Ensemble of machine learning based prediction methods for diagnosis of schizophrenia using peripheral blood gene expression profile

Vipul Vilas Wagh[1], Suchita Agrawal[2], Shruti Purohit[2], Tejaswini Pachpor[3], Leelavati Narlikar[4], Vasudeo Paralikar[2#], Satyajeet Khare[1#]

1. Symbiosis School of Biological Sciences, Symbiosis International (Deemed University), Pune 412115, India

2. Psychiatry Unit, KEM Hospital Research Centre, Pune 411011, India

3. MES Abasaheb Garware College, Pune 411004, India

4. Department of Data Science, Indian Institute of Science Education and Research, Pune 411008, India

#Correspondence:

paralikarv2010@gmail.com

satyajeetkhare@gmail.com

Running title: Ensemble learning for schizophrenia

Keywords: Schizophrenia, peripheral blood, gene expression, machine learning, ensemble learning.

Research highlights:

- Ensemble learning of Support Vector Machines (SVM) and Prediction Analysis for Microarrays (PAM) algorithms classified schizophrenia samples with higher precision.
- The pipeline developed in this analysis produced robust models with the ability to classify single microarray sample.
- Cross-platform validation of ensemble model using RNA-Sequencing data resulted in high precision.

Abstract:

The need for molecular biomarkers for schizophrenia has been well recognized. Recently, peripheral blood gene expression profiling and machine learning (ML) tools have become popular for biomarker discovery. The stigmatization associated with schizophrenia advocates the need for diagnostic models with higher precision. In this study, we propose a strategy to develop higher-precision ML models using ensemble learning. We performed a meta-analysis using peripheral blood expression microarray data. The ML models: Support Vector Machines (SVM), and Prediction Analysis for Microarrays (PAM) were developed using differentially expressed genes as features. The ensemble of SVM-radial and PAM predicted test samples with a precision of 81.33% (SD: 0.078). The precision of the ensemble model was significantly higher than SVM-radial (63.83%, SD: 0.081) and PAM (66.89%, SD: 0.097). The feature genes identified were enriched for biological processes such as response to stress, response to stimulus, regulation of the immune system, and metabolism of organic nitrogen compounds. The network analysis of feature genes identified PRF1, GZMB, IL2RB, ITGAL, and IL2RG as hub genes. Additionally, the ensemble model developed using microarray data classified the RNA-Sequencing samples with moderately high precision. The pipeline developed in this study allows the prediction of a single microarray and RNA-Sequencing sample. In summary, this study developed robust models for clinical application and suggested ensemble learning for higher diagnostic precision in psychiatric disorders.

1. Introduction:

Schizophrenia (SCZ) is a complex neuropsychiatric disorder which is characterized by a disruption in thinking and sense of self. The death rate is two times higher in schizophrenic patients, with cardiovascular diseases and suicide as the leading causes of death (Costa E Silva, 1998; Hennekens et al., 2005; Laursen et al., 2014). The lifetime prevalence of 0.2 - 0.4 % globally (GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, 2018) indicates its universal presence irrespective of cultural differences around the world. One of the significant issues in the treatment of psychiatric disorders is delayed diagnosis. The current diagnostic procedure for SCZ is based on psychological evaluation which makes it clinician dependent. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) based diagnosis for SCZ require symptoms to be persistent for the time period of six months or more (American Psychiatric Association., 2013). The delay in the treatment accounts for higher degree of years lived with disability associated with SCZ (GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, 2018). Recent studies suggest, the treatment outcome could be improved if the time elapsed before the treatment is reduced (Chan et al., 2014). Thus, having a blood test can surely strengthen and quicken the current diagnostic process for SCZ.

Molecular alterations such as gene expression changes associated with the disorder have been proposed to be used as potential biomarkers. Our previous study provides stronger evidence for the use of peripheral blood gene expression profile for biomarker discovery (Wagh et al., 2021). The recent use of machine learning (ML) tools has accelerated the biomarker discovery process for psychiatric disorders. The ML tools have already managed to provide gene expression markers with higher diagnostic potential (Hess et al., 2020; Zhu et al., 2021; Liu et al., 2022). Currently, the ML based *in silico* approaches are limited to publicly available microarray datasets. ML based *in silico* studies using biomarkers have reported higher diagnostic performance for psychiatric disorders (Yu et al., 2016; Ke et al., 2021; Zhu et al., 2021; Wu et al., 2022). However, very few studies have validated their ML models using independent test dataset (Hess et al., 2016, 2020; Liu et al., 2022). Application of the diagnostic models into clinics would need extensive validation along with appropriate data scaling methods for development of robust models. In addition to this, majority of the studies focused on accuracy and area under the curve (AUC) as evaluation parameters for the performance of ML models. However, schizophrenia is associated with higher degree of stigmatization, and it demands a diagnostic test with higher precision (True positives / (True positives + False positives)).

In this study we have attempted to develop an ensemble of ML algorithms for classification of SCZ. We selected publicly available gene expression microarray datasets for this meta-analysis. The raw data from each platform was processed independently to avoid data leakage. ML models were developed using support vector machines (SVM), and Prediction Analysis for Microarrays (PAM) algorithms. We made use of differential gene expression analysis (DGEA) for selecting features with potential diagnostic values. ML models with different set of genes were compared based on their performance in test data class prediction. The test data predictions from best performing models were ensembled to achieve higher precision. The gene ontology and networking analysis of the feature genes further highlighted the biological processes and hub genes associated with SCZ. These ensemble models were finally tested for cross-platform compatibility.

2. Materials and methods:

2.1. Identification of datasets:

Peripheral blood gene expression microarray datasets for schizophrenia (SCZ) were identified from the Gene Expression Omnibus (GEO) (Edgar et al., 2002) and ArrayExpress (Parkinson et al., 2007)

using keywords ‘Gene expression’, ‘Peripheral blood’, ‘Biomarkers’ and ‘Schizophrenia’ or ‘Schizophrenia spectrum’. A similar search was performed on databases such as PubMed and Google Scholar. Studies with immortalized cell lines, specific cell types, and custom microarray platforms were excluded. The studies in which raw data was available were retained for the analysis this study. The raw data was requested from authors when it was not available on public platforms.

2.2. Importing data and creating a meta-file:

Raw data for each dataset was imported and processed independently in R (R Core Team, 2020). Probe filtration was carried out for Illumina datasets (e.g. Illumina probes with detection P.val <0.05 in ≥ 3 samples were retained). Probe-IDs of all the arrays were mapped to HUGO Gene Nomenclature Committee (HGNC) gene symbols (Braschi et al., 2019). Gene expression values for multiple probes were averaged out for individual genes. All the datasets were combined based on the common genes to obtain a meta-file. This meta-file with raw gene expression values (raw meta-file) was processed for identification of outlier datasets and for machine learning (ML) based prediction analysis.

2.3. Identification of outlier datasets:

We made use of expression status of differentially expressed genes for identification of outlier datasets. Prior to differential gene expression analysis, samples from raw meta-file were independently quantile normalized based on the microarray platform used. Illumina, single-channel Agilent, and Affymetrix datasets were normalized using lumi (Du et al., 2008), limma (Ritchie et al., 2015), and Affy (Gautier et al., 2004) packages respectively. The normalized data was further batch corrected using ComBat (Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Zhang Y, Storey JD, 2020) and subjected to differential gene expression analysis (DGEA) using limma. The expression status of top differentially expressed genes was visualized for heterogeneity in their expression across the individual datasets using a Forest-plot (Gordon and Lumley, 2021).

2.4. Pre-processing and data scaling for machine learning:

To avoid any data leakage, the raw meta-file (raw gene expression values) was divided into train and test data before normalization and batch correction. To achieve this, samples were shuffled and then subjected to a random selection of train (90%) and test (10%) data (Figure 1A). This random selection was repeated to obtain 10 iterations of train and test datasets. Samples within the training data were quantile normalized based on the microarray platform (Figure 1B). Training data was further batch corrected independently (Figure 1B). In contrast, test data was normalized using quantile-targets (Bolstad, 2020) from train data and batch corrected using train data as reference (Figure 1B) (Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Zhang Y, Storey JD, 2020). Each iteration of normalized and batch-corrected train data was used for feature selection and development of ML models. While test data was used for the evaluation of ML models (Figure 1C).

