Journal of Politics and Ethics in New Technologies and AI

Vol 3, No 1 (2024)

Journal of Politics and Ethics in New Technologies and AI

Utilizing Synthetic Data and Artificial Neural Networks for Clinical Phenotype Prediction in Precision Medicine: A Targeted Metabolomic Analysis of Urinary Organic Acids in Autoimmune Diseases

Vasileios Fragoulakis, Athanassios Vozikis

doi: 10.12681/jpentai.38383

Journal of Politics and Ethics in New Technologies and AI
Volume 3, Issue 1 (2024)
e-ISSN: 2944-9243

https://doi.org/10.12681/jpentai.38383

RESEARCH ARTICLE

Utilizing Synthetic Data and Artificial Neural Networks for Clinical Phenotype Prediction in Precision Medicine: A Targeted Metabolomic Analysis of Urinary Organic Acids in Autoimmune Diseases

Vasileios Fragoulakis
Laboratory of Health Economics & Management, Department of Economics, University of Piraeus, 18534 Piraeus, Greece

Athanassios Vozikis
Laboratory of Health Economics & Management, Department of Economics, University of Piraeus, 18534 Piraeus, Greece

Abstract

This study aimed to create and contrast the precision of synthetic data with original data as inputs in a binary predictive feed-forward back-propagation Artificial Neural Network (ANN) for targeted analysis of urinary Organic Acids (OAs). The original dataset utilized in this analysis originated from case-control research involving 392 participants (comprising patients with autoimmune diseases and healthy individuals). Two types of synthetic data were generated using a non-parametric bootstrap replication technique and a Classification and Regression Tree (CART) model in place of the original values. Support Vector Machine (SVM) analysis was employed to pinpoint potentially crucial biomarkers for inclusion in the ANN. The accuracy of the ANN models was evaluated through the Receiver Operating Characteristic (ROC) curve, along with standard performance measurements like Sensitivity, Specificity, Positive Predicted Value, Negative Predictive Value, False Positive Rate, False Negative Rate and Overall performance. To assess the model's cross-validation and guard against overfitting, the data was randomly divided into three distinct sets: training data (50%), testing data (25%), and Holdout data (25%). The optimal architecture for all ANN models consisted of a shallow structure with one hidden layer, a hyperbolic activation function, and SoftMax as the output function. SVM analysis did not detect variations among biomarkers, indicating their equal importance. The predictive accuracy of the artificial neural network using real data was approximately 77.3%, compared to 66.6% for bootstrap-synthetic data and 51.27% for the ANN-CART model. None of the models exhibited signs of overfitting. The relatively poor performance of the ANN-CART model could be improved by adopting simpler modeling approaches and integrating alternative strategies for biomarker selection. Synthetic data quality can be enhanced through advanced statistical methodologies and may serve as a reasonable alternative for input in an ANN model while maintaining comparable accuracy in autoimmune disease prediction.

Keywords: Artificial Neural Networks, Synthetic Data, total Organic Acids, metabolomics, precision medicine

Introduction

Artificial neural networks (ANNs) are a versatile class of mathematical models commonly used to analyze non-linear data. These networks find extensive application in tasks uncovering relationships

Fragoulakis, V., & Vozikis, A. (2024). Utilizing Synthetic Data and Artificial Neural Networks for Clinical Phenotype Prediction in Precision Medicine: A Targeted Metabolomic Analysis of Urinary Organic Acids in Autoimmune Diseases. *Journal of Politics and Ethics in New Technologies and AI, 3*(1), e38383. https://doi.org/10.12681/jpentai.38383

between biological markers and disease conditions, conducting DNA analysis, and making predictions regarding genetic traits within populations. The configuration of ANNs is characterized by the presence of interconnected 'neurons' that are arranged in distinct layers, which include one or more hidden layers, an input layer (gene expression levels, metabolite concentrations, etc), and an output layer (e.g risk of disease, binary classification as “case” or “control” etc) (Paul et al., 2022). The synaptic weights connecting neurons mirror the strength of signals, drawing initially inspiration from the physiological structure of the brain. The ANN, upon receiving data, provides predictions based on initial conditions. These predictions are adjusted based on their deviation from actual values to align the model's forecasts with the true data, leading to the model's conclusion. Throughout this iterative process, an error function is utilized to assign weights to each input variable and determine the rate of adjustment of the algorithm. The primary advantage of neural networks is that they do not require a pre-established parametric correlation between the data entered into the model and the resulting outcome. Instead, the model itself processes this relationship, determining the relative weights and the nature of the correlation that exists (Tu, 1996).

To comprehend the intricate relationships among biological phenomena from a mathematical perspective, researchers often turn to advanced computational techniques and models such as non-linear activation functions, several hidden layers (the so-called “deep learning”), non-linear output functions, etc (Jawad, 2023). Although ANNs represent a cutting edge of modern applied computational sciences, the analysis of such models employs a trial-and-error approach, with the structure and methods used to approximate biological phenomena frequently based on this approach. Furthermore, the optimal relationship between the number of neurons, layers, model type, and size of available data is often empirical and ad hoc. Moreover, the ANN calculations represent a “data-hungry” procedure, and attaining optimal performance demands a considerable amount of data.

In metabolomics, data availability is generally more restricted than in genomics, thus limiting the applications of ANNs. A recent review estimated that ANNs were utilized in roughly 10% of metabolomic studies compared to genomics (25 vs 250) (Mendez et al., 2019). In such a case, the application of synthetic data can mitigate challenges associated with data availability, accessibility, and legal constraints, thereby enhancing the potential use of ANNs in metabolomic research (Giuffrè & Shung, 2023). In simple terms, synthetic datasets consist entirely of or contain a subset of, not real microdata that are artificially manufactured with or without the original data. In recent years, these data have been of great interest in the healthcare sector driving the discovery of novel scientific insights, in drug development and the mechanistic understanding of diseases.

