acta medica

ORIGINAL ARTICLE

Identification of bio-markers for insulin resistance and sensitivity through multi-omics analysis

İdil Yet[1]
ORCID: 0000-0002-1227-0716

ABSTRACT

Aims: This study aims to identify multi-omics bio-markers for insulin resistance and sensitivity using machine learning approaches on a dataset integrated from several omics.

Methods: The study included 362 patients with Insulin Resistance and Insulin Sensitivity from the Integrative Personal Omics Profiling (iPOP) database. Combining the multi-omics data from the Integrative Human Microbiome Project, this study used machine learning to reveal the relationship between insulin resistance and insulin sensitivity.

Results: Of 362 patients 186 were insulin resistance and 176 were insulin sensitivity. 11,585 features were used, including clinical features, RNA transcripts, gut microbiota, cytokines, proteins, and metabolomics. We found 21 features capable of distinguishing insulin resistance from insulin sensitivity using a well-known artificial neural network (ANN) method. The model had an area under the receiver operating characteristic (AUC) of 0.97 in the validation dataset and 0.89 in the test dataset. The ANN model's performance was compared with Random Forest model. Of the 21 new findings, two metabolites (methyl-uric acid and methylxanthine) are xenobiotics, and three RNA transcripts (*SERPINF1, SLC2A2*, and *CHL1*).

Conclusion: A small number of multi-omics features identified from 11,585 potential candidates for a machine learning model can accurately predict insulin resistance and sensitivity.

Keywords: microbiome, metabolomics, multi-omics, type II diabetes mellitus, artificial neural network.

[1] Department of Bioinformatics, Graduate School of Health Sciences, Hacettepe University, Ankara, Türkiye.

Corresponding Author: İdil Yet
E-mail: idil.yet@hacettepe.edu.tr

Received: 13 May 2024, Accepted: 23 August 2024, Published online: 30 September 2024

INTRODUCTION

Type II diabetes mellitus (T2D) affects more than 10% of the world population, and another 30% are diagnosed with prediabetes and are at risk of developing diabetes in the coming years [1, 2]. T2D is a complex disease; little is known about changes during the initial prediabetes stage, modifications in biological processes, or its alteration to T2D. Both conditions are connected with insulin resistance, which is used to investigate the earliest stages of diabetes. Innovations in next-generation sequencing (NGS) and mass spectrometry (MS) have made it possible to report novel bio-markers and pathways across several diseases, including T2D. Biological data created with NGS and MS experiments help more accurately predict health outcomes [3-5]. Massive cohort studies use data generated from NGS to identify genetic variants associated with complex diseases, such as genome-wide association studies (GWAS). GWAS-associated T2D has identified more than 300 genetic variants. However, using GWAS alone is not sufficient for a thorough understanding of complex diseases and their mechanisms as GWAS only focuses on genetic factors [6]. The past two decades have witnessed progress in the diversity of molecular data, including genomics, epigenomics, transcriptomics,

and proteomics. These multi-omics profiling approaches can be used to screen the change of molecules in diseases and examine the variation within the traits. To address these challenges and to study dynamic changes in hosts under several diseases, the Integrated Human Microbiome Project (iHMP) was established by The National Institutes of Health (NIH) [2]. Previous iHMP projects have mainly focused on the longitudinal analysis of prediabetes patients [7]. In the past decade, microbiota and metabolomics have become very popular for uncovering the associations related to health or disease conditions [8,9]. Several cohort studies try to create an atlas for biomarkers using association studies involving metabolite-wide association studies (MeWAS) and microbiome-wide association studies (MWAS) [4,10]. Prediabetes and T2D signature has also been studied at a single omic level using microbiome [11], metabolomics [12], proteomics [13], epigenomics [14]. Overmyer et al. investigated associations between the oral microbiome and metabolomics in subjects with prediabetes [15]. These studies mainly focused on prediabetes patients, but the relationships among multi-omics elements in insulin resistance were not thoroughly studied. Machine learning is a promising tool for analysing multi-omics data and identifying bio-markers for disease risk].

