International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Detecting Schizophrenia: Low vs High Dimensional Brain Imaging Features

Ayman Alfahid

Submitted: 01/02/2024 **Revised:** 10/02/2024 **Accepted:** 21/03/2024

Abstract: Early detection of schizophrenia is critical to minimize its long-term effects. This study investigates the impact of low-dimensional (Regions of Interest) and high-dimensional features (Voxel-based morphometry) on models' predictive performance for schizophrenia detection. Using the brain imaging data provided by RAMP, the study investigates the performance of a regularized linear, ensemble, and non-linear models combined with different cross-validation strategies on the low- and high-dimensional feature sets. The results show that the regularized linear model consistently outperforms the ensemble and non-linear models across both feature sets in terms of ROC-AUC, balanced accuracy, and computational efficiency. Our study also reveals that the choice of feature dimensionality does indeed impact schizophrenia detection as the low-dimensional features outperform high-dimensional features across all metrics and models. This suggests that the Regions of Interests, despite their reduced dimensionality and complexity, contain sufficient discriminative information for identifying schizophrenia, whereas the additional detail provided by the Voxel-based morphometry features does not necessarily enhance model performance. Overall, regularized linear model, combined with low-dimensional features and standard cross-validation, offers the most promising results. Using an interpretability tool, we obtained the features that have the most impact on schizophrenia detection with right pallidum grey matter volume being the most influential factor.

Keywords: *Schizophrenia; feature dimensionality; machine learning; brain grey matter; voxel-based morphometry*

1 Introduction

Schizophrenia is a complex behavioral and cognitive syndrome arising from the disruption of brain development due to genetic, environmental, or combined factors [1]. This condition often results in individuals gradually losing touch with reality, leading to delusions, hallucinations, and disordered thinking [1]. While schizophrenia can be managed through medications and psychosocial support [2], the primary challenge in effectively addressing this condition lies in its early detection and intervention [3].

In previous studies, researchers have leveraged Machine Learning techniques to detect or diagnose schizophrenia based on magnetic resonance imaging (MRI) data [4]. A study by [5] implemented a multiple kernel learning SVM model for schizophrenia detection, achieving an accuracy of 86%. A separate research [6] extracted features from the hippocampus and dorsolateral prefrontal cortex for twenty schizophrenia and healthy control cases, attaining an AUC of 0.85 with a random forest classifier. Also, [7] investigated anatomical brain connectivity for 144 schizophrenia and 154 healthy control cases using spatial statistics and an ML approach, reaching an accuracy of 79.3%. Another investigation [8] analyzed MRI and fMRI data from 296 schizophrenia and 452 healthy control subjects, finding that functional connectivity outperformed structural data, and combining both types of data led to the highest accuracy. Lastly, [9] reported that SVM and logistic regression (LR) produced the most accurate results in detecting schizophrenia.

The aim of this study is to investigate the impact of low and high-dimensional feature sets, various machine learning algorithms and cross-validation strategies in predicting schizophrenia from brain grey matter measurements. The specific research objectives are as follows:

1. To evaluate the impact of low-dimensional and high-dimensional

Department of Information Systems, College of Computer and Information Sciences, Majmaah University, Majmaah 11952, Saudi Arabia

e.alfahed@mu.edu.sa

features on models' predictive performance for schizophrenia detection.

2. To investigate the performance of a regularized linear model, an ensemble model, and a non-linear model in predicting schizophrenia.

2 Related Work

2.1 Integration of Machine Learning in Schizophrenia Diagnosis

The integration of Machine Learning (ML) in the diagnosis of schizophrenia has shown promising results, with various studies employing different ML techniques and neuroimaging modalities. A study by [1] effectively applied classification models to event-related potentials (ERPs) of patients and healthy subjects, achieving a high sensitivity and specificity with support vector machines (SVM). This aligns with [3] who explored the impact of various classification models and feature selection techniques on schizophrenia diagnosis, using structural magnetic resonance imaging data. The findings underscored the importance of selecting appropriate features and models, such as support vector machines with Gaussian kernels, to enhance diagnostic accuracy.