The aim of the present analysis was twofold: a) To resynthesize the dataset based on the available observations set and b) to develop an alternative or even improved predictive algorithm with these synthetic data. In this light, we investigated the association of the presence of Autoimmune Diseases (Ads) with selected Urinary Organic Acids (OAs) and other demographic parameters through the integration of metabolomics and artificial intelligence (AI).

Methods

The original data of the present work came from a previously published study by Tsoukalas et al. (2020). The interested reader can find the rationale, design, inclusion criteria, exclusion criteria, and the main clinical findings of the study in detail in the literature. In short, a case-control study was undertaken to explore the distinct expression patterns and predictive potential of organic acids in individuals with ADs compared to healthy controls, with the underlying hypothesis that such differences exist between the two groups. Autoimmune diseases encompass a wide array of persistent conditions such as rheumatoid arthritis (RA), Hashimoto's thyroiditis (HT), psoriasis (PSO), vitiligo (VIT), and inflammatory bowel diseases (IBD). These diseases stem from a breakdown in immune tolerance towards self-components, affecting approximately 5-10% of the populace at present (Global Autoimmune Institute, 2024). The study was retrospective and was undertaken based on 392 participants.[1] The Organic Acids included in the analysis were: Citric acid, Isocitric acid, 2-ketoglutaric acid, Succinic acid, Malic acid, 3-hydroxy3-methyl glutaric acid, Lactic acid, Pyruvic acid, 3-hydroxybutyric acid, Pyroglutamic acid, 3-hydroxyisovaleric acid, Methylmalonic acid, Homovanillic acid, 5-HIAA, 4 Hydroxyphenylacetic acid, Orotic acid, 2-Hydroxyglutaric acid, Glycolic acid, Oxalic acid, Glyceric acid, 2-hydroxy isobutyric acid, 2-hydroxy butyric acid, Ethylmalonic acid, Methylsuccinic acid, Suberic acid, Methylcitric acid and 4HPPA.

The reconstruction of the data was based on the straightforward resampling technique of nonparametric bootstrapping. The term “nonparametric” indicates that unlike inference methods, which rely upon on parametric assumptions, the nonparametric approach uses computationally intensive methods to provide inferential results and 95% confidence intervals (CIs). These types of techniques, in contrast with analytical methods, are probabilistic and do not produce fixed results in each set of experiments. Specifically, from the raw dataset—containing data for all the relevant OAs—1,000 new datasets were drawn using random sampling -extracted by the uniform distribution- with replacement with the same number of observations. Mean values for the parameters of interest were obtained from each dataset

[1] All the data concerning the baseline characteristics and the level of Organic Acids included in the analysis are located at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7764183/, and the supplementary materials at: https://www.mdpi.com/2218-1989/10/12/502/s1.

and then were used to construct a new matrix with 1,000 rows. Bootstrapping was conducted separately for each group. We treated and transformed categorical variables (such as Gender, Smoking Status, etc.) as continuous variables in the new dataset of synthetic data. As an alternative approach, the reconstruction of the data was based on the well-established CART (Classification and Regression Tree) algorithm. CART algorithm represents a non-parametric technique that develops a hierarchical tree structure with homogenous sub-divisions estimating the conditional distribution of a univariate outcome given multivariate predictors (Abedinia & Seydi, 2024). As a second step, a non-strict preselection process was undertaken to extract the most important variables for a predictive neural network model. To this end, we used the Support Vector Machine (SVM) algorithm which, due to its performance, has been characterized as one of the most prominent machine learning techniques for classification (Guido et al., 2024). SVM identifies the hyperplane that optimally separates data into distinct classes ("cases"/" control" in this work) with the greatest margin. The basic characteristic of the hyperplane is that it maximizes the distance between the nearest data points of each class, and it was originally designed for binary classification problems. All data used in the SVM algorithm were normalized by median, log-transformed, and auto-centered.

As a second step, a feed-forward back-propagation supervised ANN (Chandra et al., 2020) was employed as a predictive model based on statistically significant OAs found in the previous model. Several architectures were investigated with shallow (one hidden layer) or medium (two hidden layers) models. For investigational purposes, we compared the ANN results with the original and the synthetic data. The available data were randomly divided into "training data set" (50.0%), "testing data set" (25.0%), and "holdout data set" (25.0%). In short, the training data set produced the model parameters, the testing data calculated the predictive errors, and the holdout data set assessed the generalization error (i.e. overfitting) of the final model. As an activation and output function, we used the sigmoid function, but also other non-linear functions or combinations in the sensitivity analysis were employed (Softmax) by researchers or by default settings run by the neural network software. Sigmoid is used for models built to forecast a probability as an outcome and might be a plausible approach for the question at hand. In our ANN models, data was trained using the Batch algorithm. The batch size signifies the number of samples utilized in a single iteration before updating the parameters of the model. We used the Scaled Conjugate Gradient (SCG) as an optimization method. In short, SCG initially computes the gradient, determines a step size (Initial Lambda: 0.0000005 in our case), and updates the calculations (Initial Sigma: 0.00005) to improve efficiency. The interval offset was set at $\pm$0.5. The main outputs of the ANN models were a) the architecture of the model, b) the confusion matrix, c) the ROC curves, d) and the normalized importance of the main variables. An ROC curve

(Receiver Operating Characteristic curve) is a graph that illustrates the performance of a binary classification model across all possible classification thresholds plotting the True Positive and False Positive Rate. As a last step, we implemented synaptic weight-based scoring on the synthetic dataset (2,000 observations) plus the real dataset (392 observations) and subsequently applied it to the real dataset, aiming to distinguish inconsistencies among models. To assess their effectiveness, we employed metrics including Sensitivity, Specificity, Positive Predictive Value (PPV), False Positive Rate (FPR), False Negative Rate (FNR), and Overall Accuracy (OA).