Additionally, the system biology field has been moving from only generating data to effectively analysing this high-dimensional data using many machine learning techniques [17-20]. These studies mostly try to predict metastasis or help clinicians effectively in cancer diagnosis, prognosis, and treatment selection [18,19]. However, high-dimensional models with different multi-omics elements make it challenging to develop accurate models and lead to overfitting problems. This study aims to identify potential biomarkers for insulin resistance (IR) and insulin sensitivity (IS) using machine learning. Combining data from multiple omics including laboratory features, gut microbiota, RNA transcripts, metabolomics, cytokines and proteins, we investigated 11,585 potential features for predicting IR and IS. We identified 21 potential biomarkers that can make accurate predictions of IS and IR using feature selection approaches.

METHODS

The iPOP Project omics data was used (http://hmp2-data.stanford.edu/). iHMP Type II Diabetes Mellitus Data were obtained from iPOP [21]. Each omics data is downloaded separately from the data portal and merged using the visiting ID of samples. Ethics approval is not needed, and the Declaration of Helsinki's ethical rules and principles were followed in all procedures. We designed a cross-sectional study ignoring the longitudinal data of prediabetes patients. The samples were selected according to the steady-state plasma glucose level (SSPG) by iPOP. Individuals with an SSPG greater than 150 mg/dL were logged as IR, and below the same threshold were logged as IS [2]. The data consisted of 186 individuals classified as IR and 176 as IS. In total, 11,585 features were used (302 proteins from plasma, 66 cytokines, 51 clinical laboratory features, 96 gut microbiota, 10,346 RNA transcripts, and 724 metabolomics) (Figure 1).

Features that have missing values were excluded and data was scaled with z-score normalisation. The data was analysed using two machine learning methods: artificial neural networks (ANN) and Random Forests. To develop a model for classifying IR and IS, subjects were divided randomly in an 8:2 model training dataset to test-validation dataset ratio. The test-validation dataset was used only to verify the model performance and was randomly divided into a 5:5 ratio to obtain the test and validation datasets. A Multi-Layer Perceptron classifier from the scikit-learn library in Python programming language was used for ANN [22]. Model parameter optimisation was performed using Grid search, and the best parameters were selected for each model. All the predefined parameters are fitted with the Adaptive Movement Estimation (Adam) algorithm to adjust the learning rate dynamically, sigmoid for calculating predictions, and the remaining layers are activated with the Rectified Linear Unit (ReLU) function. The Sequential Feature Selector function was used with forward selection (Sequential Forward Selection (SFS)) for feature selection. A Random Forest Classifier from the scikit-learn library in Python programming language was used for Random forest [22]. The gini function to measure the quality

