AIMS Bioengineering

AIMS Bioengineering, 3(4): 552-565.
DOI: 10.3934/bioeng.2016.4.552
Received: 06 August 2016
Accepted: 28 November 2016
Published: 14 December 2016

http://www.aimspress.com/journal/Bioengineering

Research article

Development and validation of a skin fibroblast biomarker profile for schizophrenic patients

Marianthi Logotheti [1,2], Eleftherios Pilalis [2,3], Nikolaos Venizelos [1], Fragiskos Kolisis [4], and Aristotelis Chatziioannou [2,3,*]

[1] Neuropsychiatric Research Laboratory, Faculty of Medicine and Health, School of Health and Medical Sciences, Örebro University, Örebro, Sweden
[2] Metabolic Engineering and Bioinformatics Group, Institute of Biology, Medicinal Chemistry and Biotechnology, National Hellenic Research Foundation, Athens, Greece
[3] e-NIOS Applications PC, Athens, Greece
[4] Laboratory of Biotechnology, School of Chemical Engineering, National Technical University of Athens, Athens, Greece

* **Correspondence:** Email: achatzi@eie.gr; Tel: +30-210-727-3751.

Abstract: Gene expression profiles of non-neural tissues through microarray technology could be used in schizophrenia studies, adding more information to the results from similar studies on postmortem brain tissue. The ultimate goal of such studies is to develop accessible biomarkers. Supervised machine learning methodologies were used, in order to examine if the gene expression from skin fibroblast cells could be exploited for the classification of schizophrenic subjects. A dataset of skin fibroblasts gene expression of schizophrenia patients was obtained from Gene Expression Omnibus database. After applying statistical criteria, we concluded to genes that present a differential expression between the schizophrenic patients and the healthy controls. Based on those genes, functional profiling was performed with the BioInfoMiner web tool. After the statistical analysis, 63 genes were identified as differentially expressed. The functional profiling revealed interesting terms and pathways, such as mitogen activated protein kinase and cyclic adenosine monophosphate signaling pathways, as well as immune-related mechanisms. A subset of 16 differentially expressed genes from fibroblast gene expression profiling that occurred after Support Vector Machines Recursive Feature Elimination could efficiently separate schizophrenic from healthy controls subjects. These findings suggest that through the analysis of fibroblast based gene expression signature and with the application of machine learning methodologies we might conclude to a diagnostic classification model in schizophrenia.

Keywords: schizophrenia; peripheral biomarker; fibroblasts; machine learning; classification

1. Introduction

Schizophrenia (SZ) is a chronic psychiatric disorder with mean lifetime prevalence of almost 1% [1]. The definition of the disease has changed over the last century, but its causal pathophysiology remains obscure [2]. SZ is diagnosed on the basis of specific criteria presented in the fifth edition of Diagnostic and Statistical Manual of Mental Disorders (DSM-V) [3] or the tenth revision of International Statistical Classification of Diseases and Related Health Problem [4]. Among others, these criteria include the appearance of the positive, negative and cognitive symptoms of SZ. Positive symptoms include psychotic symptoms such as delusions and hallucinations with the characteristic lack of contact with the reality. The negative symptoms include alogia, social withdrawal, avolition and anhedonia and the cognitive symptoms include cognitive abnormalities such as working memory impairments [1].

The etiology of the disease seems to be heterogeneous affected both by environmental and genetic factors. Generally in all psychiatric disorders there is a lack of strict boundaries between the normal and the diseased condition and there is an overlap of symptoms among different disorders [5]. It is also a fact that SZ begins long before the typical diagnosis of the disease takes place. For this reason it is important to explore approaches for the identification of subjects being at risk and subjects being in the early stages of the disease [1]. These facts result to the need for identification of reliable and objective diagnostic means such as biomarkers for the discrimination among the psychiatric disorders and among the healthy and the diseased state [5].

1.1. Machine learning methodologies towards possible biomarkers

It is very important to discover biomarkers in the field of psychiatry in order to early diagnose the psychiatric disorders. It will certainly prove beneficial to detect preliminary psychotic symptoms even in prodromal phases on the maximum possible objective basis. High dimensional microarray data can be utilized for the development of biomarkers, since they pinpoint different molecular contributions to the disease pathophysiology [6]. Gene expression datasets usually include small number of samples and thousands of genes. In such high throughput studies, it remains a big challenge to select those biomarker genes that are crucial for distinguishing the sample classes [7]. Supervised learning algorithms utilize samples with known class for the training, in order to develop a model for classification of new examples that were not taken into consideration during the classification task [8].