2.5. Feature selection and development of ML models:

We used differential gene expression analysis (DGEA) as feature selection method. The top differentially expressed genes were used as features for model building. Support vector machines (SVM) (David Meyer et al., 2021) with different kernels (linear, polynomial, radial and sigmoid) and Prediction Analysis for Microarrays (PAM) (Hastie et al., 2019) algorithms were used for development of diagnostic models for SCZ. The ML models built from each iteration of the train data was cross validated (k=10) and were used for class prediction of respective test data samples (Figure 1C). The ML models were evaluated based on accuracies and were selected for ensemble learning.

2.6. Ensemble learning and evaluation:

In order, to improve the precision of class prediction, we used ensemble learning of ML models of SVM and PAM. In brief, samples predicted as cases by both ML algorithms were labelled as cases in ensemble models. The performance of ensemble model was evaluated in comparison with individual models using parameters such as precision, accuracy, sensitivity and specificity.

2.7. Functional enrichment and networking analysis:

A functional enrichment analysis for the genes of interest was carried out using g:profiler. A search tool for the retrieval of interacting genes/proteins (STRING) (Szklarczyk et al., 2019) based protein-protein interaction network (PPI) for these genes was established in cytoscape (Shannon et al., 2003). The essential nodes (hub-genes) of the PPI network were identified by maximal clique centrality (MCC) method of cytohubba plugin (Chin et al., 2014). A network of hub-genes and their first-stage nodes were later visualized in cytoscape.

2.8. Establishment of case-control study:

Protocol for this study was approved by KEM Hospital Research Centre Ethics Committee (KEMHRC ID No. 2001) and Symbiosis International (Deemed University) Independent Ethics Committee (SIU/IEC/99). We recruited 20 participants of the age group 18-65 years from Psychiatry Unit K.E.M hospital, Pune. Consent from all the participants was obtained before recruitment. Consent for participants with suspected case of SCZ was supported by the consent of first-degree relatives.

The common, exclusion criteria for control (CNT) and schizophrenia (SCZ) groups were the presence of a) acute or chronic infections, b) coronary heart disease, c) metabolic disorders d) arrhythmia, e) heart disorders, f) hyper and hypothyroidism, g) inflammatory bowels disease and h) multiple sclerosis. Female participants with polycystic ovary syndrome, pregnant and lactating mothers, and women on in-vitro fertilization (IVF) treatment at the time of recruitment, were also excluded from the study. The participants with a suspected case of schizophrenia and schizophrenia spectrum disorder were considered for recruitment under the SCZ group. Age and gender-matched participants with no history of psychiatric disorders were considered for recruitment under the control (CNT) group. All the participants irrespective of the group were subjected to diagnosis.

2.9. Clinical interview and diagnosis:

SCZ diagnosis was carried out using structured clinical interview for DSM-5 research version (SCID-5-RV) (version 1.0.0) (First MB, Williams JB, Karg RS, 2015). The SCID-5-RV was administered by a trained psychiatrist and a psychologist. The SCZ diagnosis was later confirmed by a senior psychiatrist from the team. SCZ-diagnosed participants were also administered with positive and negative syndrome scale (PANSS) (Kay et al., 1987). The absence of any psychiatric disorder in the control group participants was confirmed by administering SCID-5-RV. Age, gender, family history for psychiatric disorders, medical history, and medication status were recorded for all the participants.

2.10. Blood collection and RNA extraction:

Random (non-fasting) venous blood samples were collected in K2EDTA vacutainers and processed on the same day of recruitment. Blood cell count (hemogram) was performed on the samples collected. Peripheral Blood Mononuclear Cells (PBMCs) were isolated using Ficoll-Pacque (Sigma, Catalogue: GE17-5442-02) density gradient centrifugation and re-suspended in TRIZOL (ThermoFischer Scientific, Catalogue: 15596026). Samples were subjected to RNA-sequencing (RNA-Seq) using commercial services. In brief, ribo-depleted RNA samples were sequenced using NovaSeq 6000 system - Illumina to obtain a minimum of 60 million paired end reads of 150 nucleotide length.

2.11. Pre-processing of RNA-Sequencing data:

The quality of each sample was confirmed using FastQC (Andrews, 2010). The sequences were aligned to human genome (GENECODE hg38) (Frankish et al., 2019) using HISAT2 (Kim et al., 2019). The aligned files were subjected to gene assignment using featureCounts (Liao et al., 2014) to create a count matrix. The gene expression values were locally normalized by converting the raw count value to transcript per million (TPM). The TPM matrix was further quantile normalized and batch corrected in reference to microarray training datasets. The quantile normalized and batch corrected RNA-Seq test data was used to validate the cross-platform performance of ML models developed using microarray data.

2.12. Statistical analysis:

Data obtained from processed microarray datasets and patient samples were analysed using Microsoft Excel- Real Statistics (Zaiontz, 2013) and PAST (Hammer et al., 2001). Microarray datasets processed using PAM, SVM and ensemble approach were tested for normality using the Shapiro Wilk test. For data with normal distribution, one-way ANOVA with an alpha of 0.05 followed by Tukey's post hoc test with Dunn–Šidák correction was performed. For clinical samples, age and blood cell count data were tested for normality as mentioned previously. The difference between control and case groups was studied using an unpaired t-test for normal datasets and Mann-Whitney U test for non-normal data. The gender for control and case groups were compared using Chi-square test respectively.

3. Results:

3.1. Microarray datasets show variability in gene expression with no outliers:

We identified 7 peripheral blood expression array datasets for SCZ (Table 1). Participants from these datasets belonged to different ethnic groups. Most of the datasets identified had medicated or a mix population of SCZ participants. For Kumarasinghe et al., paired study, only the 'before treatment' samples were considered for the analysis to avoid over representation of the same samples. Only the genes common to all seven datasets were retained for the meta-analysis. The resulting metafile contained in 449 samples with 6775 genes.

The differential gene expression analysis (DGEA) of the quantile normalized and batch corrected meta-file (Supplementary figure 1) resulted in 1988 DEGs in SCZ samples with respect to controls (adj. P val <0.05) (Supplementary figure 2). Heterogeneity among the datasets was observed by Forest-plot of two up-regulated (*CLEC5A* and *EIF1AY*; LFC >1) (Supplementary figure 3A and 3B) and two down regulated (*EOMES* and *EHMT2*; LFC <-1) DEGs (Supplementary figure 3C and 3D). Expression pattern of the selected DEGs varied across all datasets with respect to the mean expression status; however no specific dataset could be identified as an outlier. Hence, all the datasets were retained for the analysis.

3.2. Ensemble learning results in higher precision for schizophrenia diagnosis:

The pre-processing of raw meta-file resulted in normalized (Supplementary figure 4) and batch corrected (Supplementary figure 5) train and test datasets. The DGEA of each train data iteration resulted in identification of DEGs as features (Supplementary figure 6). Machine learning models were built using these feature genes from training datasets.

We used two different ML models, support vector machines (SVM) and Prediction Analysis for Microarrays (PAM) for classification of test data samples. Performance of these ML models was evaluated based on the mean test data prediction accuracy. SVM models with kernels such as "linear", "polynomial", "radial", and "sigmoid" kernels did not show any significant difference in the test

data prediction accuracy (data not shown). We chose SVM-radial for further analysis. A comparison of SVM-radial models with different number of features (top5, top25, top100, top400, top1600, and all genes) revealed that performance of SVM-radial drops with features more than 400 DEGs (Figure 2A). PAM models with different number of features did not have an impact on the test data prediction accuracy (Figure 2B). SVM-radial and PAM models were further used for combinatorial analysis.