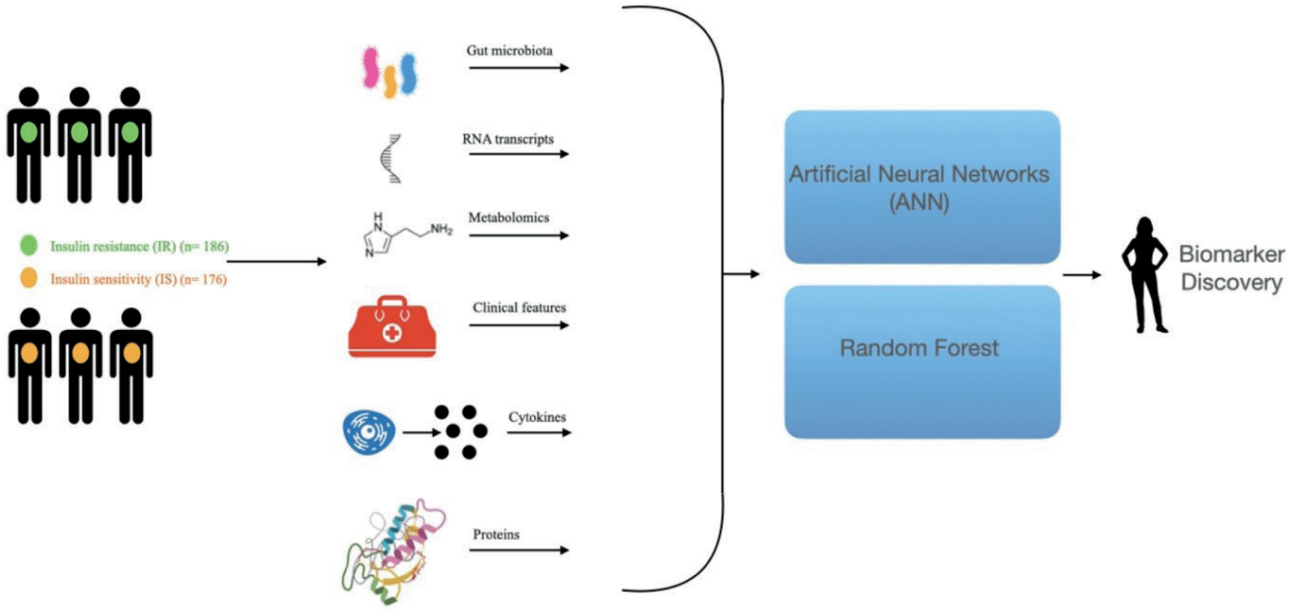

Figure 1. Summary of the multi-omics study design

of a split was selected. It ensures that each tree makes effective splits that contribute to accurate predictions when combined with others in the forest. The comparisons of ANN and Random Forest models were based on the area under the receiver operating characteristic (AUC) curve, accuracy, precision, recall and f1-score. AUC was plotted with ggroc function from ggplot library using R programming language [23]. The performance of the ANN algorithm model was compared with the Random Forest algorithm [24].

RESULTS

The study consisted of 186 patients with IR and 176 patients with IS. To develop a model capable of distinguishing IR from IS, subjects were divided randomly in an 8:2 model training to test-validation dataset ratio. We set the hidden_layer_size inside the Multi-Layer Perceptron Classifier to (6,6,6,6) in ANN model. This means we add four hidden layers with six hidden units in each, including the Adam algorithm, the ReLU function, and 500 iterations. The model resolved with 21 features with the forward selection function from 11,585 features from the data (Table 1).

The first ANN models of IR and IS data were built using training sets. Then, the model was evaluated with validation and test datasets. The same data-splitting procedure is applied to the Random Forest model. We set the n_estimators (number of trees in the forest) as "1000" in Random Forest Classifier. The function to measure the quality of a split is selected as "gini". The prediction performances of ANN model were also evaluated and compared with Random Forest model (Table 2).

Table 1. Selected 21 features with forward selection in the ANN algorithm model

Bio-marker	Omic	Total	p-value
IGHM	Proteomics	3	<0.001
APOE			<0.001
LPA			0.01
LEPTIN	Cytokines	6	0.007
SCF			<0.001
GMCSF			<0.001
MCP1			<0.001
FASL			0.014
IL7			<0.001
HDL	Clinical Laboratory	6	0.006
Monoab			0.05
MCV			<0.001
CR			<0.001
TGL			<0.001
EOTAXIN			<0.001
genus_Coprococcus	Gut Microbiota	1	0.015
methyluric acid	Metabolomics	2	<0.001
methylxanthine			<0.001
SERPINF1	Transcriptomics	3	<0.001
SLC2A2			<0.001
CHL1			0.05

Table 2. Model performance metrics

Models	Performance	Accuracy	Precision	Recall	f1-score
ANN	AUC=0.89	0.89	0.89	0.85	0.87
Random Forest	AUC=0.94	0.94	0.94	0.89	0.91

Receiver operating characteristic (ROC) curve of the ANN and Random Forest models are plotted in Figure 2.

