Identification of Children At Risk of Schizophrenia via Deep Learning and EEG Responses

David Ahmedt-Aristizabal, Tharindu Fernando, Simon Denman, Jonathan E. Robinson, Sridha Sridharan
Patrick J. Johnston, Kristin R. Laurens, Clinton Fookes

***Abstract*—The prospective identification of children likely to develop schizophrenia is a vital tool to support early interventions that can mitigate the risk of progression to clinical psychosis. Electroencephalographic (EEG) patterns from brain activity and deep learning techniques are valuable resources in achieving this identification. We propose automated techniques that can process raw EEG waveforms to identify children who may have an increased risk of schizophrenia compared to typically developing children. We also analyse abnormal features that remain during developmental follow-up over a period of $\sim$4 years in children with a vulnerability to schizophrenia initially assessed when aged 9 to 12 years. EEG data from participants were captured during the recording of a passive auditory oddball paradigm. We undertake a holistic study to identify brain abnormalities, first by exploring traditional machine learning algorithms using classification methods applied to hand-engineered features (event-related potential components). Then, we compare the performance of these methods with end-to-end deep learning techniques applied to raw data. We demonstrate via average cross-validation performance measures that recurrent deep convolutional neural networks can outperform traditional machine learning methods for sequence modeling. We illustrate the intuitive salient information of the model with the location of the most relevant attributes of a post-stimulus window. This baseline identification system in the area of mental illness supports the evidence of developmental and disease effects in a pre-prodromal phase of psychosis. These results reinforce the benefits of deep learning to support psychiatric classification and neuroscientific research more broadly.**

***Index Terms*—Early stages of psychosis, Abnormal brain activity, Convolutional neural networks, Recurrent neural networks.**

I. Introduction

DISCRIMINATING abnormalities of brain activity in individuals who are at-risk of mental illness may be pivotal for early intervention [1]. In schizophrenia, traditional methods of identifying at-risk individuals on the basis of a positive family history of the disorder have been supplemented more recently by the development of methods to identify individuals who are putatively in the premorbid [2] and/or prodromal phases of illness [3] that precede the onset of frank psychosis. These methods typically identify such individuals based on clinical signs and symptoms of disease, but event-related potential (ERP) recordings of brain function have identified a variety of functional abnormalities among these individuals at risk for schizophrenia (RSz) that were discovered previously in adults with chronic schizophrenia [4]. One such ERP component, the amplitude of the mismatch negativity (MMN) potential, reflects an automatic process that detects any deviation of an incoming stimuli from the sensory memory trace established by the preceding stimuli, and is typically elicited during passive listening to an auditory "oddball" paradigm in which an infrequent deviant (*e.g.* which varies in duration or frequency) is presented against a background of frequent standard tones. The MMN correlates with disease severity and appears to be sensitive to disease state, providing a useful biomarker for the evaluation of abnormal brain function in children at risk of developing schizophrenia [5], [6].

Several studies have demonstrated the potential of traditional machine learning (ML) techniques to detect abnormalities in brain structure and function for schizophrenia. For example, Gould et al. [7] used neuroanatomical features and support vector machines to classify cognitive subtypes of schizophrenia, while Huang et al. [8] proposed a model to classify schizophrenia using a random forest operating on importance scores from multifrequency bands of functional magnetic resonance images. Regression trees were used for prediction of psychotic relapse in [9], and structural neuroimaging features coupled with support vector machines and k-nearest neighbours were also used to detect the first psychosis episode [10], [11]. In these previous studies, the participants used for experiments have been adults with a diagnosis of schizophrenia. The evaluation of brain function during the early stages of the disease that precedes clinical diagnosis has not been investigated sufficiently. The majority of prior ML research is based on imaging data evaluation and has not considered the study of EEG-derived ERP recordings to explore whether abnormalities of auditory information processing can be detected in the premorbid phase of illness, among RSz children. Additionally, prior approaches have been centered around extracting hand-engineered features which have limitations relating to expert knowledge being required.

In recent years, deep learning (DL) techniques have revolutionised computer vision and the medical domain including the evaluation of brain signals [12] and mental health disor-

The experimental procedures involving human subjects were approved by the Joint South London and Maudsley and Institute of Psychiatry National Health Service Research Ethics Committee. K. Laurens was supported by an Australian Research Council Future Fellowship (FT170100294). Funding for data collection was provided by a National Institute for Health Research (UK) Career Development Fellowship (CDF/08/01/015) and BIAL Foundation Research Grants (36/06 and 194/12).