1.2. *Human skin fibroblasts as a cell model*

Human skin fibroblast cells have been used as an experimental model for studying the pathophysiology of psychiatric disorders such as SZ [9]. Since there are indications that SZ has a strong genetic background, the genetic contribution of the disease is probably preserved in cell cultures, such as fibroblast cell cultures. Fibroblasts share some common molecular characteristics with human brain microvascular endothelial cells [10]. Additionally, they have the advantage of being easily obtained from the subjects under study. Moreover, the medication status of the patients does not greatly affect the skin fibroblast cells [11].

No validated biological test for the discrimination of schizophrenic patients from healthy controls has been adopted in clinical practice so far. Gene expression differences between schizophrenic patients and healthy controls can be used towards the identification of biomarkers that can distinguish the patient and the healthy control classes. The model of human skin fibroblast cells could greatly contribute to the identification of brain disease biomarkers. Blood-based studies have been already used for the development of biological classifiers in bipolar disorder and SZ [6,12]. In this study, we examine the possibility of developing a classifier based on skin fibroblast gene expression of schizophrenic patients. The fibroblast cell model has some disadvantages like the fact that the age of the donors may affect the cell viability and the fact that they do not include physiological, *in vivo* signals [9]. Further confirmatory examination was performed in this study so that the diseased tissue will be also examined. More specifically, in this study it was examined if the genes identified to result in separating schizophrenic and healthy subjects based on the fibroblast gene expression profiling, could be also used for the discrimination of SZ and the healthy control postmortem brain samples.

2. Materials and Methods

2.1. *Dataset*

The dataset that was used in this study contains data collected by Cattane et al. [11], accessible at National Center For Biotechnology Information (NCBI) [13] Gene Expression Omnibus (GEO) database [14], with the accession number GSE62333. The dataset includes the gene expression of skin fibroblasts from 20 patients affected by SZ (mean age ± SD, 44.60 ± 12.67; 50% males) and 20 healthy control subjects (mean age ± SD, 48.40 ± 12.20; 45% males). The study involved patients who satisfied the DSM-IV criteria for SZ. The platform that was used for the microarray experiment was Affymetrix Human Gene 1.1 ST Array, which interrogates more than 28,000 well-annotated genes through UniGene or via Reference Sequence. This dataset was utilized for developing a classifier to distinguish the two classes.

2.2. *Data preprocessing and analysis*

In this study the raw data files were imported into R studio and the normalized values were obtained after applying the robust multi-array average (RMA) algorithm, which includes the background correction on the Affymetrix perfect match data, quantile normalization and summarization using the median polish method [15]. In order to deal with the problem of multiple probes for the same gene, for each gene the average value of its probes was used.

A gene was considered present, if its mean expression value was above 5 $\log_2$ intensity in 60% of the replicates in at least one condition. In order to identify the differentially expressed (DE) genes, limma Bioconductor package was used and linear model fitting and the empirical Bayes method were applied [16]. DE genes (SZ samples versus healthy control samples) were defined based on the following criteria: (i) ±0.3 or greater fold change (FC) in the mean expression level ($\log_2$ scale), (ii) p-value ≤ 0.05 after correction of multiple testing with the False Discovery Rate (FDR) adjustment method of Benjamini and Hochberg [17].

2.3. Functional analysis

The genes that were identified to be DE in the study were imported into BioInfoMiner tool [18]. After applying statistical enrichment tests, the imported genes were associated to the Reactome Pathways Database [19], as well as to three different ontology databases: Gene Ontology (GO) [20], the Human Phenotype Ontology [21] and the MGI Mammalian Ontology [22].

2.4. Classification algorithms and parameter optimization

In this study the tested classification algorithms were Support Vector Machines (SVM), Extremely Randomized Trees (Extra Trees), Random Forest (RF), AdaBoost and k-nearest neighbors (*k*NN). The SVM algorithm tries to find a hyperplane that separates the two classes with the maximum margin between the nearest training data points of the different classes [23]. The *k*NN algorithm is based on the idea of giving a class label to a new sample based on the class label of the predefined amount of closest samples [24]. RF and Extra Trees are two averaging algorithms on the basis of randomized decision trees [25,26]. In these methods an ensemble of classifiers is developed, and randomness is introduced in the sampling of the features and the samples used to build the classifiers. The prediction of the ensemble occurs from the average prediction of the included classifiers. In RF each tree results to a classification and the forest finally makes a selection of the classification model with the most votes [27]. Extra Trees and RF use a random subset of features, but in Extra Trees, thresholds are drawn at random (in comparison to RF, where the most discriminative threshold is chosen) for each feature and the best of these thresholds is chosen as the classification rule [26,28]. AdaBoost is another ensemble classification algorithm [29]. The principle behind this algorithm is to construct a classifier based on a combination of many weak classifiers [30]. The best parameter set of each algorithm was selected after Cross Validation (CV) grid search, which selects the parameters that result to the highest score of the CV [31]. All of the machine learning methods were implemented in scikit-learn of Python [32].