The AUC was 0.97 in the validation set and 0.89 in the test dataset. The accuracy, precision, recall and f1-score for the validation dataset are 0.91, 0.95, 0.87, and 0.91, respectively. The test dataset's accuracy, precision, recall and f1-score are 0.89, 0.89, 0.85, and 0.87, respectively. The Random Forest algorithm method was used to validate the ANN model. The AUC was 0.93 in the validation set and 0.94 in the test dataset. Of the 11,585 features, 21 features were chosen in the final model. There were three proteins (IGHM, APOE, LPA), six cytokines (LEPTIN, SCF, GMCSF, MCP1, FASL, IL7) and six clinical laboratory features (HDL, Monoab, MCV, CR, TGL, EOTAXIN), one gut microbiota (*genus_Coprococcus*), three RNA transcripts from RNAseq (*SERPINF1*, *SLC2A2*, and *CHL1*), and two from metabolites (methyl uric acid and methylxanthine. The 21 features for the final model are listed and can be found in Table 1.

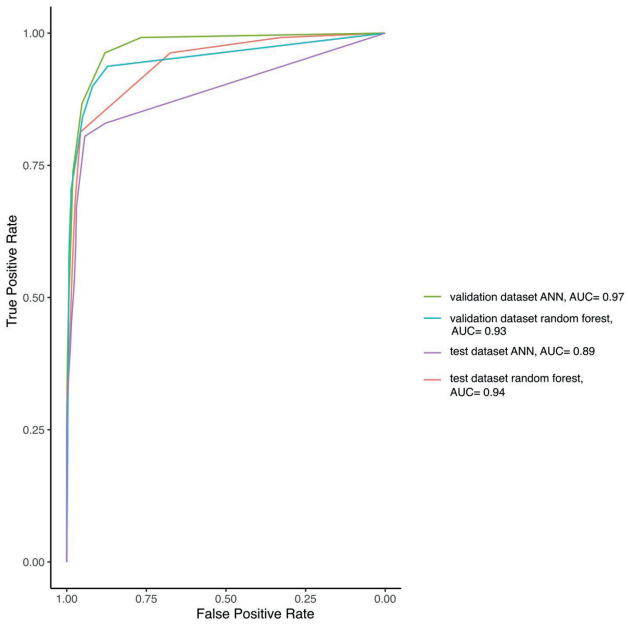

Figure 2. ROC curve of the two models (ANN and random forest) with feature selection. The models are built using 11,585 features. AUC scores of the two models using the validation and test datasets

DISCUSSION

Human transcriptomics, epigenomics, proteomics, metabolomics, and microbiome play an important role in health, and there is a strong indication that omics can be used as predictors of diseases. Arneth et. al. [12] conducted a meta-analysis on the metabolomics of Type I and Type II Diabetes mellitus, reporting several significant metabolites. Pinna et. al. [11] reported 16 operational taxonomic units (OTUs) enriched in almost 500 subjects with prediabetes. Huth et. al. [13] examined protein markers associated with T2D and prediabetes. Juvinao-Quintero et. al. [14] identified 77 differentially methylated regions associated with T2D in a meta-analysis. Overmyer et. al.'s [15] multi-omics study investigated the oral microbiome and metabolomics in (n = 97) subjects with prediabetes and found various associations. However, analyzing these omics causes challenges due to their high-dimensional profiles with the help of feature selection, like the multi-dimensional datasets can be easily compressed into low-dimensional features. In our study, when the model was generated with selected features (21 features), the model performance was improved compared to the without feature selection. The ANN model could classify the samples as IR or IS with 21 features, but no single feature could do so. The ANN model was compared with the Random Forest model, both showing similar metric and performance results (AUC), with the Random Forest exhibiting slightly better performance, showcasing its robustness as a method. On the other hand, the Sequential Feature Selector function coupled with forward selection feature selection in ANN model enabled us to focus on specific key bio-markers which was the main focus of the study. Of the 21 features, 9 of them (SCF, LPA, GMCSF, IL7, CR, APOE, MONOAB, TGL, and IGHM) reported in previous studies to be direct and inverse relationship with insulin resistance and 7 of them (HDL, MCV, EOTAXIN, LEPTIN, MCP1, FASL1 and *genus_Coprococcus)* have a positive correlation between prediabetes and diabetes groups. With this study, we proposed 21 features to become distinguishing bio-markers