Our analyses were carried out with an 8-core desktop PC with a 64-bit operating system, Intel(R) Core (TM) i7-10700 CPU, 2.90 GHz microprocessor, and 16,0 GB of RAM. Statistical analyses for this manuscript were conducted using IBM SPSS 28 (IBM Corp., Armonk, N.Y., USA), licensed via the University of Patras, the free online r-project software[2], the "Metaboanalyst 6.0" web-based platform[3] and the use of VBA for Microsoft Excel; Version 19.

Results

The absolute concentrations of OAs for both arms for the original and the (bootstrap) synthetic data are shown in Table 1. The comparison between the original and the synthetic data revealed that the relative difference between them was estimated - on average- at 0.18% and 1.75% for case and control, respectively. As anticipated, the bootstrap standard error was significantly smaller than the original ones, as the bootstrap method empirically replicates the entire sampling distribution for each variable. Results based on the SVM model are presented in graphs 1 and 2. Graph 1 indicates the recursive R-SVM, with the use of linear kernel transformation. R-SVM performs classification recursively using different feature subsets. Features are selected based on their relative contribution to the classification using cross-validation error rates. The least important features are eliminated in the subsequent steps. This process creates a series of SVM models (levels). Per the best model, the classification error rate was estimated close to 25% including 33 (all) variables. Germane to this, the variables integrated into the SVM model display comparable degrees of importance, indicating again the necessity of including all of them in the ANN model (Figure 2).

Results of the ANN model with bootstrapped synthetic data are presented in Figures 3-5. The total number of synthetic observations was divided as follows: 1,050 observations were included in the training pool, 452 observations in the testing pool, and 474 in the holdout pool. 24 synthetic observations were excluded by the analysis. The classification success rate for the synthetic data was

[2] See https://www.r-project.org
[3] See https://www.metaboanalyst.ca/home.xhtml

estimated to be close to 100%. We estimated two types of synthetic data: pure synthetic and hybrid (synthetic data plus real observations) but the overall performance and the synaptic weights were estimated as almost identical in both versions.

Synaptic weights were estimated to be used as a score for the prediction of the real data set of 392 patients used in the original analysis. Figure 3 indicates the most important variables which contributed to the model. In particular, 3hydroxibutiric acid was estimated as the most important variable compared with the other biomarkers. The best model had as an activation function the hyperbolic and as an output function the Softmax. The classification table for the ANN with the original data is presented in Table 2. The model had a predicted accuracy of 76.6% for both groups and it was not prone to overfitting. Based on normalized importance, the 2-hydroxybutyric acid was the most important contributing factor in the model. ROC analysis (figure 5) indicated that the area under the curve was estimated at 0.800.

Synthetic data based on the CART model had relatively low comparability based on graphical methods (not shown, available on request), or based on specific metrics. Table 3 shows the results. A reduced Synthetic Mean Squared Error (SMSE) signifies that the model employed for synthetic data generation exhibits a decreased relative error in comparison to the variance present in the authentic data. This implies a more accurate correspondence between the synthetic and real data. Table 4 presents the final performance comparison among three models: the Artificial Neural Network (ANN) model utilizing real data, the ANN model employing bootstrapped synthetic data, and the model utilizing Classification and Regression Trees (CART) synthetic data.

Discussion

To the best of our knowledge, this is the first attempt to use synthetic data with ANNs in a case-control study conducted in Greece. In the present analysis, we attempted to create synthetic data based on previously published original data which were then employed as inputs in an AI routine (Artificial Neural Network) to predict the presence or absence of an autoimmune disease for a specific cohort. In this context, we employed an SVM model to identify the most significant metabolites for potential biomarker use. However, the SVM did not provide a clear separation pattern among the metabolites, leading to the conclusion that all metabolites might serve as predictors for the AI model. It is known in related literature, that SVM, despite being powerful, may deviate severely from the optimal solution since it is sensitive to outliers (Karamizadeh et al., 2014) and probably was also the case in the research at hand. In the original work, all biomarkers were clinically assessed but no deletions for potential outliers were undertaken since the sample was considered representative. Nonetheless, based on the

fact that dietary intake patterns and lifestyle changes are reflected strongly in metabolomic profiles (Guasch-Ferré et al., 2018), the metabolic biomarker's values might be compassionate and unpredictable (Tsoukalas et al., 2019; Tsoukalas et al., 2019).

Although the clinical implications of this outcome are beyond the scope of this study, the computational relevance of this finding is crucial, particularly for the development and optimization of synthetic data algorithms. An initial noteworthy observation to make is that their use was, until recently, limited due to the complex process of their production (Nowok et al., 2016). Indeed, the synthetic data-generating process requires expertise in statistics and data analysis, as well as software availability. The general idea of these synthetic data algorithms was to replace the original values with high-quality artificial ones. From a statistical standpoint, the included variables, categorical or continuous, could be synthesized one by one using sequential modeling. Hence, replacements were generated by drawing from conditional distributions fitted to the original data using parametric or classification and regression tree models.

A technical description of all these parametric or non-parametric algorithms and their procedure is out of the scope of this work and the interested reader can find more details and the mathematical background in the literature and its cited references (Lu et al., 2024). The main issue here is that the use of all metabolites in the synthetic data process demanded a high number of observations and increased the computational burden, propelling the overfitting, and creating higher variance (Nußberger et al., 2021). A plausible alternative here would be to consider using simpler generative models with a reduced number of variables based on clinical recommendations or considering another statistical procedure for biomarker selection. Some alternatives might include logistic regression, a Fold-Change analysis, a Volcano Plot analysis, a Wilcoxon Sign rank test, or o Random Forest algorithm. For some of these techniques, a Bonferroni correction or a False Discovery Rate analysis would have been applied to deal with false positive issues. It has to be highlighted that in the original paper, the authors (both of them are also authors in this work) used the abovementioned statistical procedures to reduce the size of potential biomarkers. Since our work here was investigational, we preferred to avoid overlapping with the previous work.