2.5. Feature selection

Each of the tested classifiers included genes that resulted from a Support Vector Machines Recursive Feature Elimination (SVM-RFE) with cross-validated selection of the best number of features [33]. This feature selection method is based on an iterative process of removing features with lowest calculated weights for each estimator until the subset of features with the best performance is achieved. The genes that occurred after the feature selection were used for training classifiers based on their gene expression values in the skin fibroblast dataset and based on their gene expression values in an independent cohort, in order to evaluate the discrimination efficiency of those genes and to further evaluate them as possible biomarkers.

2.6. Model Evaluation

In order to evaluate the performance of each classification model, we utilized the method of 4-fold CV and the evaluation measure of receiver operating characteristic (ROC) area under the curve (AUC) score. In 4-fold CV the dataset is split in 4 random, mutually exclusive subgroups Di (D1, D2,…, D4) called folds. 4-fold CV is performed since the study includes few samples, and it is better to split them into 4 subgroups. The classification model is then trained using three out of the four subgroups and the remaining subgroup is used for testing.

ROC curves are typically used in binary classification for evaluating the output of a classification algorithm after using CV. The Y axis of the curve depicts the true positive rate, and the X axis shows the false positive rate. A larger AUC, which means a larger area between the X axis and the ROC curve corresponds to a better classification performance. For each fold in CV a ROC curve is constructed and finally a mean AUC from all the curves is calculated. The feature selection and the model evaluation were also performed using scikit-learn machine learning in Python [32]. CV was performed in the skin fibroblast dataset in the three following ways: in the first 4-fold CV each fold included the 63 DE genes; in the second 4-fold CV each fold included feature selection with the method of SVM-RFE. The third 4-fold CV included in each fold the 16 genes from the feature selection.

2.7. Data collection and analysis of the independent dataset

The second dataset of this study was used as an independent group of samples. The dataset includes postmortem brain samples derived from superior temporal cortex (Broadmann Area 22) of 23 schizophrenic (mean age ± SD, 72.2 ± 16.9) and 19 healthy control subjects (mean age ± SD, 67.7 ± 22.2) [34]. Patients were diagnosed with SZ according to DSM-III criteria. The SZ samples included 13 males and 10 female samples and the heathy control group included 11 male and 8 female samples. The accession number in GEO for the second dataset is GSE21935 and its platform for the microarray hybridization is Affymetrix Human Genome U133 Plus 2.0 Array. The normalized expression values were generated using RMA as described above. The subset of 16 genes that occurred from the SVM-RFE feature selection of the skin fibroblast dataset, were used in order to test if their corresponding expression values in the postmortem brain dataset can discriminate SZ samples from healthy control samples.

Additionally, the DE genes of this dataset were also identified using linear model fitting and the Empirical Bayes method from the BioConductor limma package were used to test the differences in mRNA levels. Linear model fitting and the empirical Bayes method were applied, and DE genes were derived with the following criteria: ± 0.3 or greater FC in the mean expression level (in $\log_2$ scale) and (ii) unadjusted p-value ≤ 0.05. No significant genes were obtained after correction of multiple testing with the FDR adjustment method of Benjamini and Hochberg. The DE genes from this dataset were also imported into BioInfoMiner for functional analysis. The DE genes and their enriched terms were used for comparison with those resulted from the skin fibroblast dataset.

3. Results

3.1. DE genes and functional analysis

The list of the 63 DE genes of the skin fibroblast cells according to the criteria set (see Materials and Methods) is presented in Supplementary Table 1. The DE genes were imported into the BioInfoMiner tool. The top GO terms are presented in Supplementary Table 2 and the top Reactome Pathways, Human Phenotype Ontology terms, and MGI Mammalian Ontology terms are presented in Supplementary Figures 1, 2 and 3 respectively.

3.2. Classifier optimization and performance, AUC scores and feature selection

The best parameter set of each algorithm after CV grid search is presented in Supplementary Table 3. The mean AUC of the ROC curve shows that the schizophrenic samples can be distinguished from the healthy control samples based on the gene expression of the 16 genes that resulted from SVM-RFE (Supplementary Table 4). SVM and AdaBoost outperform Extra Trees Classifiers, Random Forest and *k*NN, with an AUC of 0.99 (Supplementary Table 4), although these performances may be likely inflated, due to the fact that the feature selection was not applied in each training subgroup of the CV. This feature selection was applied in order to derive the smallest subset of features, which was further utilized on an independent dataset, so an estimation of the performance of the classification models based on these 16 genes would be included. Another CV was also performed, in order to estimate the performance of the classification algorithms. For this scope, the feature selection method of SVM-RFE was applied for each fold and consequently for each training subgroup of the 4-fold CV. The AUC scores of each classification model resulting from the 4-fold CV with and without feature selection (the trained model included all the 63 DE genes) are presented in Table 2.

