Diagnosis of Schizophrenia: A Comprehensive Evaluation

M. Tanveer, Jatin Jangir, M. A. Ganaie, Iman Beheshti, M. Tabish, and Nikunj Chhabra

Abstract—Machine learning models have been successfully employed in the diagnosis of Schizophrenia disease. The impact of classification models and the feature selection techniques on the diagnosis of Schizophrenia have not been evaluated. Here, we sought to access the performance of classification models along with different feature selection approaches on the structural magnetic resonance imaging data. The data consist of 72 subjects with Schizophrenia and 74 healthy control subjects. We evaluated different classification algorithms based on support vector machine (SVM), random forest, kernel ridge regression and randomized neural networks. Moreover, we evaluated T-Test, Receiver Operator Characteristics (ROC), Wilcoxon, entropy, Bhattacharyya, Minimum Redundancy Maximum Relevance (MRMR) and Neighbourhood Component Analysis (NCA) as the feature selection techniques. Based on the evaluation, SVM based models with Gaussian kernel proved better compared to other classification models and Wilcoxon feature selection emerged as the best feature selection approach. Moreover, in terms of data modality the performance on integration of the grey matter and white matter proved better compared to the performance on the grey and white matter individually. Our evaluation showed that classification algorithms along with the feature selection approaches impact the diagnosis of Schizophrenia disease. This indicates that proper selection of the features and the classification models can improve the diagnosis of Schizophrenia.

Index Terms—Classification, machine learning, Schizophrenia.

I. INTRODUCTION

SCHIZOPHRENIA is a severe mental disorder that affects millions of people worldwide. Schizophrenia makes people slowly lose contact with reality, leading to hallucinations, delusions, and extremely disordered thinking. Patients report hearing voices or seeing things that are not there; they also tend to develop fixed and false beliefs. Suicidal tendency is also a common trait among Schizophrenia patients. Moreover, patients inflicted with Schizophrenia are 2–3 times more likely to die than the general public due to patients not seeking aid for preventable physical diseases [1]. Luckily, Schizophrenia is treatable with medicines and psycho-social support, and these methods have proven successful [2]. Thus, the central blockade in eradicating Schizophrenia is lack of its early detection.

Several attempts have been made to remedy this problem. Many studies show promising results but, Machine Learning (ML) has seen little use in clinical practice for Schizophrenia. This can perhaps be credited to the unreliability or stability of some machine learning models; this influences the consensus that ML solutions are not dependable and cannot be trusted, especially for a medical job. Even though doctors sometimes make mistakes themselves, people still trust them. Developing this same level of trust for a machine would be an arduous task. Machine learning should be seen as a research tool to advance the field of study, not the be-all and end-all. In the past few years, extensive research has been done on various classification algorithms and their improved versions have been proposed like for SVM, it’s extensions like twin support vector machine (TWSVM) [3], twin bounded SVM (TBSVM) [4] etc., are proposed to improve the performance of SVM algorithm. Other methods such as k-nearest neighbour (KNN) [5], random forest (RaF) [6] have also been thoroughly studied. Interested readers can refer to the comprehensive review on TWSVM [7].

Our intent with this study is to perform a comprehensive evaluation and bring awareness towards using different classification algorithms and their variants and extensions for diagnosing schizophrenia disease. In this study, we evaluate single-modal methods using exclusively Structural MRI (sMRI) scans to train and validate them. We use the same dataset for all the algorithms. The results of this study will help choose a suitable classification algorithm and feature selection technique based on the requirement. The rest of the paper is organized as follows: In Section III, we discuss about subjects, 3D MRI processing and give a brief description of various classification algorithms and feature selection techniques used and also about validation.
and experimental setup. Performance of various classification algorithms on white matter, grey matter and integrated matter is discussed in Section IV. Analysis and summarizing of results is done in Section V. Conclusions and future works are discussed in Section VI. Please note, Figures and Tables referenced from the attached Supplementary Paper have been suffixed with an “S-”.

