Supervised machine learning models and protein-protein interaction network analysis of gene expression profiles induced by omega-3 *polyunsaturated fatty acids*

Sergey Shityakov (shityakoff@hotmail.com)
Würzburg University https://orcid.org/0000-0002-6953-9771
Jane Pei-Chen Chang
China Medical University
Ching-Fang Sun
China Medical University Hospital
David Ta-Wei Guu
China Medical University
Thomas Dandekar
Julius-Maximilians-Universitat Wurzburg
Kuan-Pin Su
China Medical University Hospital

Research article

Keywords: gene expression, polyunsaturated fatty acids, supervised machine learning, protein-protein interaction networks, clustering

Posted Date: August 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-49619/v2

Version of Record: A version of this preprint was published at Current Chinese Science on January 12th, 2022. See the published version at https://doi.org/10.2174/2210298102666220112114505.

Abstract

Background

Omega-3 polyunsaturated fatty acids (PUFAs), such as eicosapentaenoic (EPA) and docosahexaenoic (DHA) acids have beneficial effects on human health but their effect on gene expression in elderly individuals (age ≥ 65) is largely unknown. To examine this, the gene expression profiles were analyzed in the healthy subjects (n = 96) at baseline and after 26 weeks of supplementation with EPA+DHA to determine up-regulated and down-regulated differentially expressed genes (DEGs) triggered by PUFAs. The protein-protein interaction networks were constructed by mapping these DEGs to a human interactome and linking them to the specific pathways.

Results

The results revealed that up-regulated DEGs were associated with neurotrophin/MAPK signaling, whereas the down-regulated DEGs were linked to the cancer, acute myeloid leukemia, and long-term depression pathways. Additionally, machine learning (ML) approaches were able to cluster the EPA/DHA-treated and control groups by the logistic regression algorithm performing the best.

Conclusion

Overall, this study highlights the pivotal changes in DEGs induced by PUFAs and provides the rationale for the implementation of ML algorithms as predictive models for this type of biomedical data.

Background

The beneficial effects of long-chain n-3 polyunsaturated fatty acids (PUFAs), such as eicosapentaenoic acid (EPA) and docosahexaenoic (DHA) are well known to promote human health [1]. The usage of PUFAs as food supplements demonstrated the ability to target inflammation and other underlying pathogenic factors found in depression and cancer [2-4]. On the other hand, the PUFAs impact on gene expression profiles in health and disease is not completely understood. Some researchers implemented the whole genome transcriptomics analysis to examine the genetic outcomes of a high EPA+DHA intake, resulting in a decreased gene expression involved in inflammatory and atherogenic-related pathways [5]. Moreover, PUFAs could modulate the cellular functions through specific changes in gene expression via DNA methylation mechanisms or binding and subsequent activation of peroxisome proliferator-activated receptors [6]. However, results from such studies are inconsistent, providing the basis for the development of alternative techniques, such as machine learning (ML) algorithms. These approaches can be used for the analysis of high-throughput deep sequencing data due to their computational efficiency and robustness in finding common patterns obtained from a small sample size [7]. In particular, a multivariate ML analysis, known as Gaussian process classification, confirmed that baseline fatty acids predicted response to treatment in the ω-3 PUFA group with high degrees of sensitivity, specificity, and accuracy [8]. In another study, researchers found that the ML algorithm based on the support-vector machines

classifier yielded good agreement with the certified/reference values for the prediction of EPA and DHA in the reference standard using vibrational spectroscopic data [9]. Moreover, some authors have also explored the interactions between serum free fatty acids, including PUFAs, and fecal microbiota in obese patients through the regression tree modeling, showing a non-obese-linked profile with serum EPA > 0.235 µg/mL [10]. Therefore, in the present study we aimed to identify differentially expressed genes (DEGs) associated with the particular metabolic or disease pathways in the healthy subjects and to assess the ability of different ML algorithms to predict and cluster the EPA/DHA- or high-oleic acid sunflower oil (HOSF)-specific gene expression profiles.

Results And Discussion

To obtain DEGs expressed in healthy individuals at baseline and after 26 weeks of treatment with 1.8 g of EPA+DHA mixture and HOSF as a negative control, publicly available microarray data set (GSE12375) was retrieved from the GEO repository. The unnormalized gene expression data (Supplementary Figure 1 [A, B]) was tested for quality assessment purposes calculating normalized unscaled standard error (NUSE). This quality control analysis, designed only for comparing arrays within one dataset, has revealed no signs of low-quality arrays with significantly elevated NUSE values (Supplementary Figure 2 [A, B]). Subsequently, data preprocessing was performed to eliminate the effect of background noise and to normalize and summarize the gene expression values per each probe of the database (Supplementary material 3 [A, B]) according to the standard protocol published elsewhere [15].

To evaluate the overall structure of the RNA-seq dataset, we performed the unsupervised dimensionality reduction of gene expression data (23,941 genes) at baseline and after 26 weeks of treatment with EPA+DHA mixture using PCA. As a result, no clustering was observed between EPA+DHA and HOSF samples (Figure 2 [A, B]) at these particular time points, indicating the possible limitation of the PCA methodology, which might be due to the effect size of the biological signal as well as on the fraction of samples containing this signal [16].

Next, the total number of DEGs was evaluated (Table 1 and Figure 3 [A]), showing a significant increase in the number of these genes after 26 weeks of treatment (1,805) with EPA+DHA mixture compared to the baseline (779). Similar patterns were observed in the study of Bouwens and coauthors where the EPA/DHA and HOSF consumptions using the same conditions and time points resulted in gene expression changes of 1,040 genes induced by PUFAs and 298 genes induced by the sunflower oil. Of these genes, 140 were overlapping between the groups, and 900 were belonged to the unique genes in the EPA/DHA group [5]. We identified 847 and 312 up-regulated unique genes and 753 and 262 unique down-regulated DEGs at baseline and after 26 weeks of PUFA treatment (Figure 3 [B, C]), where the number of up-regulated genes is slightly dominated over the down-regulated ones except for the overlapping (common) genes (97.5 ± 0.71 vs. 107.5 ± 0.71), showing the inverse dynamics (Figure 3 [C]).

To further investigate the functional impact of up-regulated and down-regulated DEGs upon the treatment with EPA+DHA mixture vs. control at different time points, these genes were mapped onto the DAVID

database to identify the specific metabolic or disease pathways. This database is known to be one of the most popular tools in the field of high-throughput functional annotation bioinformatics and microarray analysis [17]. Subsequently, the most significantly enriched KEGG pathways were linked to the neurotrophin and mitogen-activated protein kinase (MAPK) signaling, complement/coagulation cascades, and axone guidance for the up-regulated DEGs and cancer or acute myeloid leukemia pathways, long-term depression, fructose/mannose and arachidonic acid metabolism for the down-regulated genes (Table 2). Further, all the genes related to these pathways were subjected to the PPI network analysis to observe their involvement in the interconnections of different pathways (Figure 4 [A-D]).