3.3. Independent postmortem brain dataset

For the postmortem brain dataset that was used for validation of the 16 gene model, maximal classification performance was achieved by the RF classifier, with mean AUC score of 0.82 after applying 4-fold CV. The mean AUC scores of all tested classification methods are presented in Supplementary Table 4.

In the postmortem brain dataset 125 genes were DE according to the statistical criteria presented in Materials and Methods section. Additionally, after comparison of the DE genes in the two datasets, only two genes were identified to be common: CNTN3 and TPBG. Several term clusters from the functional analysis of the DE genes analysis of the skin fibroblast dataset were also identified from the respective analysis of the postmortem brain dataset (Table 3).

Table 1. Genes that occurred after the SVM-RFE feature selection on the skin fibroblast samples and could discriminate the schizophrenic and healthy control subjects.

Gene Symbol	Gene title	Fold Change (log_2)	Corrected p-value
RANBP3L	RAN binding protein 3-like	1.271	0.0377
JUN	jun proto-oncogene	1.056	0.0028
NRP2	neuropilin 2	0.844	0.0168
PPP1R15A	protein phosphatase 1, regulatory subunit 15A	0.631	0.0217
TMEM255B	transmembrane protein 255B	0.538	0.0487
SLC44A1	solute carrier family 44 (choline transporter), member 1	0.537	0.0371
THBS3	thrombospondin 3	0.502	0.0221
MCOLN1	mucolipin 1	0.414	0.0164
TSC22D1	TSC22 domain family, member 1	0.362	0.0406
TUBG1	tubulin, gamma 1	0.344	0.0406
MECP2	methyl CpG binding protein 2	0.333	0.0149
ANKRD39	ankyrin repeat domain 39	-0.307	0.0496
CST3	cystatin C	-0.387	0.0236
RPLP2	ribosomal protein, large, P2	-0.423	0.0236
CRLF1	cytokine receptor-like factor 1	-0.592	0.0406
SNORA37	small nucleolar RNA, H/ACA box 37	-1.149	0.0091

Table 2. Mean performance estimation values of different classification algorithms after applying 4-fold CV, either with no feature selection or with feature selection of SVM-RFE included in each fold.

Classification models	Mean AUC (no feature selection)
*k*NN	0.98
SVM	0.95
RF	0.93
EXTRA TREES	0.90
AdaBoost	0.88
Classification models	**Mean AUC (with feature selection)**
*k*NN	0.83
SVM	0.88
RF	0.74
EXTRA TREES	0.83
AdaBoost	0.68

4. Discussion

The top ranked biological processes resulting from the BioInfoMiner tool include response to cyclic adenosine monophosphate (cAMP) (FOSB, FOS, JUN and JUND) (Supplementary Table 2). The pathway analysis detected among the statistical significant REACTOME pathways, mitogen activated protein kinase (MAPK) related pathways (Supplementary Figure 1). Both MAPK and cAMP signaling pathways involve responses to extracellular stimuli, modification of the functionality of receptors and finally they affect the cell survival and neuroplasticity. Abnormalities in those pathways have been observed in frontal cortical areas of schizophrenic patients [35]. Mitogen activated protein kinases are expressed in the central nervous system and the extracellular signal-regulated kinases are related to long-lasting neuronal plasticity [36]. There are also previous studies on second messenger signal transduction system of cAMP that exhibit abnormal function in postmortem brain, in cerebrospinal fluid and blood platelets of SZ samples. Additionally neuroleptic treatment in animal models resulted in the opposite pattern of cAMP metabolism in schizophrenic patients, indicating that cAMP signal transduction may be a target of this medication [37].

Table 3. Common clusters of enriched terms between the skin fibroblast dataset and the postmortem brain dataset derived from the functional analysis.

