HHS Public Access

Author manuscript

Schizophr Res. Author manuscript; available in PMC 2023 July 01.

Published in final edited form as:

Schizophr Res. 2022 July ; 245: 116–121. doi:10.1016/j.schres.2021.03.020.

Development of a Computerized Adaptive Diagnostic Screening Tool for Psychosis

Robert D. Gibbons, PhD[1,2],

Ishanu Chattopadhyay, Ph.D.[1],

Herbert Meltzer, MD[3],

John M. Kane, MD[4,5,6],

Daniel Guinart, MD[4,5,6]

[1]Center for Health Statistics, Department of Medicine, and the Committee on Quantitative Methods, University of Chicago, Chicago, Illinois

[2]Departments of Public Health Sciences (Biostatistics), Psychiatry, Comparative Human Development, University of Chicago, Chicago, Illinois

[3]Northwestern University, Department of Psychiatry, Chicago, Illinois

[4]The Zucker Hillside Hospital, Department of Psychiatry Research, New York

[5]Center for Psychiatric Neuroscience, Feinstein Institute for Medical Research, Manhasset, NY, USA

[6]The Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Manhasset, New York

Abstract

We develop a two-stage diagnostic classification system for psychotic disorders using an extremely randomized trees machine learning algorithm. Item bank was developed from clinician-rated items drawn from an inpatient and outpatient sample. In stage 1, we differentiate schizophrenia and schizoaffective disorder from depression and bipolar disorder (with psychosis). In stage 2 we differentiate schizophrenia from schizoaffective disorder. Out of sample classification accuracy, determined by area under the receiver operator characteristic (ROC) curve,

Corresponding author: Robert D. Gibbons, PhD, Department of Public Health Sciences, The University of Chicago, 5841 South Maryland Ave MC2000, Chicago, IL 60637 - 1447.

Disclosures

Dr. Gibbons has been an expert witness for the US Department of Justice, Merck, Glaxo-Smith-Kline, Pfizer and Wyeth and is a founder of Adaptive Testing Technologies, which distributes the CAT-MH™ battery of adaptive tests in which CAT-Psychosis is included. Dr. Meltzer has been a consultant and/or advisor for or has received honoraria from Eli Lily, Gershon Lehrman Group, Guidepoint Global, LB Pharmaceuticals, Quincy Bioscience, Sumitomo Dianippon Pharma, and is a board member and shareholder in Acadia Pharmaceuticals. Dr. Kane has been a consultant and/or advisor for or has received honoraria from Alkermes, Allergan, LB Pharmaceuticals, H. Lundbeck, Intracellular Therapies, Janssen Pharmaceuticals, Johnson and Johnson, Merck, Minerva, Neurocrine, Newron, Otsuka, Pierre Fabre, Reviva, Roche, Sumitomo Dainippon, Sunovion, Takeda, Teva and UpToDate and is a shareholder in LB Pharmaceuticals and Vanguard Research Group. Dr. Guinart has been a consultant for and/or has received speaker honoraria from Otsuka America Pharmaceuticals and Janssen Pharmaceuticals. Dr. Chattopadhyay's has nothing to disclose.

was outstanding for stage 1 (Area under the ROC curve (AUC)=0.93, 95% confidence interval (CI) = 0.89, 0.94), and excellent for stage 2 (AUC=0.86, 95% CI = 0.83, 0.88). This is achieved based on an average of 5 items for stage 1 and an average of 6 items for stage 2, out of a bank of 73 previously validated items.

Introduction

Serious mental illnesses, particularly disorders that present with psychotic symptoms, are common and are associated with significant functional impairment and disability worldwide (James et al., 2018). Psychosis is a distinguishing characteristic of schizophrenia spectrum disorders, including schizoaffective disorder, and a common but variable characteristic of severe mood disorders, such as bipolar disorder or major depressive disorder (American Psychiatric Association, 2013).

Presence and severity of psychotic symptoms is a significant contributor to disability across all aforementioned conditions (Morgan et al., 2014; Rabinowitz et al., 2013) and should accordingly be considered as an important target for assessment and evaluation during all stages of the illness, ranging from early intervention to prevention of relapse. Duration of untreated psychosis, which is the time that elapses between onset of psychotic symptoms and onset of treatment, is associated with poor general symptomatic outcome, poorer social functioning, and worse global outcomes. Further, preventing relapse is also important, because relapse interferes with the social and vocational development of individuals, impacting long-term outcomes. (Fusar-Poli et al., 2017)

Psychotic symptoms include hallucinations, delusions and disorganized thinking and behavior. Other relevant symptoms of psychotic disorders include catatonia, affective flattening, social isolation, and anhedonia (Picchioni and Murray, 2007). Formulating a specific diagnosis when psychotic symptoms are involved is challenging, because the clinical features are relatively nonspecific, therefore the need for an accurate early diagnosis is critical. Clinical guideline recommendations vary if the patient's presentation suggests an affective rather than schizophrenia spectrum psychosis, which has a clear impact on treatment choice (National Institute for Health and Care Excellence (NICE), 2014). This is also relevant because prospective diagnostic stability after a first episode both for schizophrenia spectrum psychoses and for affective spectrum psychoses is quite high, and in fact comparable to other clinical diagnoses across all of medicine (Fusar-Poli et al., 2016). Conversely, retrospective diagnostic stability is rather low, indicating that many patients who receive a non-specific diagnosis of psychosis will eventually develop either schizophrenia or related psychoses or affective psychoses (Fusar-Poli et al., 2016). Therefore, being able to determine a valid diagnosis of either schizophrenia spectrum or affective spectrum psychotic disorders as soon as possible has significant clinical impact (Catts and O'Toole, 2016; Fusar-Poli et al., 2017).

Unfortunately, there is a lack of brief yet valid diagnostic tools in mental health that can be used in regular clinical practice. Current evidence-based diagnostic tools (First, Williams and Karg, 2016; Sheehan et al., 1998; Wing et al., 1990) require a trained interviewer and consist of lengthy assessments that can be done only in a few select

research-centered facilities, far beyond the reach of many academic hospitals, let alone smaller community clinics, where the majority of mental healthcare takes place. While some first-stage screening tools are capable of accurately detecting a clinically relevant early presence of psychotic symptoms, they are not disorder-specific (Loewy et al., 2011; Phalen et al., 2018), or they need to be administered by a clinician (Miller et al., 2003).

Recently, we developed and validated a computerized adaptive testing (CAT) tool based on multidimensional Item Response Theory (MIRT) for psychosis (CAT-Psychosis) that can accurately and reliably measure severity of psychotic symptoms, both administered by a clinician, but also as a self-report, requiring only a very brief administration time and a median of 12 adaptive items (Guinart et al., 2020). Using this same tool, we also found that the CAT-Psychosis can discriminate psychosis from healthy controls, both when administered by a clinician (Area Under the ROC Curve (AUC)=0.965, 95% Confidence Interval (CI): 0.945–0.98) and as a self-report (AUC= 0.850, 95% CI: 0.807–0.894). Thus, the self-report version of the CAT-Psychosis makes population-level screening for psychotic disorders possible.