Previously, some PUFAs, such as arachidonic acid, were discovered to activate MAPK in rat liver epithelial WB cells by a protein kinase C-dependent mechanism [18]. Moreover, it has been already demonstrated in many *in vitro*, *in vivo* and in clinical studies that PUFAs can downregulate cancer-related cellular proteins and modify signaling to inhibit tumor growth and metastatic rate and to extend patients' survival duration [19, 20]. Furthermore, PUFAs can promote the axonal outgrowth by translational regulation of Tau and collapsin response mediator protein 2 expression and probably for axonal guidance as well by the modulation of lipid rafts, which are essential for this particular pathway [21-23]. EPA and DHA are also well-known in alleviating symptoms of different mental illnesses starting from anxiety and depression to bipolar disorder and schizophrenia [24]. Recently, the PUFA effect on preoperative bleeding has been estimated in the randomized placebo-controlled clinical trial to be associated with a lower risk of bleeding at higher fish oil levels [25]. This information might support our findings that PUFAs could potentially trigger the upregulation of complement/coagulation cascade genes, reconsidering current recommendations not to use fish oil before cardiac surgery [26]. The negative impact of some PUFA dietary components on sugar metabolism has been also examined in the study, where the authors evaluated them for the prevention and treatment of type 2 diabetes [26]. The evidence might suggest that consumption of fish oil supplements at high doses could lead to a further worsening of glucose metabolism [27]. Another metabolic change caused by EPA and DHA is most likely related to the observed effects of marine n-3 PUFAs on eicosanoid profiles, resulting in a decreased arachidonic acid metabolism via the inhibition of phospholipase A2 and cyclooxygenase [28].

To construct the PPI network and maximize the genome coverage, the human interactome was implemented as a nonredundant and undirected binary data set comprising 16,018 unique HGNC-curated protein IDs and 299, 018 interactions compiled from the literature [15]. As an outcome, the PPI networks belonging to cancer/leukemia pathways and MAPK/neurotrophin signaling were identified as the subnetworks with a high number of nodes and edges, where the less expressed proteins were located on the periphery of the network. Interestingly, the gene expression of neurotrophins and their target receptors were determined to be differentially up-regulated by n-3 PUFAs via an age-dependent mechanism in the C57BL/6 mice cerebral cortex [29]. On the contrary, DHA has previously been shown to inhibit the induction and progression of acute myeloid leukemia cell lines (KG1a and HL-60) via the DHA-induced apoptosis, increasing the expression of the pro-apoptotic Bax protein [30, 31].

Before applying supervised ML algorithms, the inferential statistics, such as one-way ANOVA, was performed on DEGs comprising the analyzed subnetworks, to show that most of the gene expression data were statistically significantly different (p-value < 0.05) except for the plakoglobin (JUP), lymphoid enhancer-binding factor 1 (LEF1) genes, and MYC proto-oncogene from cancer/leukemia pathways (Table 3). The logistic regression, naïve Bayes, and DNN models were implemented to solve the binary (EPA/DHA vs. HOSF) classification problem together with the receiver operating characteristics (ROC) analysis. The ROC analysis is considered to be valuable in evaluating predictive models, including gene expression pipelines and molecular docking experiments, since it captures the trade-off between sensitivity and specificity over a continuous range [32-34]. The ROC curves were plotted for the ML models (Supplementary Figure 4 and 5 [A-C]) to assess the true and false positive rates as the total number of correct positive results predicted among all the positive samples and the total number of incorrect positive predictions among all negative samples in the dataset. Besides, the area under the ROC curve (AUC) was calculated as a numerical value that can be used to compare the logistic regression and naïve Bayes models getting maximal AUC values for the former model (Tables 4 and 5). Ideally, the best prediction model produces a curve on the top-left corner (0, 1) indicating perfect classification (100% sensitivity and specificity), whereas our ROC curves occurred in the top half of the graph, giving only a better-than-average model prediction. In our case, the ROC/AUC analysis cannot be used for screening and diagnostic purposes as is it lacks a minimum sample size (≈ 300) required to estimate the ML performance more precisely [35].

The DNN classifier with three hidden layers provided the worst performance on all the datasets with the lowest accuracy and harmonic mean of the precision and recall (F-score) values (Table 6) due to the small sample size (n = 96). Indeed, logistic regression models for binary classification can provide better performance than DNN because they are less prone to overfitting and not so difficult to train [36, 37]. Overall, the models exhibited their best performance values for the MAPK signaling pathway except for the Naïve Bayes model classifier, where the highest performance values were detected for cancer/leukemia pathway. Moreover, the “refined” model performances for this pathway were significantly improved after the exclusion of the JUP, LEF1, and MYC genes (p-value ≥ 0.05), which has contributed by decreasing the accuracy and F-score values, especially in the performance of DNN model. Finally, the decision surface analysis was utilized using the top two feature importance determinants for DEGs that contributed towards classifying the EPA/DHA and HOSF samples (Supplementary Figures 6 and 7 [A-C]). From Figures 5 and 6, it is clear that the logistic regression and naïve Bayes methods were able to separate the majority of the EPA/DHA samples from the HOSF ones, learning the underlying patterns quite well based on the most important DEG features. Even though the naïve Bayes classifier is considered to be one of the most popular classifiers for class prediction or pattern recognition from microarray gene expression data, it is highly sensitive to any outliers with the classical estimates of the location and scale parameters [38]. In fact, a discrepancy was observed between the Naïve Bayes model performance and data clustering for the cancer/leukemia pathway before the model “refinement”, where the good model performance produced no data clustering (Figure 6 [C]). This phenomenon might be explained by the interference of high DEG number (n = 12), small relative importance scores for the top

feature contributors, and the presence of statistically insignificant genes. In addition, the outlier analysis for the logistic regression and naïve Bayes models confirmed by the previous results on the model performances, where the former model outperformed the latter one by a decreasing number of outliers observed in cancer/leukemia pathway clustering (Table 7). Indeed, the outlier detection and removal procedure previously published elsewhere was able to improve the ML classification accuracy from 63% to 76% by reducing the variance of the training data and matching the accuracy of clinical judgment of medical experts [39]. Given that there is a limited number of medical facilities providing highly specialized and complex health care, this approach may be useful to improve the diagnostic process for patients without access to these facilities.

Conclusions

The present study analyzed the gene expression profiles and the PPI interaction pathways that may be triggered by the PUFA supplements, such as EPA and DHA in elderly healthy individuals using comprehensive machine learning methodology. To achieve this goal, the GSE12375 Affymetrix data were extracted from the GEO database to identify the up-regulated and down-regulated DEGs and to construct metabolism or disease-specific PPI networks, including cancer, acute myeloid leukemia, and long-term depression. In the next step, the ML techniques were able to cluster the EPA/DHA and HOSF groups, providing the best performance by using the logistic regression modeling. However, more research is needed to confirm these observations by improving the ML accuracy or performing clinical trials to pinpoint the PUFA effects on gene expression in healthy and sick individuals.