COMMON CLUSTERS	GO TERM	Number of DE genes		GENES/ TERM
		GSE62333	GSE21935	
Immune response	GO:0038095: Fc-epsilon receptor signaling pathway	4	-	132
	GO:0050778: Positive regulation of immune response	-	11	545
	GO:0045087: Innate immune response	-	9	448
	GO:0002253: Activation of immune response	-	10	407
	GO:0006954: Inflammatory response	-	10	434
	GO:0050729: Positive regulation of inflammatory response	-	6	116
	REACTOME PATHWAY			
Innate Immune System	R-HSA-166658 : Complement cascade	-	4	53
	R-HSA-166663 : Initial triggering of complement	-	4	31
	R-HSA-2871796 : FCERI mediated MAPK activation	5	-	246
	MGI MAMMALIAN PHENOTYPE			
Abnormal nervous system morphology	MP:0000774 : decreased brain size	3	-	97
	MP:0000953 : abnormal oligodendrocyte morphology	-	3	20
	MP:0002182 : abnormal astrocyte morphology	-	4	42

Finally, the REACTOME pathway analysis resulted to terms classified to innate immune system (CALM3, JUN, FOS, PSMB4, SPRED2). A study including RNA sequencing, resulted to SZ genes that are enriched in immune related pathways [38]. Other epidemiological, genomic and transcriptomic studies on postmortem brain and peripheral samples also indicate abnormalities of the innate and the adaptive immune system of the disease. Still, the exact mechanism by which the immune system aberrations confer to the disease phenotype remains elusive. Several missing links exist concerning the chain of events that link the observed immune disturbances with the disease manifestation [39].

According to the human phenotype ontology, the resulted phenotypic abnormalities related to the nervous system are of great interest in SZ (Supplementary Figure 2). More specifically, the resulted term of abnormality of the autonomic nervous system is in accordance to abnormal autonomic nervous system activity reported in patients with SZ [40].

Concerning the statistical significant terms of the mammalian phenotype it is worth mentioning that there are genes related to abnormal brain morphology and more specifically to abnormal hippocampus neuron morphology and decreased brain size (Supplementary Figure 3). Even after many years of brain scanning, the brain structural abnormalities in SZ are not completely understood. There are many indications, that here is a decrease of the brain and intracranial size of schizophrenic patients [41]. Reductions of grey and white matter as well as of whole-brain in MRI-scans of schizophrenic patients in comparison to healthy controls have been reported [1]. ENIGMA SZ Working Group is a consortium that performs an effort to collect and analyze neuroimaging data of severe mental illnesses [42]. In the context of ENIGMA a meta-analysis on brain MRI scans from 2028 patients affected by SZ and 2540 healthy controls showed that SZ patients have smaller hippocampus, amygdala, thalamus and intracranial volumes in comparison to healthy controls [43]. Hippocampus is involved in SZ through neuropsychological defects. In summary, the hippocampal pathophysiology in SZ is involved at a morphological, molecular and functional level, resulting to alterations in the structures and functions of hippocampal neuron circuits and subsequently leading to disturbances in glutamatergic neurons [44].

Decreased fibroblast proliferation resulted from the MGI mammalian phenotype (Supplementary Figure 3) is also in accordance with previous studies concerning skin biopsies that resulted to decreased fibroblast growth and abnormal morphology of fibroblast cultures of SZ samples as compared to fibroblast cell cultures of healthy controls [45,46].

Many other findings resulting from the functional analysis have not been related to SZ, so it would be interesting to further investigate them. Still, a more detailed examination of these findings is beyond the scope of this study, which actually focuses on the use of skin fibroblast cells as a possible model for developing classification models in SZ. For this reason the molecular abnormalities resulting from the functional analysis and were related to previous findings in the SZ research, can be considered a validation of the skin fibroblast cell model in the study of SZ.

So far, biomarker development in the psychiatric field is in its infancy as most studies lack validation through independent cohorts. The need for a test involving peripheral tissues (with blood being the most studied tissue type) that could help in the prediction and diagnosis of SZ is an important issue in the studies of the disease [47]. Similar studies that exploit the gene expression profiling of non-neural samples through machine learning methods for the development of diagnostic classification methods have been already performed in psychiatric research. For example, an artificial neuron networks classification method performed on gene expression signature of whole peripheral blood could afford to develop classification models in SZ with good performance [6]. Peripheral blood gene expression through the use of microarray technology has been used for the identification of biomarker genes that can differentiate bipolar disorder upon the mood state of the patients [48].

The biomarker panel from human skin fibroblasts presented in this study is validated through an independent postmortem brain schizophrenic dataset. Ultimately, this 16-gene model of possible biomarkers could be further developed with the goal of finally developing a diagnostic tool for SZ with clinical use. Among the genes resulting from the feature selection, genes that have been previously implicated in SZ are RPLP2 and SLC44A1. In a study that examined protein expression alterations in postmortem brain samples of SZ subjects, altered expression of ribosomal proteins including the gene RPLP2 has been reported [49]. In a translational convergent functional genomics study for the identification and prioritization of genes involved in SZ, the gene SLC44A1 is included in the resulted genes [50].