However, while certain cut-off points on an underlying continuous measure can, in some cases, yield high sensitivity and specificity (Gibbons et al., 2013), screening for and eventually establishing a diagnosis within types of psychosis is not the primary objective of the CAT-Psychosis, which is designed to assess the dimensional measurement of severity of psychotic symptoms. In other words, a diagnostic tool needs to be designed specifically to ascertain the possibility of a condition, such as a specific disorder, rather than obtaining a dimensional score and/or grading the severity of such disorder.

An alternative approach to lengthy interviews for traditional diagnostic assessment is computerized adaptive diagnosis (CAD), in which individuals answer a series of symptom questions until there is high probability that they do or do not meet a threshold for a predetermined diagnosis, such as schizophrenia. CAD is based on decision-theoretic models such as decision trees and random forests (ensembles of decision trees) (Geurts et al., 2006) in which the next administered symptom item is conditional on the answers to the previously administered symptom items. The objective of CAD is to administer the fewest number of items to make a diagnosis with a specified level of confidence. As in the case of CAT, the adaptive part of the algorithm selects the next most informative item to administer based on the responses to the items that have been previously administered, dramatically reducing the number of items needed to reach a decision, thus minimizing administration time and patient burden. In addition, these decision trees can include items related to other disorders as well, potentially generating a predictive probability score for each disorder of interest, i.e. the likelihood that a patient suffers from schizophrenia or schizoaffective disorder versus depression or bipolar disorder. There are three fundamental differences between CAD and CAT. First, CAD produces a binary classifier of whether someone meets criteria for diagnosis, whereas CAT produces a dimensional severity score. Second, CAD includes a criterion, in this case a gold standard diagnosis based on a structured clinical interview to train the CAD, whereas CAT is criterion free. We use the criterion diagnosis to validate the screening properties of CAT, but the CAT does not use diagnosis to develop a dimensional severity score. Both CAD and CAT produce uncertainty estimates

respectively for the classification and score, which further distinguish them from classical approaches to classification and measurement. Third, CAD is based on logistic regression or machine learning algorithms, whereas CAT is based on unidimensional or multidimensional Item Response Theory (IRT) (the CAT-Psychosis is based on multidimensional IRT (MIRT) (Guinart et al., 2020)). Previous studies document the success of CAD. Gibbons and colleagues (Gibbons et al., 2013) developed of a CAD for major depressive disorder (MDD) that reduces a clinician's time to make the diagnosis from 90 minutes (based on the Structured Clinical Interview (SCID) for DSM-5) to less than a minute of the patient's time (without clinician input) yet maintained sensitivity of 0.95 and specificity of 0.87 for the DSM diagnosis. Heretofore, there has been no CAD-based diagnostic tool for diagnosing psychosis. Thus, this study developed a CAD-based tool for psychosis which can minimize the time needed to obtain a diagnosis of schizophrenia or schizoaffective disorder.

Material and Methods

Item Bank Development and Patient Sample

The first step in the development process was to generate an item bank from which questions are drawn. The item bank for this study was constructed from clinician-rated items on the Brief Psychiatry Rating Scale (BPRS)(Overall and Gorham, 1988), the Schedule for Affective Disorders and Schizophrenia (SADS) (Endicott and Spitzer, 1978), the Scale for the Assessment of Positive Symptoms (SAPS) (Andreasen, 1984), and the Scale for the Assessment of Negative Symptoms (SANS)(Andreasen, 1983). This yielded 144 items for which existing clinician ratings were available for a sample of 649 subjects (535 schizophrenia, 43 schizoaffective, 54 depression, and 17 bipolar) drawn from inpatients and outpatients in the Psychobiology Clinic at Case Western Reserve University. The item bank was then reduced to 73 items that best discriminated high and low levels of psychosis in our previous study (Guinart et al., 2020).

Classification Algorithm

Next, a classification algorithm was generated that was able to adaptively select the most predictive question based on the responses to previous symptom queries. Unlike IRT-based CAT, the goal of which is to efficiently measure a latent construct like psychosis and then validate it using an external criterion (e.g., SCID DSM-5 clinician diagnosis) (Gibbons et al., 2016), CAD uses the diagnostic information (i.e., external criterion) to derive a classifier based on a subset of the symptoms in the item bank that maximizes the association between the items and the diagnosis (Gibbons et al., 2013).

An important distinction in the learning and modeling task here is the constraint on the number of features that may be used per sample. While the model can draw from a relatively large set of features (i.e., symptoms) in training, when evaluating a specific test sample, it is restricted to a pre-specified number of features. This constraint directly arises from the need for limiting the items per test sample, minimizing patient burden while maintaining high classification accuracy. By using decision trees or ensembles of decision trees (i.e., a random forest), the number of estimators (trees) in each sample and the maximum depth (i.e., number of items administered) of each tree is constrained. This constraint will incur

a performance cost. It has been established that random forests and its related architectures generally produce significant improvement over classical decision trees in most applications (Breiman, 2001). Here, with this particular constraint, we build an ensemble with a limited number of estimators, each of which have a constrained depth. A taxonomy of machine learning classification methods (CADs) can be found in Figure 1. A weighted combination of the component estimators is the final model, using weights optimized from training data. Intuitively, the optimization of these weights introduces additional degrees of freedom for the learning algorithm to operate in, that boosts the performance of this constrained ensemble over a single estimator with constrained depth used in the development of the CAD-MDD (Gibbons et al., 2013).

A search of possible classification algorithms indicated that the extra-trees algorithm performs best, i.e., maximizes AUC under the constraint described above, while allowing for the generation of hundreds of distinct test sets. The Extra-Trees method (extremely randomized trees) was proposed in (Geurts et al., 2006) with the objective of further randomizing tree building in the context of numerical input features, where the choice of the optimal cut-point is responsible for a large proportion of the variance of the induced tree. With respect to random forests, the method drops the idea of using bootstrap copies of the learning sample, and instead of trying to find an optimal cut-point for each one of the K randomly chosen features at each node, it selects a cut-point at random. This idea is productive in the context of many problems characterized by a large number of numerical features varying more or less continuously: it leads often to increased accuracy thanks to its smoothing and at the same time significantly reduces computational burdens linked to the determination of optimal cut-points in standard trees and random forests. From a statistical point of view, dropping the bootstrapping idea leads to an advantage in terms of bias, whereas the cut-point randomization reduces variance. From a functional point of view, the Extra-Tree method produces piece-wise multilinear approximations, rather than the piece-wise constant ones of random forests.