Methods

The transcriptional profile of GSE12375 was obtained from the Gene Expression Omnibus database [5], which is based on the Affymetrix NuGO array (human) using the NuGO-Hs1a520180 annotation data. In particular, fasting venous blood samples were collected at baseline (n = 48) and after 26 weeks (n = 48) of supplementation with either 1.8 g EPA/DHA mixture or HOSF. 4 ml blood was collected for the isolation of peripheral blood mononuclear cells (PBMC), using BD Vacutainer cell preparation tubes with sodium citrate (BD, Breda, The Netherlands). Immediately after blood collection, PBMCs were isolated according to the manufacturer's manual. Total RNA was isolated from all PBMC samples using the Qiagen RNeasy Micro kit (Qiagen, Venlo, The Netherlands), labeled using a one-cycle cDNA labeling kit (MessageAmpTM II-Biotin Enhanced Kit, Ambion, Inc.), and hybridized for custom-designed NuGO GeneChip arrays. Prior to DEG analysis, the probe cell intensity data (CEL files) were converted into the gene expression values, and the background correction was performed by the robust multi-array average (RMA) algorithm within the Bioconductor environment. The principal component analysis (PCA) was utilized for a clustering of the RMA-normalized gene expression data together with the variance proportions for the most contributing principal components [11]. The LIMMA (Linear Models for Microarray Data) algorithm was implemented to identify relevant DEGs at baseline and after 26 weeks of supplementation with either 1.8 g EPA/DHA mixture or HOSF with p-value < 0.05. These DEGs were subsequently submitted to the DAVID web server

to identify and construct the enriched KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways [12]. To demonstrate the potential protein-protein interactions (PPI), the DEGs were mapped on the compiled dataset of a human interactome for the PPI network construction using the Cytoscape v2.8 software platform. The human interactome was obtained from the laboratory of Cell Trafficking and Signal Transduction (University of Verona, Verona, Italy). The one-way analysis of variance (ANOVA) was used to compare two means from the EPA/DHA and HOSF groups using the F-distribution and to cluster the samples based on the gene expression data. Finally, the construction of ML models, including logistic regression, naïve Bayes, and deep neural networks (DNN) were implemented for the analyzed DEGs associated with the specific pathways according to the workflow (Figure 1) using the TensorFlow v1.12 and Keras v2.24 algorithms within the Python environment [13]. The DNN model was trained for 100 iterative cycles (epochs) over all the samples according to the protocol published elsewhere [14]. All the figures were prepared using the Meta-Chart online graphing tool, GraphPad Prism v7.0 software for Windows, R Graphics package, and matplotlib Python library. The R and bash scripts together with the Bioconductor and Cytoscape output files are provided in the Supplementary material.

Declarations

Supplementary information

The R, bash and Python scripts together with the Bioconductor and Cytoscape output files for supervised machine learning models and protein-protein interaction network analysis of gene expression profiles induced by omega-3 *polyunsaturated fatty acids* are provided in the Supplementary material.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All experimental and theoretical data are available as Supplementary material

Project name: AI models of PUFA-induced gene expression

Project home pages: https://github.com/virtualscreenlab/Virtual-Screen-Lab

Operating system(s): Linux

Programming language: Python, R

Other requirements: TensorFlow, Keras, matplotlib

License: GNU GPL

Any restrictions to use by non-academics: none

Competing interests

The authors declare that they have no competing interests

Funding

The authors of this work were supported by the following grants provided to K.P.S: MOST 106-2314-B-039-027-MY3, 108-2320-B-039-048, 108-2813-C-039-133-B, and 108-2314-B-039-016 from the Ministry of Science and Technology, Taiwan.

Authors' Contributions

S.S. initiated and executed the study, data analysis, and drafted the manuscript. J.C., C.F.S, and D.G. contributed to data analysis and drafted the manuscript. K.P.S. and T.D. initiated the study, initiated the analysis framework and contributed to drafting the manuscript. All authors have read and approved the manuscript.

Acknowledgments

Not applicable

Authors' Information

Sergey Shityakov is a researcher in bioinformatics, machine learning, genomic sequence analysis, and neuroscience; Jane Pei-Chen Chang is a researcher in molecular and clinical psychiatry and nutrition; Ching-Fang Sun is a researcher in molecular and clinical psychiatry and nutrition; David Ta-Wei Guu is a researcher in molecular and clinical psychiatry and nutrition; Thomas Dandekar is a professor in bioinformatics, machine learning, and genomic sequence analysis; Kuan-Pin Su is a professor in molecular and clinical psychiatry and nutrition.

Abbreviations

PUFAs: Polyunsaturated fatty acids; EPA: Eicosapentaenoic acid; DHA: Docosahexaenoic acid; DEGs: Differentially expressed genes; ML: Machine learning; HOSF: High-oleic acid sunflower oil; PCA: Principal component analysis; ANOVA: Analysis of variance; DNN: Deep neural networks; ROC: Receiver operating characteristics; AUC: Area under the ROC curve; PPI: Protein-protein interactions; PBMC: Peripheral blood mononuclear cells; MAPK: Mitogen-activated protein kinase

References

1. Abete P, Testa G, Galizia G, Della-Morte D, Cacciatore F, Rengo F. PUFA for human health: diet or supplementation? Curr Pharm Des. 2009;15(36):4186-90.
2. Yang B, Ren X-L, Li Z-H, Shi M-Q, Ding F, Su K-P, et al. Lowering effects of fish oil supplementation on proinflammatory markers in hypertension: results from a randomized controlled trial. Food Funct. 2020;11(2):1779-89.
3. Guu T-W, Mischoulon D, Sarris J, Hibbeln J, McNamara RK, Hamazaki K, et al. International Society for Nutritional Psychiatry Research Practice Guidelines for Omega-3 Fatty Acids in the Treatment of Major Depressive Disorder. Psychother Psychosom. 2019;88(5):263-73.
4. Brown I, Lee J, Sneddon AA, Cascio MG, Pertwee RG, Wahle KWJ, et al. Anticancer effects of n-3 EPA and DHA and their endocannabinoid derivatives on breast cancer cell growth and invasion. Prostaglandins Leukot Essent Fatty Acids. 2019:102024.
5. Bouwens M, van de Rest O, Dellschaft N, Bromhaar MG, de Groot LC, Geleijnse JM, et al. Fish-oil supplementation induces antiinflammatory gene expression profiles in human blood mononuclear cells. The American journal of clinical nutrition. 2009;90(2):415-24. Epub 2009/06/12.
6. Maktoobian Baharanchi E, Moradi Sarabi M, Naghibalhossaini F. Effects of Dietary Polyunsaturated Fatty Acids on DNA Methylation and the Expression of DNMT3b and PPARalpha Genes in Rats. Avicenna J Med Biotechnol. 2018;10(4):214-9.
7. Piles M, Fernandez-Lozano C, Velasco-Galilea M, Gonzalez-Rodriguez O, Sanchez JP, Torrallardona D, et al. Machine learning applied to transcriptomic data to identify genes associated with feed efficiency in pigs. Genet Sel Evol. 2019;51(1):10.
8. Amminger GP, Mechelli A, Rice S, Kim SW, Klier CM, McNamara RK, et al. Predictors of treatment response in young people at ultra-high risk for psychosis who received long-chain omega-3 fatty acids. Transl Psychiatry. 2015;5:e495.
9. Karunathilaka SR, Yakes BJ, Choi SH, Bruckner L, Mossoba MM. Comparison of the performance of partial least squares and support vector regressions for predicting fatty acids/fatty acid classes in marine oil dietary supplements using vibrational spectroscopic data. J Food Prot. 2020.
10. Fernandez-Navarro T, Diaz I, Gutierrez-Diaz I, Rodriguez-Carrio J, Suarez A, de Los Reyes-Gavilan CG, et al. Exploring the interactions between serum free fatty acids and fecal microbiota in obesity through a machine learning algorithm. Food Res Int. 2019;121:533-41.
11. Tsunoda T, Koh Y, Koizumi F, Tsukiyama S, Ueda H, Taguchi F, et al. Differential gene expression profiles and identification of the genes relevant to clinicopathologic factors in colorectal cancer selected by cDNA array method in combination with principal component analysis. Int J Oncol. 2003;23(1):49-59.
12. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003;4(5):P3.
13. Rampasek L, Goldenberg A. TensorFlow: Biology's Gateway to Deep Learning? Cell systems. 2016;2(1):12-4. Epub 2016/05/03.

14. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001;7(6):673-9.
15. Shityakov S, Dandekar T, Forster C. Gene expression profiles and protein-protein interaction network analysis in AIDS patients with HIV-associated encephalitis and dementia. HIV AIDS (Auckl). 2015;7:265-76.
16. Lenz M, Muller F-J, Zenke M, Schuppert A. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. Sci Rep. 2016;6:25696.
17. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.
18. Hii CS, Ferrante A, Edwards YS, Huang ZH, Hartfield PJ, Rathjen DA, et al. Activation of mitogen-activated protein kinase by arachidonic acid in rat liver epithelial WB cells by a protein kinase C-dependent mechanism. J Biol Chem. 1995;270(9):4201-4.
19. Gu Z, Shan K, Chen H, Chen YQ. n-3 Polyunsaturated Fatty Acids and their Role in Cancer Chemoprevention. Curr Pharmacol Rep. 2015;1(5):283-94.
20. Vaughan VC, Hassing MR, Lewandowski PA. Marine polyunsaturated fatty acids and cancer therapy. Br J Cancer. 2013;108(3):486-92.
21. Mita T, Mayanagi T, Ichijo H, Fukumoto K, Otsuka K, Sakai A, et al. Docosahexaenoic Acid Promotes Axon Outgrowth by Translational Regulation of Tau and Collapsin Response Mediator Protein 2 Expression. J Biol Chem. 2016;291(10):4955-65.
22. Stillwell W, Shaikh SR, Zerouga M, Siddiqui R, Wassall SR. Docosahexaenoic acid affects cell signaling by altering lipid rafts. Reprod Nutr Dev. 2005;45(5):559-79.
23. Guirland C, Zheng JQ. Membrane lipid rafts and their role in axon guidance. Adv Exp Med Biol. 2007;621:144-55.
24. Su KP, Shen WW, Huang SY. Effects of polyunsaturated fatty acids on psychiatric disorders. Am J Clin Nutr. 2000;72(5):1241.
25. Akintoye E, Sethi P, Harris WS, Thompson PA, Marchioli R, Tavazzi L, et al. Fish Oil and Perioperative Bleeding. Circ Cardiovasc Qual Outcomes. 2018;11(11):e004584.
26. Bedi HS, Tewarson V, Negi K. Bleeding risk of dietary supplements: A hidden nightmare for cardiac surgeons. Indian Heart J. 2016;68 Suppl 2:S249-S50.
27. Brown TJ, Brainard J, Song F, Wang X, Abdelhamid A, Hooper L, et al. Omega-3, omega-6, and total dietary polyunsaturated fat for prevention and treatment of type 2 diabetes mellitus: systematic review and meta-analysis of randomised controlled trials. Bmj. 2019;366:l4697.
28. Calder PC. Omega-3 polyunsaturated fatty acids and inflammatory processes: nutrition or pharmacology? Br J Clin Pharmacol. 2013;75(3):645-62.
29. Balogun KA, Cheema SK. The expression of neurotrophins is differentially regulated by omega-3 polyunsaturated fatty acids at weaning and postweaning in C57BL/6 mice cerebral cortex.

Neurochem Int. 2014;66:33-42.

30. Yamagami T, Porada CD, Pardini RS, Zanjani ED, Almeida-Porada G. Docosahexaenoic acid induces dose dependent cell death in an early undifferentiated subtype of acute myeloid leukemia cell line. Cancer Biol Ther. 2009;8(4):331-7.
31. Chiu LCM, Wong EYL, Ooi VEC. Docosahexaenoic acid modulates different genes in cell cycle and apoptosis to control growth of human leukemia HL-60 cells. Int J Oncol. 2004;25(3):737-44.
32. Parodi S, Muselli M, Fontana V, Bonassi S. ROC curves are a suitable and flexible tool for the analysis of gene expression profiles. Cytogenet Genome Res. 2003;101(1):90-1.
33. Shityakov S, Forster C. In silico predictive model to determine vector-mediated transport properties for the blood-brain barrier choline transporter. Adv Appl Bioinform Chem. 2014;7:23-36.
34. Shityakov S, Forster C. In silico structure-based screening of versatile P-glycoprotein inhibitors using polynomial empirical scoring functions. Adv Appl Bioinform Chem. 2014;7:1-9.
35. Bujang MA, Adnan TH. Requirements for Minimum Sample Size for Sensitivity and Specificity Analysis. J Clin Diagn Res. 2016;10(10):YE01-YE6.
36. Kim Y, Kim H-G, Li Z, Choi H-J. Avoiding Overfitting in Deep Neural Networks for Clinical Opinions Generation from General Blood Test Results. Stud Health Technol Inform. 2017;245:1274.
37. Hu Y, Luo S, Han L, Pan L, Zhang T. Deep supervised learning with mixture of neural networks. Artif Intell Med. 2020;102:101764.
38. Ahmed MS, Shahjaman M, Rana MM, Mollah MNH. Robustification of Naive Bayes Classifier and Its Application for Microarray Gene Expression Data Analysis. Biomed Res Int. 2017;2017:3020627.
39. Li W, Mo W, Zhang X, Squiers JJ, Lu Y, Sellke EW, et al. Outlier detection and removal improves accuracy of machine learning approach to multispectral burn diagnostic imaging. J Biomed Opt. 2015;20(12):121305.

Tables

Table 1: Gene expression characteristics determined for healthy elderly individuals (n = 96) at baseline and after 26 weeks of treatment with 1.8 g of EPA+DHA mixture and HOSF as a negative control.

Time	Total genes	DEGs				
		Total	Common	Unique	Up-regulated	Down-regulated
Baseline	23941	779	205	574	408	371
26 week	23941	1805	205	1600	944	861

Table 2: Enriched KEGG pathways (p-value < 0.05) of the total, up-regulated and down-regulated DEGs in samples of healthy elderly individuals (n = 96) at baseline and after 26 weeks of treatment with 1.8 g of EPA+DHA mixture and HOSF as a negative control.