Finally, only two genes were found to be commonly DE both in skin fibroblast cell samples and in postmortem brain samples of schizophrenic subjects. Some of the resulted terms of the functional analysis in the two examined datasets (GSE62333, GSE21935) are clustered into the common superclasses of Table 3, such as immune response related terms. These common clusters of terms resulting from DE genes in the two datasets may imply that the skin fibroblast cell model could capture molecular abnormalities that take place in the brain tissue of SZ samples, thus reflecting underlying similarities in the exercised molecular mechanisms.

The mean AUC scores after the 4-fold CV shows that SZ samples can be distinguished from the healthy control samples based on the gene expression of the 16 genes that resulted from SVM-RFE (Supplementary Table 4). SVM and AdaBoost outperform Extra Trees, Random Forest and *k*NN, with an AUC of 0.99 (Supplementary Table 4), although these performances are likely inflated, due to the fact that the feature selection was applied on the whole training dataset and the CV does not correctly mimic the application of the classifier to a completely independent test set, since these predictors 'have already seen' the left out samples. Nevertheless, SVM-RFE feature selection was applied in order to derive the smallest subset of features, which will have the ability to classify schizophrenic subjects with an unknown label. Through the application of SVM-RFE, the minimum subset of genes that achieve the highest AUC according to an embedded CV in the feature selection process is selected. This study aimed to find informative subsets of genes that will be further utilized in other independent datasets. The 16 genes from the feature selection were also used as a basis for training and testing classification models on an independent dataset of another tissue (which in this case is the diseased postmortem brain tissue). In other words, in this study it was studied if a gene signature obtained from skin fibroblast cells can be used for separating healthy control samples from SZ samples derived from postmortem brain tissue. For this reason, an estimation of the performance of the classification models both in the skin fibroblast dataset and in the independent postmortem brain dataset of these 16 genes was performed. The CV in the postmortem brain dataset resulted to a satisfactory mean AUC score of 0.82 with the use of a RF classifier. The feature selection is also important, since through the dimensionality reduction, covariance is reduced and irrelevant attributes are excluded. Another CV was also performed, in order to estimate the performance of the classification algorithms, without taking into consideration the 16 genes from the SVM-RFE. For this scope, the feature selection method of SVM-RFE was applied for each fold. The AUC scores of each classification model resulting from the 4-fold CV with and without feature selection (the trained model included all the 63 DE genes) are presented in Table 2. This CV was applied for the estimation of the performance of the classification methods used in this study, including the SVM-RFE feature selection method in each fold, resulting to worse, though satisfactory mean AUC scores of the classification models (Table 2).

There are also some other classification approaches in SZ research, such as machine learning algorithms for pattern classification with the use of neuroimaging as a clinical diagnostic or prognostic tool [51], as well as neurocognitive test batteries [52]. These classification approaches utilize physiological and cognitive data. It would be interesting to combine data from all these different levels and to develop composite classification models. Still, the development of classification models based on gene expression data can be proved very important in classification tasks, if we consider that there are already clinically applied assays, deriving from machine learning methods based on gene expression data. In 2002, a 70-gene classifier or the Mammaprint® assay has been developed based on the expression of 70 genes in tumors biopsied from women with breast cancer and has been applied in clinical practice [53].

There are some limitations concerning this study. Concerning the medication effects in the expression of skin fibroblast cells, it is assumed that because of the fact that the skin fibroblast cells have been subcultured for at least five passages, any effects from *in vivo* exposure to factors such as hormones, and drugs are minimized [54]. Additionally, limited sample size can be considered another limitation of this study, but concerning the dataset of SZ skin fibroblast cells, we were limited by the content of the available datasets. The approach of this study to use independent datasets and different tissues bears several pitfalls and this might be reflected by the fact that only two genes were DE in both samples. However the postmortem brain dataset is useful as a validation dataset in a study based on skin fibroblast cells, due to the fact that postmortem brain is a model that includes the diseased tissue in SZ. Furthermore, lacking a validation of the classifier itself can be considered a limitation of this study. However, through the methodological approach used in this study, the aim was to identify a subset of genes with a biological interest in SZ. The validation performed in the independent dataset from postmortem brain cells, mostly concerns the validation of the selection of the best subset of genes as a potential diagnostic signature in SZ. Additionally, a classifier trained with expression data of skin fibroblast cells, would be better validated as a classifier, with the use of an independent dataset containing gene expression data from skin fibroblasts. Continuing this work, with additional skin fibroblast cells from schizophrenic patients and healthy controls for model development as well as for the validation on completely independent test cohorts, may help to develop and define the true utility of the skin fibroblast based classification models.