II. RELATED WORKS

The current diagnosis scheme for Schizophrenia is to rule out other mental disorders and then employ psychiatric and physical screening. The studies [8], [9] evaluated magnetic resonance imaging (MRI) scans for the detection of Schizophrenia. A review of MRI findings in schizophrenia [8] discusses brain abnormalities due to Schizophrenia. Thus, there is significant evidence that by using MRI scans, one can exploit ML techniques to automate and improve the detection of Schizophrenia. Several studies [10], [11] have already attempted to do so with varying degree of success. A basic summary for most of the studies is: process the MRI scan into a usable format, apply a feature extraction algorithm on the MRI scan to select the appropriate features, then finally use a classification algorithm for the diagnosis of schizophrenia. The classification of schizophrenia patients and healthy controls from sMRI scans in two large independent samples was studied in [12]. The authors used whole-brain grey matter densities from MRI scans with SVM as the classifier and concluded that SVM models trained with less than 130 samples results in an unstable model. The key difference of the study [12] from previous similar studies [10], [13], was utilizing a large dataset and using an entirely separate dataset to perform the validation. Additionally, noting that typical schizophrenia medications affect the striatum (part of the brain), they masked it out, ensuring the model doesn’t relate medication effects to Schizophrenia detection. Each imaging technique provides a different view of the brain functioning. To get the benefit of different imaging techniques, a multimodal classification model [14] combined 3 different data types: resting state functional MRI (rs-fMRI), Diffusion tensor imaging (DTI) and sMRI. While the idea proposed in [14] don’t have the main focus on classification but to design and evaluate a multivariate method which can find cross-information in more than two data types. In [15], multi-set canonical correlation analysis (MCCA) was used to combine the data of rs-fMRI, Electroencephalogram (EEG) and sMRI and proposed ensemble feature selection approach which resulted in very high prediction performance approaching 100% by utilizing the additional modalities. Though they also concluded that combining multiple modalities does not always result in an enhanced result. A similar study [16] which used rs-fMRI and sMRI with a similar outcome of increased accuracy when compared to single modalities. It can easily be realized what the major downfall of this multi-modal training scheme is: lack of data. Some datasets combined from various sources reach the 250 marks, like the one used in [12] or in but most datasets sit at 80 samples.

Gaining insight from the previous studies about the feasibility and reliability of individual classification based on the sMRI, a novel machine-learning(5,7),(995,993)
whole-brain data (i.e., total of 29,852 voxel values per modality) as raw features for classification tasks.

C. Classification Algorithms

The classification algorithms evaluated in this study for schizophrenia are explained below. Detailed information on the algorithms is available in the Supplementary file.

1) Random Forest (RaF) [6]: Random Forests proposed by Leo Breiman et al. in 2001 is a collection of tree predictors (also called tree-structured classifiers, which at their core are nested if-else statements used to vote for classes) where each tree is generated via independent and identically distributed random vectors. It has been shown that a sufficiently large forest always converges and a forest generated using random features generally produces better accuracy than a single tree classifier. RaF combined both the concepts of bagging and random subspace which improved its generalisation performance.

2) Oblique RaF (MPRaF-T, MPRaF-P and MPRaF-N) [20]: Oblique RaF was proposed to handle multiclass classification with an improved geometric property. Multisurface Proximal Support Vector Machine (MPSVM) [21] is used to generate clustering hyperplanes at the non-terminal nodes of a decision tree. Now, RaF is implemented using the MPSVM-based decision trees and then subsequently using various regularisation methods. It was shown that Oblique RaF performs better than RaF and have significantly less variance and bias. MPRaF-T, MPRaF-P, and MPRaF-N represent the MPSVM-based RaFs with Tikhonov, axis-parallel, and NULL space regularization, respectively.

3) Heterogeneous RaF [22]: As noted by [20], RaF’s data splitting leads to axis-parallel decision boundaries, which can lead to poor utilization of the geometric property of the data. But [22] noted that even though Oblique RaF allows for oblique splits, it is sub-optimal. The Heterogeneous RaF uses diverse linear-classifiers at the tree’s nodes and searches for the best split at every node by optimizing the impurity criteria. Heterogeneous RaF Forests are shallower and faster to train than RaF. For the decision trees, each split is rated based on impurity criterion. All the splits at each non-leaf nodes are linked with an impurity measure. The one which is having the maximum value is the selected split for that particular node. The six different classifiers which have been employed are SVM, MPSVM, Linear Discriminant Analysis (LDA), Least Squares SVM (LSSVM), Ridge Regression (RR) and Logistic Regression (LR) as they have performed well in several domains [23].

