Machine Learning Approaches for the Prediction of Obesity using Publicly Available Genetic Profiles

Casimiro Aday Curbelo Montañez, Paul Fergus,
Abir Hussain, Dhiya Al-Jumeily, Basma
Abdulaimma, Jade Hind
Department of Computer Science
Liverpool John Moores University
Liverpool, United Kingdom

Naeem Radi
Al Khawarizmi International College
Abu Dhabi, United Arab Emirates

Abstract—**This paper presents a novel approach based on the analysis of genetic variants from publicly available genetic profiles and the manually curated database, the National Human Genome Research Institute Catalog. Using data science techniques, genetic variants are identified in the collected participant profiles and then indexed as risk variants in the National Human Genome Research Institute Catalog. Indexed genetic variants or Single Nucleotide Polymorphisms are used as inputs in various machine learning algorithms for the prediction of obesity. Body mass index status of participants is divided into two classes, Normal Class and Risk Class. Dimensionality reduction tasks are performed to generate a set of principal variables - 13 SNPs - for the application of various machine learning methods. The models are evaluated using receiver operator characteristic curves and the area under the curve. Machine learning techniques including gradient boosting, generalized linear model, classification and regression trees, k-nearest neighbours, support vector machines, random forest and multilayer perceptron neural network are comparatively assessed in terms of their ability to identify the most important factors among the initial 6622 variables describing genetic variants, age and gender, to classify a subject into one of the body mass index related classes defined in this study. Our simulation results indicated that support vector machine generated the highest area under the curve value of 90.5%.**

Keywords—Data Science; Feature Selection; Genetics; Machine Learning; Obesity; Receiver Operating Characteristic Curve; Single Nucleotide Polymorphisms

I. INTRODUCTION

The availability and evolution of high-throughput genomics has improved biomedical research and moved biologist into the big data domain [1], [2]. This has enabled scientists to uncover genetic variations associated with the risk of diseases in a more accurate and reliable way. Researchers of this discipline are dealing with enormous data sets, encountering challenges with handling, processing and moving information that were once the domain of other areas of research such as astronomy [3]. However, this rises another issue, the extraction of useful knowledge and quality interpretation of the data, which could be translated into valuable information for doctors and patients [4]. The last one, is one of the most important challenges in bioinformatics [2], [5].

Genome-wide association studies (GWAS) were aimed at revealing variants at genomic loci that are associated with complex traits in the population. In these studies, a large number of genetic variants or Single Nucleotide Polymorphisms (SNPs) are tested for association with the trait of interest or common diseases, such as obesity, diabetes or coronary artery disease among others [6]–[9]. Associations between the investigated trait and a SNP are termed genome-wide significant. Currently, associations of common variants should reach P-Values threshold levels of $P \leq 5 \times 10^{-8}$ to be considered significant [10], [11]. As the price of genome-wide genotyping has dropped, the number of studies utilizing GWAS has increased dramatically [12]–[16]. Numerous genetic variants that reached genome-wide significant levels in GWAS have been gathered in databases. An example of this type of database is the National Human Genome Research Institute (NHGRI) Catalog. It is a manually curated and publicly available database of SNP-trait association data, discovered via genetic association studies and aimed at finding variants correlated with disease risk [17]. The importance of GWAS is advancing scientific understanding of disease mechanisms as well as providing starting points and potential opportunities for researchers to improve the development of medical treatments or preventing therapies [18], [19].

There is growing evidence that genetic variation plays an important role in determining individual susceptibility to complex disease traits. Complex diseases, such as obesity, require a greater implication of the scientific community to help counteract this global phenomenon [20]. Currently, obesity is considered a worldwide epidemic [21]–[23]. It is associated with multiple conditions, including Type 2 Diabetes (T2D), cardiovascular disease and certain types of cancer [24], [25], making obesity one of the largest global health problems. Several studies have been conducted with the aim of studying obesity aetiology [26]. The advance of GWAS has successfully identified multiple polymorphisms associated with the risk of obesity and higher body mass index (BMI). Common variations such as those in the fat mass and obesity-associated (FTO) and melanocortin 4 receptor (MC4R) genes have been associated with obesity and BMI [27], [28]. These genetic variants have been manually collected to create databases such as the NHGRI Catalog [17] or the Genome-wide Association Studies database (GWASdb), which includes less significant genetic variants in addition to genome-wide significant ones [29]. The large volume of SNPs data motivates the use of data mining and pattern recognition techniques in disease risk prediction and classification [30].

This paper introduces a genetic profile predictive study using machine learning algorithms, in which SNP arrays for a set of subjects are used to predict the future risk of developing complex diseases such as obesity, based on BMI status and SNP profile. In addition, feature selection algorithms are used to find a set of SNPs relevant to the above-mentioned prediction. Consequently, we have developed a methodology which includes: publicly available genetic profiles data collection, data processing - data cleaning, exploratory data analysis and NHGRI Catalog indexation and feature selection. Furthermore, various machine learning models are used for the prediction of obesity.