DEGs		Pathway	Genes	P-value
Total	Common	–	–	–
	Unique (baseline)	Complement/coagulation cascade Long-term depression Neurotrophin signaling	6 6 8	0.027 0.027 0.033
	Unique (26 week)	Pathways in cancer Acute myeloid leukemia	38 11	0.0061 0.0095
Up-regulated	Common (baseline)	–	–	–
	Common (26 week)	–	–	–
	Unique (baseline)	Neurotrophin signaling Complement/coagulation cascade	7 5	0.0047 0.011
	Unique (26 week)	Axon guidance MAPK signaling	11 17	0.022 0.04
Down-regulated	Common (baseline)	Arachidonic acid metabolism	3	0.045
	Common (26 week)	–	–	–
	Unique (baseline)	Long-term depression	4	0.042
	Unique (26 week)	Acute myeloid leukemia Pathways in cancer Fructose mannose metabolism	10 20 5	0.00023 0.031 0.035

Table 3: One-way ANOVA statistics for enriched KEGG pathways, using DEGs from samples of healthy elderly individuals (n = 96) at baseline and after 26 weeks of treatment with 1.8 g of EPA+DHA mixture and HOSF as a negative control.

Pathways in cancer			MAPK signaling			Neurotrophin signaling		
DEGs	F	p-value	DEGs	F	p-value	DEGs	F	p-value
CDK4	5.36	0.025	DUSP1	4.66	0.036	RAP1A	6.94	0.011
IGF1R	7.36	0.009	FOS	3.97	0.052	RPS6KA1	5.88	0.019
JUP	3.98	0.052	GNG12	4.73	0.035	IRS4	9.48	0.003
LEF1	3.92	0.054	MAP2K6	5.62	0.022	CSK	7.58	0.008
MTOR	4.41	0.041	MAPK1	4.24	0.045	RPS6KA4	5.08	0.029
MYC	3.88	0.054	NFATC2	5.69	0.021	IRAK3	8.68	0.005
NFKB1	7.16	0.01	NTF3	6.18	0.016	NTRK3	4.67	0.036
PML	7.78	0.008	NTRK2	5.15	0.028			
STAT3	8.14	0.007	PAK1	5.66	0.021			
STAT5A	6.89	0.012	PRKCG	11.74	0.001			
STAT5B	6.58	0.013						
TRAF1	4.04	0.05						

Table 4: Logistic regression model performance for enriched KEGG pathways, using DEGs from samples of healthy elderly individuals (n = 96) at baseline and after 26 weeks of treatment with 1.8 g of EPA+DHA mixture and HOSF as a negative control.

Pathway	Genes	Accuracy	Precision	F-score	AUC
Neurotrophin signaling	7	0.67	0.74	0.68	0.78
MAPK signaling	10	0.73	0.85	0.74	0.80
Pathways in cancer	12	0.6	0.81	0.59	0.82
Pathways in cancer*	8	0.73	0.85	0.74	0.84

*: refined model

Table 5: Naïve Bayes model performance for enriched KEGG pathways, using DEGs from samples of healthy elderly individuals (n = 96) at baseline and after 26 weeks of treatment with 1.8 g of EPA+DHA mixture and HOSF as a negative control.

Pathway	Genes	Accuracy	Precision	F-score	AUC
Neurotrophin signaling	7	0.53	0.67	0.53	0.58
MAPK signaling	10	0.6	0.7	0.61	0.78
Pathways in cancer	12	0.73	0.85	0.74	0.86
Pathways in cancer*	8	0.8	0.88	0.81	0.80

*: refined model

Table 6: DNN model performance for enriched KEGG pathways, using DEGs from samples of healthy elderly individuals (n = 96) at baseline and after 26 weeks of treatment with 1.8 g of EPA+DHA mixture and HOSF as a negative control.

Pathway	Genes	Accuracy	Precision	F-score
Neurotrophin signaling	7	0.6	0.7	0.61
MAPK signaling	10	0.67	0.83	0.67
Pathways in cancer	12	0.4	0.79	0.3
Pathways in cancer*	8	0.53	0.81	0.5

*: refined model

Table 7: Outliers for the logistic regression and naïve Bayes models of enriched KEGG pathways, using DEGs from samples of healthy elderly individuals (n = 96) at baseline and after 26 weeks of treatment with 1.8 g of EPA+DHA mixture and HOSF as a negative control.

Pathway	Logistic regression		Naïve Bayes	
	EPA/DHA	HOSF	EPA/HDA	HOSF
Neurotrophin signaling	3	4	2	5
MAPK signaling	1	4	0	5
Pathways in cancer	4	6	3	12
Pathways in cancer*	3	5	4	2

