effects of genetic and clinical variables respectively.

In the clinical analysis and the analysis for the joint effects of genetic and clinical data the AUC for RF attained the best results (72.41% for clinical, 73.96% for the joint of genetic and clinical) in comparison to other models. The reason for this is that RF algorithm is a randomized decision tree-based ensemble [30]. RF trees are typically grown deeply (hundreds to thousands of trees) and each tree is grown using bootstrap aggregating or bagging to the training algorithm. The prediction of unseen data is based on the majority voting for classification. The RF algorithm is generally favoured in the genomic domain as deep trees promote low bias, while bootstrap aggregation improves the performance of the final model because bootstrap sampling is able to de-correlate the trees so that it reduces variance [30].

Our genetic-based prediction analysis showed little predictive power when employing SNPs found to have genome-wide significance as inputted features. In future work to optimize predictive accuracy; it would be interesting to drop the suggestive association threshold of $p < 1\times10^{-5}$ to increase the number of SNPs utilize in the classification analysis as demonstrated in [31][32]. Wei *et al.* [31] and Gul *et al.* [32] found that much higher predictive accuracy is obtained when increasing the number of SNPs, and comparatively poorer performance is attained when including only SNPs above genome-wide significance threshold. However, increasing the number of SNPs introduces an additional layer of complexity in machine learning modelling construction including the problem of multicollinearity [33], and the problem of dimensionality [34]. Consequently, an alternative approach such as deep learning needs to be investigated. The motivation of considering the application of deep learning is that the ability to transform big data into valuable knowledge through its characteristic of automatic feature learning in which performing feature extraction at multiple levels of abstraction that allow a system to