D. Ahmedt-Aristizabal is with CSIRO, Data61, Canberra, Australia (e-mail: david.ahmedtaristizabal@csiro.au.)

D. Ahmedt-Aristizabal, T. Fernando, S. Denman, S. Sridharan, and C. Fookes are with the Image and Video Research Lab, SAIVT, Queensland University of Technology, Australia.

K. Laurens, P. Johnston and J. Robinson are with the School of Psychology and Counselling and the Institute of Health Biomedical Innovation (IHBI), Queensland University of Technology, Australia. (e-mail: kristin.laurens@qut.edu.au.)

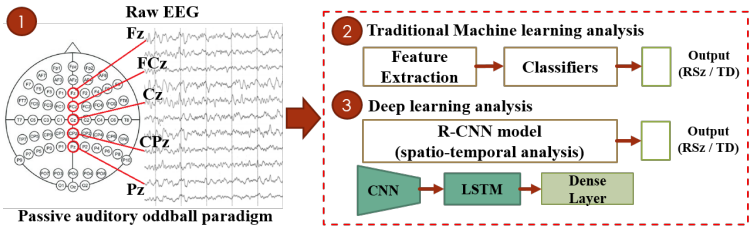

Fig. 1. Overview of the proposed learning algorithms for identifying abnormal brain activity in RSz children. *Dataset:* **1.** EEG data from RSz and TD children are recorded during the presentation of a passive auditory oddball paradigm. A total of 1,600 trials of 300 ms length from the combined standard and deviant stimuli represent the data analysed for each participant. A subset of five channels from the midline are selected for analysis. *Learning algorithms:* **2.** Hand-engineering features are extracted, comprising early negative and mid-latency positive component mean amplitudes. Then, traditional classifiers are used to identify EEG abnormalities. **3.** A hybrid deep learning model (R-CNN) is trained to extract spatial and temporal features from raw data. The output of the model is represented by the classification accuracy of each group.

ders [13]. The major advantage of DL compared to traditional ML is that feature engineering is not required, and the model itself discovers the optimal features directly from the data. These techniques were proposed to learn representations from neuroimaging [14] and EEG recordings [15] to identify patients with schizophrenia. However, investigation into how DL can analyse complex conditions of brain response to auditory stimulus changes in children have been limited so far. Analysis of ERP components elicited by deviant auditory stimuli have been described in premorbid [5] and prodromal [6] phases of schizophrenia, but an automated classification of EEG recordings was not considered.

In this paper, we compare the brain electrical activity of RSz children relative to their typically developing (TD) peers through an initial assessment (A1) conducted at 9-12 years and two reassessments completed approximately 24 months (A2) and 48 months (A3) later. We first use ML algorithms with hand-engineered features extracted from EEG recordings to classify physiological signals. Then, we introduce DL models such as convolutional (CNN) and recurrent neural networks (RNN) to process data without feature transformation Of the original EEG recordings, in each assessment phase. We are also motivated to identify features that remain in the developmental trajectory of RSz children. We aim to explore if the robustness of the biomarker and model change with age and putative disease progression. To the best of our knowledge, this is the first work to apply deep learning to EEG data obtained at multiple time points from individuals at risk for schizophrenia. We envision that DL algorithms such as memory networks could provide the basis for ongoing significant breakthroughs in distinguishing vulnerability to psychiatric disorders such as schizophrenia.

Our technical contributions are summarised as follows:

1) We introduce and compare traditional machine learning and deep learning architectures to classify EEG-based vulnerability for schizophrenia.
2) We demonstrate that a hybrid deep learning network is capable of identifying children at-risk of schizophrenia, can provide the location of relevant features, and is able to determine the impact of each recording channel for the identification task.

II. Materials and Methods

In this paper, we evaluate different approaches that quantify brain electrical activity to identify children at risk of schizophrenia, by using either traditional ML or DL architectures. In the ML approach, we use hand-crafted features defined as the early negative and mid-latency positive component mean amplitudes of the EEG response to each stimulus. Then, we compare the classification performance with traditional classifiers. For the DL approach, we design and train from scratch an architecture based on hybrid networks such as recurrent convolutional neural networks (R-CNN) [16]. This model extracts spatio-temporal representations directly from the raw data without pre-processing, and performs classification. An abstract view of these approaches is given in Fig. 1. For evaluation, we first classify trials from RSz and TD children at each of the three assessment phases. Then, we determine whether a test participant may exhibit vulnerability for schizophrenia based on the total number of trials per recording with functional abnormalities according to a threshold of acceptance. Finally, we compare and detect abnormal patterns that remain in RSz children across the developmental trajectory. Details of each strategy are described in the following subsections.