*: refined model

Figures

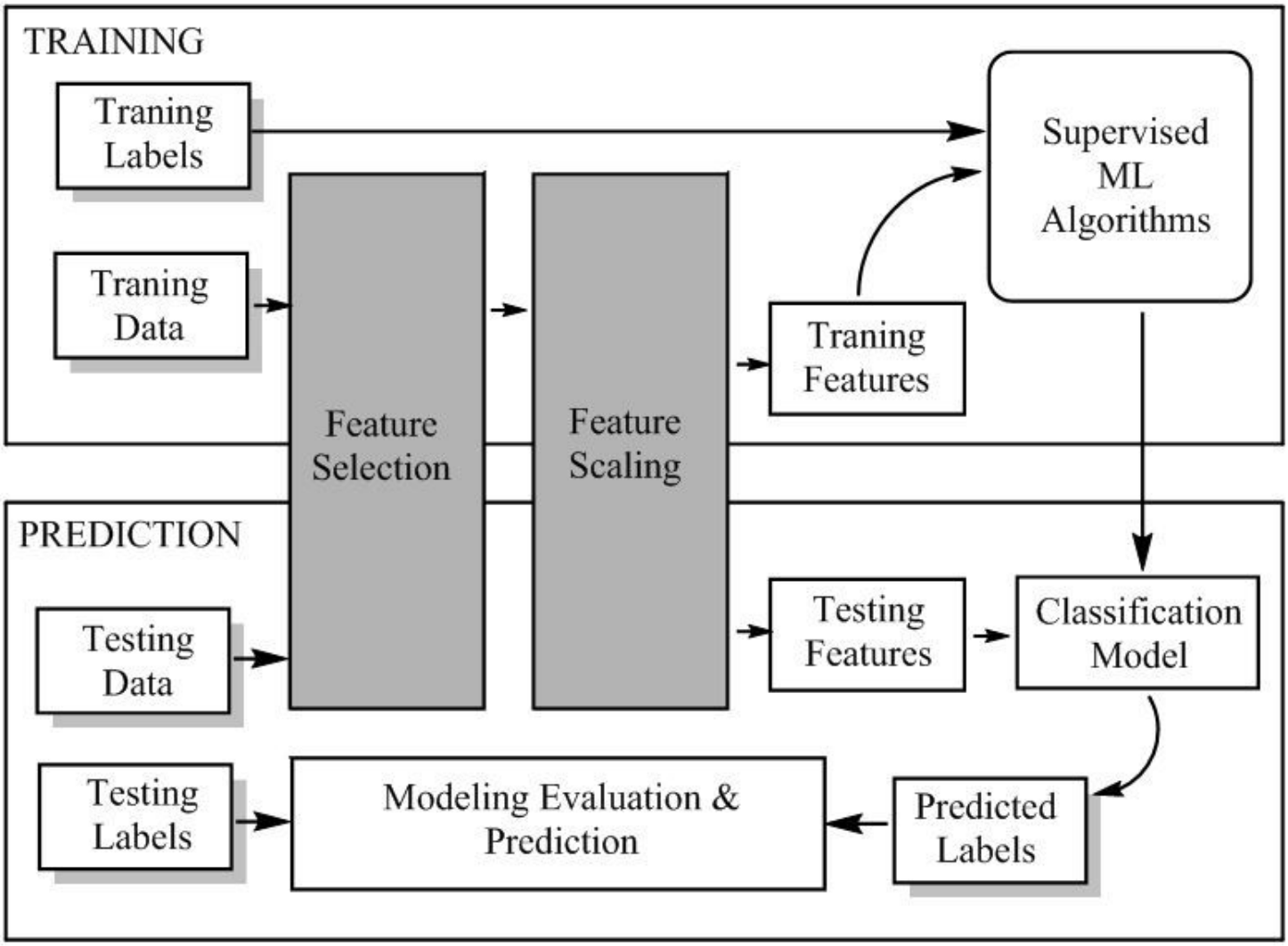

Figure 1

Supervised ML workflow for the PUFAs vs. control (HOSF) classification system.

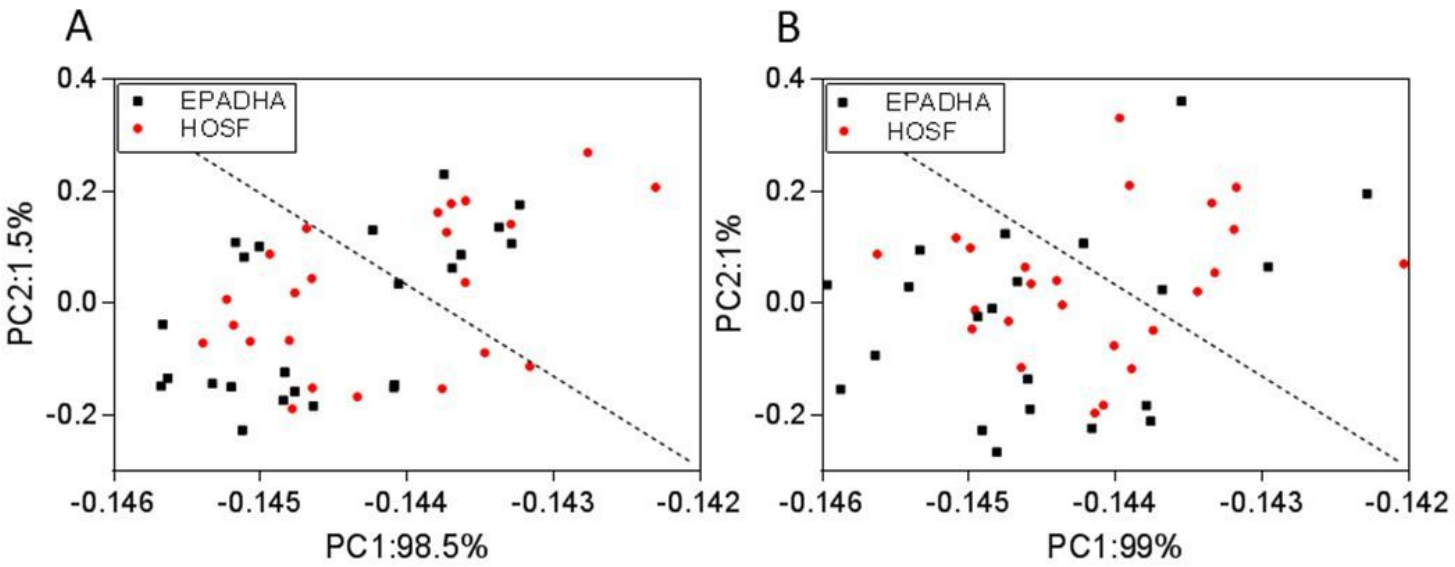

Figure 2

The PCA analysis of RMA-normalized gene expression data for healthy elderly individuals (n = 48) at baseline (A) and after 26 weeks (B) of treatment with 1.8 g of EPA+DHA mixture and HOSF as a negative control. Proportions of variance for principal components 1 (PC1) and 2 (PC2) are shown in percentage.