A deeper understanding of the biology of genomes is necessary to decipher, interpret, and optimize the clinical utility of the variation in the human genome.

The remainder of this paper is organised as follows. Section II reviews the potential application of using artificial intelligence in obesity prediction. Section III introduces the database utilised as well as the steps necessary to identify the genetic variants of interest. In addition, details about feature selection and machine learning model comparison adopted are presented. The results are reported in Section IV, while the findings are discussed in Section V. In Section VI the paper is concluded.

II. Background

Prevention and management of obesity in humans is a complex and challenging task. This condition has a complex aetiology, produced by the combined action of interactions between genes, environmental factors and behaviour [31]. Literature survey indicates that in industrialized countries, between 60% and 70% of the variation in obesity-related phenotypes corresponds to hereditary factors [32]. Consequently, the genetic components that affect the susceptibility of obesity are becoming important elements in determining an individual's risk for this disease [8]. Nonetheless, the translation of these advances in obesity therapies has proven to be a complicated task [26].

To date, numerous studies using SNPs discovered in GWAS for predicting complex diseases have been conducted [33]–[35]. Dominique et al [36] built several multi locust genetic risk indicators for obesity. Genetic risk scores (GRS) were designed from SNP sets identified as genome-wide significant in their corresponding studies. In their analysis, the authors discussed the effects of cumulative genetic risk on body-mass phenotypes in white and black young adults. Young black and white adults with high genetic risk gained more weight and were more likely to become obese compared to those with a lower genetic risk. However, associations in the black samples were smaller in magnitude and not statistically significant.

In a comparable study, Hung et al [37] investigated the effect of GRS combining multiple BMI risk variants for the prediction of obesity in patients with Major Depressive Disorder (MDD). The individual risk of each SNP provided a modest and limited effect in the prediction of obesity. Hence, the authors developed GRS based on 32 well-defined common SNPs, to investigate the association of these GRS with BMI and to further help predicting obesity. Linear and logistic regression models were utilised to predict BMI and to examine the relationship between GRS and obesity, in addition to age, sex, ancestry, and depression status. The results showed that the combination of non-genetic risk factors, GRS and depression, produced results close to the conventional threshold for clinical use and enhanced prediction of obesity. Furthermore, their results suggest that GRS may better predict obesity in depressed patients than in healthy controls.

Numerous studies have used different mechanisms to compare the predictive ability against diseases using solely genetic variants as features. Uhmn et al [34] applied various machine learning techniques including support vector machine (SVM), decision tree, decision rule and k-nearest neighbour algorithm (K-NN), to predict susceptibility to chronic hepatitis using SNPs data. Using several SNPs selected from possible candidate genes which may cause hepatitis, the authors applied a feature selection approach that involved several models, to select a smaller subset of SNPs as features for classification. Subsequently, a number of machine learning techniques were assessed as a tool to diagnose chronic hepatitis. The results suggested that decision rule with backward elimination and backward elimination with backtracking provided the highest accuracy for chronic hepatitis [34]. However, the highest score for sensitivity and specificity was achieved by decision rule with backward elimination and decision tree with backward elimination, respectively. Additionally, both the decision tree and rule based system provided potential for the diagnosis of chronic hepatitis.

Genetically based risk assessment is still in its infancy in which known variants are not decisive when explaining the risk of disease occurrence as predictors in clinical environments [38]. The findings presented in the following sections of this paper, will help the development of new strategies to mitigate the effects of obesity in the global population.

III. Materials and Methods

This section presents the genetic data used in our predictions, the pre-processing steps and feature selection approaches to identify the most significant SNPs for the detection of obesity. Seven machine learning algorithms are comparatively used for the prediction and their performances are evaluated using the Area Under the Curve (AUC), sensitivity (SE) and specificity (SP).

Our solution is based on identifying risk variants present in participants of Direct to Consumer Genetic Testing (DTCGT) services, whose genetic profiles are extracted from the Personal Genome Project (PGP) using web scraping techniques. Indexation of SNPs identified as risks in the NHGRI Catalog and in participants' genetic profiles from the PGP Catalog is performed - considering information extracted from the risk.allele and genotype fields from the NHGRI Catalog and participants' genetic profiles respectively, in addition to the SNP identification as key between data frames

[39]. These identified SNPs and their corresponding diseases or traits are utilised as features. The results indicated that the number of features (variables) obtained, exceeded the number of participants (observations) which posed a dimensionality problem. Thus, random forest was used for features reduction [40]–[42]. Furthermore, a set of 8 SNPs determined in our previous experiments as possible candidates for the prediction of obesity was also considered. A total of 13 SNPs associated with obesity and T2D related traits, and prostate cancer are used as inputs for our prediction. Seven machine learning models simulated for the prediction of obesity were used, including: gradient boosting [43], generalised linear model [44], classification trees [45], k-nearest neighbours (KNN) [46], support vector machine (SVM) [47], random forest (RF) [48] and multilayer perceptron (MLP) neural network [49] trained using backpropagation.