We ensemble the select ML models of SVM-radial and PAM for this combinatorial analysis. In brief, only the samples identified as case by both algorithms were classified as case to reduce false positives in test data class prediction. Ensemble of SVM-radial and PAM with top 400 DEGs (ensemble-400) had highest precision of 81.33% (SD: 0.078) compared to all other ensemble models (Supplementary figure 7A). Also, accuracy of ensemble-400 (68.59%, SD: 0.055) and ensemble-all genes (69.48%, SD: 0.048) was significantly high compared to ensemble-5 (60.37%, SD: 0.057) (Supplementary figure 7B). However, no significant difference was observed in sensitivity and specificity of the ensemble models (Supplementary figure 7C and 7D). We chose ensemble-400 based on absolute value (expressed in percentage) of precision for further analysis. The precision of ensemble-400 was significantly higher compared to the SVM-radial: 66.83 (SD: 0.081) and PAM: 66.89% (SD: 0.097) (Figure 2C). The ensemble learning achieved higher precision without a significant drop in accuracy when compared to individual models (Figure 2D). Interestingly, the decrease in sensitivity of ensemble-400 was not significant when compared to SVM-radial and PAM (Supplementary figure 8A and 8B).

3.3. Functional enrichment and network analysis identifies biological processes, pathways and hub genes associated with schizophrenia:

Top 400 DEGs as features in combinatorial analysis predicted SCZ samples with the highest precision (lesser false positives). We identified common genes (n: 207) between top 400 DEGs from all the 10 iterations of training datasets. Functional enrichment analysis of the common genes identified apoptosis and natural killer cell mediated cytotoxicity as the top two enriched Kyoto encyclopaedia of genes and genomes (KEGG) pathways (adj.P.val <0.05). The majority of DEGs were enriched for the biological processes associated with immune function. However, the top biological process enriched were response to stimulus and response to stress (Figure 3A). Metabolic processes such as organonitrogen compound metabolic process and regulation of nitrogen compound metabolic process were also among the significantly enriched processes. We identified top 5 hub genes (*PRF1*, *GZMB*, *IL2RB*, *ITGAL*, and *IL2RG*) from the PPI network of common genes (Figure 3B).

3.4. Cross-platform validation of ensemble models:

A case-control study was established with 20 participants of Indian origin (SCZ: 10, CNT: 10). There was no significant difference between age, gender, and blood cell counts between case and control group participants (Table 2). All the SCZ affected participants recruited in this study were on antipsychotic medication. Ribonucleic acid (RNA) isolated from the peripheral blood samples of the participants were subjected to RNA-Sequencing analysis. The sixteen samples with RNA-Sequencing data were used for cross platform validation of the models (SCZ: 8 and CNT: 8). SVM-radial (55.63%, SD: 0.11), PAM (55.63%, SD: 0.05) and ensemble with top400 DEGs (58.75%, SD: 0.11) predicted test samples with low accuracy (Figure 4A). However, the ensemble-400 was able to classify SCZ samples with moderate precision (71.25%, SD: 0.14) which was significantly higher than PAM (17.19%, SD: 0.06) and higher than SVM-radial (65.00%, SD: 0.20), though the difference was not significant (Figure 4B). The higher precision of ensemble models was accompanied with higher specificity and lower sensitivity when compared to the individual models (data not shown). Overall performance of PAM was poor when compared to SVM-radial and ensemble model (Figure 4A and 4B).

4. Discussion:

A neuropsychiatric disorder such as schizophrenia is associated with stigmatization. Hence, a diagnostic test with higher precision is desirable. In this study, we attempted to develop an ensemble model with higher precision for classifying SCZ samples. The ensemble model resulted in the test data precision of 81.33% (SD: 0.078), without any significant drop in accuracy. We developed a unique strategy of pre-processing microarray data to build robust model that can be applied to single sample level for clinical application.

Microarray data was pre-processed for building ML models. Seven different datasets with participants from varied ethnicity were included in this analysis. The pre-processing of data involved shuffling of samples, division of samples into train and test data, normalization, and batch correction. Shuffling of samples ensured the representation of each dataset in train and test data. The ten-fold split of samples into train and test data before normalization prevented data leakage. The train data was pre-processed independently whereas test data was normalized and batch corrected using train data as reference (Figure 1B). This pre-processing of test data removed the existing technical differences, and allowed us to process and predict each test sample independently. The pre-processing resulted in the development of robust models with better chance of survival in clinics.

We used differential gene expression analysis (DGEA) for feature selection, which can be important for application of models into clinical settings. Feature selection methods perform better than extraction methods for their “explainability” in clinical settings (Bhandari et al., 2022). In addition, the simple filter methods such as DGEA are computationally less intensive as compared to embedded methods of feature selection. The comparison of ML models with different sets of DEGs revealed that the number of features does not affect the performance of PAM models, unlike SVM for inexplicable reasons (Figure 2A and 2B). Further, we combined the test data predictions from SVM and PAM models to develop ensemble learning for higher precision. Both the ML algorithms use different logic for classification. Thus, a consensus between the models resulted in higher precision without any significant drop in accuracy compared to the individual models (Figure 2C and 2D). The increase in precision was associated with increased specificity and an expected drop in sensitivity (Supplementary figure 8A and 8B). Each individual datasets used in this analysis showed a variation in prediction accuracy when tested using ensemble model. Unavailability of medication status for each sample did not allow us to assess the impact of medication status on test data prediction. It should be noted, that the antipsychotic treated samples from GSE18312 showed an intermediate accuracy when compared to the drug naive samples from Kumarasinghe et al. and GSE54913 datasets. Similarly, the mixed medication status samples from GSE27383, GSE38481, and GSE38484 did not show a variation in overall accuracy or precision. A precision on the higher side of 60% for each individual dataset suggested robustness of our prediction algorithm (Supplementary figure 9).

We could identify very few studies using ML models for SCZ class predictions. *In silico* analysis by Jonathan Hess et al., reported an area under the curve (AUC) of 0.72 to 0.77 for SCZ Vs CNT and 0.607 for bipolar disorder (BD) Vs SCZ in two separate studies with independent test datasets (Hess et al., 2016, 2020). A recent study reported a much higher AUC of 0.993 with 10 fold cross validation for classification of SCZ samples from that of controls (Zhu et al., 2021). However, the higher performance of the models in this study can be attributed to uniform population with fewer confounding factors such as ethnicity. We did not come across a study with the aim to develop ML models with higher precision using ensemble learning. The previous studies with multi-modal approach suggests the need to integrate biological and clinical information for better performance of ML models (Fernandes et al., 2020; Ke et al., 2021). However, the unavailability of clinical information for each sample in the publicly available GEO datasets restricted this analysis to only gene expression markers.

Ensemble of SVM-radial and PAM with top 400 DEGs (Ensemble-400) classified test data samples with highest precision (Figure 2C). We identified common genes (n: 207) between the top 400 DEGs in the 10 iterations of training datasets. Enrichment analysis of these common genes highlighted key pathways such as apoptosis, and natural killer cell mediated cytotoxicity (Figure 3A) which have been known to be dysregulated in SCZ (Yovel et al., 2000; Parellada and Gassó, 2021). The biological processes related to immune function and organonitrogen compound metabolic process have also been associated with SCZ earlier (Van Kesteren et al., 2017; Dmitrieva et al., 2022). In addition to biological processes and pathways we also identified the key regulators (hub genes) of the protein-protein interaction network (Figure 3B). These hub genes were involved in top enriched biological processes such as response to stress and response to stimulus. The hub genes identified in this study have been previously reported in association with neuropsychiatric disorders including SCZ (Fallin et al., 2005; Ghazaryan et al., 2014; Ibrahim et al., 2017; Gao et al., 2018). The genome wide association studies have also associated *PRF1*, *GZMB*, and *IL2RB* with SCZ (Ripke et al., 2014; Pardiñas et al., 2018). Further, *IL2RB*, *ITGAL* and *IL2RG* are known to be differentially expressed in peripheral blood of SCZ affected individuals (Ghazaryan et al., 2014; Leirer et al., 2019). Interestingly, only *ITGAL* and *IL2RB* have been reported to be differentially expressed in the first episode SCZ affected individuals (Leirer et al., 2019). Of these 207 DEGs *MAP4K1*, *GOT2*, *MCM3*, *SIGIRR*, *SRPK1*, *TIPARP*, *RPRD1A*, *ATIC*, *NKG7*, and *SCAP* were also highlighted in our previous study for their association with SCZ (Wagh et al., 2021)