To provide out-of-sample validation, we create an empirical distribution for AUC, which are shown as individual black lines on Figures 2a and 2b, depicting uncertainty in the estimated ROC curves (see results). Each individual line is generated from a random split of the dataset into training and validation subsets. This approach is similar to “leave-p out cross-validation”, except we set a dataset fraction instead of a fixed p (validation fraction=0.5, training fraction=0.5). We repeat this process and generate an empirical distribution for AUC. We terminate the iteration when we get no significant change in the AUC distribution under a Kolmogorov-Smirnov (KS) test. Since this split can be modeled as a random draw from the complete dataset, confidence bounds for AUC are computed directly from percentiles of the empirical distribution.

In our dataset, the number of cases by group is heavily tilted toward schizophrenia. To ensure that this class imbalance does not negatively impact our prediction results, the implementation of the classification algorithm used in the study (Extremely Randomized trees from the scikit-learn Python library) explicitly takes account of class imbalance, by inversely weighting samples according to their class frequency. We investigated different variations of the class weighting, including using static weights inversely proportional to

class frequencies, and where the sample weights are computed based on the bootstrap sample for every tree grown. The imbalance in the dataset is therefore expected to have little or no impact on the reported results.

The restriction of the complexity of the decision trees to a maximum number of 12 items per test (which is different from the actual number of items being used for each subject) is a design decision, motivated by the number of questions that subjects are comfortable in answering. We tested the validity of this choice by studying the gains from increasing that number from 6–18 items.

The relative feature importances of the items used in the two classification scenarios is assessed via Gini Importance or Mean Decrease in Impurity (MDI) (Louppe et al, 2013), which calculates each feature importance as the sum over the number of splits (across all trees inferred for each classification scenario) that include the feature, proportionally to the number of samples it splits. We studied the top 15 items ranked by their importances.

Diagnostic Screening

To optimize the screening process, while minimizing patient burden, we have developed a two-stage screener. In stage 1, we differentiate schizophrenia and schizoaffective disorder from other disorders (in our sample bipolar disorder and depression with psychotic features). While one can stop there and conclude that the patient is suffering from a primarily psychotic disorder, we can also proceed to the second stage of the classification process and determine whether the patient has schizophrenia or schizoaffective disorder. The purpose of this study was to determine whether we could develop classifiers for the first stage and the second stage that have high classification accuracy and minimal patient burden. To this end, for each classifier, we have computed out of sample AUC and the corresponding 99% confidence interval.

Results

Based on the extra-trees algorithm with a maximum tree depth of 12 items (combination of two 6 node trees) for each of the two classifiers, and eliminating redundant items, the average number of items administered was 5 for the first classifier (schizophrenia and schizoaffective disorder versus bipolar disorder and depression) and 6 for the second classifier (schizophrenia versus schizoaffective disorder). A total of 29 extra-tree forms were generated for the first classifier and 13 forms for the second classifier. The average number of unique items administered per form (across both entire trees) was 27 for the first classifier and 31 for the second classifier, which illustrates excellent coverage of the 73-item bank for the two diagnostic classifiers.

Classification accuracy for the first stage screener of schizophrenia and schizoaffective disorder versus bipolar disorder and depression had AUC = 0.93, 99% confidence interval (CI) (0.89, 0.94). The average ROC curve and random splits into training and validation datasets are displayed in Figure 2a. Classification accuracy for the second stage screener of schizophrenia versus schizoaffective disorder had AUC = 0.86, 99% CI=0.83, 0.88. The

average ROC curve and random splits into training and validation datasets are displayed in Figure 2b.

To provide a comparison, we repeated the analyses using a Random Forest, with the same 12 item maximum tree depth for each classifier. Classification accuracy for the first stage screener of schizophrenia and schizoaffective disorder versus bipolar disorder and depression had AUC = 0.92, 99% CI=0.91, 0.93, which was similar to the extra-trees algorithm (AUC=0.93). Classification accuracy for the second stage screener of schizophrenia versus schizoaffective disorder had AUC = 0.80, 99% CI= 0.78, 0.82, which was lower than the extra-trees algorithm (AUC=0.86).

We have also compared our extra-trees results to the best possible performance which would be obtained using a random forest using all 73 items. Classification accuracy for the first stage screener of schizophrenia and schizoaffective disorder versus bipolar disorder and depression had AUC = 0.97, 99% CI=0.96, 0.98, versus the 12 item (max) extra-trees AUC=0.93 obtained for the extra-trees algorithm limited to a max of 12 items. Classification accuracy for the second stage screener of schizophrenia versus schizoaffective disorder had AUC = 0.92, 99% CI= 0.90, 0.93, versus the 12 item (max) extra-trees AUC=0.86. As noted above, eliminating redundancy, the unique number of items administered is approximately 6 items per screener, with results approaching the classification accuracy of the entire 73 item bank.

Figures 3a and 3b compare classification accuracy as measured by AUC for maximum tree depths of 6, 12 and 18 items. For both classifiers, the extra-trees results for maximum depth of 6 items is significantly inferior to 12 and 18 items, but no significant difference found between 12 and 18 items. This validates our selection of a maximum tree depth of 12 items.

Figures 4a and 4b display symptom importance in the two classification functions listed by item number, where the item text is available in the Supplement for each item. For the classification of schizophrenia and schizoaffective disorder versus depression and bipolar disorder, the top 5 items in terms of importance were: (1) Experienced auditory hallucinations, (15) Severity of delusions of any type, (54) Insight (aware of psychotic symptoms), (29) Remarks have special meaning, and (61) Ability to communicate. For the classification of schizophrenia versus schizoaffective disorder, the top 5 items in terms of importance were: (30) Feel that people can read your mind, (15) Severity of delusions of any type, (25) Duration and severity of hallucinations, (6) Thought broadcasting, and (29) Remarks have special meaning.

Discussion

In this study, we report for the first time the development of a highly accurate, two-stage computerized adaptive screener for schizophrenia and schizoaffective disorder. Classification accuracy for both the first and the second stage screeners represent “outstanding” and “excellent” classification accuracies, respectively (Hosmer and Lemeshow, 2000), and are obtained after administration of a total of 11 items.

A timely and accurate diagnosis of schizophrenia spectrum disorders such as schizophrenia or schizoaffective disorder is critical due to the long-term implications of these disorders, including a significant clinical, familial and economic burden (Chong et al., 2016; Holm et al., 2020) as well as a significant decrease in life expectancy (Olfson et al., 2015; Tanskanen et al., 2018). Early detection and subsequent intervention can reduce the duration of untreated illness and improve treatment outcomes for individuals with psychosis (Srihari et al., 2014). However, clinical diagnosis poses a significant challenge due to a remarkable number of overlapping symptoms between disorders as well as a variable number of overlapping features with other disorders that can also present with psychotic symptoms such as major depressive disorder and bipolar disorders. Nonetheless, the need for an accurate early diagnosis is critical, as treatment choice varies between an affective psychosis and/or schizophrenia spectrum psychosis.