As a second approach, we also created a bootstrap synthetic data set with replacement based on empirical distribution. The main disadvantage of this method was that does not produce integers for categorical variables but only rational numbers (with decimals) for these types of data. Since our first aim was to recreate numbers able to produce similar predictive results to the original one, we did not make any transformations for categorical variables, but a plausible refinement might be the scope of

feature research on this topic. A reasonable initial approach to improve the quality of generated data would have been to ignore the decimal part of the bootstrap replications and to keep only the integer part of the synthetic data. An experimental approach is necessary to either confirm or refute the effectiveness of the suggested method. The overall performance of the bootstrapped model can be considered adequate and reasonable, while refinements could further improve their general use with other data or even other model's predictability (Michelucci & Venturini, 2021). Similar attempts with bootstrapped data have also been implemented with promising results. As a last step, we created 3 different ANNs. Results of the original model indicated that a concise system was constructed with a balanced sensitivity and specificity of close to 80% in both training and test sets. The optimum architecture was a small, swallow model that converged quickly and avoided overfitting.

Since the small (raw) data sets are commonplace in metabolomics, the development of adequate neural networks is a very crucial issue. This type of model, namely the multiplayer perceptron, represents one of the most common types of algorithms in this field and might be able to produce similar predictive accuracy if an adequate fine-tuning approach is followed. Despite some ethical considerations (Offenhuber, 2024), there are promising results that create opportunities for the generalized use of generative synthetic data not only for ANNs but also for other statistical models. This is especially important for the healthcare sector since this data is sensitive and has to be used to assist healthcare research and clinical practice.

References

Abedinia, A., & Seydi, V. (2024). Building semi-supervised decision trees with semi-cart algorithm. *International Journal of Machine Learning and Cybernetics*, 1-18. https://doi.org/10.1007/s13042-024-02161-z

Chandra Sekhar, C., Panda, N., Ramana, B.V., Maneesha, B., Vandana, S. (2021). Effectiveness of Backpropagation Algorithm in Healthcare Data Classification. In: Sharma, R., Mishra, M., Nayak, J., Naik, B., Pelusi, D. (Eds). *Green Technology for Smart City and Society. Lecture Notes in Networks and Systems*, vol 151. Springer, Singapore. https://doi.org/10.1007/978-981-15-8218-9_25

Giuffrè, M., & Shung, D. L. (2023). Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digital Medicine, 6*(1), 186. https://doi.org/10.1038/s41746-023-00927-3

Global Autoimmune Institute. (2024). *The Global Landscape of Autoimmune Disease*. Available at: https://www.autoimmuneinstitute.org/articles/the-global-landscape-of-autoimmune-disease/

Guasch-Ferré, M., Bhupathiraju, S. N., & Hu, F. B. (2018). Use of Metabolomics in Improving Assessment of Dietary Intake. *Clinical chemistry, 64*(1), 82–98. https://doi.org/10.1373/clinchem.2017.272344

Guido, R., Ferrisi, S., Lofaro, D., & Conforti, D. (2024) An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. *Information, 15*(4), 235. https://doi.org/10.3390/info15040235

Jawad, E. (2023). The Deep Neural Network-A Review. *IJRDO - Journal of Mathematics, 9*(9), 1-5. https://doi.org/10.53555/m.v9i9.5842

Karamizadeh, S., Abdullah, S.M., Halimi, M., Shayan, J., & Rajabi, M.J. (2014). Advantage and drawback of support vector machine functionality. *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, 63-65.

Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., & Wei, W. (2023). Machine learning for synthetic data generation: a review. Available at: https://arxiv.org/html/2302.04062v6/#S1

Mendez, K. M., Broadhurst, D. I., & Reinke, S. N. (2019). The application of artificial neural networks in metabolomics: a historical perspective. *Metabolomics: Official journal of the Metabolomic Society, 15*(11), 142. https://doi.org/10.1007/s11306-019-1608-0

Michelucci, U., & Venturini, F. (2021). Estimating Neural Network's Performance with Bootstrap: A Tutorial. *Machine Learning and Knowledge Extraction, 3*(2), 357-373. https://doi.org/10.3390/make3020018

Nowok, B., Raab, G.M., & Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, **74**(11), 1–26.

Nußberger, J., Boesel, F., Lenz, S., Binder, H. & Hess, M. (2021). Synthetic observations from deep generative models and binary omics data with limited sample size. *Briefings in Bioinformatics, 22*(4), bbaa226. https://doi.org/10.1093/bib/bbaa226

Offenhuber, D. (2024). Shapes and frictions of synthetic data. *Big Data & Society, 11*(2). https://doi.org/10.1177/20539517241249390

Paul, A. K. & Prasad, A. & Kumar, A. (2022). Review on Artificial Neural Network and its Application in the Field of Engineering. *Journal of Mechanical Engineering, 1*(1), 53-61.

Tsoukalas, D., Alegakis, A. K., Fragkiadaki, P., Papakonstantinou, E., Tsilimidos, G., Geraci, F., ... & Tsatsakis, A. (2019). Application of metabolomics part II: Focus on fatty acids and their metabolites in healthy adults. *International Journal of Molecular Medicine, 43*(1), 233-242. https://doi.org/10.3892/ijmm.2018.3989

Tsoukalas, D., Fragoulakis, V., Papakonstantinou, E., Antonaki, M., Vozikis, A., Tsatsakis, A., Buga, A. M., Mitroi, M., & Calina, D. (2020). Prediction of Autoimmune Diseases by Targeted Metabolomic Assay of Urinary Organic Acids. *Metabolites*, *10*(12), 502. https://doi.org/10.3390/metabo10120502

Tsoukalas, D., Fragoulakis, V., Sarandi, E., Docea, A. O., Papakonstaninou, E., Tsilimidos, G., ... & Calina, D. (2019). Targeted metabolomic analysis of serum fatty acids for the prediction of autoimmune diseases. *Frontiers in Molecular Biosciences, 6*, 120. https://doi.org/10.3389/fmolb.2019.00120

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, *49*(11), 1225–1231.