We also performed cross-platform validation of the ensemble model on RNA-Sequencing data. To achieve this, we established a case-control study with age and gender matched participants. Cross-platform validation involved pre-processing of RNA-Sequencing data to generate transcript per million (TPM) values. This step was important since normalize counts and counts per million reads mapped (CPM) values generated from RNA-sequencing data are proportional to the length of the gene. This can cause serious problems when normalizing with quantile targets from microarray data. TPM values on the other hand are normalized for the gene length. Hence the use of TPM values for cross-platform validation seemed plausible. The TPM values were further batch corrected in reference to microarray train data. Similar to microarray data, the ensemble model predicted SCZ with significantly better precision as compared to individual models (Figure 4B). SVM-radial performed relatively better when compared to the PAM in cross-platform data prediction. However, overall low accuracies of individual models suggest a need for the development of better cross-platform normalization techniques (Figure 4A).

5. Strengths and limitations of the study:

To the best of our knowledge, this is the first study that makes use of ensemble learning for SCZ diagnosis with higher precision. The higher precision offered by ensemble learning, even with the existing diversity in the samples with respect to ethnicity, age, gender, and medication status, indicates the robustness of the models. The pre-processing of raw microarray data in this analysis ensures no data leakage and allows the prediction of a single microarray test sample. Notably, cross-platform validation confirms that transcript-per-million (TPM) normalization of RNA-Sequencing data results in high precision. The limitation of the study includes a relatively smaller sample size for model training. The performance of ML models in cross-platform validation was relatively poor suggesting a scope for development of better normalization methods. Also, the analysis was restricted to SCZ and did not include other related neuropsychiatric disorders, such as bipolar disorder and major depressive disorder for multiclass classification. The unavailability of clinical information for each sample restricted its use in developing multi-modal ML models.

6. Conclusions and future directions:

In conclusion, we provide a proof of concept for developing robust predictive models with higher precision for diagnosing SCZ. The current strategy effectively deals with the problems like data leakage and pre-processing of single microarray samples. The feature genes and biological pathways identified in this study can be pursued to explore their potential role in the disorder. Most importantly, this study attempted cross-platform class prediction using RNA-Sequencing data as test samples. However, a relatively poor cross-platform performance indicates the need for better cross-platform normalization techniques. In addition, the availability of data from other high throughput genome wide studies may create novel avenues for development of multi-modal learning. Specifically, the multi-omic approach integrating genomic, transcriptomic, and proteomic data will surely result in precise diagnosis of psychiatric disorders.

Declarations:

Funding:

The study was funded by an intramural research grant (MjRP/19-20/1516) from Symbiosis Centre for Research & Innovation (SCRI), SIU, Pune, India. VVW received the research fellowships from UGC, New Delhi.

Compliance with ethical standards:

KEM Hospital Research Centre Ethics Committee (KEMHRC ID No. 2001) and Symbiosis International (Deemed University) Independent Ethics Committee (SIU/IEC/99) approved the study protocol. Informed consent was obtained from all the participants. The consent for participants with schizophrenia was supported by the consent of a first-degree relative. Clinical interviews were administered by a trained psychiatrist and a psychologist. The diagnosis was confirmed by a senior psychiatrist. All the participants were compensated for their travel and time.

Acknowledgements:

The authors would like to thank all the participants, their parents, relatives, and caretakers for their time and generous participation in making this project possible. We want to thank Deepa Raut, a phlebotomist, for her valuable assistance in blood collection. We would also like to thank Paul Tooney (Associate professor, New Castle University, Australia) for sharing data on request.

Data availability:

The RNA sequencing data will be available in NCBI SRA upon publication

Core availability:

The R scripts used for the analysis are available on GitHub (https://github.com/Vpsy8/SCZ-)

Conflict of interest:

The authors declare no conflict of interest.

References:

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders.* (5th ed.). American Psychiatric Publishing.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Bhandari, N., Walambe, R., Kotecha, K., and Khare, S. P. (2022). A comprehensive survey on computational learning methods for analysis of gene expression data. *Front. Mol. Biosci.* 9, 907150. doi:10.3389/fmolb.2022.907150.

Bolstad, B. (2020). preprocessCore: A collection of pre-processing functions.

Braschi, B., Denny, P., Gray, K., Jones, T., Seal, R., Tweedie, S., et al. (2019). Genenames.org: The HGNC and VGNC resources in 2019. *Nucleic Acids Res.* 47, D786–D792. doi:10.1093/nar/gky930.

Chan, M. K., Gottschalk, M. G., Haenisch, F., Tomasik, J., Ruland, T., Rahmoune, H., et al. (2014). Applications of blood-based protein biomarker strategies in the study of psychiatric disorders. *Prog. Neurobiol.* 122, 45–72. doi:10.1016/j.pneurobio.2014.08.002.

Chin, C.-H., Chen, S.-H., Wu, H.-H., Ho, C.-W., Ko, M.-T., and Lin, C.-Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* 8 Suppl 4, S11. doi:10.1186/1752-0509-8-S4-S11.

Costa E Silva, J. A. (1998). Nations for Mental Health: An Action Programme on Mental Health for Underserved Populations.

David Meyer, Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2021). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R. Available at: https://cran.r-project.org/package=e1071.

Dmitrieva, E., Smirnova, L., Seregin, A., Zgoda, V., Semke, A., and Ivanova, S. (2022). Proteomic profile of serum from patients with schizophrenia spectrum disorders. *PeerJ* 10, e13907. doi:10.7717/peerj.13907.

Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: A pipeline for processing Illumina microarray. *Bioinformatics* 24, 1547–1548. doi:10.1093/bioinformatics/btn224.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–10. doi:10.1093/nar/30.1.207.

Fallin, M. D., Lasseter, V. K., Avramopoulos, D., Nicodemus, K. K., Wolyniec, P. S., McGrath, J. A., et al. (2005). Bipolar I disorder and schizophrenia: a 440-single-nucleotide polymorphism screen of 64 candidate genes among Ashkenazi Jewish case-parent trios. *Am. J. Hum. Genet.* 77, 918–36. doi:10.1086/497703.

Fernandes, B. S., Karmakar, C., Tamouza, R., Tran, T., Yearwood, J., Hamdani, N., et al. (2020). Precision psychiatry with immunological and cognitive biomarkers: a multi-domain prediction for the diagnosis of bipolar disorder or schizophrenia using machine learning. *Transl. Psychiatry* 10. doi:10.1038/S41398-020-0836-4.

First MB, Williams JB, Karg RS, S. R. (2015). Structured clinical interview for DSM-5-Research version. *Arlington, VA Am. Psychiatry Assoc.* SCID-5, 1–94.

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 47, D766–D773. doi:10.1093/nar/gky955.

Gao, L.-J., Zhao, X., Li, J.-G., Xu, Y., and Zhang, Y. (2018). [Bioinformatics analysis of genes related to pathogenesis of major depression disorder]. *Sheng Li Xue Bao* 70, 361–368. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/30112561.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi:10.1093/bioinformatics/btg405.

GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, G. 2017 D. and I. I. and P. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet (London, England)* 392, 1789–1858. doi:10.1016/S0140-6736(18)32279-7.

Ghazaryan, H., Petrek, M., and Boyajyan, A. (2014). Chronic schizophrenia is associated with over-expression of the interleukin-2 receptor gamma gene. *Psychiatry Res.* 217, 158–162. doi:10.1016/j.psychres.2014.03.020.

Gordon, M., and Lumley, T. (2021). Package “forestplot” Title Advanced Forest Plot Using “grid” Graphics.