The comprehensive overview by [4] of AI techniques, particularly ML and deep learning, in schizophrenia diagnosis using MRI modalities, further supports these findings. This study provided a detailed comparison between conventional ML and deep learning, revealing the superiority of AI-based computer-aided schizophrenia diagnosis systems (CADS) over traditional methods. Meanwhile, [5] investigated the application of Multiple Kernel Learning classifiers in conjunction with the Boruta feature selection method. This study achieved high classification accuracy, highlighting the utility of advanced ML techniques in distinguishing schizophrenia patients from controls using neurophysiological biomarkers.

On the other hand, [6] approached the diagnosis of schizophrenia from a molecular perspective, using machine learning to analyze variations in neuroactive amino acids and synaptic elements. The study's multivariable hypothesis-driven analyses revealed significant discriminative molecular signatures, indicating the potential of ML in uncovering novel biological insights related to schizophrenia. In contrast, [7] and [8] focused on anatomical connectivity and multimodal neuroimaging data, respectively. Leveraging machine learning techniques, [7] distinguished between schizophrenia and bipolar disorder based on white matter integrity patterns, while [8] integrated structural and functional MRI data to achieve a high classification accuracy for schizophrenia.

2.2 Challenges in Schizophrenia Diagnosis Using Machine Learning

Despite the progress, there are considerable challenges in diagnosing schizophrenia using machine learning techniques. The heterogeneity of schizophrenia, as [21] pointed out, poses a significant obstacle due to the complex nature of the human brain and the variability in clinical manifestations. Similarly, [18] emphasized the importance of addressing heterogeneity in machine learning approaches to improve predictions and understanding of schizophrenia's neurobiological background.

Another challenge is the accuracy and reliability of these techniques. According to [9], while classifiers like SVM and Logistic regression performed well in classifying schizophrenia, autism, and other disorders, there were challenges in consistency and reliability, particularly when applied to ultra-high risk and first-episode psychosis subjects. This is corroborated by [20], who found that machine learning did not provide substantial added value over traditional statistical methods in predicting schizophrenia, indicating a need for further optimization of these models.

Furthermore, [6] identified the need for a multi-level approach in studying the glutamatergic alterations in schizophrenia, highlighting the complexity of integrating various molecular data into machine learning models for accurate diagnosis. In a similar vein, [7] also encountered difficulties in differentiating schizophrenia from bipolar disorder using ML, despite identifying specific white matter integrity patterns, underscoring the challenge of distinguishing between similar psychiatric disorders with ML techniques.

3 Methods

3.1 Dataset

The dataset used for this task is already pre-processed and was made available by RAMP as

part of a RAMP challenge. The input data consists of Regions Of Interest (ROIs) of Grey Matter (GM) scaled for the Total Intracranial Volume (TIV), with 284 features; these ROI features form the low-dimensional features in this study. In addition, the dataset includes Voxel-based morphometry GM 3D images. Masking the brain provides a flattened input of 331,695 features (voxels) for each participant; these VBM features form the high-dimensional features used in this study [10].

The training set has 410 samples, with 257 samples having gender 0 and 153 samples having gender 1. There are two targets with the ' healthy control' class having 222 samples and the 'schizophrenia' class having 188. The age distribution in the training set reveals that most of the participants (169) fall in the 20-29 years' age group. On the other hand, the test set has 103 samples, with 65 samples having gender 0 and 38 samples having gender 1. The 'healthy control' class has 55 samples while the 'schizophrenia' class has 48 samples. Like the training set, the age distribution in the test set also shows that most participants (41) fall in the 20-29 years' age group.

3.2 Model Selection

The three machine learning models selected for this study include logistic regression (linear model), random forest (ensemble model), and SVM (non-linear model).

1. Linear Model - Logistic Regression

This is a linear model that is well-suited for classification tasks. Logistic Regression was chosen for its simplicity, interpretability, and robustness to noise. Also, its computational efficiency means it can be trained faster on both high-dimensional and low-dimensional data. This study uses a regularized logistic regression to prevent overfitting and ensure that the model generalizes to the test set [16]. The regularization primarily adds a penalty term to the objective function, which discourages the LR model from giving too much importance to any single feature [11]. In Lasso regularization (L1), the penalty term is the absolute value of the coefficients, while in Ridge regularization (L2), it is the square of the coefficients [12]. After carrying out a GridSearch, the L1 regularization was selected as the most appropriate for this study; it can be represented as follows:

$$Objective\ function = -\log likelihood + \lambda * \Sigma|\beta\_i|$$

where λ represents the regularization parameter and β_i represents the coefficients of the model.