Statistical computing and graphics were performed using the software environment R [50]. The Caret Package [51] was used for the development of the machine learning model testing, while Boruta package [39-40] was used for feature selection.

A. *Identification of Informative Genetic Variants*

In this paper, we compare 164 genetic profiles of DTCGT users participating in the PGP with GWAS results indexed in the NHGRI Catalog. The participant profiles contain SNPs and associated variables which include rsid (SNP ID), chromosome, position, and genotype. We focus on identifying the frequency of the variants of interest - identified as risk alleles - in the 164 examined profiles and if they are indexed in the NHGRI Catalog. This mapping allows us to link SNPs in our samples with risk alleles indexed in the NHGRI Catalog [39]. Hence, this process resulted in a reduced set of SNPs per participant, constructed by previously identified risk SNPs reported in the NHGRI Catalog - using metadata from the NHGRI Catalog ("ebicat37").

This search process for risk variants in the samples, generated a data frame with 6620 variables (SNPs) that have been associated with numerous conditions, including obesity related ones. Therefore, these risk genetic variants are considered as possible candidate features for the classification analysis. Since the number of variables is considerably larger than the number of observations, dimensionality reduction is performed using feature selection techniques.

B. *Feature Selection*

Obesity relates to the total fat mass of an individual which is preferably measured by direct fat measurement methods using, for example, imaging techniques [8]. However, surrogate measurements such as BMI or waist circumference are commonly used for practical and economic reasons. BMI was calculated for all subjects, then a Status feature from participant's BMI was generated. Following the World Health Organisation (WHO) classification for BMI [52], 5 standard weight status categories associated with BMI ranges for adults were derived: Underweight, Normal range, Overweight, Obese and Extremely obese. Extremely obese is commonly divided into Obese I, Obese II and Obese III, but we grouped then into one category for convenience. Participants in this study are not distributed evenly in the 5 status levels recognised by WHO, representing a data balance issue. The problem of data imbalance was solved by creating two classes: Normal class and Risk class. The former class includes underweight and standard range BMI status, whereas the latter class includes the overweight, obese and extremely obese BMI status. Of the total participants, 57% belonged to Normal class whilst the remaining 43% fitted into the Risk class.

In addition to the genetic profiles, other information such as age, gender, height and weight were also collected. The BMI and Status variables are described in Table I. All these features were taken into consideration for obesity prediction processes.

All the risk variants identified were unlikely to be applicable for classification purposes. Hence, only a subset of the most relevant SNPs was selected in our prediction task. Random forest algorithm was used to identify the highest score features and for dimensionality reduction purposes.

A set of 13 SNPs was used as features for the classification process. The reduced set of features was determined by the RF algorithm and a set of 8 SNPs recommended in our experimental research as possible candidates for the prediction of obesity. These 8 SNPs were manually selected after filtering participant's profiles by obesity and T2D related disease-traits SNPs from the NHGRI Catalog. Exploratory data analysis was performed to identify the SNPs that are potentially associated with the BMI status of the participants. The final set of genetic variants as features, is composed of 8 SNPs proposed in our research experiments: rs12567355, rs3001167, rs586688, rs11241130, rs7525133, rs2076529, rs12970134 and rs10195252. While, the remaining 5 SNPs were identified by the random forest algorithm: rs17104630, rs1447295, rs4242382, rs10090154 and rs12978500. Table II shows the final set of SNPs and related information such as chromosome number where the SNP is located, associated reported gene and disease-trait.

C. *Classification Models (Prediction of BMI Status)*

In this study, prediction of obesity is identified by discriminating between normal and risk participants. Cross-validation technique that repeatedly split the data into training and test sets is used during the modelling to overcome the problem of overfitting.

Machine learning algorithms were designed and evaluated using appropriate training and testing sets. The selection of hyperparameters to establish an approximately optimal configuration for each classifier is addressed using Caret for

TABLE I
INFORMATION EXTRACTED FROM THE PGP PROFILES

Variables	Description
Age	Age
Gender	Male = 0, Female = 1
Height	Height in meters
Weight	Weight in Kg
BMI	$\frac{\text{Weight (Kg)}}{(\text{Height(m)})^2}$
Status	Underweight, normal range, Overweight, Obese and Extremely obese

Detailed description of the recorded clinical features extracted from the PGP.