Hammer, Ø., Harper, D. A. T., and Ryan, P. D. (2001). PAST: PALEONTOLOGICAL STATISTICS SOFTWARE PACKAGE FOR EDUCATION AND DATA ANALYSIS.

Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G. (2019). pamr: Pam: Prediction Analysis for Microarrays. R package. Available at: https://cran.r-project.org/package=pamr.

Hennekens, C. H., Hennekens, A. R., Hollar, D., and Casey, D. E. (2005). Schizophrenia and increased risks of cardiovascular disease. *Am. Heart J.* 150, 1115–1121.

Hess, J. L., Tylee, D. S., Barve, R., de Jong, S., Ophoff, R. A., Kumarasinghe, N., et al. (2016). Transcriptome-wide mega-analyses reveal joint dysregulation of immunologic genes and transcription regulators in brain and blood in schizophrenia. *Schizophr. Res.* 176, 114–124. doi:10.1016/j.schres.2016.07.006.

Hess, J. L., Tylee, D. S., Barve, R., de Jong, S., Ophoff, R. A., Kumarasinghe, N., et al. (2020). Transcriptomic abnormalities in peripheral blood in bipolar disorder, and discrimination of the major psychoses. *Schizophr. Res.* 217, 124–135. doi:10.1016/j.schres.2019.07.036.

Ibrahim, E. C., Guillemot, V., Comte, M., Tenenhaus, A., Zendjidjian, X. Y., Cancel, A., et al. (2017). Modeling a linkage between blood transcriptional expression and activity in brain regions to infer the phenotype of schizophrenia patients. *NPJ Schizophr.* 3, 25. doi:10.1038/s41537-017-0027-3.

Kay, S. R., Fiszbein, A., and Opler, L. A. (1987). The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophr. Bull.* 13, 261–276. doi:10.1093/schbul/13.2.261.

Ke, P.-F., Xiong, D.-S., Li, J.-H., Pan, Z.-L., Zhou, J., Li, S.-J., et al. (2021). An integrated machine learning framework for a discriminative analysis of schizophrenia using multi-biological data. *Sci. Rep.* 11, 14636. doi:10.1038/s41598-021-94007-9.

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi:10.1038/s41587-019-0201-4.

Laursen, T. M., Nordentoft, M., and Mortensen, P. B. (2014). Excess Early Mortality in Schizophrenia. *Annu. Rev. Clin. Psychol.* 10, 425–448.

Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Zhang Y, Storey JD, T. L. (2020). Package “sva” Title Surrogate Variable Analysis. Available at: https://git.bioconductor.org/packages/sva [Accessed January 28, 2021].

Leirer, D. J., Iyegbe, C. O., Di Forti, M., Patel, H., Carra, E., Fraietta, S., et al. (2019). Differential gene expression analysis in blood of first episode psychosis patients. *Schizophr. Res.* 209, 88–97. doi:10.1016/j.schres.2019.05.011.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–30. doi:10.1093/bioinformatics/btt656.

Liu, S., Lu, T., Zhao, Q., Fu, B., Wang, H., Li, G., et al. (2022). A machine learning model for predicting patients with major depressive disorder: A study based on transcriptomic data. *Front. Neurosci.* 16. doi:10.3389/fnins.2022.949609.

Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., et al. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* 50, 381–389. doi:10.1038/s41588-018-0059-2.

Parellada, E., and Gassó, P. (2021). Glutamate and microglia activation as a driver of dendritic apoptosis: a core pathophysiological mechanism to understand schizophrenia. *Transl. Psychiatry* 11, 271. doi:10.1038/s41398-021-01385-9.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., et al. (2007). ArrayExpress - A public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* 35, 747–750. doi:10.1093/nar/gkl995.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K. H., Holmans, P. A., et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–427. doi:10.1038/nature13595.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47–e47. doi:10.1093/NAR/GKV007.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–504. doi:10.1101/gr.1239303.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi:10.1093/NAR/GKY1131.

Van Kesteren, C. F. M. G., Gremmels, H., De Witte, L. D., Hol, E. M., Van Gool, A. R., Falkai, P. G., et al. (2017). Immune involvement in the pathogenesis of schizophrenia: A meta-analysis on postmortem brain studies. *Transl. Psychiatry* 7. doi:10.1038/tp.2017.4.

Wagh, V. V., Vyas, P., Agrawal, S., Pachpor, T. A., Paralikar, V., and Khare, S. P. (2021). Peripheral Blood-Based Gene Expression Studies in Schizophrenia: A Systematic Review. *Front. Genet.* 12. doi:10.3389/fgene.2021.736483.

Wu, X., Zhu, L., Zhao, Z., Xu, B., Yang, J., Long, J., et al. (2022). Application of machine learning in diagnostic value of mRNAs for bipolar disorder. *Nord. J. Psychiatry* 76, 81–88. doi:10.1080/08039488.2021.1937311.

Yovel, G., Sirota, P., Mazeh, D., Shakhar, G., Rosenne, E., and Ben-Eliyahu, S. (2000). Higher natural killer cell activity in schizophrenic patients: the impact of serum factors, medication, and smoking. *Brain. Behav. Immun.* 14, 153–69. doi:10.1006/brbi.1999.0574.

Yu, J. S., Xue, A. Y., Redei, E. E., and Bagheri, N. (2016). A support vector machine model provides an accurate transcript-level-based diagnostic for major depressive disorder. *Transl. Psychiatry* 6, e931. doi:10.1038/tp.2016.198.

Zaiontz, C. (2013). The data analysis for this paper was generated using the Real Statistics Resource Pack software. Available at: www.real-statistics.com.

Zhu, L., Wu, X., Xu, B., Zhao, Z., Yang, J., Long, J., et al. (2021). The machine learning algorithm for the diagnosis of schizophrenia on the basis of gene expression in peripheral blood. *Neurosci. Lett.* 745. doi:10.1016/j.neulet.2020.135596.

Figure legends:

Figure 1: The workflow depicts the steps involved in sample processing and class prediction analysis. A) A raw meta-file with 6775 common genes (rows) across all the datasets and 449 samples (columns) was used for the analysis. Samples from the raw meta-file were shuffled and divided into train and test data (9:1). Random selection of was repeated to produce ten iterations of train and test data. B) Each iteration of train data was independently quantile normalized. The quantile targets for each batch in the train data were used to normalize respective test data. Normalized train data was batch corrected independently, while test data was batch corrected in reference to the respective train data. C) The pre-processed train data was used for feature selection and ML model. The pre-processed test data was used for model testing.

Figure 2: Test data prediction accuracy and precision of SVM-radial, PAM and ensemble models with differentially expressed genes (DEGs) as features. Machine learning models with different number of features were compared using one-way ANOVA followed by Tukey's post hoc test with Dunn–Šidák correction. A) No significant difference was observed between SVM-radial models with different number of features. B) Similar comparison of PAM models also resulted in no significant difference. The ensemble of SVM-radial and PAM models with top 400 DEGs outperformed the individual models. C) The test data classification precision for ensemble model with top 400 DEGs was significantly high compared to the SVM-radial and PAM. D) The drop in the accuracy of ensemble model was not significant.

Figure 3: Functional enrichment and networking analysis was performed using the common genes from top 400 DEGs of ten iterations of training sets. The enriched top 10 biological processes (adj.P.val <0.05) and KEGG pathways were visualized using a bar-plot. A STRING based protein-protein network was established for common genes using Cytoscape. The maximal clique centrality (MCC) method of cytohubba plugin identified top 5 hub genes. The network of 5 hub genes and their first-degree nodes was visualized. The intensity of node colour represents the MCC rank. i.e., PRF1 ranked as number one followed by GZMB, IL2RB, ITGAL, and IL2RG respectively.

Figure 4: Cross-platform validation of SVM-radial, PAM and ensemble models with top 400 differentially expressed genes. Microarray based ensemble learning was tested for their ability to classify schizophrenia samples from RNA-Sequencing platform. A comparative analysis between the models was performed using one-way ANOVA followed by Tukey's post hoc test with Dunn–Šidák correction. A) No significant differences were observed in the prediction accuracies of SVM-radial, PAM and ensemble model with top 400 DEGs. B) However, ensemble models classified test samples with significantly higher precision compared to the individual models.