5. Conclusions

In this study it was examined if gene expression of skin fibroblast cells could be exploited through supervised machine learning for the development of classification models that can discriminate SZ from healthy control samples. A subset of genes that could discriminate the schizophrenic and the healthy control samples based on the gene expression of skin fibroblast cells could also sufficiently discriminate SZ and healthy control subjects based on the expression values of those genes in postmortem brain samples. These are indications that skin fibroblast cells could be used for the identification of potential biomarkers in SZ. Concerning the classification models that have been developed based on the skin fibroblasts gene expression, achieving AUC scores over 0.9, other independent skin fibroblast samples of schizophrenic patients should be used, in order to further blindfold validate the generalization ability of the developed classification models.

Conflict of Interest

All authors declare no conflicts of interest in this paper.

References

1. Kahn RS, Sommer IE, Murray RM, et al. (2015) Schizophrenia. *NRDP* 1: 15067.
2. Tandon R, Gaebel W, Barch DM, et al. (2013) Definition and description of schizophrenia in the DSM-5. *Schizophr Res* 150: 3–10.
3. American Psychiatric Association (2013) Diagnostic and statistical manual of mental disorders (DSM-5). Arlington: American Psychiatric Publishing.
4. World Health Organization (2010) International Classification of Diseases, Tenth Revision.
5. Alawieh A, Zaraket FA, Li JL, et al. (2012) Systems biology, bioinformatics, and biomarkers in neuropsychiatry. *Front Neurosci* 6: 187.
6. Takahashi M, Hayashi H, Watanabe Y, et al. (2010) Diagnostic classification of schizophrenia by neural network analysis of blood–based gene expression signatures. *Schizophr Res* 119: 210–218.
7. Yousef M, Najami N, Abedallah L, et al. (2014) Computational approaches for biomarker discovery. *JILSA* 6: 153–161.
8. Kotsiantis SB (2007) Supervised Machine Learning: A Review of Classification Techniques. Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies Emerging Artificial Intelligence Applications in Computer Engineering. Greece: IOS Press. 3–24.
9. Kalman S, Garbett KA, Janka Z, et al. (2016) Human dermal fibroblasts in psychiatry research. *Neuroscience* 320: 105–121.
10. Vumma R, Johansson J, Lewander T, et al. (2011) Tryptophan transport in human fibroblast cells-a functional characterization. *Int J Tryptophan Res* 4: 19–27.
11. Cattane N, Minelli A, Milanesi E, et al. (2015) Altered gene expression in schizophrenia: findings from transcriptional signatures in fibroblasts and blood. *Plos One* 10: e0116686.
12. Clelland CL, Read LL, Panek LJ, et al. (2013) Utilization of never-medicated bipolar disorder patients towards development and validation of a peripheral biomarker profile. *Plos One* 8: e69082.
13. Dooley EE (2004) National Center for Biotechnology Information. *Environ Health Perspect* 112: A674.
14. Clough E, Barrett T (2016) The Gene Expression Omnibus Database. *Methods Mol Biol* 1418: 93–110.
15. Irizarry RA, Hobbs B, Collin F, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249–264.
16. Ritchie ME, Phipson B, Wu D, et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43: e47.
17. Reiner A, Yekutieli D, Benjamini Y (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19: 368–375.
18. Koutsandreas T, Pilalis E, Vlachavas EI, et al. (2015) Making sense of the biological complexity through the platform-driven unification of the analytical and visualization tasks. IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE). Serbia: IEEE Computer Society. 1–6.
19. Croft D, O'Kelly G, Wu G, et al. (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* 39: D691–697.
20. Ashburner M, Ball CA, Blake JA, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25–29.

21. Groza T, Kohler S, Moldenhauer D, et al. (2015) The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. *Am J Hum Genet* 97: 111–124.

22. Smith CL, Eppig JT (2012) The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mamm Genome* 23: 653–668.

23. Cortes C, Vapnik V (1995) Support-vector networks. *Machine Learning* 20: 273–297.

24. Dhawan M, Selvaraja S, Duan ZH (2010) Application of committee kNN classifiers for gene expression profile classification. *Int J Bioinform Res Appl* 6: 344–352.

25. Díaz-Uriarte R, Alvarez de Andrés S (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7: 3.

26. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Machine Learning* 63: 3–42.

27. Lou W, Wang X, Chen F, et al. (2014) Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection Using Random Forest and Gaussian Naïve Bayes. *Plos One* 9: e86703.