4) Kernel Ridge Regression (KRR) [24]: One of the kernel-based methods is the Kernel Ridge Regression (KRR). The KRR has a closed-form solution which lends it to faster training. Despite being relatively straightforward than other members of kernel-based methods such as SVM, it can produce comparable results. The kernel ridge regression method is based on Ridge Regression and Ordinary Least Squares.

5) K Nearest Neighbours (KNN) [5]: K nearest neighbours algorithm assigns the label depending upon the similarity of the point with its neighbours. A constant $K$ is first chosen for the algorithm. The Euclidean distance of the given point is calculated and the $K$ nearest members are selected from it. The number of data points is counted and new data points are assigned to the category for which there are maximum number of neighbours. The number of nearest neighbours, $K$, in our case is 5.

6) Neural Networks [25]: Neural networks are network of node layers comprising of an input layer, multiple hidden layers and an output layer. Each layer has multiple number of nodes and the nodes of each layer are interconnected with the other layers. The output of each of the layer is calculated through an activation function and the output of activation layer of the last layer is the final output. Adam optimization technique [25] has been used in order to tune the parameters.

7) Random Vector Functional Link Network (RVFL) [26]: RVFL is the randomized version of the functional link neural network. It shows that from the input layer to the hidden layer, the value of weights can be generated randomly in a suitable domain and fixed in the learning stage. The closed-form based RVFL obtains the output weights in a single-step and exhibits a higher efficiency than the iterative method.

8) Random Vector Functional Link Network With Auto Encoder (RVFLAE) [27]: Autoencoder is an unsupervised learning model for which the output and input layers share the same neurons in order to reconstruct its own inputs. In this method, we adopt a sparse autoencoder to learn appropriate network parameters of RVFL, which are developed via $l_1$ norm optimization instead of the usual $l_2$ norm retaining more informative features.

9) Support Vector Machine (SVM) [28]: SVM is a binary classification algorithm which classify the labelled data in two classes, SVM generates an optimal hyperplane using data to separate the classes. Since there may be more than one hyperplane possible for that, SVM finds the optimal hyperplane to do the classification by solving a Quadratic Programming Problem (QPP). Thus, a new data point can be classified based on the optimal hyperplane formed by SVM. And if the given data is not linearly separable, SVM do the task using kernel method.

10) Twin Support Vector Machine (TWSVM) [3]: Inspired from SVM, TWSVM is a classification algorithm which classifies the given labelled data into two classes by generating two non-parallel hyperplanes. TWSVM solves two smaller sized QPPs, unlike SVM, each for one class. The two required planes are formed by solving a problem which minimize the distance of points of corresponding class to the plane and keep it as far as possible from another class. Then, a new data point is assigned a class by calculating its distance from the planes.

11) Twin Bounded Support Vector Machine (TBSVM) [4]: TBSVM is an improved version of TWSVM which modified the optimizations problem in TWSVM to give better performance and thus making the classification more accurate. It added an extra regularization term in the formulation of TWSVM which applied the structural risk minimization principle in the model.

12) Least Square Twin Support Vector Machine (LSTSVm) [29]: LSTSVm is the least squares version of TWSVM. The formulation of LSTSVm leads to fast and simple algorithm to generate the two non-parallel hyperplanes for binary classification. The two primal problems used to find the required hyperplanes are formulated in least squares sense
i.e. using the equality constraints instead of the inequality constraints. The problem in LST SVM can be solved very easily and simply by solving a system of two linear equations.

13) Robust Energy Based Least Square Twin Support Vector Machine (RELST SVM) [30]: Robust energy based LST SVM, proposed by Tanveer et al., is another extension of TWSVM which adds a maximum margin regularization term in primal problem and moreover uses an energy parameter in the constraints, which helped in lessening the effect of noise in the data. According to a recent study [31], RELST SVM model leads to better classification performance among the TWSVM models.

14) Pinball General Twin Support Vector Machine (PinGTSVM) [32]: Pinball general twin support vector machine also generates non-parallel hyperplane for classification, similar to TWSVM, but uses pinball loss function in place of hinge loss without affecting the computational complexity of the algorithm. The use of pinball loss function makes it less sensitive to noise in classification of data and make it more stable for re-sampling of data.