Table 1: Schizophrenia Microarray gene expression datasets

Dataset	Platform	Control /SCZ	Female (%)	Medication status	Genes analysed	Ethnicity or origin
GSE18312 (51)	Affymetrix Human Exon 1.0 ST Array	8+13	33	Medicated	17131	San Diego and Taiwan
GSE27383 (52)	Affymetrix Human Genome U133 Plus 2.0 Array	29+43	NA	Mix	21826	Multi ethnic groups
GSE38481 (53)	Illumina HumanRef-8_V3 beadchip	22+15	27	Mix	12647	Denmark and Netherland
GSE38484 (53)	Illumina HumanHT-12_V3 beadchip	96+106	42	Mix	17233	Denmark and Netherlands
GSE48072 (54)	Illumina HumanHT-12_V4 beadchip	31+35	53	NA	15155	Finland, Sweden, Caucasian
GSE54913 (55)	Arraystar Human LncRNA microarray V2.0	12+18	NA	Treatment naive	13003	Han Chinese
Kumarasinghe et al. 2013 (56)	Illumina HumanHT-12_V3 beadchip	11+10	38	Treatment naive	10544	Sinhalese

Note: The “Mix” medication status indicates that the participants with and without drug treatment were part of the study. While, “NA” indicates unavailability of information on medication status.

Table 2: Demographics and selected features of cell types in case-control study

Characteristics	**SCZ**	**Control**	***P* value**
Age (Year)* (n: 20)	33.0 (9.754)	36.6 (6.04)	0.4988
Gender (M/F)# (n: 20)	3/5	4/6	0.9200
Neutrophil (103/µL)#	4.8 (2.04)	3.9 (1.26)	0.2810
Lymphocyte (103/µL)#	2.3 (0.68)	2.2 (0.43)	0.8623
Monocyte (103/µL)#	0.5 (0.13)	0.6 (0.11)	0.3617
Eosinophil (103/µL)$	0.2 (0-0.3)*	0.2 (0.1-0.7)*	0.2369
Basophil (103/µL)$	0.025 (0-0.1)*	0.040 (0-0.1)*	0.9650

Note: SCZ: Schizophrenia, CNT: Control, * Chi-squared test, # Unpaired t-test (s.d.), $ Mann-Whitney U test (min-max). Note: Since cell counts for two samples were not available the comparison of blood cell counts between SCZ and CNT was performed on 18 samples.

Figure 1

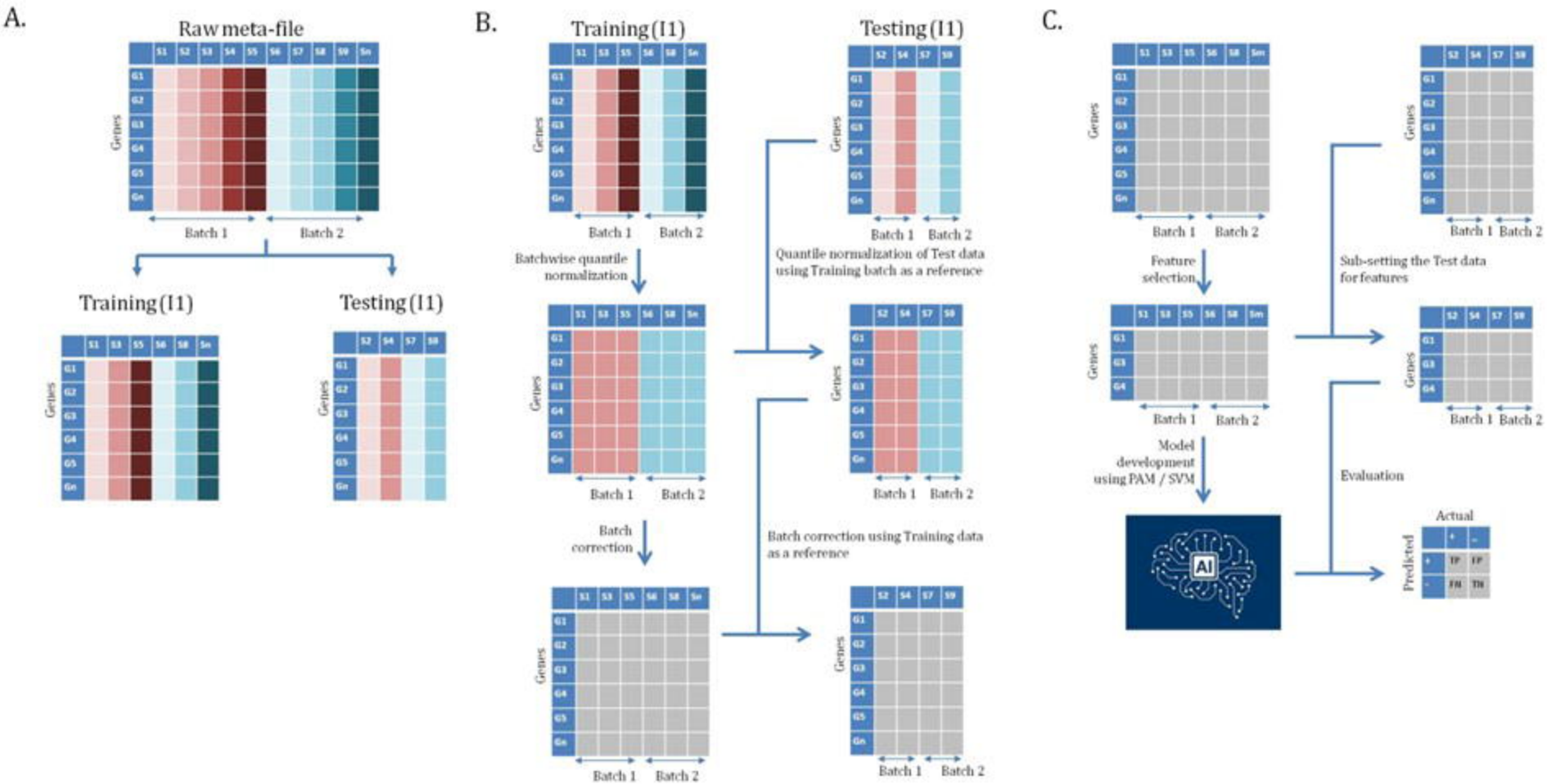
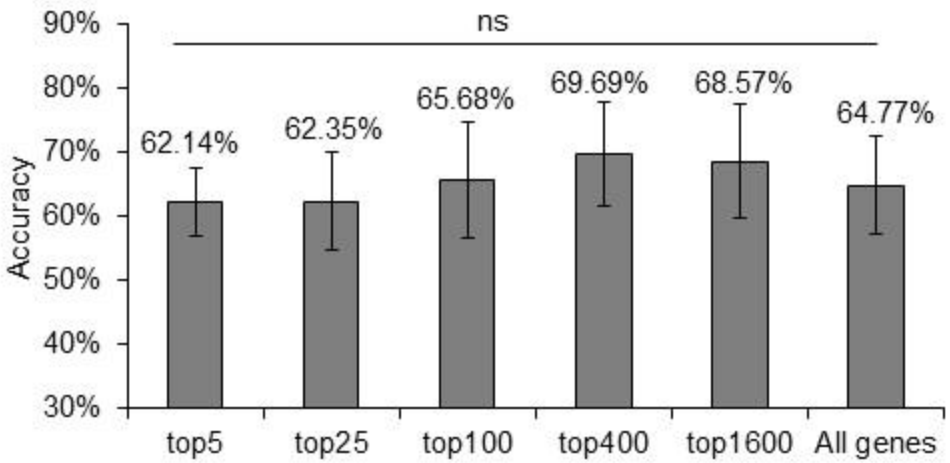
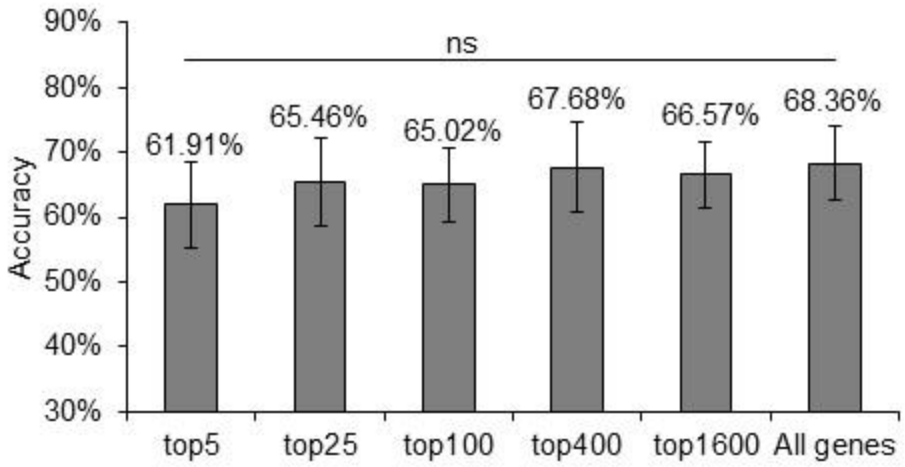