for prediabetes. Zhao et. al. [25] detected that prediabetes patients had reduced secretion of methyl uric acid and methylxanthine, which are xenobiotics in tea and coffee. Zhou et.al. [7] used the subset of this data longitudinal to analyse the dynamics of microbiomes in prediabetes with more straightforward methods like logistic regression without selecting the significant features. Notably, using the data cross-sectional helped us increase the sample size, and feature selection enabled us to focus on substantial bio-markers. We identified three bio-marker genes, including *SERPINF1, SLC2A2,* and *CHL1*. These three genes linked to diabetes have been studied in a T2D study [26,27]. Results suggested that *SLC2A2* mutation is an autosomal recessive cause of neonatal diabetes mellitus [26]. *SERPINF1* is related to obesity and changes leptin levels in populations at risk of T2D [26]. *CHL1* encodes a protein, and its expression has indicated a decrease in T2D [27]. We can conclude that this study is the first to establish a separation between IR and IS and these five biomarkers: three RNA transcripts (SERPINF1, SLC2A2, and CHL1), and two metabolites (methyl uric acid and methylxanthine).

Our study has some limitations. One limitation of machine learning algorithms is their overfitting problem. To overcome this problem, we used cross-validation for ANN model with the Random Forest model, and feature selection is coupled with ANN for the multi-site variables. Another limitation is the statistical power. In a systems biology study, it is essential to have a sufficient number of samples to get enough power. Although the sample size for multi-omics profiling has increased over the past ten years, the number of samples can still be lower when selecting a stringent significance level required to correct for multiple testing. To detect more bio-markers, the sample size needs to be increased. Another limitation is the multi-dimensional in-person data. The studies should generate system biology data for individuals on multiple biological platforms across different technologies and tissues. An additional limitation of this study is that, while the model was validated using test and validation datasets derived from the original data, further validation on an independent dataset is necessary to fully assess its generalizability and robustness. Overall, the insulin-resistant and insulin-sensitive subjects differed, and multi-omics elements enabled us to explore the early signs of disease development individually. Future studies will help to develop additional information on how the multi-omics elements affect disease development.

CONCLUSION

Multi-omics analyses of IR and IS cross-sectional profiling demonstrated insight into disease aetiology. We found 21 features characterising IR from IS using the artificial neural network method with a high AUC measure. Future work is required to assess the bio-markers we propose in this study and applies to other IR and IS cases. Overall, the frequency of T2D is increasing, and the problems it brings are also growing. With this study, we contributed to the literature about the assessment of IR and IS by measuring the 21 significant features using the ANN model.

Author contribution

Study conception and design: IY; data collection: IY; analysis and interpretation of results: IY; draft manuscript preparation: IY. The author reviewed the results and approved the final version of the manuscript.

Ethical approval

The study was carried out with published data. The data is open access and freely available on iPOP Project Data Portal (http://hmp2-data.stanford.edu/). No ethical approval was needed.

Funding

The author declare that the study received no funding.

Conflict of interest

The author declare that there is no conflict of interest.

REFERENCES

[1] Çetintepe SP YA, Kılıçarslan A. An Overview of Diabetes Mellitus and Work Life. Acta Medica. 2016;47(2):54-60.

[2] Integrative HMPRNC. The Integrative Human Microbiome Project. Nature. 2019;569(7758):641-8.

[3] Sardaraz M, Tahir M, Ikram AA. Advances in high throughput DNA sequence data compression. J Bioinform Comput Biol. 2016;14(3):1630002.

[4] Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. Nat Genet. 2014;46(6):543-50.

[5] Lightbody G, Haberland V, Browne F, Taggart L, Zheng H, Parkes E, et al. Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. Brief Bioinform. 2019;20(5):1795-811.