TABLE II
LIST OF FEATURES SELECTED USING THE RF ALGORITHM

Feature	Chrom	Reported gene/s	Strongest.SNP.risk.allele	Disease-trait
rs12567355	1	CD53	rs12567355-A	Obesity-related traits
rs3001167	1	EEF1A1P14	rs3001167-G	Obesity-related traits
rs586688	1	NAV1	rs586688-A	Obesity-related traits
rs7525133	1	RHBG	rs7525133-A	Visceral adipose tissue adjusted for BMI
rs17104630	6	NKX2-1	rs17104630-G	Height
rs12970134	10	MC4R	rs12970134-A	Type 2 diabetes, BMI, Weight, Waist circumference
rs12978500	11	C2CD4C	rs12978500-C	Obesity-related traits
rs10195252	12	COBLL1, GRB14	rs10195252-C and -T	Triglycerides and Waist-hip ratio
rs11241130	18	NREP	rs11241130-G	Obesity-related traits
rs2076529	19	BTNL2	rs2076529-C	Waist-hip ratio
rs1447295	21	MYC, intergenic	rs1447295-A	Prostate cancer
rs4242382	21	intergenic	rs4242382-A	Prostate cancer
rs10090154	21	NR	rs10090154-T	Prostate cancer

Final set of 13 SNPs (features) selected for the classification implementation. SNP associated metadata extracted from the NHGRI Catalog is also included.

automated parameter tuning. Each model was automatically tuned using Caret's Package default search grid, and evaluated using 3-fold Cross Validation (CV) [53]. Tuning parameters shown in Table III produced the models with the best ROC values. Although 3-fold was used, commonly used 10-fold [54] CV was also considered. However, through empirical analysis, the results did not show better improvement in comparison to those reported when using 3-fold CV. Sensitivity, specificity and area under receiver operating characteristics curve were utilized to evaluate the different predictive models. The discriminative power of the models was measured using AUC.

Sensitivity and specificity values are commonly used to measure the predictive capabilities of classifiers. Sensitivities refer to the true positive rate or recall rate (Risk class). Specificities measure the proportion of true negatives (Normal class). Sensitivities are considered higher priority than Specificities in this paper as it is important to predict a risk case so that the health specialist can design appropriate therapies.

The Area Under the Curve (AUC) is an accepted performance metric that provides a value equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. This is a suitable evaluation method for binary classification.

Seven well known machine learning algorithms [44], [48], [55], [56] are used and constructed, as listed in Table IV.

IV. RESULTS

Initial experiments considering the preliminary 6622 features - 6620 SNPs, plus the age and gender - revealed very poor results, practically random. The classifiers were unable to discriminate appropriately from the two classes. Machine learning models suffer from decrease in performance when the number of features is excessively high. Consequently, feature selection mechanisms are used to identify the optimal set of features capable of improving the performance of the classifiers. In this section, the results obtained when the classifiers are fed with the optimal set of features proposed in Table II are presented.

TABLE III
TUNING PARAMETERS SELECTED

Object class	Parameters	Best

gbm	n.trees
	interaction.depth
	shrinkage
	n.minobsinnode
		n.trees = 100
		interaction.dedepth = 1
		shrinkage = 0.1
		n.minobsinnode = 10

glmnet	alpha
	lambda
		α = 0.1
		λ = 0.02768717

rpart	cp	cp = 0
knn	k	k = 7

svmRadial	C
	sigma
		C = 0.5
		σ = 0.04797839

rf	mtry	mtry = 2

nnet	size
	decay
		size = 3
		decay = 0.1

Best tuning parameters using Caret automated parameter tuning [51].

TABLE IV
OBJECT CLASSES SELECTED

Object class	Classifier	Category
gbm	Stochastic Gradient Boosting.	Nonlinear
glmnet	Lasso and Elastic-net regularized generalized linear models.	Linear
rpart	CART (Classification and Regression Trees).	Nonlinear
knn	K-Nearest Neighbor.	Nonlinear
svmRadial	Radial Basis Function Kernel Support Vector Machine.	Nonlinear
rf	Random Forest.	Nonlinear
nnet	Backpropagation Neural Network.	Nonlinear

Methods specified when applying the different classifiers in R.

The results in Table V show that sensitivities for most of the classifiers are lower than the corresponding specificities. This means that most tested models can classify normal cases better than risk ones. However, SVM is the only algorithm that shows higher sensitivity than specificity. This result is encouraging given that predicting risk classes is more important than those that are normal. The AUC values are relatively low for GBM and CART with slightly higher values for GLMNET, MLP, RF and KNN. SVM produced the highest result with marginally over 90% AUC. These results clearly deviate from randomness in contrast with the results obtained in previous experiments when using all the features without applying feature selection techniques.

Even though all the 13 features selected are necessary to obtain the best results for SVM, the variable importance varied among the various algorithms. The results indicate that there is at least a feature or more with no relevance for some classifiers. Fig. 1 contains the variable importance for each algorithm. Feature rs10195252 was considered not relevant in

TABLE V

USING PROPOSED SET OF FEATURES

Classifier	Sensitivity	Specificity	AUC
GBM	0.5882	0.7826	0.7366
GLMNET	0.7059	0.8261	0.8031
CART	0.6471	0.8696	0.7417
KNN	0.6471	0.9130	0.8862
SVM	0.8824	0.8696	0.9054
RF	0.5294	0.9565	0.8798
MLP	0.7059	0.8261	0.8517

Sensitivity, Specificity and AUC values for each classifier when predicting the two classes in the test data.

four out of the seven models, including GBM, GLMNET, CART and RF. Furthermore, the features rs7525133, rs12567355 and rs17104630 did not contribute with the results when GBM was used. Similarly, the feature rs17104630 was not considered by CART. Conversely, all the features were utilised by KNN, SVM and MLP models, which also reported the best AUC values along with RF.