To our knowledge, this is the first application of the extremely randomized trees algorithm in mental health research. The algorithm performed well, outperforming a more conventional random-forest machine learning algorithm with the same item depth. The extra-trees algorithm using an average of 6 unique items administered per screener performed almost as well as a random forest based on administration of all 73 items. Breaking the classification algorithm into two parts, schizophrenia, and schizoaffective disorder vs other and then schizophrenia and schizoaffective disorder, also produced statistical efficiencies, since in many applications (e.g., large-scale screening or case-finding) only the first classifier would be used.

There are multiple advantages of adaptive test administration. First, it mimics the clinician's clinical interview, by basing the selection of the next symptom query based on the responses to the previous symptom queries. Second, we can dramatically reduce the length of the interview and administer the test in settings where skilled psychiatric clinicians are unavailable, opening a window of opportunity for expansion of evidence-based medicine beyond the limits of big hospitals and academic centers.

In future steps, this newly developed tool will need to be embedded into a user-friendly interface and tested in an independent sample for validation against gold standard screening and/or diagnostic tools. Additionally, further studies could include clinical high-risk and/or at-risk patients to test the ability of these tools too to detect very early symptoms of psychosis. Given the growing evidence that smartphones hold great potential to help diagnose, monitor, and treat psychiatric disorders (Melbye et al., 2020), facilitation of remote administration on multiple devices will be necessary. Additionally, it is possible that different routes to a diagnosis taken by any individual could be informative for data-driven classification, potentially mirroring phenotypical and/or neurobiological subgroups within psychotic disorders.

A potential limitation of this tool relates to the fact that symptom severity is not measured, but that can be easily overcome with the administration of other brief, self-assessed computerized adaptive tools (Guinart et al., 2020). An additional limitation is that other diagnoses that can present with psychosis or psychotic features are not included, such as delusional disorder, drug-induced psychosis, or psychosis secondary to other medical

conditions, albeit they are less frequent. Thus, if such diagnoses are suspected, additional diagnostic tests would be needed. Finally, we have developed the CAD-Psychosis based on clinician ratings of the 73 symptoms in the item bank. We have previously developed a patient self-report version of the item-bank (Guinart et al., 2020), which can be used to develop a patient self-report version of the CAD-Psychosis as we have done for the CAT-Psychosis (Guinart et al., 2020). This would involve the collection of a new dataset with the self-report items and diagnostic interviews. It may be that the same algorithm applied to the self-report data would result in the same high classification accuracy, or alternatively unique self-report versions of the two classifiers could be constructed.

Conclusion

Using an extremely randomized trees method, we developed a brief, two-stage computerized adaptive screener for psychotic disorders, with potential for self-administration using digital devices and a significant reduction in administration time, while maintaining high diagnostic accuracy. Future studies in independent samples based on patient self-report, are needed to further validate this tool against gold standard structured clinical interviews.

Acknowledgments

Funding Source

The National Institute of Mental Health, through its support of the *Bipolar and Schizophrenia Network for Intermediate Phenotypes*: NIMH, MH077851 to RDG.

References

American Psychiatric Association, 2013. DSM-5: Diagnostic and Statistical Manual of Mental Disorders, (5th Ed.). ed. Washington, DC.

Andreasen NC, 1984. The Scale for the Assessment of Positive Symptoms (SAPS). Iowa City, Iowa Univerisity Iowa.

Andreasen NC, 1983. The Scale for the Assessment of Negative Symptoms (SANS). Iowa City, Iowa Univerisity Iowa.

Breiman L, 2001. Random Forests. Mach. Learn 45, 5–32. 10.1023/A:1010933404324

Catts SV, O'Toole BI, 2016. The treatment of schizophrenia: Can we raise the standard of care? Aust. N. Z. J. Psychiatry 50, 1128–1138. 10.1177/0004867416672725 [PubMed: 27821411]

Chong HY, Teoh SL, Wu DB-C, Kotirum S, Chiou C-F, Chaiyakunapruk N, 2016. Global economic burden of schizophrenia: a systematic review. Neuropsychiatr. Dis. Treat 12, 357–373. 10.2147/NDT.S96649 [PubMed: 26937191]

Endicott J, Spitzer RL, 1978. A diagnostic interview: the schedule for affective disorders and schizophrenia. Arch. Gen. Psychiatry 35, 837–844. 10.1001/archpsyc.1978.01770310043002 [PubMed: 678037]

First MB, Williams JBW, Karg RS, S.R., 2016. Structured Clinical Interview for DSM-5 Disorders, Clinician Version (SCID-5-CV). Arlington, VA, Am. Psychiatr. Assoc

Fusar-Poli P, Cappucciati M, Rutigliano G, Heslin M, Stahl D, Brittenden Z, Caverzasi E, McGuire P, Carpenter WT, 2016. Diagnostic Stability of ICD/DSM First Episode Psychosis Diagnoses: Meta-analysis. Schizophr. Bull 42, 1395–1406. 10.1093/schbul/sbw020 [PubMed: 26980142]

Fusar-Poli P, McGorry PD, Kane JM, 2017. Improving outcomes of first-episode psychosis: An overview. World Psychiatry 16, 251–265. 10.1002/wps.20446 [PubMed: 28941089]

Geurts P, Ernst D, Wehenkel L, 2006. Extremely randomized trees. Mach. Learn 63, 3–42. 10.1007/s10994-006-6226-1

Gibbons RD, Hooker G, Finkelman MD, Weiss DJ, Pilkonis PA, Frank E, Moore T, Kupfer DJ, 2013. The computerized adaptive diagnostic test for major depressive disorder (CAD-MDD): a screening tool for depression. J. Clin. Psychiatry 74, 669–674. 10.4088/JCP.12m08338 [PubMed: 23945443]

Gibbons RD, Weiss DJ, Frank E, Kupfer D, 2016. Computerized Adaptive Diagnosis and Testing of Mental Health Disorders. Annu. Rev. Clin. Psychol 12, 83–104. 10.1146/annurev-clinpsy-021815-093634 [PubMed: 26651865]

Guinart D, de Filippis R, Rosson S, Patil B, Prizgint L, Talasazan N, Meltzer H, Kane JM, Gibbons RD, 2020. Development and Validation of a Computerized Adaptive Assessment Tool for Discrimination and Measurement of Psychotic Symptoms. Schizophr. Bull Nov 9, Epub ahead of print.