Conflict of Interest:

Non-Declared

Funding Information:

Non-Declared

Patients Consent:

Not Applicable

Authors' Contributions:

V.F. conceived and designed the study to fulfill the requirements of his post-doctoral accreditation research in the Department of Economic Science, University of Piraeus, supervised by Prof. Athanassios Vozikis. V.F. analyzed the study's database and wrote the manuscript of this paper. Prof. A.V. reviewed, corrected, and provided valuable comments throughout the manuscript. V.F. and A.V. read and approved the final version of the paper. V.F. is the guarantor for the overall content.

Acknowledgments:

This work was supported by the "European Institute of Molecular Medicine". The authors would like to thank all the administrative, technical, and medical staff of the "Metabolomic Medicine Clinic", for their dedicated involvement in this study and the provision of the research database.

The authors would also like to acknowledge the developers of freely available Artificial Intelligence (AI) online platforms "SciSpace," "ChatGPT," "PI," "Perplexity," and the "DeepL" AI Translator for their contribution to the research process and their assistance with the refinement of English text.

Tables & Figures

Table 1. Comparative organic acids analysis in the ADs group compared to control for original and synthetic data

	Autoimmune diseases	Control Group	Autoimmune diseases	Control Group
	Original Data		Bootstrapped Synthetic Data	
Citric acid	88.45±66.17	96.2+75.7	87.73+4.27	95.89±6.32
Isocitric acid	5.04±4.99	5.21+3.76	4.99+0.32	5.19±0.31
2-ketoglutaric acid	11.99±11.54	15.86+16.57	12.03+0.74	15.75±1.36
Succinic acid	3.07±7.27	4.91+13.82	2.68+5.9	4.87±1.11
Fumaric acid	0.04±0.27	0.07+0.31	0.03+0.02	0.07±0.02
Malic acid	0.40±0.86	0.66+0.63	0.4+0.05	0.66±0.05
3-hydroxy 3-methylglutaric acid	2.17±1.75	2.19+2.13	2.16+0.12	2.18±0.18
Lactic acid	7.88±9.63	16.81+75.43	7.88+0.62	17.05±6.23
Pyruvic acid	7.76±6.04	8.61+6.4	7.78+0.4	8.6±0.53
3-hydroxybutyric acid	9.14±54.47	5.44+16.3	9.14+3.6	5.42±1.3
Pyroglutamic acid	19.04±16.90	23.97+16.29	19.02+1.13	23.9±1.38
3 hydroxyisovaleric acid	10.25±10.52	13.98+15.29	10.34+0.69	14.02±1.25
Methylmalonic acid	0.63±0.97	0.95+0.87	0.63+0.06	0.94±0.07
Homovanillic acid	2.12±1.63	2.57+2.4	2.04+1.54	2.55±0.19
5-HIAA	2.69±3.01	3.51+5.52	2.67+0.2	3.52±0.46
4 Hydroxyphenylacetic acid	11.40±13.41	10.96+8.86	11.36+0.89	10.87±0.7
Orotic acid	0.01±0.16	0.01+0.11	0.01+0.01	0.01±0.01
2-Hydroxyglutaric acid	2.53±1.71	1.95+4.23	2.52+0.11	1.93±0.34
Glycolic acid	22.68±17.91	26.86+23.11	22.85+1.15	26.81±1.86
Oxalic acid	4.66±3.55	5.95+4.54	4.61+0.22	5.96±0.39
Glyceric acid	2.04±7.56	1.52+4.08	1.96+0.48	1.52±0.33
2-hydroxy isobutyric acid	4.75±2.81	2.79+3.94	4.78+0.18	2.79±0.32
2-hydroxy butyric acid	0.16±0.77	0.39+0.96	0.16+0.05	0.39±0.08
Ethylmalonic acid	1.64±2.26	1.9+1.9	1.62+0.15	1.9±0.15
Methylsuccinic acid	0.34±0.86	0.17+0.54	0.34+0.05	0.17±0.05
Suberic acid	0.08±0.55	0.1+0.39	0.08+0.04	0.1±0.03
Methylcitric acid	0.11±0.31	0.27+0.45	0.11+0.02	0.27±0.04
4HPPA acid	0.55±0.88	0.79+0.76	0.55+0.06	0.79±0.06

Concentrations of organic acids are expressed as mmol/mol Creatinine; 5-HIAA: 5- Hydroxyindoloacetic acid, 4-HPPA: 4-Hydroxyphenypyruvic acid; Synthetic data were produced via 1,000 non-parametric bootstrap replications