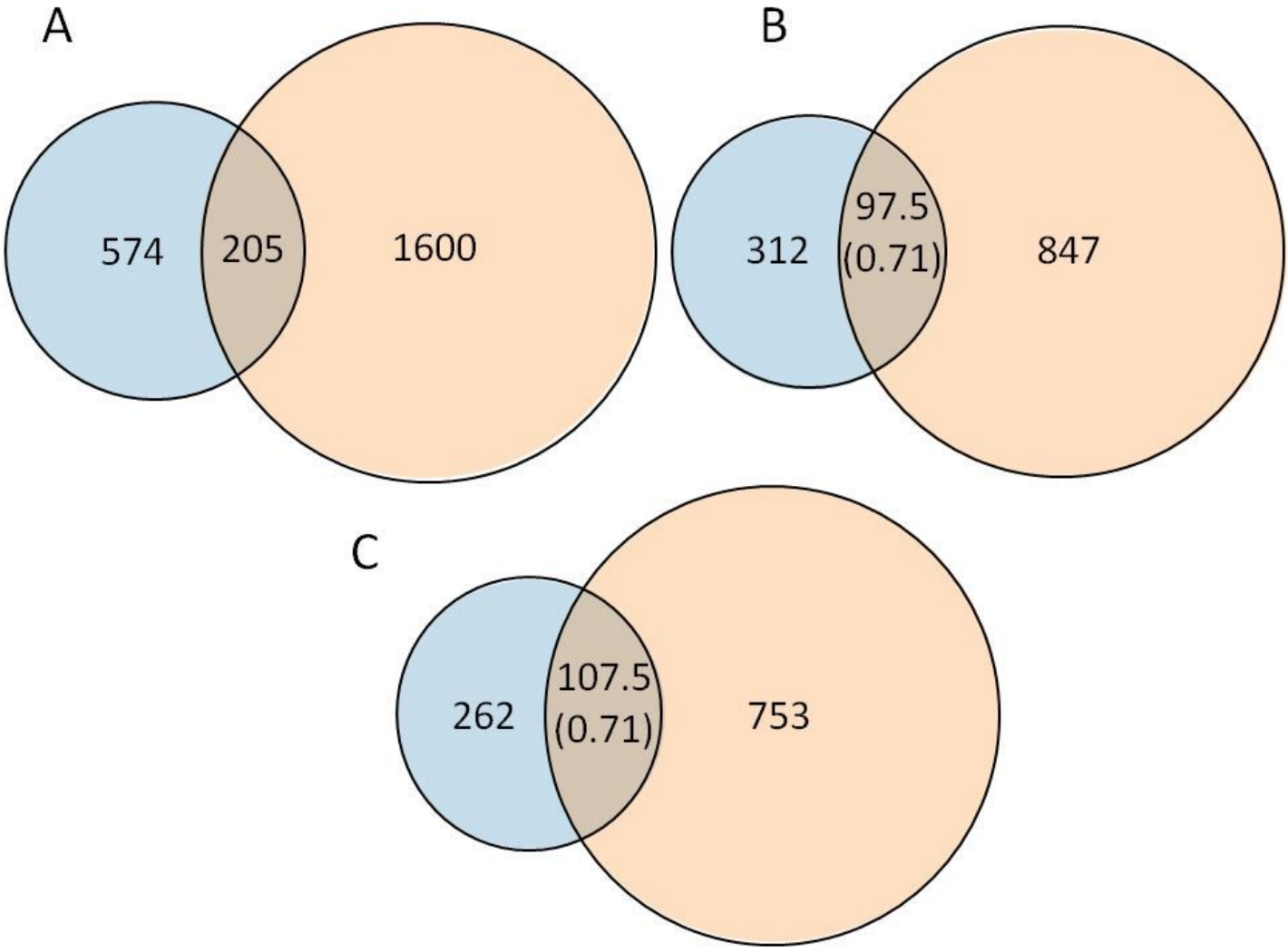

Figure 3

Venn diagram of total DEGs (A) and their up-regulated (B) and down-regulated (C) fractions determined for healthy elderly individuals (n = 96) at baseline and after 26 weeks of treatment with 1.8 g of EPA+DHA mixture and HOSF as a negative control. The up-regulated and down-regulated common DEGs are present as mean ± (SD).

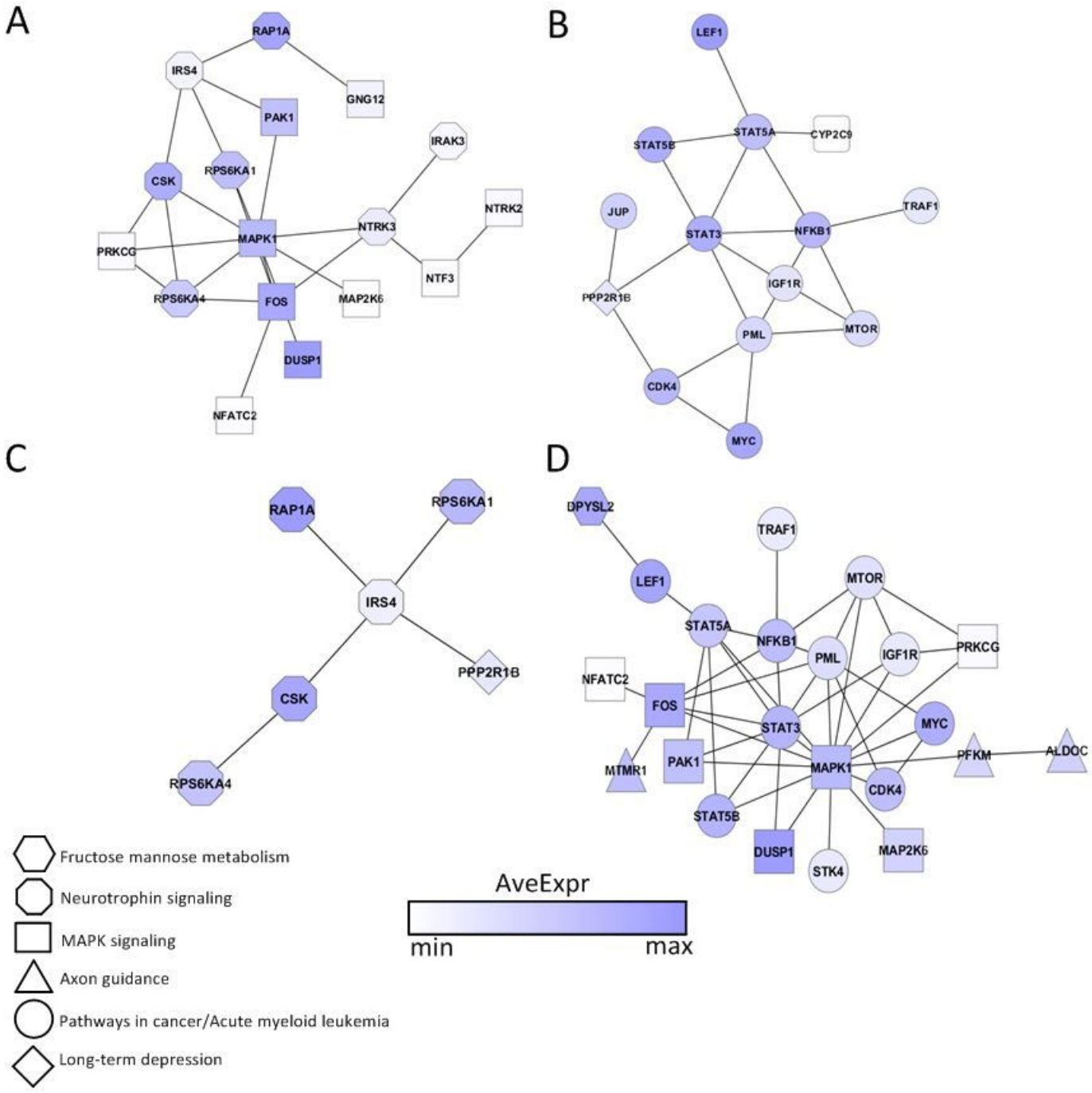

Figure 4

PPI networks linked to the specific pathways constructed by mapping the up-regulated (A) and down-regulated (B) DEGs together with DEGs determined at baseline (C) and after 26 weeks (D) of treatment with 1.8 g of EPA+DHA mixture and HOSF as a negative control. The nodes are colored according to the gene average expression (AveExpr) values. All PPI networks are displayed using the unweighted force-directed layout.