[6] Vujkovic M, Keaton JM, Lynch JA, Miller DR, Zhou J, Tcheandjieu C, et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. Nat Genet. 2020;52(7):680-91.

[7] Zhou W, Sailani MR, Contrepois K, Zhou Y, Ahadi S, Leopold SR, et al. Longitudinal multi-omics of host-microbe dynamics in prediabetes. Nature. 2019;569(7758):663-71.

[8] Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Erratum: Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017;551(7679):256.

[9] Beaumont M, Goodrich JK, Jackson MA, Yet I, Davenport ER, Vieira-Silva S, et al. Heritable components of the human fecal microbiome are associated with visceral fat. Genome Biol. 2016;17(1):189.

[10] Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, Garg N, et al. Microbiome-wide association studies link dynamic microbial consortia to disease. Nature. 2016;535(7610):94-103.

[11] Pinna NK, Anjana RM, Saxena S, Dutta A, Gnanaprakash V, Rameshkumar G, et al. Trans-ethnic gut microbial signatures of prediabetic subjects from India and Denmark. Genome Med. 2021;13(1):36.

[12] Arneth B, Arneth R, Shams M. Metabolomics of Type 1 and Type 2 Diabetes. Int J Mol Sci. 2019;20(10).

[13] Huth C, von Toerne C, Schederecker F, de Las Heras Gala T, Herder C, Kronenberg F, et al. Protein markers and risk of type 2 diabetes and prediabetes: a targeted proteomics approach in the KORA F4/FF4 study. Eur J Epidemiol. 2019;34(4):409-22.

[14] Juvinao-Quintero DL, Marioni RE, Ochoa-Rosales C, Russ TC, Deary IJ, van Meurs JBJ, et al. DNA methylation of blood cells is associated with prevalent type 2 diabetes in a meta-analysis of four European cohorts. Clin Epigenetics. 2021;13(1):40.

[15] Overmyer KA, Rhoads TW, Merrill AE, Ye Z, Westphall MS, Acharya A, et al. Proteomics, Lipidomics, Metabolomics, and 16S DNA Sequencing of Dental Plaque From Patients With Diabetes and Periodontal Disease. Mol Cell Proteomics. 2021;20:100126.

[16] Jihad M, Yet I. Multiomics Integration at Single-Cell Resolution Using Bayesian Networks: A Case Study in Hepatocellular Carcinoma. OMICS. 2023;27(1):24-33.

[17] Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Brief Bioinform. 2022;23(1).

[18] Albaradei S, Thafar M, Alsaedi A, Van Neste C, Gojobori T, Essack M, et al. Machine learning and deep learning methods that use omics data for metastasis prediction. Comput Struct Biotec. 2021;19:5008-18.

[19] Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med. 2021;13(1):152.

[20] Joshi A, Rienks M, Theofilatos K, Mayr M. Systems biology in cardiovascular disease: a multiomics approach. Nat Rev Cardiol. 2021;18(5):313-30.

[21] Snyder M. iPOP and its role in participatory medicine. Genome Med. 2014;6(1):6.

[22] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825-30.

[23] Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York2016.

[24] Sarica A, Cerasa A, Quattrone A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Front Aging Neurosci. 2017;9:329.

[25] Zhao X, Fritsche J, Wang J, Chen J, Rittig K, Schmitt-Kopplin P, et al. Metabonomic fingerprints of fasting plasma and spot urine reveal human pre-diabetic metabolic traits. Metabolomics. 2010;6(3):362-74.

[26] Huang Y, Cai L, Liu X, Wu Y, Xiang Q, Yu R. Exploring biomarkers and transcriptional factors in type 2 diabetes by comprehensive bioinformatics analysis on RNA-Seq and scRNA-Seq data. Ann Transl Med. 2022;10(18):1017.

[27] Taneera J, Lang S, Sharma A, Fadista J, Zhou Y, Ahlqvist E, et al. A systems genetics approach identifies genes and pathways for type 2 diabetes in human islets. Cell Metab. 2012;16(1):122-34.