2. Ensemble Model - Random Forest

Random Forest learns a model M by generating several decision trees, $\{T_1, \ldots, T_k\}$, on various subsets of a given dataset that includes input variables $\{X_1, \ldots, X_n\}$ and outcome variables Y. Predictions are made by each individual tree, and the overall prediction results from a collective aggregation, typically through majority voting, of these individual predictions.

The choice of random forest is justified by its ability to handle complex patterns in data and parallelization [13] which offers faster training and feature importance estimation.

3. Non-linear Model - Multi-layer Perceptron (MLP)

Multilayer Perceptron (MLP) is a neural network that consists of multiple layers of interconnected artificial neurons. The model learns the optimal weights for these connections by minimizing an error function, typically using the backpropagation algorithm [14]. The choice of MLP is justified by its ability to model complex, non-linear relationships in the data and its suitability for high-dimensional data [14]. MLPs can capture intricate patterns in data, making them suitable for capturing the relationships between brain grey matter and schizophrenia.

3.3 Model Pipeline

The pipelines built for the three models involve data preprocessing, hyperparameter tuning, cross-validation, and model fitting based on the pre-extracted ROI and VBM features. In the data preprocessing step, a StandardScaler was used to standardize the features. The transformation is defined as:

$$z = (x - \mu)/\sigma$$

where z is the standardized feature, σ is the standard deviation, x is the original feature, and μ is the mean of the feature. This ensures consistent scales for all features and mitigates the potential adverse effects that disparate scales may have on the model learning process [15].

To find the best hyperparameters for each model, StratifiedKFold cross-validation with five

splits was used for hyperparameter tuning with GridSearchCV. The optimal hyperparameters found were C=0.1 and penalty=l1 for Logistic Regression; max_depth=10, min_sample_split=2, and n_estimators=200 for Random Forest; as well as hidden_layer_sizes=(200, 150, 100, 50, 25), alpha=0.0001, activation='relu', solver='adam', max_iter=1000, and learning_rate_init=0.001 for MLP.

For model evaluation, two cross-validation strategies were used i.e. a standard cross-validation using StratifiedKFold (with 5 splits) and a group stratified cross-validation using GroupKFold (with 2 splits), considering sex as the group. The standard cross-validation ensured consistent proportion for each class across all folds, thus reducing the possibility of biased estimates. Meanwhile, group stratified cross-validation maintained each distinct sex group in a single fold. The number of splits for GroupKFold cross-validation was set to 2, given that any split that exceeds the number of groups would compromise the required distinct separation and result in an error.

3.4 Evaluation Metrics

The models were evaluated using both performance and computational metrics. For performance evaluation, the area under the ROC curve (ROC-AUC) and balanced accuracy were used. ROC-AUC was chosen because it measures the ability of the model to discriminate between the schizophrenia patients and healthy controls classes, even when the classes are imbalanced [17]. Meanwhile, balanced accuracy was chosen because it provides a measure of the model's overall performance while accounting for class imbalance.

Furthermore, three metrics were used for computational cost evaluation: training time, inference time, and model size. The training time captures the time taken for a model to learn from the training data; it was selected because it indicates the model's efficiency. Meanwhile, inference time measures the average time taken for a model to make a single prediction on the test data, making it a critical metric for real-world application where prompt predictions are required. The model size indicates the memory requirements of each model; this is crucial for deployment in resource-constrained settings.

4 Results

The results obtained for the low-dimensional features (ROIs) are shown in Table 1.1 while the results obtained for the high-dimensional features (VBM) are shown in Table 1.2. In Table 1.1, the results for low-dimensional features show that Logistic Regression outperforms the other models in terms of ROC-AUC and balanced accuracy for the two cross-validation strategies, with ROC-AUC scores of 0.835438 and 0.786470 for standard cross-validation and group cross-validation, respectively. Meanwhile, MLP and Random Forest models record similar performance, with MLP having slightly higher ROC-AUC and balanced accuracy scores. In terms of computational cost, Logistic Regression has the lowest inference time while the MLP model has the highest inference time.