28. Breiman L (2001) Random Forests. *Machine Learning* 45: 5–32.

29. Freund Y, Schapire R (1999) A Short Introduction to Boosting. *JSAI* 14: 771–780.

30. Mozos OM, Stachniss C, Burgard W (2005) Supervised Learning of Places from Range Data using AdaBoost. IEEE International Conference on Robotics and Automation. Spain. 1742–1747.

31. Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. *J Mach Learn Res* 13: 281–305.

32. Pedregosa F, Varoquaux G, Gramfort A, et al. (2011) Scikit-learn: Machine Learning in Python. *JMLR* 12: 2825–2830.

33. Zhang X, Lu X, Shi Q, et al. (2006) Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics* 7: 197.

34. Barnes MR, Huxley-Jones J, Maycox PR, et al. (2011) Transcription and pathway analysis of the superior temporal cortex and anterior prefrontal cortex in schizophrenia. *J Neurosci Res* 89: 1218–1227.

35. Funk AJ, McCullumsmith RE, Haroutunian V, et al. (2012) Abnormal Activity of the MAPK- and cAMP-Associated Signaling Pathways in Frontal Cortical Areas in Postmortem Brain in Schizophrenia. *Neuropsychopharmacology* 37: 896–905.

36. Impey S, Obrietan K, Storm DR Making New Connections. *Neuron* 23: 11–14.

37. Muly C (2002) Signal transduction abnormalities in schizophrenia: the cAMP system. *Psychopharmacol Bull* 36: 92–105.

38. Xu J, Sun J, Chen J, et al. (2012) RNA-Seq analysis implicates dysregulation of the immune system in schizophrenia. *BMC Genomics* 13 Suppl 8: S2.

39. Horvath S, Mirnics K (2014) Immune system disturbances in schizophrenia. *Biol Psychiatry* 75: 316–323.

40. Toichi M, Kubota Y, Murai T, et al. (1999) The influence of psychotic states on the autonomic nervous system in schizophrenia. *Int J Psychophysiol* 31: 147–154.

41. Ward KE, Friedman L, Wise A, et al. (1996) Meta-analysis of brain and cranial size in schizophrenia. *Schizophr Res* 22: 197–213.

42. Thompson PM, Stein JL, Medland SE, et al. (2014) The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav* 8: 153–182.

43. van Erp TG, Hibar DP, Rasmussen JM, et al. (2016) Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Mol Psychiatry* 21: 585.

44. Harrison PJ (2004) The hippocampus in schizophrenia: a review of the neuropathological evidence and its pathophysiological implications. *Psychopharmacology Berl* 174: 151–162.

45. Mahadik SP, Mukherjee S, Laev H, et al. (1991) Abnormal growth of skin fibroblasts from schizophrenic patients. *Psychiatry Res* 37: 309–320.
46. Mukherjee S, Mahadik SP, Schnur DB, et al. (1994) Abnormal growth of cultured skin fibroblasts associated with poor premorbid history in schizophrenic patients. *Schizophr Res* 13: 233–237.
47. Chan MK, Krebs MO, Cox D, et al. (2015) Development of a blood-based molecular biomarker test for identification of schizophrenia before disease onset. *Transl Psychiatry* 5: e601.
48. Le-Niculescu H, Kurian SM, Yehyawi N, et al. (2008) Identifying blood biomarkers for mood disorders using convergent functional genomics. *Mol Psychiatry* 14: 156–174.
49. English JA, Fan Y, Focking M, et al. (2015) Reduced protein synthesis in schizophrenia patient-derived olfactory cells. *Transl Psychiatry* 5: e663.
50. Ayalew M, Le-Niculescu H, Levey DF, et al. (2012) Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction. *Mol Psychiatry* 17: 887–905.
51. Iwabuchi SJ, Liddle PF, Palaniyappan L (2013) Clinical Utility of Machine-Learning Approaches in Schizophrenia: Improving Diagnostic Confidence for Translational Neuroimaging. *Frontiers in Psychiatry* 4: 95.
52. Irani F, Brensinger CM, Richard J, et al. (2012) Computerized Neurocognitive Test Performance in Schizophrenia: A Lifespan Analysis. *Am J Geriatr Psychiatry* 20: 41–52.
53. Sinn P, Aulmann S, Wirtz R, et al. (2013) Multigene Assays for Classification, Prognosis, and Prediction in Breast Cancer: a Critical Review on the Background and Clinical Utility. *GebFra* 73: 932–940.
54. Akin D, Manier DH, Sanders-Bush E, et al. (2004) Decreased Serotonin 5-HT2A Receptor-Stimulated Phosphoinositide Signaling in Fibroblasts from Melancholic Depressed Patients. *Neuropsychopharmacology* 29: 2081–2087.