Figure 1. Recursive Support Vector Machine Classification

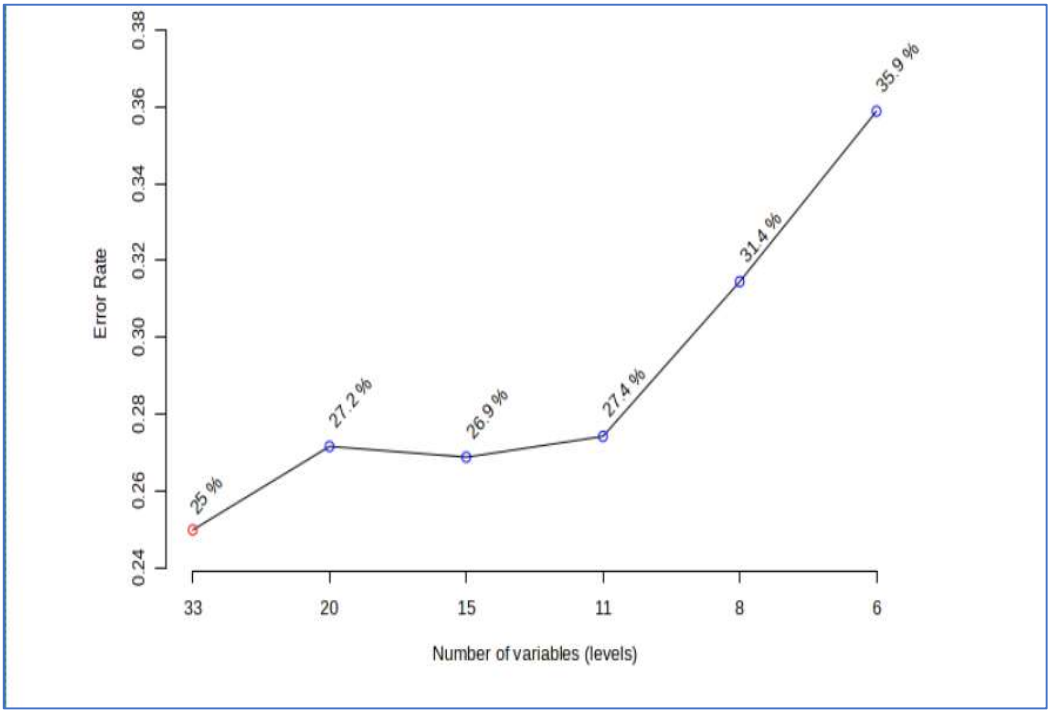
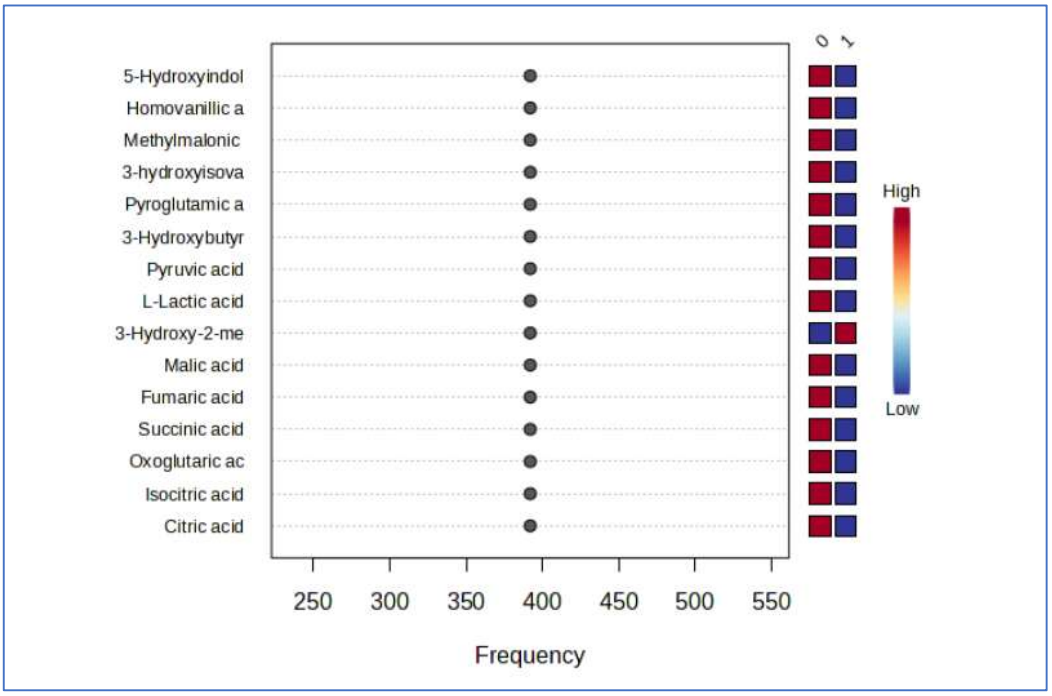


Figure 2. Support Vector Machine, Variable Importance



"O" represents control; "1" represents cases

Figure 3. Contribution of Biomarkers and Factors to the Predicted Accuracy of the ANN with Bootstrapped Synthetic Data