D. Feature Selection Methods

The feature selection methods evaluated in this study for schizophrenia are explained below:

1) RankFeatures() Function and Its Various Criterions: The rankfeatures() is a MATLAB [33] function which ranks key features by class separability criterion. It uses various independent evaluation criteria to assess the significance of features. The criterion here refers to an objective function that minimises the overall feasible feature subset. The rankfeatures() uses the following feature independent criteria:

a) T-test [34]: The “ttest” is the default, criteria used by the rankfeatures() function. The T-test ranks the features based on an absolute value, two-sample T-test with pooled variance estimate. The “ttest” criterion assumes that the classes are normally distributed.

b) Entropy [35]: The “entropy” criterion uses the relative entropy, also known as Kullback–Liebler distance or divergence. The “entropy” criterion assumes that the classes are normally distributed.

c) Bhattacharyya [34]: The “Bhattacharyya” criterion uses the minimum attainable classification error, or Chernoff bound to rank features. The “Bhattacharyya” criterion assumes that the classes are normally distributed.

d) ROC [34]: The “ROC” criterion uses the area between the empirical receiver operating characteristic (ROC) curve and the random classifier slope to rank features. The “ROC” criterion is a non-parametric test.

e) Wilcoxon [36]: The “Wilcoxon” criterion uses the absolute value of the standardised U-statistic of a two-sample unpaired Wilcoxon test, also known as Mann-Whitney, to rank features. The “Wilcoxon” criterion is also a non-parametric test.

2) Minimum Redundancy Maximum Relevance (MRMR) algorithm [37]: The MRMR algorithm is a sequential feature selection method that finds an optimal set of mutually and maximally dissimilar features. The MRMR performs this by maximising the relevance of the feature set to the response variable and minimising the redundancy of a feature set.

3) Neighborhood Component Analysis (NCA) [38]: NCA is a non-parametric feature selection method used explicitly for regression and classification algorithms. The feature weights (importance of a feature) are obtained using a gradient ascent technique to maximise the expected leave-one-out classification accuracy with a regularisation term.

E. Validation, Experimental Setup

This study used Matlab R2021a [33] to implement all the required code for the different methods. The functions used were (but not limited to): rankfeatures(), fscmrmr(), fscnca(). In all experiments, 10-fold cross-validation was used. To study the variation of accuracy with increasing number of features and to obtain the minimum or the optimal number of features, we experimented with 100–1300 (with a step size of 100) selected features. The various hyper parameter ranges used for various methods have been tabulated in Table I. The classification accuracies corresponding to different classification models versus feature selection approaches for the combined matter at 500 features are available in Table II. The results of 500 features are presented as maximum accuracy for Integrated GM and WM occurs at the same. The Tables corresponding to WM and GM are available in the Supplementary file as Table S-2 and S-3.

IV. RESULTS

The performance of the classification models varies with different feature extraction methods and the number of features selected. We discuss the performance of the models with GM, WM and the integration of GM and WM data. When discussing classification models, accuracy means the average accuracy across both feature extraction techniques and the feature range, unless otherwise stated. Similarly, the accuracy assigned to a given feature extraction technique represent the average accuracy across the different classification models. Tables of other metrics i.e. AUC, Sensitivity, Specificity, Precision, F-Measure and G-Mean are available in the Supplementary file.

1) White Matter: Across the entire feature range (i.e. 100–1300), the TWSVM based classification models showed the highest average accuracy compared to the rest of the classification models, as can be seen in Fig. S-1a. The rest

| Parameter Name                  | Symbol | Range/Value          |
|--------------------------------|--------|----------------------|
| Penalty parameters             | $c_i^{1}$ | $\{10^{0} | i = 5 : 5\}$ |
| Non-linear kernel parameter    | $\gamma$ | $\{2^{|i| = -10 : 10\}$ |
| NCA regularisation term        | $\lambda$ | $\{2^{|i/N^{|i| = 1 : 20\}$ |
| RELST SVM parameter            | $\rho$ | $\{0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ |
| RVFL & RVFL-AE parameters     | $C$ | $\{-5 : 1 : 14\}$ |
|                              | $N$ | $\{3 : 20 : 203\}$ |
| Ensemble size of the trees    |        | 100                  |
| for RaP methods               |        |                      |
| pinGTSVM parameter            | $\epsilon$ | 0.05                |

1) $i = 1, 2, 3, \ldots$, Number of samples.
of the families, i.e. RaF, Neural Networks, KNN and KRR, stay together through the feature range. Among SVM-based classifiers, the non-linear RELSTSVM (75.16%) and non-linear TBSVM (74.30%) achieved the highest average accuracy. Non-linear TWSVM followed them with 73.24% accuracy, closely followed by non-linear LSTSVM at 72.47%. The linear TBSVM and RELSTSVM showed ~70% accuracy. The lowest-performing models are pinGTSVM and RVFLAE with ~60.7% accuracy. Heterogeneous-RaF achieves the maximum accuracy of 84.04% for WM with 900 features selected using Wilcoxon feature selection.