Holm M, Taipale H, Tanskanen A, Tiihonen J, Mitterdorfer-Rutz E, 2020. Employment among people with schizophrenia or bipolar disorder: a population-based study using nationwide registers. Acta Psychiatr. Scand 10.1111/acps.13254

Hosmer DW, Lemeshow S, 2000. Applied logistic regression, 2nd Editio. ed. John Wiley & Sons, Inc., New York, NY, US. 10.1002/0471722146

James SL, Abate D, Abate KH, et al. , 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 392, 1789–1858. 10.1016/S0140-6736(18)32279-7 [PubMed: 30496104]

Loewy RL, Pearson R, Vinogradov S, Bearden CE, Cannon TD, 2011. Psychosis risk screening with the Prodromal Questionnaire--brief version (PQ-B). Schizophr. Res 129, 42–46. 10.1016/j.schres.2011.03.029 [PubMed: 21511440]

Louppe G, Wehenkel L,Sutera A,Geurts P 2013. Understanding variable importances in forests of randomized trees. Proceedings of the 26th International Conference on Neural Information Processing Systems – Vol.1, 431–439, Lake Lahore, NV.

Melbye S, Kessing LV, Bardram JE, Faurholt-Jepsen M, 2020. Smartphone-Based Self-Monitoring, Treatment, and Automatically Generated Data in Children, Adolescents, and Young Adults With Psychiatric Disorders: Systematic Review. JMIR Ment. Heal 7, e17453. 10.2196/17453

Miller TJ, McGlashan TH, Rosen JL, Cadenhead K, Cannon T, Ventura J, McFarlane W, Perkins DO, Pearlson GD, Woods SW, 2003. Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: predictive validity, interrater reliability, and training to reliability. Schizophr. Bull 29, 703–715. 10.1093/oxfordjournals.schbul.a007040 [PubMed: 14989408]

Morgan C, Lappin J, Heslin M, Donoghue K, Lomas B, Reininghaus U, Onyejiaka A, Croudace T, Jones PB, Murray RM, Fearon P, Doody GA, Dazzan P, 2014. Reappraising the long-term course and outcome of psychotic disorders: the AESOP-10 study. Psychol. Med 44, 2713–2726. 10.1017/S0033291714000282 [PubMed: 25066181]

National Institute for Health and Care Excellence (NICE), 2014. Psychosis and schizophrenia in adults: prevention and management. [WWW Document]. URL www.nice.org.uk (accessed 11.20.20).

Olfson M, Gerhard T, Huang C, Crystal S, Stroup TS, 2015. Premature Mortality Among Adults With Schizophrenia in the United States. JAMA psychiatry 72, 1172–1181. 10.1001/jamapsychiatry.2015.1737 [PubMed: 26509694]

Overall JE, Gorham DR, 1988. The Brief Psychiatric Rating Scale (BPRS): recent developments in ascertainment and scaling.,,. Psychopharmacol Bull 24, 97–99.

Phalen PL, Rouhakhtar PR, Millman ZB, Thompson E, DeVylder J, Mittal V, Carter E, Reeves G, Schiffman J, 2018. Validity of a two-item screen for early psychosis. Psychiatry Res. 270, 861–868. 10.1016/j.psychres.2018.11.002 [PubMed: 30551336]

Picchioni MM, Murray RM, 2007. Schizophrenia. BMJ 335, 91–95. 10.1136/bmj.39227.616447.BE [PubMed: 17626963]

Rabinowitz J, Berardo CG, Bugarski-Kirola D, Marder S, 2013. Association of prominent positive and prominent negative symptoms and functional health, well-being, healthcare-related quality of life and family burden: a CATIE analysis. Schizophr. Res 150, 339–342. 10.1016/j.schres.2013.07.014 [PubMed: 23899997]

Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, Hergueta T, Baker R, Dunbar GC, 1998. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J. Clin. Psychiatry 59 Suppl 2, 22–57.

Srihari VH, Tek C, Pollard J, Zimmet S, Keat J, Cahill JD, Kucukgoncu S, Walsh BC, Li F, Gueorguieva R, Levine N, Mesholam-Gately RI, Friedman-Yakoobian M, Seidman LJ, Keshavan MS, McGlashan TH, Woods SW, 2014. Reducing the duration of untreated psychosis and its impact in the U.S.: the STEP-ED study. BMC Psychiatry 14, 335. 10.1186/s12888-014-0335-3 [PubMed: 25471062]

Tanskanen A, Tiihonen J, Taipale H, 2018. Mortality in schizophrenia: 30-year nationwide follow-up study. Acta Psychiatr. Scand. 138, 492–499. 10.1111/acps.12913 [PubMed: 29900527]

Wing JK, Babor T, Brugha T, Burke J, Cooper JE, Giel R, Jablenski A, Regier D, Sartorius N, 1990. SCAN. Schedules for Clinical Assessment in Neuropsychiatry. Arch. Gen. Psychiatry 47, 589–593. 10.1001/archpsyc.1990.01810180089012 [PubMed: 2190539]