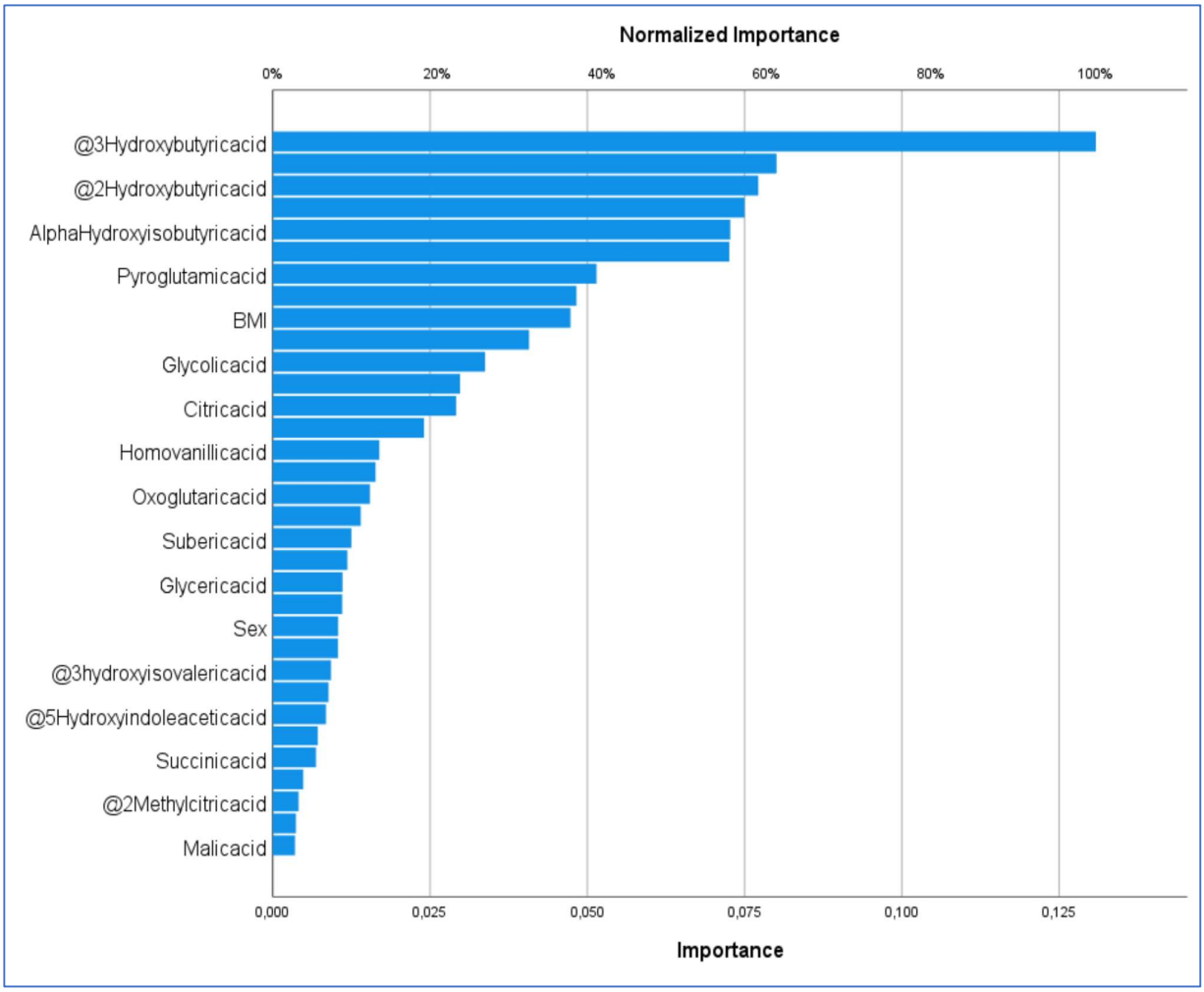


Figure 4. The architecture of the original Model with real data*

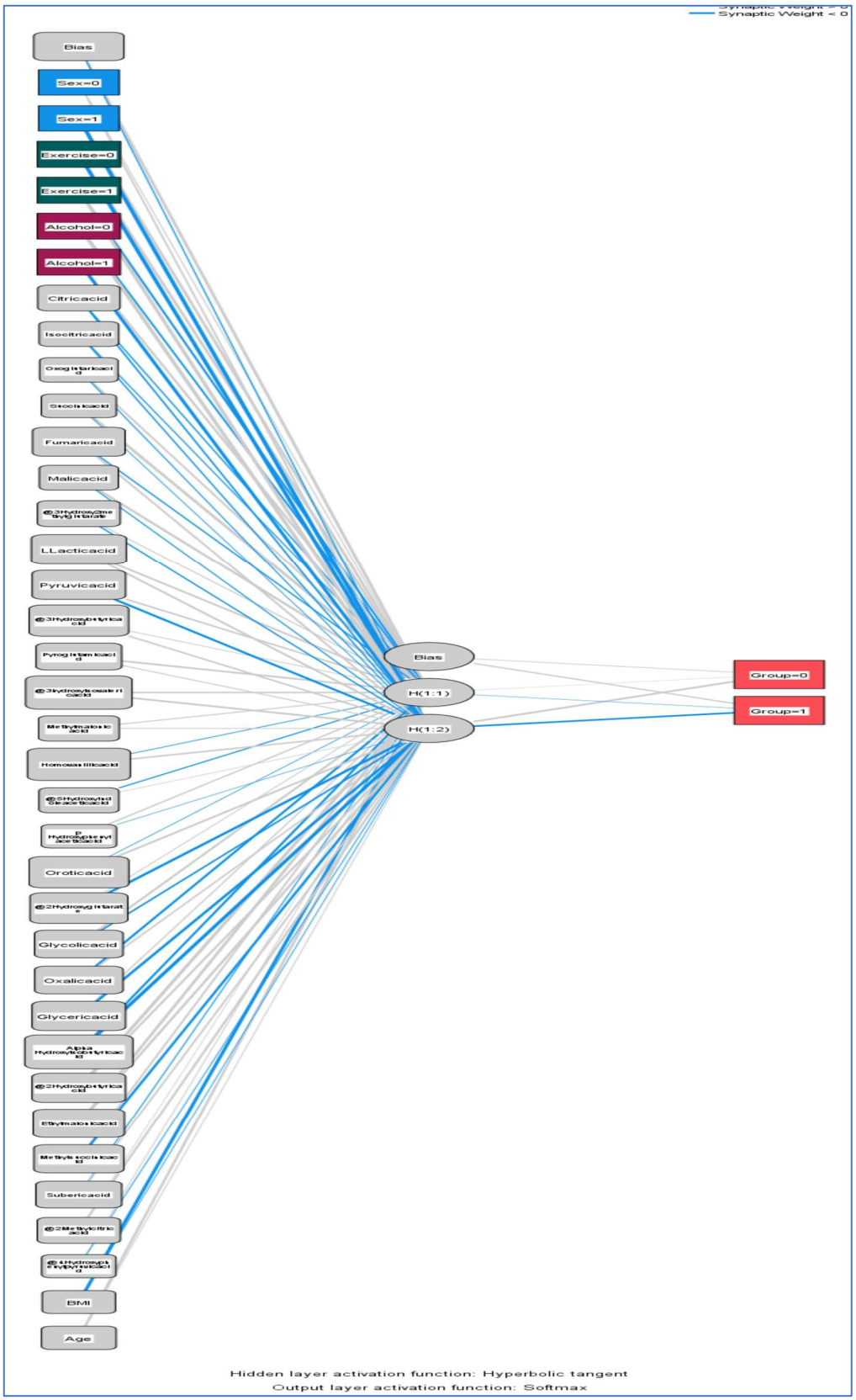
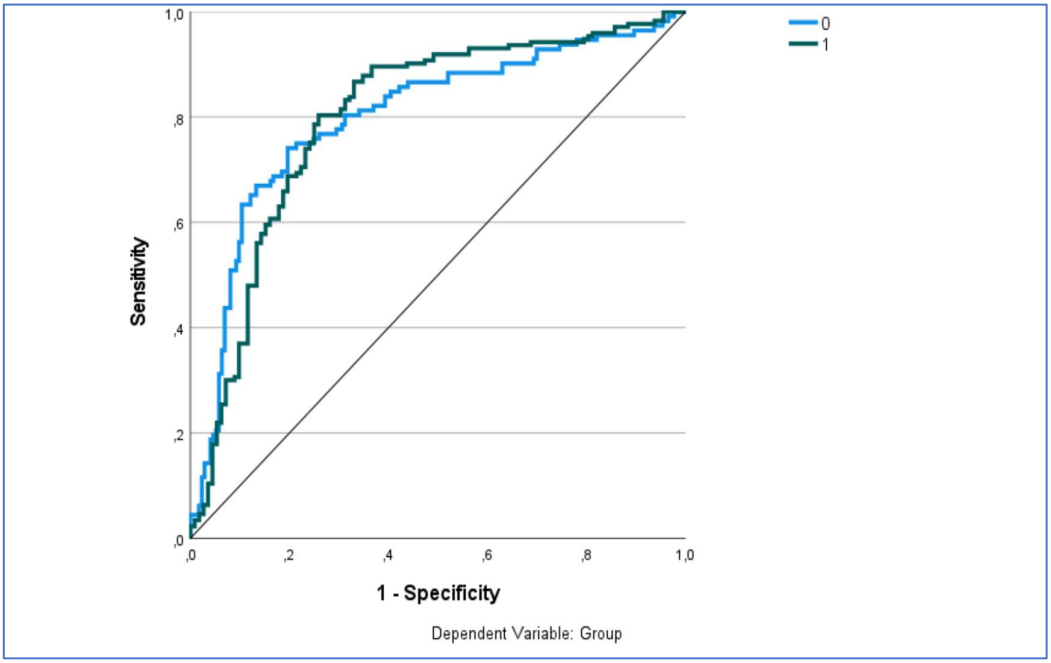