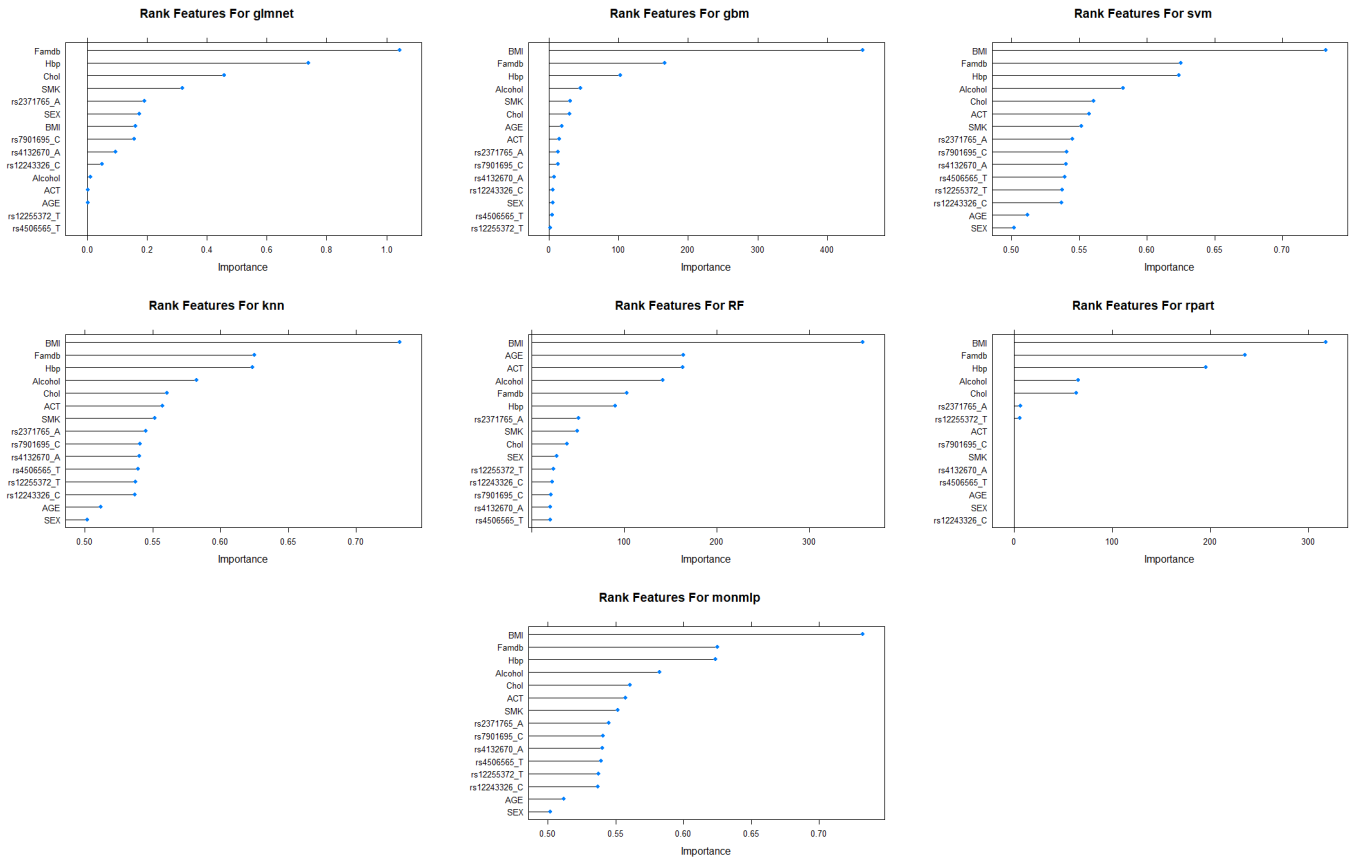

Fig. 5. Variable Importance Plots for Each Model.

learn complex functions that mapping the input features to the output directly from raw data. In addition to artificial neural networks of multiple nonlinear layers, the hierarchical representations of data can be explored with increasing levels of abstraction [35].

V. Conclusion

We investigated the contribution of genotypic risk factors and conventional risk factors including clinical, and sociodemographic factors for the classification of T2D in case-control cohorts. This study used existing datasets provided by the Genotypes and Phenotypes (dbGap) database. Various stringent quality control assessment steps followed by logistic regression association analysis are performed to find the top-ranked significant SNPs associated with T2D. Seven supervised machine learning algorithms are used to conduct three analyses considering genomic data only, the combination of genetic and clinical data, and lastly clinical data only. The simulation results revealed that genetic data analysis achieved the predictive performance of 57.84% for K-Nearest Neighbor. While for clinical data analysis and the joint effects of genetic and clinical data analysis, Random Forest obtained the best predictive accuracy of 72.41% and 73.96%, respectively. Using genotype variables alone significantly reduced the predictive classification accuracy in comparison to the joint effects of genetic and clinical variables analysis. The interpretation of the results suggested that the improvement of classification prediction accuracy for all classifiers are entirely due to clinical variables, with no predictive value emerging from genotype variables alone.

Acknowledgments

The authors would like to thank Al Khawarizmi International College for the financial support for this research. The dataset(s) utilized 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).

References

[1] World Health Organization, “Global Report on Diabetes,” Isbn, vol. 978, p. 88, 2016.

[2] C. D. Mathers and D. Loncar, “Projections of global mortality and burden of disease from 2002 to 2030,” PLoS Med., vol. 3, no. 11, pp. 2011–2030, 2006.

[3] 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.

[4] 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.

[5] 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.

[6] P. Z. Gatzeva-topalova, L. R. Warner, A. Pardi, and M. Carlos, "Analysis of Candidate Genes on Chromosome 20q12-13.1 Reveals Evidence for BMI Mediated Association of PREX1 with Type 2 Diabetes in European Americans," vol. 18, no. 11, pp. 1492–1501, 2011.

[7] M. Hattori, "Finishing the euchromatic sequence of the human genome," Tanpakushitsu Kakusan Koso., vol. 50, no. 2, pp. 162–168, 2005.

[8] D. Altshuler, E. Lander, and L. Ambrogio, "A map of human genome variation from population scale sequencing," Nature, vol. 476, no. 7319, pp. 1061–1073, 2010.

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

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

[11] E. E. Korkiakangas, M. A. Alahuhta, and J. H. Laitinen, "Barriers to regular exercise among adults at high risk or diagnosed with type 2 diabetes: A systematic review," Health Promot. Int., vol. 24, no. 4, pp. 416–427, 2009.

[12] A. J. Cooper, S. J. Sharp, M. A. Lentjes, R. N. Luben, K. T. Khaw, N. J. Wareham, and N. G. Forouhi, "A prospective study of the association between quantity and variety of fruit and vegetable intake and incident type 2 diabetes," Diabetes Care, vol. 35, no. 6, pp. 1293–1300, 2012.