Among the RaF based models, Heterogeneous-RaF shows the best performance, with an average accuracy of 67.70%. In contrast, MPRAf-T shows the lowest average accuracy with 62.93%. Among the variants of neural networks, the standard neural network and randomized based neural network show ~65% average accuracy. The RVFLAE model showed the lowest average accuracy (59.06%) among the neural network models. Also, the non-linear kernel-based KRR model (accuracy 69%) is better than the linear kernel-based KRR model (accuracy 63.67%) in terms of average accuracy.

Discussing the feature selection methods we can refer to S-2b, the Wilcoxon is the best choice across the entire feature range with an average accuracy of 75.21%. The NCA again is very unstable but performs very well with an average accuracy of 68.08%. Entropy and Bhattacharya both perform terribly at lower features (dipping as low as 52%) but approach the best performing Wilcoxon at higher features.

Wilcoxon feature selection performs better than all other methods for the entire feature range. Thus, we can compare classifiers based on Wilcoxon feature selection. When comparing w.r.t. Wilcoxon, the linear RELSTSVM (78.85%) perform just marginally lower than the non-linear variation (79.13%). The neural network (78.28%) performs significantly better than when using the average, but it is not at the top.

3) Integrated GM and WM: The combined matter (i.e. integrated GM and WM data) achieves better results than individual grey matter and white matter. From Fig. 1, it can be inferred that the SVM based classifiers perform significantly better than the rest of the families for the entire feature range. The RaF based methods come next, followed by KNN and KRR. Neural networks again perform the worst. Non-linear TBSVM and non-linear RELSTSVM achieved the highest 78.47% and 77.11% average accuracy, respectively, followed by non-linear
Fig. 1. Average performance of classifiers families w.r.t Integrated WM and GM (Combined matter).

Fig. 2. Average performance of feature selection methods w.r.t Integrated WM and GM (Combined matter).

Fig. 3. Performance of linear and non-linear kernels for combined matter.

TWSVM (77.62%) non-linear LSTSM (76.32%). The lowest-performing models are RVFLAE, pinGTSVM and KNN with ~66% accuracy. The standard neural network achieves the maximum accuracy of 86.71% for combined matter with 500 features selected using Wilcoxon feature selection.

Among the RaF based models, the Heterogeneous-RaF method shows the best performance, with an average accuracy of 73.31%. In contrast, MPRaF-T shows the lowest average accuracy with 68.92%. Among the variants of neural networks, the standard neural network and randomized based neural network show 73% and 70.77% average accuracy, respectively. The RVFLAE model showed the lowest average accuracy (65.85%) among the neural network models. Also, the non-linear kernel-based KRR model (accuracy 72.13%) is better than the linear kernel-based KRR model (accuracy 71.15%) in terms of average accuracy.

Discussing the feature selection methods we can refer to Fig. 2, the Wilcoxon is the best choice across the entire feature range with an average accuracy of 77.12%. The NCA is unstable and performs poorly with an average accuracy of 69.78%. Entropy and Bhattacharya both start out being the worst performing methods at lower features (dipping as low as 60%) but approach the best performing Wilcoxon at higher features. ROC and T-test perform better than MRMR, which is at 70.71%. ROC and T-test achieve an average accuracy of ~72%.

Wilcoxon feature selection performs better than all other methods for the entire feature range. Thus, we can compare classifiers based on Wilcoxon feature selection. When comparing w.r.t. Wilcoxon, the standard neural network comes out on top (83.98%), becoming the best classifier. The Heterogeneous-RaF becomes the second-best classifier at 82.36%. The linear TBSVM (80.61%) and linear RELSTSVM (79.75%) performs slightly better than non-linear TBSVM (80.56%) and non-linear RELSTSVM (79.62%).