Table 2. Classification Table for Artificial Neural Network with Real Data

		Predicted		
		Case	Control	% Correct
Training	Case	58	16	78.4%
	Control	20	93	82.3%
	Overall Percent	41.7%	58.3%	80.7%
Testing	Case	24	14	63.2%
	Control	14	46	76.7%
	Overall Percent	38.8%	61.2%	71.4%
Holdout	Case	26	13	66.7%
	Control	12	56	82.4%
	Overall Percent	35.5%	64.5%	76.6%

Figure 5. ROC curve for the ANN with real data



"O" represents control; "1" represents cases

Table 3. Comparison of Synthetic data based on CART Model vs the Real Data

	MSE	S_MSE	df
Group	0.0000	0.0000	1
Citric.acid	0.0005	0.8119	4
Isocitric.acid	0.0009	1.4517	4
Oxoglutaric.acid	0.0010	1.5054	4
Succinic.acid	0.0003	0.6680	3
Fumaric.acid	-	-	0
Malic.acid	0.0002	1.3152	1
X3.Hydroxy.2.methylglutarate	0.0010	1.5098	4
L.Lactic.acid	0.0002	0.3019	4
Pyruvic.acid	0.0008	1.2574	4
X3.Hydroxybutyric.acid	0.0001	0.1716	2
Pyroglutamic.acid	0.0002	0.3293	4
X3.hydroxyisovaleric.acid	0.0002	0.2881	4
Methylmalonic.acid	0.0002	0.6034	2
Homovanillic.acid	0.0017	2.6462	4
X5.Hydroxyindoleacetic.acid	0.0010	1.6442	4
p.Hydroxyphenylacetic.acid	0.0014	2.2592	4
Orotic.acid	0.0006	1.3368	3
X2.Hydroxyglutarate	0.0020	3.1111	4
Glycolic.acid	0.0007	1.0447	4
Oxalic.acid	0.0010	1.5080	4
Glyceric.acid	0.0000	0.0299	2
Alpha.Hydroxyisobutyric.acid	0.0004	0.8380	3
X2.Hydroxybutyric.acid	-	-	0
Ethylmalonic.acid	0.0005	0.7632	4
Methylsuccinic.acid	-	-	0
Suberic.acid	-	-	0
X2.Methylcitric.acid	0.0003	1.0738	2
X4.Hydroxyphenylpyruvic.acid	0.0000	0.0922	1
Sex	-	-	1
Age	0.0012	1.8856	4
BMI	0.0009	1.4305	4
Exercise	0.0014	8.4738	1
Alcohol	0.0000	0.0989	1

MSE: Mean Square Error; SMSE: Standardized Mean Square Error; DF: Degree of Freedom

Table 4. Comparative Performance Indicators of Artificial Neural Network Models

	Real Data	Bootstrapped Synthetic	CART Synthetic Data
Sensitivity	84.00%	79.00%	72.40%
Specificity	77.84%	76.30%	60.60%
PPV	84.90%	83.70%	72.60%
NPV	76.50%	64.30%	60.60%
FPR	22.20%	23.70%	39.40%
FNR	23.40%	35.70%	39.20%
OA	77.30%	66.60%	51.27%

PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate; FNR: False Negative Rate; OA: Overall Performance