Fig. 2 illustrates that GBM and CART performed poorly in comparison with the other classifiers. GLMNET performed slightly better than GBM and CART. However, SVM, KNN, RF and MLP classifiers produced the best results in that order, which reflect the sensitivity, specificity and AUC values in Table V.

V. DISCUSSION

We present a predictive study using publicly available participants' profiles, as an attempt to understand, and effectively identify predisposition to obesity.

Our study shows the possibility of predicting obesity susceptibility based on BMI status and GWAS results, using SNPs as features.

Thirteen most relevant features were utilised after performing feature selection using random forest algorithm as well as using a set of SNPs recommended from our experimental analyses. These features were employed as inputs in the selected machine learning models. While commonly used 10-fold CV was considered in our analysis, 3-fold was ultimately used since it provided the best performance results. Among the tested models, SVM showed the best overall results with SE=88.24%, SP=86.96% and AUC=90.5%. Sensitivity analysis indicated that the SVM model is robust to the number of selected SNPs.

As shown in Table II, of the total number of 13 features selected, 10 are related to obesity diseases-traits - including rs12970134 related to T2D too -, whilst the outstanding 3 features, rs1447295, rs4242382 and rs10090154 are associated with prostate cancer. Some studies suggest that obesity protects against localised prostate cancer but increases the risk of advanced cancer [57].

The feature rs10195252, considered less relevant in 4 of the algorithms, is associated with triglycerides and waist-hip ratio according to the information extracted from the NHGRI Catalog. The feature rs4242382, when evaluated in SVM and KNN, was the least important. However, rs1447295 was

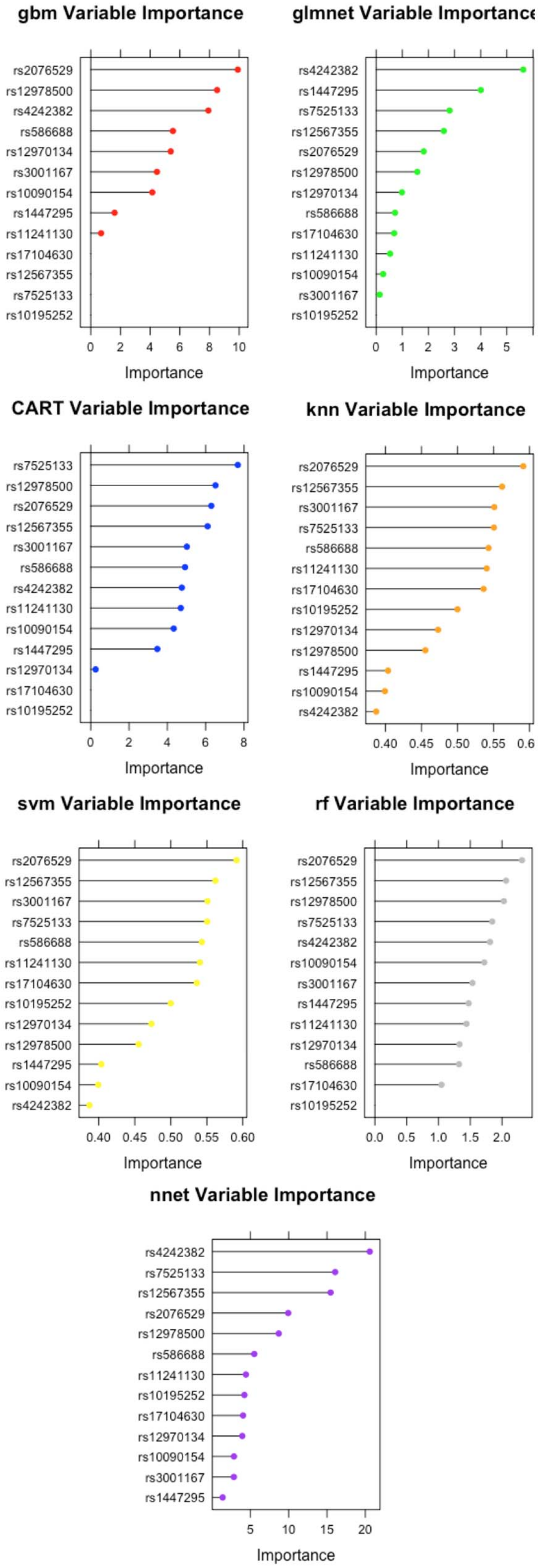

Fig. 1. Variable importance plots for the seven models. The results were obtained with the proposed set of 13 features. Each feature or SNP represents how important it was for the algorithm to help in the discrimination of the two classes.