V. DISCUSSION

This study aimed to perform a comprehensive evaluation of modern classification techniques and feature selection methods for schizophrenia classification. We have presented a basic overview of the different classification methods and evaluated them against different feature selection approaches on the same dataset. Selection of the optimal features and using the best available classification technique is essential for an MRI-based machine learning system aimed at early diagnosis. Our findings indicate that using twin SVM-based methods such as RELSTSVM or TBSVM performs best for nearly all matter types. The random forest based methods generally perform mediocrely. The worst performing classification models are the pinGTSVM, RVFL-AE, KRR, KNN and MPRaF-T.

Fig. 3 is constructed by averaging only the methods that have both linear and non-linear variants (i.e. SVM, KRR, TWSVM, TBSVM, LSTWSVM and RELSTSVM). Inferring from Fig. 3, a critical observation can be made for the performance of non-linear kernel functions against linear ones. In nearly all studied methods (pinGTSVM and standard SVM being the exception), the non-linear kernel function performs better than the linear variation. This observation follows suit with the result that a linear kernel is the degenerate version of the non-linear (RBF or Gaussian) kernel [39]. Thus, an adequately tuned non-linear kernel consistently out-performs the linear kernel. But, an observation can be made that with an increasing number of features, the advantage of using a non-linear kernel diminishes. This diminished performance can be attributed to the fact that at a
higher number of features, one may not need to map features to a higher dimension [40]. Thus, the time penalty to tune the kernel function (in case of non-linear) becomes outweighed by the rapid computation of the linear kernel when both kernels provide relatively similar accuracy. Thus, with a large number of features, one is better off using a linear kernel.

The feature selection methods significantly impact the performance of classification models, as one might expect. The Wilcoxon was the all-around best feature selection method, performing the best for all the different matter types and across the entire feature range. This observation is supported by previous studies [41] and [42]. The exceptional performance of Wilcoxon is especially prevalent when sample sizes are small or the data doesn’t resemble Normal distribution. Entropy and Bhattacharya are fascinating methods. At lower feature numbers, they perform equally terribly, but at a higher number of features, they approach the best performance. This behaviour can be seen in Fig. S-2b and Fig. 2, especially in integrated and grey matter. The MRMR, ROC and T-Test all perform mediocre, varying based on what matter type is used. At present, using our specific dataset, our findings indicate that we can classify schizophrenia patients with a maximum of 86.71% accuracy when using a standard neural network with 500 features from combined matter, selected using Wilcoxon. The main advantages of twin SVM based models like TWSVM, TBSVM, RELST SVM, LST SVM etc. over standard SVM is that they give competitive performance in terms of accuracy and reduce the computational complexity of SVM because these models generate two non-parallel hyperplanes instead of a single hyperplane in SVM which leads to solving two smaller sized Quadratic Programming Problems (QPPs) instead of one larger QPP in SVM. The paper [31] also concludes that twin SVM based models performs better than other family of classifiers.

Referring to Fig. 4, our results strongly suggest that using both grey matter (GM) and white matter (WM), i.e. integrated matter, leads to improved performance for the classification of schizophrenia patients. The GM performs better than WM after reaching a threshold number of features (in our case 700), but the results shoot up by a substantial margin (~4%) when the combined matter is used. It can be seen from Fig. 4 that this is the case for all the evaluated classification techniques and that there are no exceptions for this observation.

VI. CONCLUSION

In this study, we comprehensively evaluated various classification models to identify the best available machine learning model for the classification of schizophrenia subjects. We assessed 24 classification models involving the variants of support vector machines, twin support vector machines, random forest, Kernel ridge regression and neural networks. Additionally, we evaluated 7 feature selection methods: Wilcoxon, MRMR, ROC, Entropy, T-Test, Bhattacharyya and NCA. Moreover, these evaluations were conducted on the features based on grey matter (GM), white matter (WM) and the integrated GM and WM data.

The contributions from this paper are four-fold. First, we underlined how different families of machine learning algorithms perform with the schizophrenia dataset. We found that, for the most part, the Non-linear twin SVM-based family of classifiers outperform all other classifiers. This family includes (in the order of gradually worsening performance) RELST SVM, TBSVM, TWSVM and LST SVM. However, the pinGTSVM is the worst-performing family member, ranking the lowest across all the classifiers. On the other hand, the KNN, KRR, MPRaF-T and non-linear SVM are the lowest performers. Most RaF-based methods occupy the middle of the spectrum. An additional observation is that, for the most part, the non-linear variant of a method outperforms the linear variation.