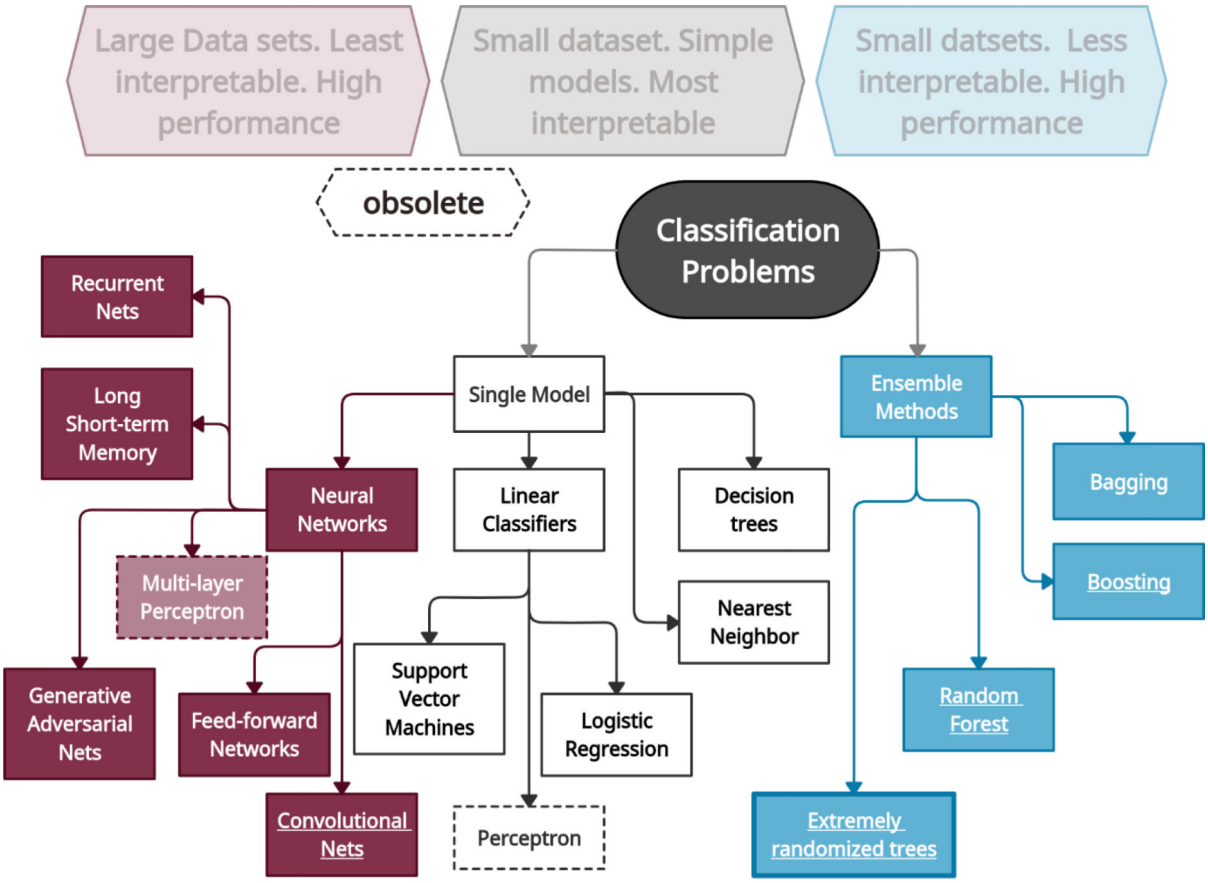

Figure 1:

A non-exhaustive taxonomy of major classification approaches in current machine learning practice. Large to very large datasets are typically better solved via Neural Networks and their many variants, while smaller datasets generally achieve high classification performance with ensemble methods, such as Random Forests, and general bagging and boosting algorithms. Maximum transparency or interpretability is obtained by simpler frameworks such as decision trees and support vector machines, which in general are superseded in performance by the ensemble methods (although counterexamples in specialized applications exist). Extremely randomized trees is similar to Random Forests in performance, have similar intuition behind its algorithm, and achieve further reductions in variance in some applications, which the present study is an example of.

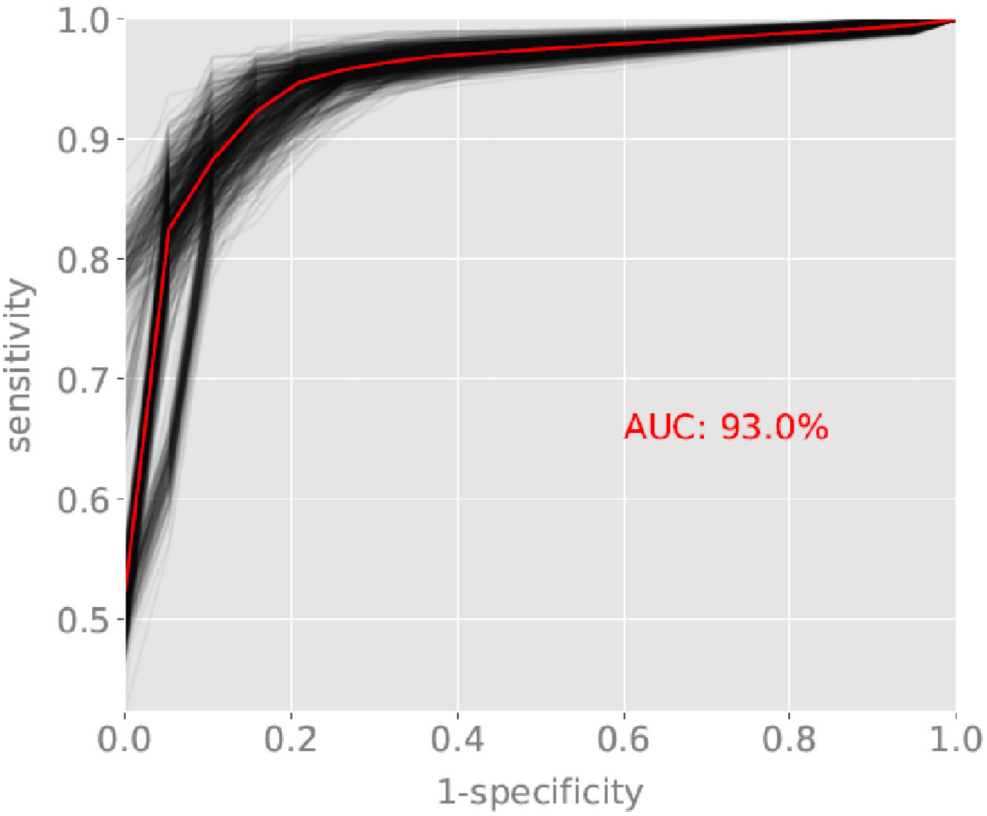

Figure 2a:

ROC curve for the classification of schizophrenia and schizoaffective disorder versus depression and bipolar disorder with psychotic features. Red line reflects the average ROC curve and AUC of 93.0%. Black lines represent sampling of individual ROC curves generated from random splits of the dataset into training and validation subsets.