A.
Raw meta-file
Genes
Batch 1
Batch 2
Training (I1)
Testing (I1)
B.
Training (I1)
Testing (I1)
Batchwise quantile normalization
Quantile normalization of Test data using Training batch as a reference
Batch correction
Batch correction using Training data as a reference
C.
Feature selection
Sub-setting the Test data for features
Model development using PAM / SVM
Evaluation
Actual
Predicted
AI

Figure 2

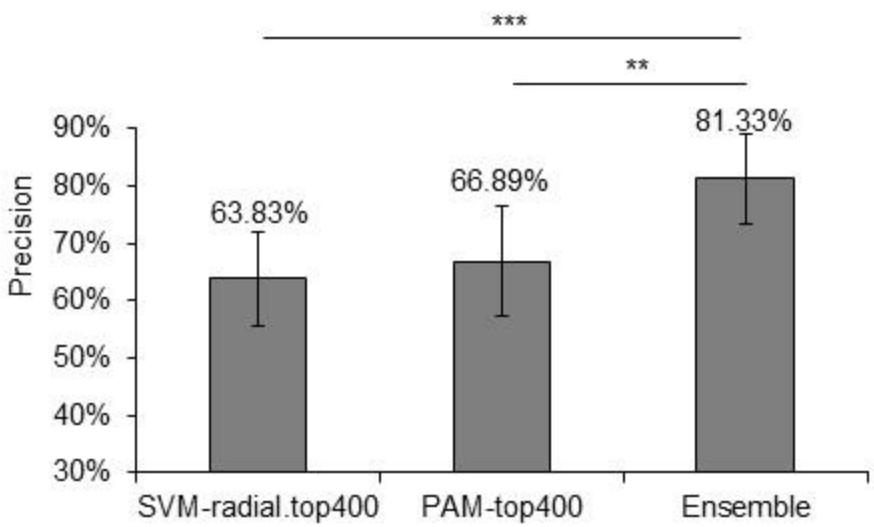
A.



B.



C.



D.

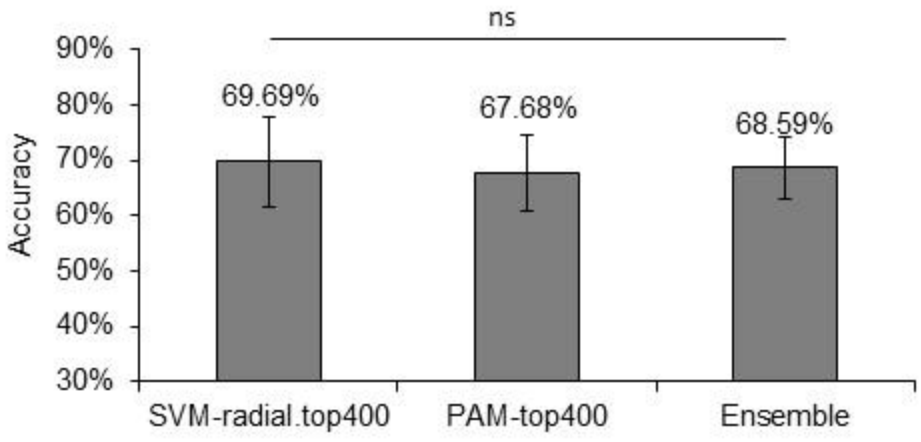
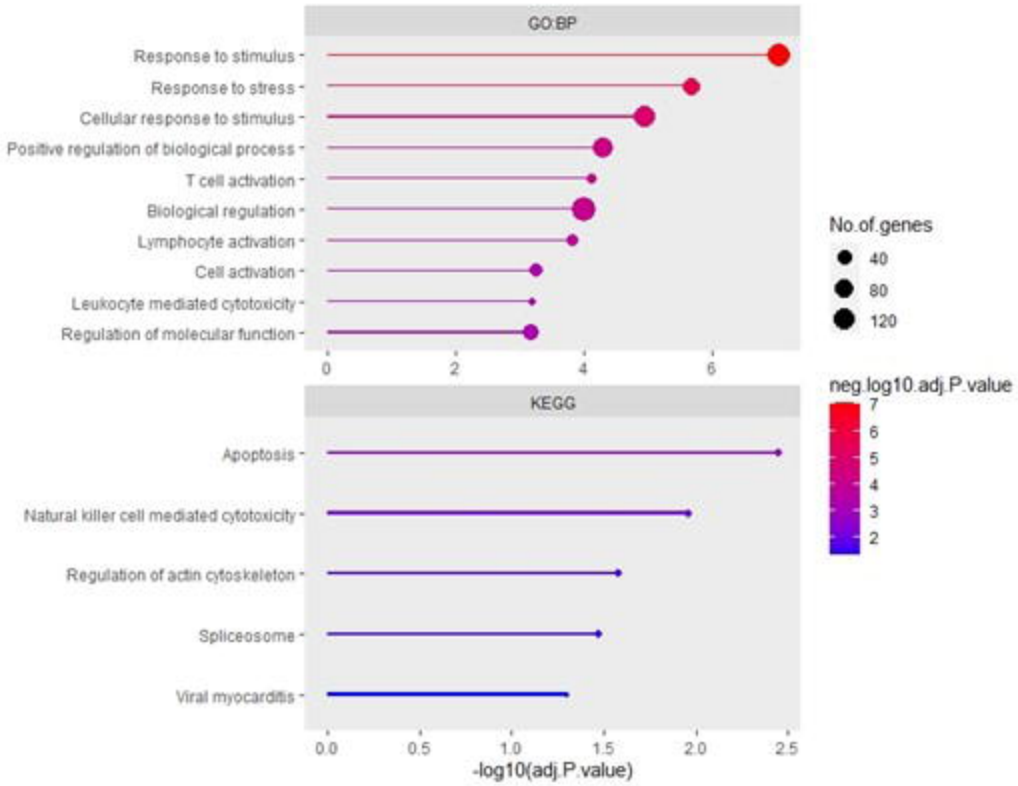


Figure 3

A.


GO:BP
Response to stimulus
Response to stress
Cellular response to stimulus
Positive regulation of biological process
T cell activation
Biological regulation
Lymphocyte activation
Cell activation
Leukocyte mediated cytotoxicity
Regulation of molecular function
0
2
4
6
KEGG
Apoptosis
Natural killer cell mediated cytotoxicity
Regulation of actin cytoskeleton
Spliceosome
Viral myocarditis
0.0
0.5
1.0
1.5
2.0
2.5
-log10(adj.P.value)
No.of.genes
40
80
120
neg.log10.adj.P.value
7
6
5
4
3
2

B.