[13] P. Fergus, A. Hussain, D. Hignett, D. Al-Jumeily, K. Abdel-Aziz, and H. Hamdan, "A machine learning system for automated whole-brain seizure detection," Appl. Comput. Informatics, vol. 12, no. 1, pp. 70–89, 2016.

[14] A. J. Hussain, P. Fergus, H. Al-Askar, D. Al-Jumeily, and F. Jager, "Dynamic neural network architecture inspired by the immune algorithm to predict preterm deliveries in pregnant women," Neurocomputing, vol. 151, no. P3, pp. 963–974, 2015.

[15] M. W. Libbrecht and W. S. Noble, "Machine learning in genetics and genomics," Nat. Rev. Genet., vol. 16, no. 6, pp. 321–332, 2017.

[16] A. K. Tanwani, J. Afridi, M. Z. Shafiq, and M. Farooq, "Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets," pp. 1–12, 2009.

[17] C. A. C. Montanez, P. Fergus, A. Hussain, D. Al-Jumeily, B. Abdulaimma, J. Hind, and N. Radi, "Machine learning approaches for the prediction of obesity using publicly available genetic profiles," Proc. Int. Jt. Conf. Neural Networks, vol. 2017–May, pp. 2743–2750, 2017.

[18] J. Hind, A. Hussain, D. Al-Jumeily, B. Abdulaimma, C. A. C. Montanez, and P. Lisboa, "A robust method for the interpretation of genomic data," Proc. Int. Jt. Conf. Neural Networks, vol. 2017–May, pp. 3385–3390, 2017.

[19] 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.

[20] "GENEVA Nurses' Health Study Type 2 Diabetes Project report," 2009.

[21] "GENEVA - Health Professionals Follow-up Study - Type 2 Diabetes project Quality control report," 2009.

[22] 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.

[23] C. a Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, P. Morris, and K. T. Zondervan, "Data quality control in genetic case-control association studies," Nat. Protoc., vol. 5, no. 9, pp. 1564–1573, 2011.

[24] R. Trevethan, "Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice," Front. Public Heal., vol. 5, no. November, pp. 1–7, 2017.

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

[26] M. Kuhn, "Building Predictive Models in R Using the caret Package," J. Stat. Softw., vol. 28, no. 5, pp. 1–26, 2008.

[27] F. Mittag, F. Büchel, M. Saad, A. Jahn, C. Schulte, Z. Bochdanovits, J. Simón-Sánchez, M. A. Nalls, M. Keller, D. G. Hernandez, J. R. Gibbs, S. Lesage, A. Brice, P. Heutink, M. Martinez, N. W. Wood, J. Hardy, A. B. Singleton, A. Zell, T. Gasser, and M. Sharma, "Use of support vector machines for disease risk prediction in genome-wide association studies: Concerns and opportunities," Hum. Mutat., vol. 33, no. 12, pp. 1708–1718, 2012.

[28] F. Dudbridge, "Power and Predictive Accuracy of Polygenic Risk Scores," PLoS Genet., vol. 9, no. 3, 2013.

[29] 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.

[30] X. Chen and H. Ishwaran, "Random Forests for Genomic Data Analysis," Genomics, vol. 99, no. 6, pp. 323–329, 2013.

[31] 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.

[32] H. Gul, Y. Aydin Son, and C. Acikel, "Discovering missing heritability and early risk prediction for type 2 diabetes: A new perspective for genome-wide association study analysis with the Nurses??? Health Study and the Health Professionals??? Follow-Up Study," Turkish J. Med. Sci., vol. 44, no. 6, pp. 946–954, 2014.

[33] S. Waaijenborg and A. H. Zwinderman, "Correlating multiple SNPs and multiple disease phenotypes: Penalized non-linear canonical correlation analysis," Bioinformatics, vol. 25, no. 21, pp. 2764–2771, 2009.

[34] N. Sharma and K. Saroha, "Study of dimension reduction methodologies in data mining," Int. Conf. Comput. Commun. Autom. ICCCA 2015, pp. 133–137, 2015.

[35] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," Brief. Bioinform., p. bbw068, Jul. 2016.