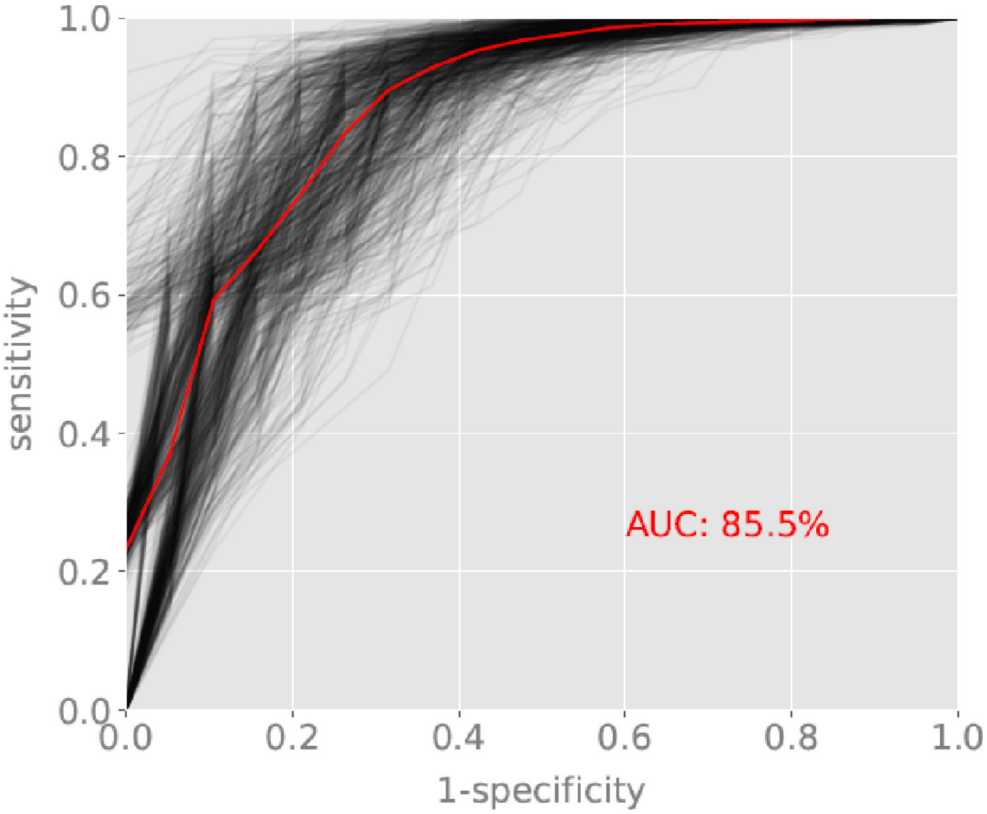

Figure 2b:
ROC curve for the classification of schizophrenia versus schizoaffective disorder. Red line reflects the average ROC curve and AUC of 85.5%. Black lines represent sampling of individual ROC curves generated from random splits of the dataset into training and validation subsets.

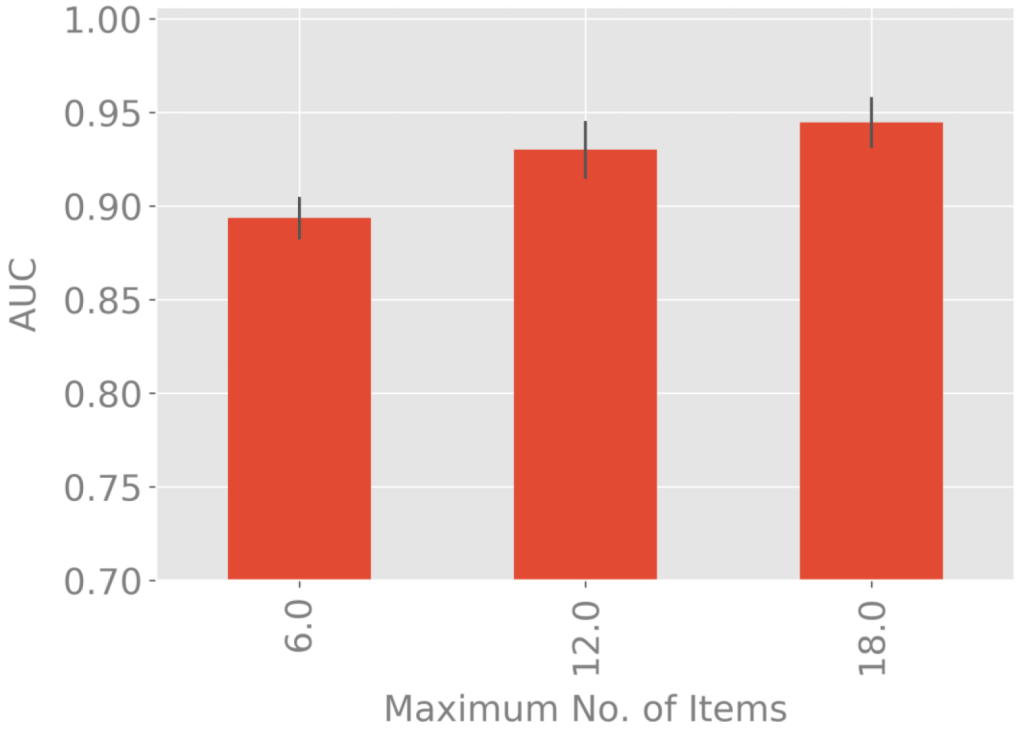

Figure 3a:
Variation of AUC with 99% confidence bounds with the maximum number of items used in the decision tree corresponding to each test for the classification of schizophrenia and schizoaffective disorder versus depression and bipolar disorder with psychotic features.

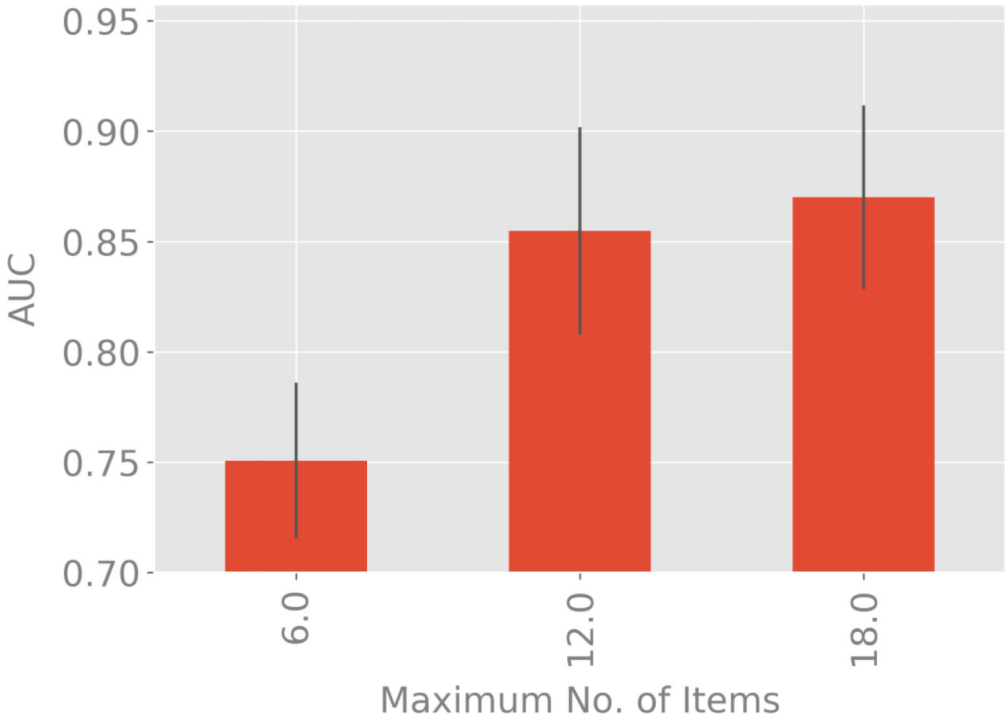

Figure 3b:
Variation of AUC with 99% confidence bounds with the maximum number of items used in the decision tree corresponding to each test for the classification of schizophrenia versus schizoaffective disorder