A. Dataset

Scalp potentials with a 10-10 configuration were captured at a sampling rate of 500Hz from 30 Ag/AgCl electrodes during the presentation of a passive auditory oddball paradigm. This task comprised 1600 tones at 1000 Hz, including 1360 standard stimuli of 25 ms duration (85%) and 240 deviant stimuli of 50 ms duration (15%), all with a rise and fall time of 5 ms. The stimulus onset asynchrony was 325 ms for each standard and 350 ms for each deviant tone, which yielded an isochronous interstimulus interval of 300 ms.

Table I summarizes the number of participants included in our dataset at each assessment. RSz children included individuals with a positive family history of schizophrenia (in at least one second-degree relative) and/or children presenting a triad of replicated antecedents of schizophrenia that included a delay or abnormality in speech and/or motor development, internalising or externalising problems, and at least one psychotic-like experience [17]. TD children were those who presented none of the antecedents of schizophrenia and no family history of schizophrenia spectrum illness in any first-, second- or third-degree relative. RSz children were also categorised according to the developmental trajectory: A1 at 0 months (aged 9-12 years), A2 at ∼24 months (aged 11-14 years), and A3 at ∼48 months (aged 13-16 years).

TABLE I
RSZ AND TD PARTICIPANTS AND TRIALS FOR EACH ASSESSMENT PHASE.

Assessment Phase	Participants		Trials	
	RSz	TD	RSz	TD
A1	65	40	104,000	64,000
A2	65	45	104,000	72,000
A3	57	42	91,200	67,200

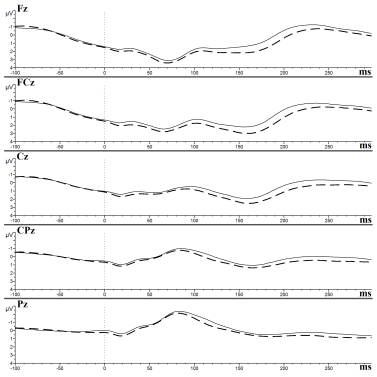

Fig. 2. Grand average of the standard and deviant waveforms from the five midline channels from the post-stimulus window in RSz (dashed line) and TD children (solid line).

Each participant is represented by a single recording session with multiple trials and each trial is defined as the post-stimulus window of 300 ms. The total number of trials per participant of each group (RSz and TD) corresponds to the combined number of standard and deviant stimulus (*i.e.* 1600 trials). We selected a subset of midline channels relevant to oddball task analysis (Fz, FCz, Cz, CPz, and Pz) [5] (see Fig. 1). Midline channels are least likely to be excluded in a participant due to recording problems/artefacts A simple pre-processing routine was performed on the raw EEG data by applying a 50 Hz notch filter [5]. Fig. 2 illustrates the window length of the average brain response elicited by the standard and deviant tones for RSz and TD children.

B. Learning algorithms

In this paper, we conduct a systematic evaluation of ML and DL techniques commonly used for sequence modeling.

1) Traditional machine learning: To compare ML classifiers, we first discuss the process of feature extraction. For each trial, we compute the early negative (mean amplitude between 80-220 ms) and mid-latency positive (mean amplitude between 160-290 ms) component amplitudes. These time-windows correspond to those typically used for the manual evaluation of the passive auditory oddball paradigm [5], [18]. These values are baselined to the average amplitude of the 100 ms window preceding the stimulus onset. Therefore, the feature space of each trial is represented by 10 features (two amplitude values for the five channels of interest), with each participant being of dimension [#trials,10]. To classify these hand-engineering features, we adopt traditional classifiers including k-nearest neighbors (KNN), decision trees (DT), and support vector machines (SVMs). These are among the most common classifiers used in previous studies in neuroimaging for psychiatric evaluations [7], [10], [11].

2) Deep learning: In our study, we compare DL models that process raw signals and eschew handcrafted features for the identification of relevant features at the early stages

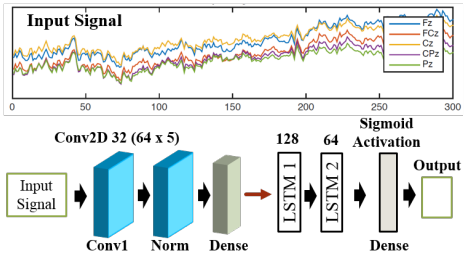

Fig. 3. Representation of the input signal (one trial from a RSz child) and the R-CNN model (2D-CNN-LSTM). A CNN is used to extract spatial features, and an LSTM is designed to extract temporal features.

of psychosis. The input representation of each raw trial is defined by [#trials,300,5], which corresponds to the post-stimulus window of interest of 300 ms and the five channels of evaluation. We implement per participant a min-max scaling of each channel. No other pre-processing or feature extraction techniques are performed.