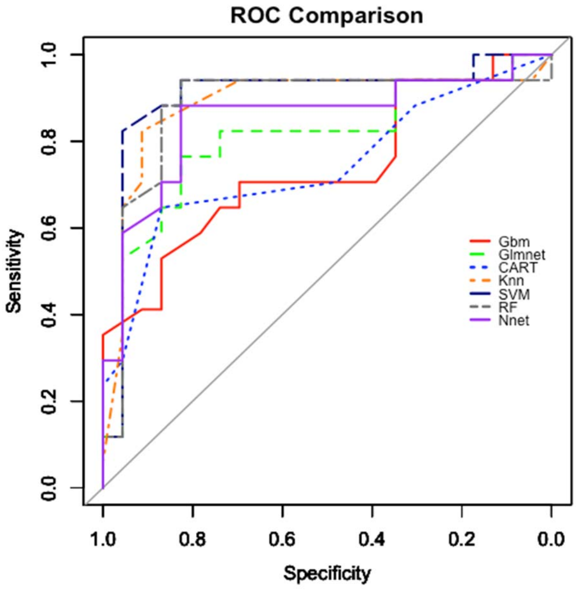

Fig. 2. ROC curve for PGP data using proposed set of features. As the curve moves away from the grey diagonal line towards top left corner of the graph, Sensitivity and Specificity increase, and at the same time, the area under the graph increases. Hence, the models with higher performance will show the ROC curve closer to the top left corner, increasing the area under curve.

considered the least re levant in MLP. Both genetic variants, rs4242382 and rs1447295, are associated with prostate cancer. Results also revealed that the features rs2076529 and rs12567355 - both associated with obesity related traits - are in the top 3 most important features of several models. In the particular case of SVM, the top 3 most important variables were associated with obesity-related traits - rs12567355 and rs3001167 - and waist-hip ratio - rs2076529.

VI. Conclusion

Using publicly available data from participants of the PGP, this paper presents a genetic profile predictive study. Based on previous works and extensive experimental research, well documented and publicly available SNPs and related metadata are used for subsequent identification of genetic risk variants in participant profiles extracted from the PGP. Random forest based feature selection algorithm and extensive experiments conducted by the authors of this paper are used to identify an optimal set of 13 features associated with obesity related disease traits and prostate cancer.

The selected set of features was used to train and test seven well-known machine learning algorithms. Quality measures including sensitivity, specificity and area under receiver operating characteristics curve were utilized to evaluate the performance of the predictive models. The analyses revealed that Support Vector Machine achieved high predictive performance among the studied models, followed by K-Nearest Neighbour.

Our study produces results deviated from randomness and demonstrated the potential of using machine learning approaches in the context of the prediction of complex diseases and personalized patient care. Despite the encouraging results reported in this paper, more in-depth research is still required. The fact that some models used all the features, and others did not, will be subject of study in future work. Hence, different feature selection techniques will be considered and compared as we only used RF based algorithm for dimensionality reduction. Additionally, models showing the best performance will be thoroughly investigated.

References

[1] V. Marx, “Biology: The big challenges of big data,” *Nature*, vol. 498, no. 7453, pp. 255–260, 2013.

[2] P. Muir *et al.*, “The real cost of sequencing: scaling computation to keep pace with data generation,” *Genome Biol.*, vol. 17, no. 1, p. 53, Dec. 2016.

[3] Z. D. Stephens *et al.*, “Big Data: Astronomical or Genomical?,” *PLOS Biol.*, vol. 13, no. 7, p. e1002195, 2015.

[4] J. H. Moore, F. W. Asselbergs, and S. M. Williams, “Bioinformatics challenges for genome-wide association studies.,” *Bioinformatics*, vol. 26, no. 4, pp. 445–55, Feb. 2010.

[5] P. Tarczy-Hornoch and M. Minie, “Bioinformatics Challenges and Opportunities,” in *Medical Informatics*, vol. 8, Boston: Springer US, 2005, pp. 63–94.

[6] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, “Five Years of GWAS Discovery,” *Am. J. Hum. Genet.*, vol. 90, no. 1, pp. 7–24, Jan. 2012.

[7] J. Hardy and A. Singleton, “Genomewide association studies and human disease.,” *N. Engl. J. Med.*, vol. 360, no. 17, pp. 1759–68, Apr. 2009.

[8] T. Fall and E. Ingelsson, “Genome-wide association studies of obesity and metabolic syndrome,” *Mol. Cell. Endocrinol.*, vol. 382, no. 1, pp. 740–757, 2014.

[9] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, and A. Duncanson, “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, Jun. 2007.

[10] O. A. Panagiotou and J. P. A. Ioannidis, “What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations,” *Int. J. Epidemiol.*, vol. 41, no. 1, pp. 273–286, Feb. 2012.

[11] J. Fadista, A. K. Manning, J. C. Florez, and L. Groop, “The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants,” *Eur. J. Hum. Genet.*, vol. 24, no. 8, pp. 1202–1205, Aug. 2016.