Table 1.1. Results obtained for low-dimensional features

	Standard CV		Group CV		Time	
	ROC AUC	Balanced ACC	ROC AUC	Balanced ACC	Train (s)	Inference (10^{-3})
LR	0.835	0.762	0.787	0.696	1.833	0.056
RF	0.804	0.735	0.773	0.689	1.794	1.788
MLP	0.806	0.735	0.768	0.685	1.373	2.066

The results for high-dimensional features (Table 1.2) exhibit similar trends to those observed for the low-dimensional features. Logistic Regression has the highest ROC-AUC for the two cross-validation strategies, with ROC-AUC scores of 0.715319 and 0.737097 for standard cross-validation and group cross-validation, respectively. However, MLP records the highest balanced accuracy score for standard cross-validation while Random Forest records the least performance across all performance metrics. In terms of computational cost, Logistic Regression remains the most efficient by having the least training time and inference time.

Table 1.2. Results obtained for high-dimensional features

	Standard CV		Group CV		Time	
	ROC AUC	Balanced ACC	ROC AUC	Balanced ACC	Train (s)	Inference (10^{-3})
LR	0.712	0.637	0.737	0.673	46.297	0.695
RF	0.618	0.592	0.667	0.537	97.588	2.650
MLP	0.669	0.643	0.706	0.658	265.587	2.244

The results in Table 1.1 and 1.2 show that the Logistic Regression model consistently performs better than Random Forest and MLP in terms of ROC-AUC, balanced accuracy, and computational efficiency not just for low-dimensional features but also for the high-dimensional features. However, the difference in performance is more significant with low-dimensional features, indicating that LR may be more suited for handling such data especially given the model's simplicity and linear nature.

RF and MLP record similar performance for low-dimensional features, however, the case is different for high-dimensional features with MLP showing superior performance to RF. This can probably be explained by the fact that MLP is a non-linear model that can capture complex relationships within high-dimensional data compared to an ensemble or tree-based model like Random Forest. Also, all the models exhibit better performance with the low-dimensional features compared to high-dimensional features. This indicates that the ROIs might contain the most relevant and discriminative information for predicting schizophrenia compared to the high-dimensional features.

4.2 Predictions on Test Data (ROIs)

Predictions were also made on the low-dimensional test data. The results equally indicate that Logistic Regression has the highest performance with a balanced accuracy of 0.7814 and an AUC-ROC of 0.8439. MLP follows with a balanced Accuracy of 0.7737 and an AUC-ROC of 0.8356 while Random Forest has the least performance with a balanced accuracy of 0.7282 and an AUC-ROC of 0.8034. These test results are consistent with the trends observed in the cross-validation results.

To understand the features contributing the most to the predictions, SHAP explainability framework was employed to generate a global explanation for the best performing model, Logistic Regression. The resulting SHAP plot (with top twenty features) as shown in Figure 1 reveals that rpal_GM_Vol ((right pallidum grey matter volume), iMedfroCbr_CSF_Vol (intermediary medial frontal cortex cerebrospinal fluid volume), and IAntIns_GM_Vol (inferior anterior insula grey matter volume) are the top three features with the greatest impact on the the model's prediction.

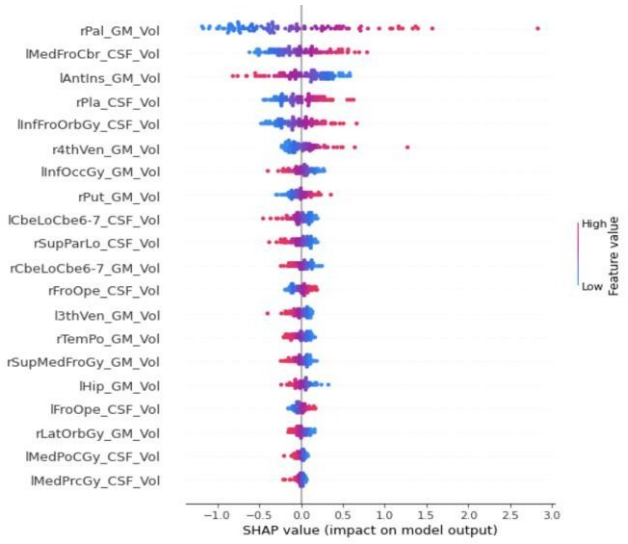

Figure 1. SHAP's global explanation of LR's predictions

5 Discussion

5.1 Low vs High Dimensional Features

The distinction between low- and high-dimensional features in our research, specifically ROIs and VBMs, significantly impacts the performance and computational efficiency of machine learning models used for schizophrenia detection. Low-dimensional features, derived from Regions Of Interest (ROIs) of Grey Matter (GM), consist of 284 features that provide a summarized view of the brain's structure. High-dimensional features, obtained from Voxel-based Morphometry (VBM) GM 3D images, offer a detailed representation with 331,695 features (voxels) for each participant. This granular view, while rich in information, presents challenges in model training and performance.