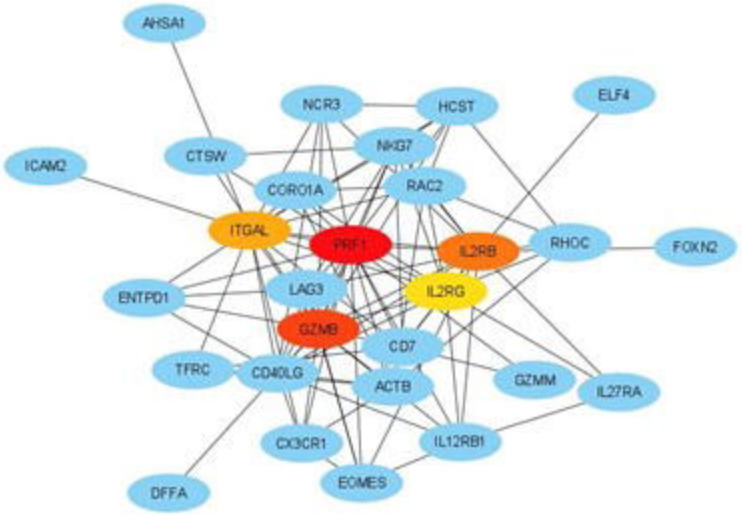
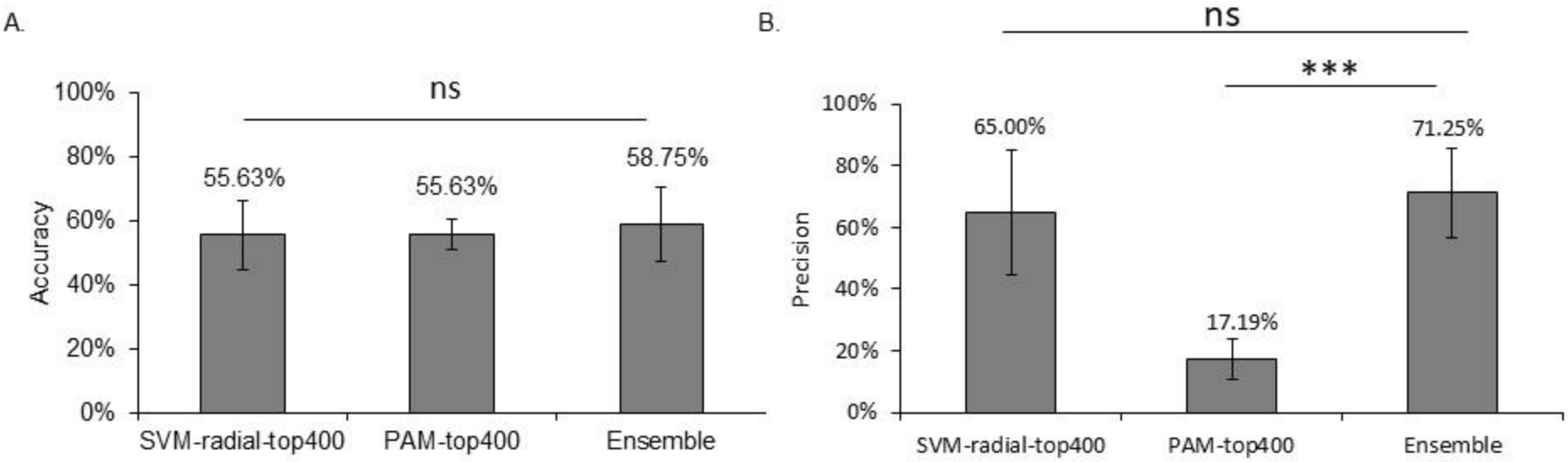
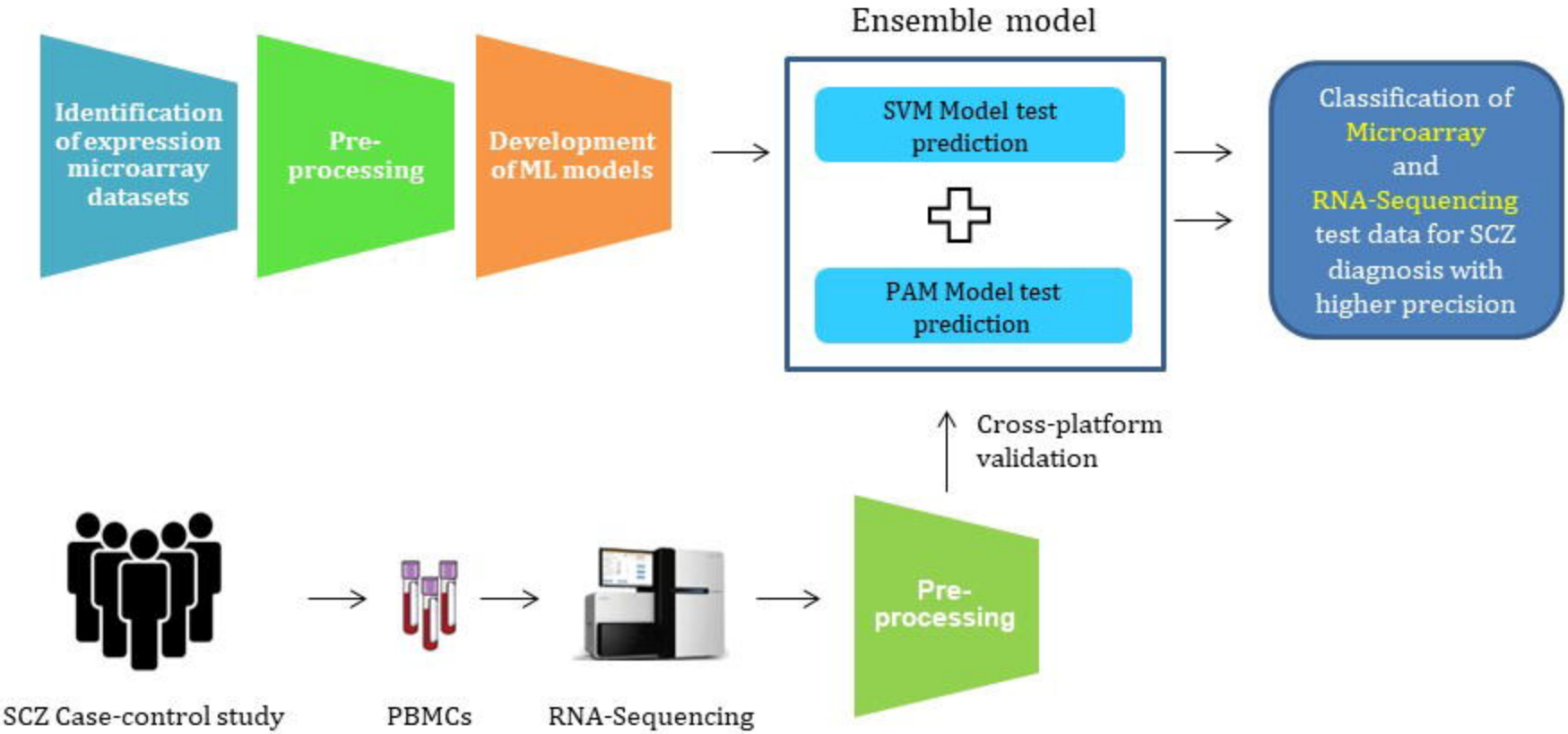

AHSA1
NCR3
HCST
ELF4
ICAM2
CTSW
NKG7
CORO1A
RAC2
ITGAL
PRF1
IL2RB
RHOC
FOXN2
ENTPD1
LAG3
IL2RG
GZMB
CD7
TFRC
CD40LG
ACTB
GZMM
IL27RA
CX3CR1
IL12RB1
DFFA
EOMES

Figure 4


A.
ns
100%
80%
60%
40%
20%
0%
Accuracy
55.63%
55.63%
58.75%
SVM-radial-top400
PAM-top400
Ensemble
B.
ns
***
100%
80%
60%
40%
20%
0%
Precision
65.00%
17.19%
71.25%
SVM-radial-top400
PAM-top400
Ensemble

Graphical abstract



Blood based SCZ diagnosis using ensemble learning for higher precision