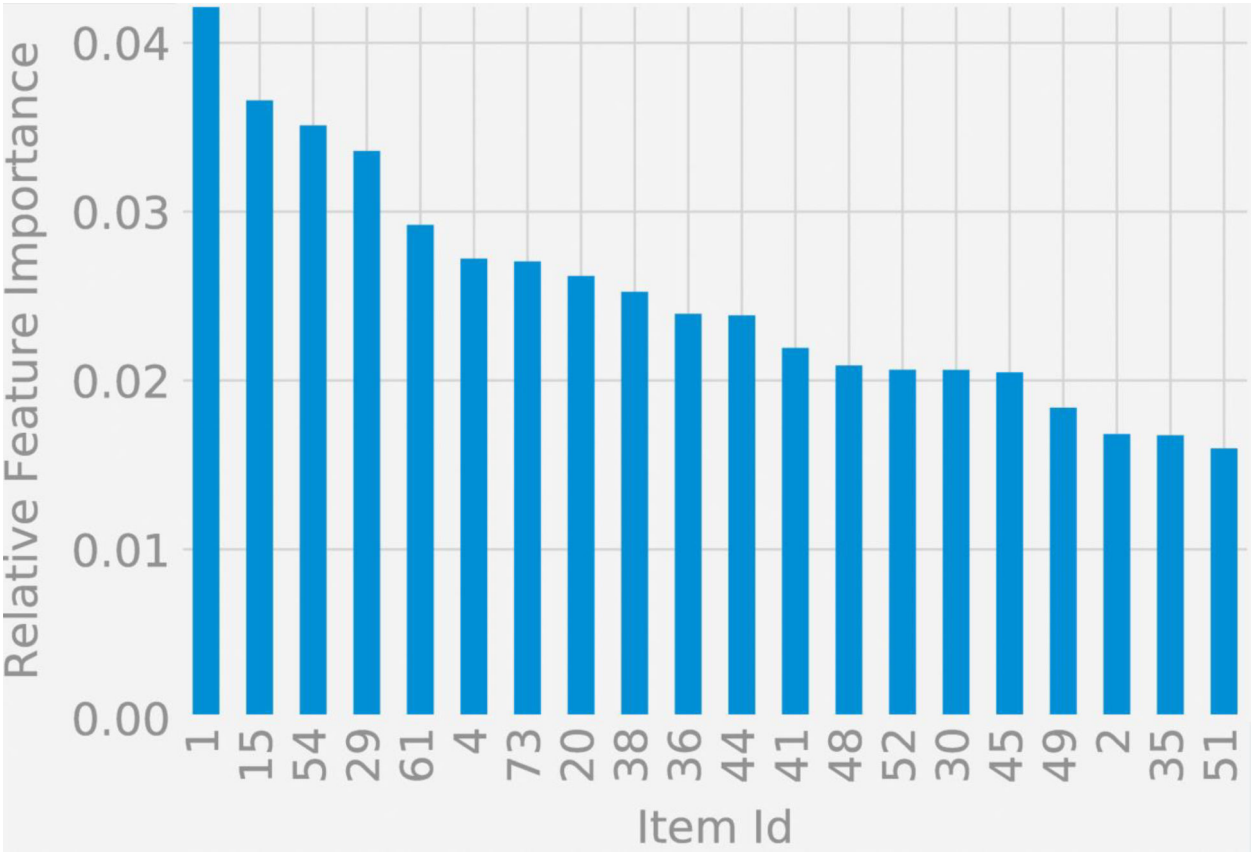

Figure 4a:
Relative feature importances for the classification of schizophrenia and schizoaffective disorder versus depression and bipolar disorder with psychotic features.

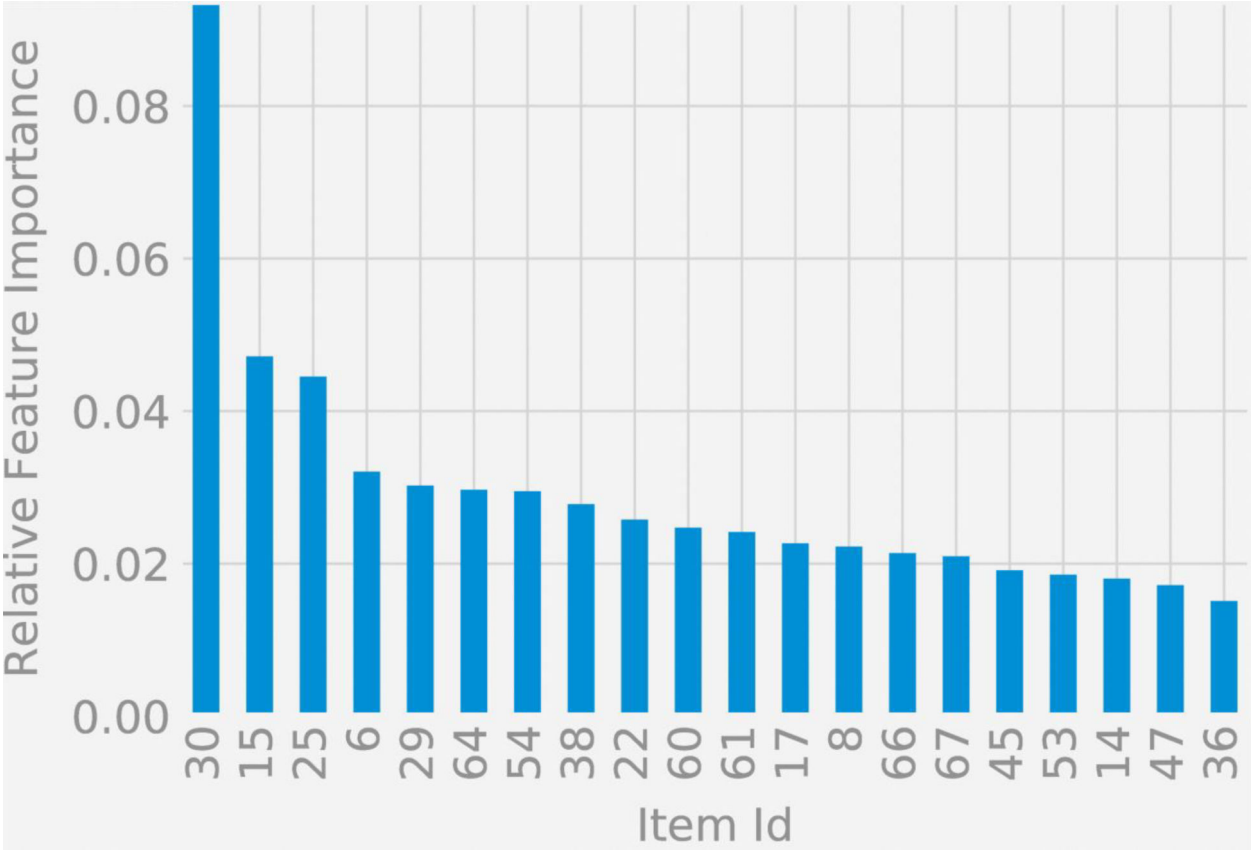

Figure 4b:
Relative feature importances for the classification of schizophrenia versus schizoaffective disorder