The second contribution is that we evaluated the performance of different feature selection methods. Our results indicate that Wilcoxon is the best performing methods with a top rank across all the matter types. Entropy and Bhattacharyya have improved performance with an increasing number of features. NCA is an unstable method, although it had a good average performance in GM. T-Test, ROC, MRMR are also reasonable choices for feature selection, but they are not recommended.

Third, we found that utilising both grey and white matter for classification yields better results than any individual matter type. In conclusion, the feature selection method, the number of features selected and the classification model should be appropriately chosen for better generalisation performance on the classification of schizophrenia subjects. This study recommends using standard neural network, RELST SVM, TWSVM, or heterogeneous-RaF as the classification model with ~700–1200 features selected via Wilcoxon feature selection method for better generalisation performance on the classification of schizophrenia datasets. We hope that the evaluation presented in this paper encourages future research to use better classification algorithms and feature selection algorithms for clinical dataset classification.

New developments in machine learning are rapid and can improve the results of previous algorithms by a significant margin. In the future, much scope remains for the development of better specific models. Therefore, these new variants or methods need to be tested on real-life datasets such as schizophrenia to grasp their viability. In the future, one can extend this study by various margins, i.e. (1) the dataset can be enhanced to utilise data from multiple sources; thus, it should be evaluated if combining data from various sources (i.e. MRI images with varying scanning parameters such as slice thickness, field of view, bandwidth,
repetition time, etc.) leads to better generalisation or if it results in a worse performance. (2) This study utilised a single feature extraction technique (i.e. DARTEL); thus, future research for the effect of different feature extraction techniques needs to be conducted. (3) This study used a single modality (i.e. sMRI) and usage of different modalities including the Functional MRI (fMRI), Electroencephalogram (EEG) should also be evaluated using a similar setup in future studies. The source code will be available at https://github.com/mtanveer1.

ACKNOWLEDGMENT

The authors would like to thank the Indian Institute of Technology Indore for providing the required facilities and support for this work.