Our findings indicate that models leveraging low-dimensional features consistently outperform those using high-dimensional features across several metrics, including ROC-AUC and balanced accuracy. This suggests that ROIs, despite their reduced complexity, contain sufficient discriminative information for identifying schizophrenia, whereas the additional detail provided by VBM features does not necessarily enhance model performance. This observation underscores the importance of feature selection in the development of effective diagnostic tools, where the goal is to maximize accuracy while minimizing computational demands. Furthermore, the computational load associated with processing high-dimensional data is notably higher, as evidenced by increased training times. This is crucial for the practical deployment of diagnostic models in clinical settings, where efficiency and timeliness are key. The preference for low-dimensional features, therefore, extends beyond their predictive performance to include considerations of computational feasibility and resource allocation.

5.2 Standard vs Group cross-validation (CV)

The performance of all the three models tend to be higher with the standard CV strategy compared to the group CV for both low and high dimensional features. This finding suggests that standard CV allows for more diverse training data which can improve models' generalization. It is however important to note that standard CV may not account for the inherent grouping structure in the dataset (i.e. sex group 0 & 1). In contrast, group CV provides a more realistic assessment of model performance as it maintains the sex group structure during validation, thus better accounting for potential biases caused by sex-related differences in brain grey matter features.

6 Conclusion

This study investigated the performance and computational efficiency of three different machine learning models in predicting Schizophrenia on both low-dimensional (ROIs) and high-dimensional (VBM) data. Logistic Regression, Random Forest, and MLP were chosen as regularized linear, ensemble, and non-linear models, respectively. The impact of different cross-validation strategies i.e. standard CV and group CV were also assessed.

From the results, Logistic Regression outperformed the other models in terms of ROC-AUC, balanced accuracy, and computational efficiency for both low-dimensional and high-dimensional features. However, the performance difference is more significant with low-dimensional features, suggesting that Logistic Regression is particularly well-suited for handling such data. In addition, all models performed better with low-dimensional features compared to high-dimensional features, implying that ROIs might already capture the most relevant and discriminative information for predicting schizophrenia. In addition, this study reveals that standard cross-validation produced better performance for all models compared to group cross-validation. However, that should be taken with a grain of salt given that group cross-validation offers a more realistic assessment of model performance by accounting for the inherent grouping structure in the data, such as sex-related differences in brain grey matter features.

Using the SHAP explainability framework, the study identified the top twenty features with the greatest impact on the model's predictions, providing insights into the most important brain regions associated with schizophrenia. This information is critical to understanding the underlying neurobiological mechanisms of schizophrenia and can guide future research.

Overall, the findings suggest that the linear model, when used with low-dimensional features and standard cross-validation, offers the most promising results in terms of accuracy and computational efficiency. Future studies integrate multimodal data to enhance predictive accuracy,

replicate findings across diverse populations to ensure generalizability, and evaluate machine learning models in clinical trials to bridge the gap between research and clinical practice.

Availability of Data and Materials: The dataset used for this study is available on RAMP (https://ramp.studio/).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding this study.

References

[1] Shanarova, Nadezhda, Marina Pronina, Mikhail Lipkovich, Valery Ponomarev, Andreas Müller, and Juri Kropotov. 2023. "Application of Machine Learning to Diagnostics of Schizophrenia Patients Based on Event-Related Potentials" *Diagnostics* 13, no. 3: 509. https://doi.org/10.3390/diagnostics13030509

[2] "World Health Organization on schizophrenia", [online] Available: https://www.who.int/news-room/fact-sheets/detail/schizophrenia.