[12] S. Gretarsdottir, A. F. Baas, G. Thorleifsson, H. Holm, M. den Heijer, and J.-P. P. M. de Vries, “Genome-wide association study identifies a sequence variant within the DAB2IP gene conferring susceptibility to abdominal aortic aneurysm,” *Nat. Genet.*, vol. 42, no. 8, pp. 692–697, Aug. 2010.

[13] E. K. Speliotes, C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson, and A. U. Jackson, “Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index,” *Nat. Genet.*, vol. 42, no. 11, pp. 937–948, Nov. 2010.

[14] K. a. Tryka *et al.*, “NCBI’s Database of Genotypes and Phenotypes: dbGaP,” *Nucleic Acids Res.*, vol. 42, no. D1, pp. D975–D979, 2014.

[15] S. Kamitsuji, T. Matsuda, K. Nishimura, S. Endo, C. Wada, and K. Watanabe, “Japan PGx Data Science Consortium Database: SNPs and HLA genotype data from 2994 Japanese healthy individuals for pharmacogenomics studies,” *J. Hum. Genet.*, vol. 60, no. 6, pp. 319–326, Jun. 2015.

[16] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol, “Human genetic variation and its contribution to complex traits,” *Nat. Rev. Genet.*, vol. 10, no. 4, pp. 241–251, Apr. 2009.

[17] D. Welter *et al.*, “The NHGRI GWAS Catalog, a curated resource of SNP-trait associations,” *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1001–D1006, Jan. 2014.

[18] P. Blank and F. Gutzwiller, “Current challenges in handling genetic data,” *Eur. J. Med. Sci.*, no. August, pp. 1–2, Aug. 2014.

[19] K. Christensen and J. C. Murray, “What Genome-wide Association Studies Can Do for Medicine,” *N. Engl. J. Med.*, vol. 356, no. 11, pp. 1094–1097, Mar. 2007.

[20] S. Vallgårda, M. E. J. Nielsen, M. Hartlev, and P. Sandøe, “Backward- and forward-looking responsibility for obesity: policies from WHO, the EU and England,” *Eur. J. Public Health*, vol. 25, no. 5, pp. 845–848, Oct. 2015.

[21] R. T. Hurt, C. Kulisek, L. a. Buchanan, and S. a. McClave, “The obesity epidemic: Challenges, health initiatives, and implications for gastroenterologists,” *Gastroenterol. Hepatol.*, vol. 6, no. 12, pp. 780–792, 2010.

[22] Y. Poloz and V. Stambolic, “Obesity and cancer, a case for insulin signaling,” *Cell Death Dis.*, vol. 6, no. 12, p. e2037, Dec. 2015.

[23] S. J. van Dijk *et al.*, “Epigenetics and human obesity,” *Int. J. Obes.*, vol. 39, no. 1, pp. 85–97, Jan. 2015.

[24] K. A. Hirko, E. D. Kantor, S. S. Cohen, W. J. Blot, M. J. Stampfer, and L. B. Signorello, “Body Mass Index in Young Adulthood, Obesity Trajectory, and Premature Mortality.,” *Am. J. Epidemiol.*, no. 31, 2015.

[25] S. Ramachandrappa and I. S. Farooqi, “Genetic approaches to understanding human obesity,” *J. Clin. Invest.*, vol. 121, no. 6, pp. 2080–2086, Jun. 2011.

[26] J. S. El-Sayed Moustafa and P. Froguel, “From obesity genetics to the future of personalized obesity therapy,” *Nat. Rev. Endocrinol.*, vol. 9, no. 7, pp. 402–413, Mar. 2013.

[27] K. Wang *et al.*, “A Genome-Wide Association Study on Obesity and Obesity-Related Traits,” *PLoS One*, vol. 6, no. 4, p. e18939, Apr. 2011.

[28] A. Scuteri, S. Sanna, W.-M. Chen, M. Uda, G. Albai, and J. Strait, “Genome-Wide Association Scan Shows Genetic Variants in the FTO Gene Are Associated with Obesity-Related Traits,” *PLoS Genet.*, vol. 3, no. 7, p. e115, 2007.

[29] M. J. Li *et al.*, “GWASdb: A database for human genetic variants identified by genome-wide association studies,” *Nucleic Acids Res.*, vol. 40, no. D1, pp. 1047–1054, 2012.

[30] S. Okser, T. Pahikkala, and T. Aittokallio, “Genetic variants and their interactions in disease risk prediction – machine learning and network perspectives,” *BioData Min.*, vol. 6, no. 1, p. 5, 2013.

[31] K. R. Rao, N. Lal, and N. V Giridharan, “Genetic & epigenetic approach to human obesity,” *Indian J. Med. Res.*, vol. 140, no. November 2014, pp. 589–603, 2015.

[32] A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, and F. R. Day, “Genetic studies of body mass index yield new insights for obesity biology,” *Nature*, vol. 518, no. 7538, pp. 197–206, Feb. 2015.

[33] J. Listgarten, “Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms,” *Clin. Cancer Res.*, vol. 10, no. 8, pp. 2725–2737, Apr. 2004.