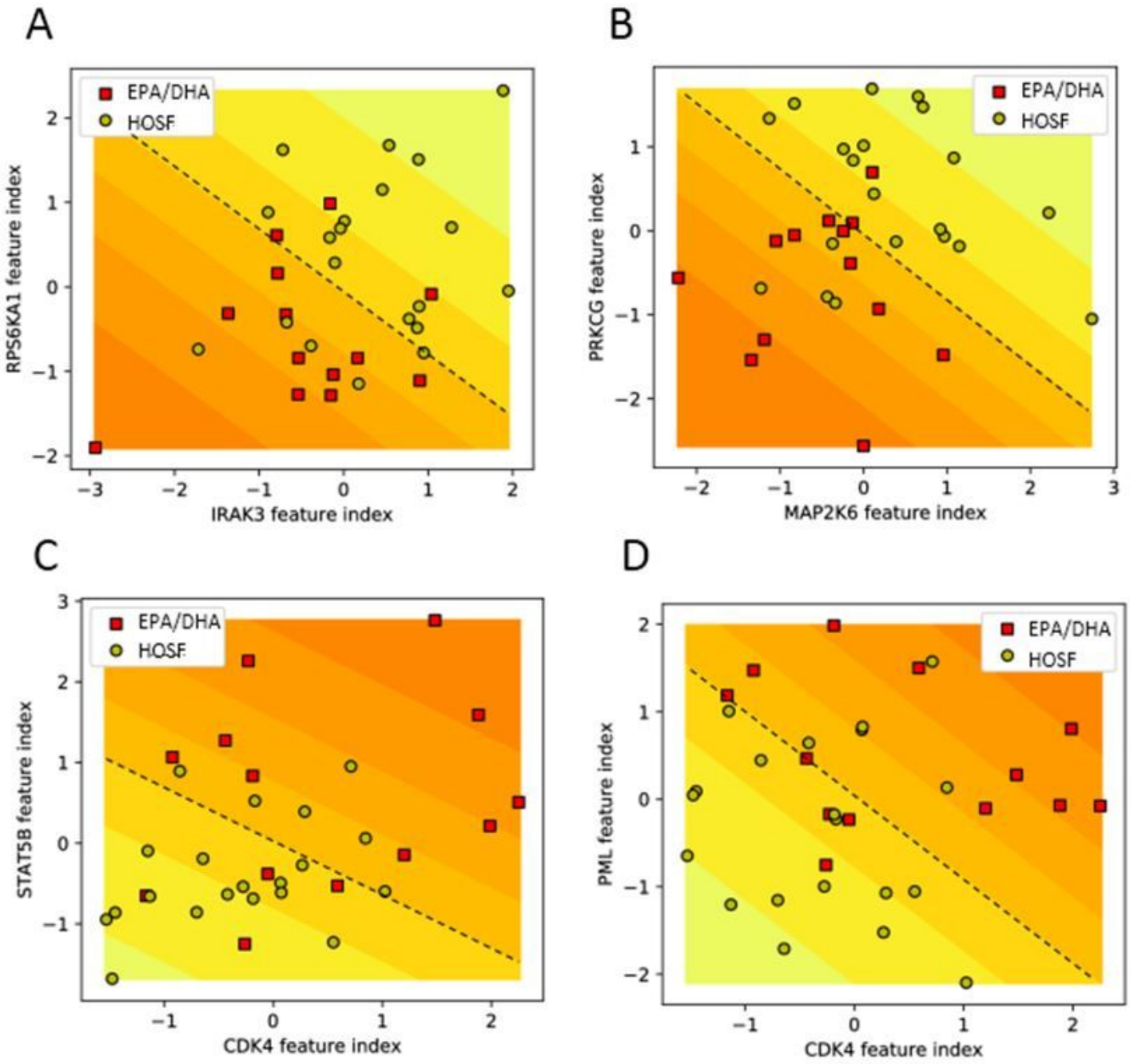

Figure 5

Decision surface of the logistic regression models using two top DEG features, which belong to neurotrophin (A) and MAPK (B) signaling, and the pathways associated with cancer and acute myeloid leukemia before (C) and after the refinement (D).

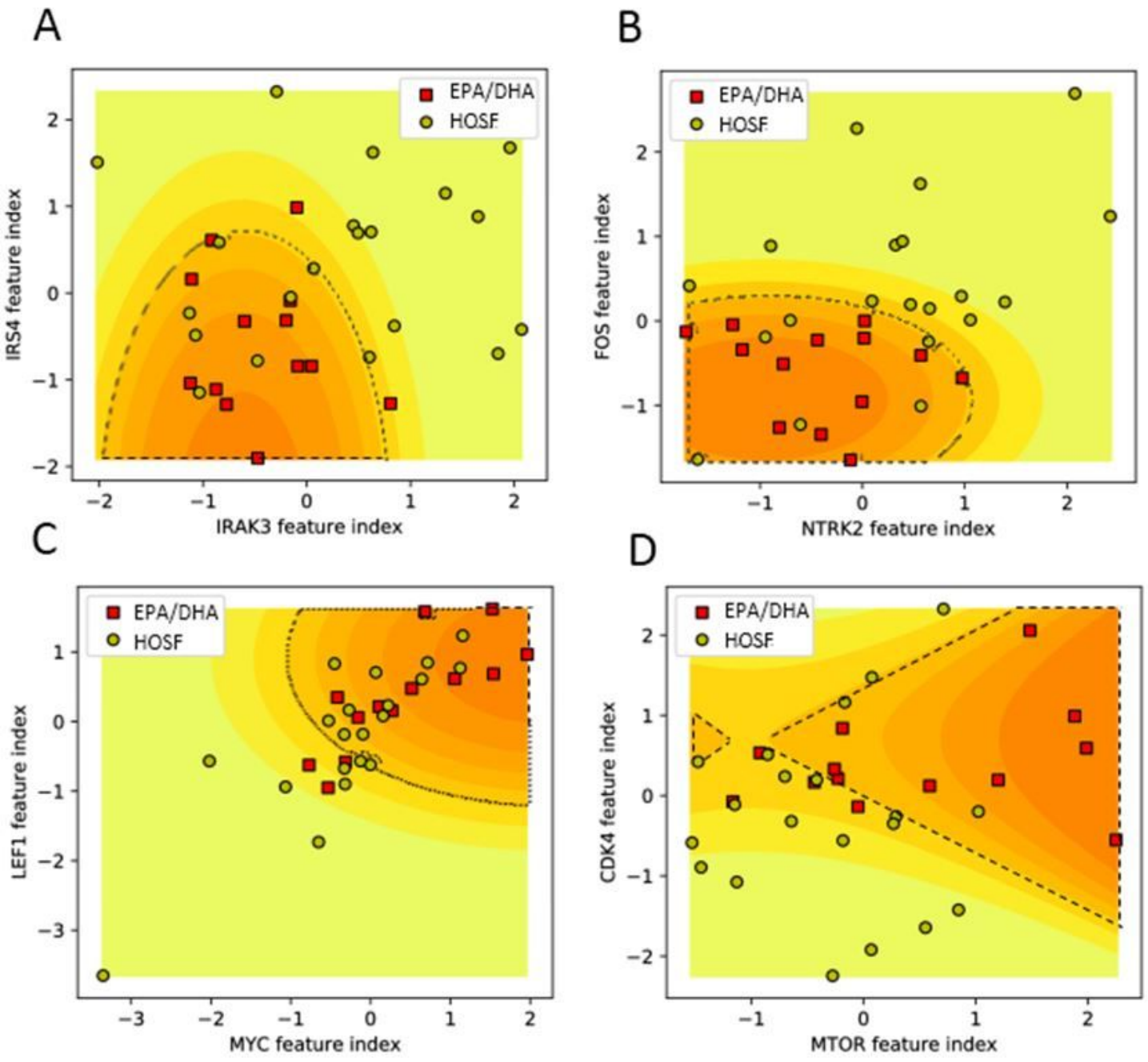

Figure 6

Decision surface of the naïve Bayes models using two top DEG features, which belong to neurotrophin (A) and MAPK (B) signaling, and the pathways associated with cancer and acute myeloid leukemia before (C) and after the refinement (D).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymetarial.zip

- GraphicalAbstractSupplementary.doc