[3] M. Tanveer, J. Jangir, M. A. Ganaie, I. Beheshti, M. Tabish and N. Chhabra, "Diagnosis of Schizophrenia: A Comprehensive Evaluation," in *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1185-1192, March 2023, doi: 10.1109/JBHI.2022.3168357.

[4] Sadeghi, D., Shoeibi, A., Ghassemi, N., Moridian, P., Khadem, A., Alizadehsani, R., & Acharya, U. R. (2022). An overview of artificial intelligence techniques for diagnosis of Schizophrenia based on magnetic resonance imaging modalities: Methods, challenges, and future works. *Computers in Biology and Medicine*, 105554.

[5] Febles ES, Ortega MO, Sosa MV, Sahli H (2022) Machine learning techniques for the diagnosis of schizophrenia based on event related potentials. medRxiv

[6] De Rosa A, Fontana A, Nuzzo T, Garofalo M, Di Maio A, Punzo D, Copetti M, Bertolino A, Errico F, Rampino A (2022) Machine learning algorithm unveils glutamatergic alterations in the post-mortem schizophrenia brain. Schizophrenia 8(1):1–16

[7] Sutcubasi B, Metin SZ, Erguzel TT, Metin B, Tas C, Arikan MK, Tarhan N (2019) Anatomical connectivity changes in bipolar disorder and schizophrenia investigated using whole-brain tract-based spatial statistics and machine learning approaches. Neural Comput Appl 31(9):4983–4992

[8] Lei D, Pinaya WH, Young J, van Amelsvoort T, Marcelis M, Donohoe G, Mothersill DO, Corvin A, Vieira S, Huang X (2020) Integrating machining learning and multimodal neuroimaging to detect schizophrenia at the level of the individual. Hum Brain Mapping 41(5):1119–1135

[9] Yassin W, Nakatani H, Zhu Y, Kojima M, Owada K, Kuwabara H, Gonoi W, Aoki Y, Takao H, Natsubori T (2020) Machine-learning classification using neuroimaging data in schizophrenia, autism, ultra-high risk and first-episode psychosis. Transl Psychiatry 10(1):1–11

[10] Edouard D., Antoine G., and Benoît D. (2023). RAMP Challenge - Predict schizophrenia using brain anatomy. Université Paris-Saclay.

[11] Arafa, A., Radad, M., Badawy, M., & El-Fishawy, N. (2021, July). Regularized logistic regression model for cancer classification. In *2021 38th National Radio Science Conference (NRSC)* (Vol. 1, pp. 251-261). IEEE.

[12] Kolluri, J., Kotte, V. K., Phridviraj, M. S. B., & Razia, S. (2020, June). Reducing overfitting problem in machine learning using novel L1/4 regularization method. In *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)* (pp. 934-938). IEEE.

[13] Liu, S., & Hu, T. (2018, November). Parallel random forest algorithm optimization based on maximal information coefficient. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)* (pp. 1083-1087). IEEE.

[14] Abinaya, S., & Devi, M. K. (2022). Enhancing crop productivity through autoencoder-based disease detection and context-aware remedy recommendation system. In *Application of Machine Learning in Agriculture* (pp. 239-262). Academic Press.

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

[16] Idowu, J. A. (2024). Debiasing Education

Algorithms. *International Journal of Artificial Intelligence in Education*, 1-31.

[17] Thölke, P., Mantilla Ramos, Y. J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., ... & Jerbi, K. (2022). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *bioRxiv*, 2022-07.

[18] Schnack, H. G. (2019). Improving individual predictions: machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). Schizophrenia research, 214, 34-42.

[19] Wang, K. Z., Bani-Fatemi, A., Adanty, C., Harripaul, R., Griffiths, J., Kolla, N., ... & De Luca, V. (2020). Prediction of physical violence in schizophrenia with machine learning algorithms. Psychiatry research, 289, 112960.

[20] Bracher-Smith, M., Rees, E., Menzies, G., Walters, J. T., O'Donovan, M. C., Owen, M. J., ... & Escott-Price, V. (2022). Machine learning for prediction of schizophrenia using genetic and demographic factors in the UK biobank. Schizophrenia research, 246, 156-164.

[21] Tandon, N., & Tandon, R. (2019). Using machine learning to explain the heterogeneity of schizophrenia. Realizing the promise and avoiding the hype. Schizophrenia Research, 214, 70-75