REFERENCES

[1] T. M. Laursen, M. Nordentoft, and P. B. Mortensen, “Excess early mortality in schizophrenia,” *Ann. Rev. Clin. Psychol.*, vol. 10, pp. 423–448, 2014.
[2] “World Health Organization on schizophrenia.” [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/schizophrenia
[3] “Combination of FMRI-SMRI-EEG data improves discrimination in individuals with schizophrenia and those at high genetic risk,” *Schizophrenia Res.*, vol. 169, no. 1, pp. 76–82, 2015.
[4] Y. Guo, J. Qiu, and W. Lu, “Support vector machine-based schizophrenia classification using morphological information from amygdaloid and hippocampal subregions,” *Brain Sci.*, vol. 10, no. 8, 2020, Art. no. 562.
[5] L. Zhang and P. N. Suganthan, “Oblique decision tree ensemble via multisurface proximal support vector machine,” *IEEE Trans. Cybern.*, vol. 45, no. 10, pp. 2165–2176, Oct. 2015.
[6] L. Mangasarian and E. W. Wild, “Multisurface proximal support vector machine classification via generalized eigenvalues,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 69–74, Jan. 2006.
[7] R. Katuraw, P. N. Suganthan, and L. Zhang, “Heterogeneous oblique random forest,” *Pattern Recogn.*, vol. 99, 2020, Art. no. 107078.
[8] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, “Do we need hundreds of classifiers to solve real world classification problems?” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133–3181, 2014.
[9] R. Katuraw and P. N. Suganthan, “An ensemble of kernel ridge regression for multi-class classification,” *Procedia Comput. Sci.*, vol. 108, pp. 375–383, 2017.
[10] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization.” [arXiv:1412.6980].
[11] L. Zhang and P. N. Suganthan, “A comprehensive evaluation of random vector functional link networks,” *Inf. Sci.*, vol. 367, pp. 1094–1105, 2016.
[12] Y. Zhang, J. Wu, Z. Cai, B. Du, and S. Y. Philip, “An unsupervised parameter learning model for RVFL neural network,” *Neural Netw.*, vol. 112, pp. 85–97, 2019.
[13] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[14] M. A. Kumar and M. Gopal, “Least squares twin support vector machines for pattern classification,” *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7535–7543, 2009.
[15] M. Tanveer, M. A. Khan, and S.-S. Ho, “Robust energy-based least squares twin support vector machines,” *Appl. Intell.*, vol. 45, no. 1, pp. 174–186, 2016.
[16] M. Tanveer, C. Gautam, and P. N. Suganthan, “Comprehensive evaluation of twin SVM based classifiers on UCI datasets,” *Appl. Soft. Comput.*, vol. 83, 2019, Art. no. 105617.
[17] M. Tanveer, A. Sharma, and P. N. Suganthan, “General twin support vector machine with pinball loss function,” *Inf. Sci.*, vol. 494, pp. 311–327, 2019.
[18] X. Li et al., “Subcortical structure alterations impact language processing in individuals with schizophrenia and those at high genetic risk,” *Schizophrenia Res.*, vol. 169, no. 1, pp. 76–82, 2015.
[19] Y. Guo, J. Qiu, and W. Lu, “Support vector machine-based schizophrenia classification using morphological information from amygdaloid and hippocampal subregions,” *Brain Sci.*, vol. 10, no. 8, 2020, Art. no. 562.
[20] L. Zhang and P. N. Suganthan, “An ensemble of kernel ridge regression for multi-class classification,” *Procedia Comput. Sci.*, vol. 108, pp. 375–383, 2017.
[21] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization.” [arXiv:1412.6980].
[22] L. Zhang and P. N. Suganthan, “A comprehensive evaluation of random vector functional link networks,” *Inf. Sci.*, vol. 367, pp. 1094–1105, 2016.
[23] Y. Zhang, J. Wu, Z. Cai, B. Du, and S. Y. Philip, “An unsupervised parameter learning model for RVFL neural network,” *Neural Netw.*, vol. 112, pp. 85–97, 2019.
[24] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[25] M. A. Kumar and M. Gopal, “Least squares twin support vector machines for pattern classification,” *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7535–7543, 2009.
[26] M. Tanveer, M. A. Khan, and S.-S. Ho, “Robust energy-based least squares twin support vector machines,” *Appl. Intell.*, vol. 45, no. 1, pp. 174–186, 2016.
[27] M. Tanveer, C. Gautam, and P. N. Suganthan, “Comprehensive evaluation of twin SVM based classifiers on UCI datasets,” *Appl. Soft. Comput.*, vol. 83, 2019, Art. no. 105617.
[28] M. Tanveer, A. Sharma, and P. N. Suganthan, “General twin support vector machine with pinball loss function,” *Inf. Sci.*, vol. 494, pp. 311–327, 2019.
[29] MATLAB, 9.10.0.1602886 (R2021a). Natick, MA, USA: The MathWorks Inc., 2021.
[30] S. Theodoridis and K. Koutroubamas, “Feature Selection,” in *Pattern Recognition*, 4th ed. S. Theodoridis and K. Koutroubamas, Eds., Boston, MA, USA: Academic Press, 2009, ch. 5, pp. 261–322.
[31] S. Kullback, *Information Theory and Statistics*, New York, NY, USA: Dover Publications, 1997.
[32] F. Wilcoxon, “Individual comparisons by ranking methods,” in *Breakthroughs in Statistics*, New York, NY, USA: Springer, 1992, pp. 196–202.
[33] C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” *J. Bioinf. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, 2005.
[34] W. Yang, K. Wang, and W. Zuo, “Neighborhood component feature selection for high-dimensional data,” *James Cash Penney*, vol. 7, no. 1, pp. 161–168, 2012.
[35] S. S. Keerthi and C.-J. Lin, “Asymptotic behaviors of support vector machines with Gaussian kernel,” *Neural Comput.*, vol. 15, no. 7, pp. 1667–1689, 2003.
[36] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, “A practical guide to support vector classification,” Taipei, 2003.
[37] C.-J. Tian, J. Lv, and X.-F. Xu, “Evaluation of feature selection methods for mammographic breast cancer diagnosis in a unified framework,” *BioMed Res. Int.*, vol. 2021, 2021.
[38] C. Liao, S. Li, and Z. Luo, “Gene selection using Wilcoxon rank sum test and support vector machine for cancer classification,” in *Proc. Int. Conf. Comput. Inf. Sci.*, 2006, pp. 57–66.