[34] S. Uhmn, D.-H. Kim, Y.-W. Ko, S. Cho, J. Cheong, and J. Kim, “A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis,” *Expert Syst.*, vol. 26, no. 1, pp. 60–69, Feb. 2009.

[35] V. Aguiar-Pulido, J. a Seoane, J. R. Rabuñal, J. Dorado, A. Pazos, and C. R. Munteanu, “Machine Learning Techniques for Single Nucleotide Polymorphism—Disease Classification Models in Schizophrenia,” *Molecules*, vol. 15, no. 7, pp. 4875–4889, Jul. 2010.

[36] B. W. Domingue, D. W. Belsky, K. M. Harris, A. Smolen, M. B. McQueen, and J. D. Boardman, “Polygenic Risk Predicts Obesity in Both White and Black Young Adults,” *PLoS One*, vol. 9, no. 7, p. e101596, Jul. 2014.

[37] C.-F. Hung, G. Breen, D. Czamara, T. Corre, C. Wolf, and S. Kloiber, “A genetic risk score combining 32 SNPs is associated with body mass index and improves obesity prediction in people with major depressive disorder,” *BMC Med.*, vol. 13, no. 1, p. 86, Dec. 2015.

[38] W. G. Feero, A. E. Guttmacher, and T. A. Manolio, “Genomewide Association Studies and Assessment of the Risk of Disease,” *N. Engl. J. Med.*, vol. 363, no. 2, pp. 166–176, 2010.

[39] C. Aday Curbelo Montañez, P. Fergus, A. Hussain, D. Al-Jumeily, B. Abdulaimma, and H. Al-Askar, “A Genetic Analytics Approach for Risk Variant Identification to Support Intervention Strategies for People Susceptible to Polygenic Obesity and Overweight,” in *Intelligent Computing Theories and Application: 12th International Conference, ICIC 2016, Lanzhou, China, August 2-5, 2016, Proceedings, Part I*, D.-S. Huang, V. Bevilacqua, and P. Premaratne, Eds. Cham: Springer International Publishing, 2016, pp. 808–819.

[40] M. B. Kursa and W. R. Rudnicki, “Feature Selection

with the Boruta Package," *J. Stat. Softw.*, vol. 36, no. 11, pp. 1–13, 2010.

[41] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta - A system for feature selection," *Fundam. Informaticae*, vol. 101, no. 4, pp. 271–285, 2010.

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

[43] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front. Neurorobot.*, vol. 7, no. DEC, 2013.

[44] J. Xing, H. Gao, Y. Wu, Y. Wu, H. Li, and R. Yang, "Generalized Linear Model for Mapping Discrete Trait Loci Implemented with LASSO Algorithm," *PLoS One*, vol. 9, no. 9, p. e106985, Sep. 2014.

[45] T. S. Kershaw *et al.*, "Using Clinical Classification Trees to Identify Individuals at Risk of STDs During Pregnancy," *Perspect. Sex. Reprod. Health*, vol. 39, no. 3, pp. 141–148, Sep. 2007.

[46] Z. Yao and W. L. Ruzzo, "A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data," *BMC Bioinformatics*, vol. 7, no. Suppl 1, p. S11, 2006.

[47] S.-H. Chen *et al.*, "A support vector machine approach for detecting gene-gene interaction," *Genet. Epidemiol.*, vol. 32, no. 2, pp. 152–167, Feb. 2008.

[48] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, Jun. 2012.

[49] L. J. Lancashire, C. Lemetre, and G. R. Ball, "An introduction to artificial neural networks in bioinformatics--application to complex microarray and mass spectrometry datasets in cancer studies," *Brief. Bioinform.*, vol. 10, no. 3, pp. 315–329, Dec. 2008.

[50] R Development Core Team, "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria, 2008.

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

[52] World Health Organization, "WHO | World Health Organization," *WHO*, 2016. [Online]. Available: http://www.who.int/en/. [Accessed: 15-Nov-2016].

[53] I. K. Valavanis, S. G. Mougiakakou, K. a Grimaldi, and K. S. Nikita, "A multifactorial analysis of obesity as CVD risk factor: use of neural network based methods in a nutrigenetics context.," *BMC Bioinformatics*, vol. 11, p. 453, 2010.

[54] T. H. Pers, A. Albrechtsen, C. Holst, T. I. a Sørensen, and T. a. Gerds, "The validation and assessment of machine learning: A game of prediction from high-dimensional data," *PLoS One*, vol. 4, no. 8, 2009.

[55] P. Larrañaga *et al.*, "Machine learning in bioinformatics," *Brief. Bioinform.*, vol. 7, no. 1, pp. 86–112, 2006.

[56] J. H. Friedman, "Stochastic gradient boosting," *Comput. Stat. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002.

[57] S. J. Lewis *et al.*, "Associations between an obesity related genetic variant (FTO rs9939609) and prostate cancer risk," *PLoS One*, vol. 5, no. 10, pp. 3–9, 2010.