We exploit many insights about suitable network architectures. We choose to design and train our network from scratch to learn representations of raw EEG signals with a low number of trainable parameters. Through extensive experiments, we compare CNNs, RNNs and R-CNNs architectures. 1D-CNNs and 2D-CNNs are multi-layered architectures where the primary idea is to train one-dimensional or two-dimensional convolution filters to extract local features at different levels of a hierarchy [19]. RNNs such as long short-term memory units (LSTM) [20] or gated recurrent units (GRU) [21] have proven to be stable in modeling dependencies in sequential data by employing an external memory cell state. R-CNNs are used to better exploit variable-length sequential data to extract spatio-temporal features and classify through an end-to-end deep learning model. R-CNN constitutes of a combination of 1D / 2D convolution layers followed by stacked recurrent units (LSTMs or GRUs). Shallow hybrid networks have shown to be suitable for analysing small amounts of training data [22] and EEG recordings in a mental load classification task [23]. Networks using LSTMs have outperformed counterparts using GRUs in the specific domain of seizure detection from EEG data [24].

We adopt an R-CNN model (2D-CNN-LSTM) which shows the best performance for trial classification on assessment A1 (see Section III-C). The chosen network architecture is displayed in Fig. 3. The input is transformed into a feature map through the first computational block consisting of (1) one 2D convolutional layer with 32 kernels with size $[64 \times 5]$ (experimentally evaluated), (2) one normalization layer, and (3) one fully connected layer. In the second computational block, the CNN output is subsequently fed to two stacked LSTM layers. The hidden state dimension of the LSTMs are determined empirically and are set to 128 and 64 units, respectively. Finally, the second hidden recurrent layer is fed into a classification layer with a sigmoid activation function. This layer provides an output of $[q, 1-q]$, where q represents the probability of an abnormal trial to be true and reciprocally $[1-q]$ describes the probability for the post-stimulus to be false.

III. Evaluation

A. Experimental setup

1) Trial classification: To evaluate and compare the performance of each proposed ML and DL model, we adopt the trial classification strategy and the recordings from the initial assessment A1. With the model that shows the best performance, we evaluate trials and participants in the reassessment phases.

We adopt a 5-cross-validation (CV) strategy [25] across participants. One fold is considered the test data set, and the remaining folds are the training data set. This ensures that per fold, all trials from one child belong to either the train or test set (*e.g.* for phase A1, 52 and 13 RSz children are adopted for training and testing, respectively). We combine all trials across children of the training set, then, we classify trials that have not been seen during training. The training set is divided into 80% for training the model and 20% for tuning the parameters. The final trial classification is computed as the average performance of each fold.

2) Participant identification: To evaluate the identification of RSz children at an individual level in each test fold, we compute the proportion of trials that were correctly classified with an abnormality to predict the group membership of each participant. Where this proportion exceeds a threshold of 60%, we consider the test participant to exhibit vulnerability for schizophrenia. This boundary level was adopted with the consideration of the total number of trials and heterogeneity of EEG responses.

3) Identification of abnormal patterns among assessments via trial classification: We aim to compare the robustness of the model for capturing abnormal brain activity across the developmental trajectories (A1, A2 and A3). In this scenario, we train the model using all trials from one assessment phase and test the model with trials from another phase. We evaluate each test phase (target phase) with 80% of trials allocated for validation and 20% for test, with a 5-fold CV scheme. Using this split, we optimize the classifier to perform well on the target distribution. To generate the final classification of each test phase, we utilise majority voting.

B. Implementation details

For traditional ML classifiers, we find hyperparameters that minimize 5-fold CV loss by using automatic optimizers such as the Bayesian optimization computed in MATLAB [26]. We select the following parameters. *KNN:* number of neighbours (6), distance (euclidean). *DT:* maximum number of splits (20), split criterion (Gini's diversity index). *SVM:* kernel function (gaussian), box constraint level (0.0115), kernel scale (0.2938).

We determine experimentally DL parameters such as the number of layers, the number of channels and the filter size. We also evaluate potential performance improvements by using intermediate normalization by batch normalization and dropout for regularization. We train the proposed DL model by optimizing the binary cross-entropy loss. We use Adam optimizer with a learning rate of 10^{-3}, and decay rates for the first and second moments of 0.9 and 0.999 respectively. We employ dropout with a probability of 35% in the LSTM architecture. Batch-size was set to 32. We train the model over 200 epochs using the default initialization parameters from Keras [27]. As this is a data with an uneven class distribution, we balance the training data at the trial level using class weight parameters [27].

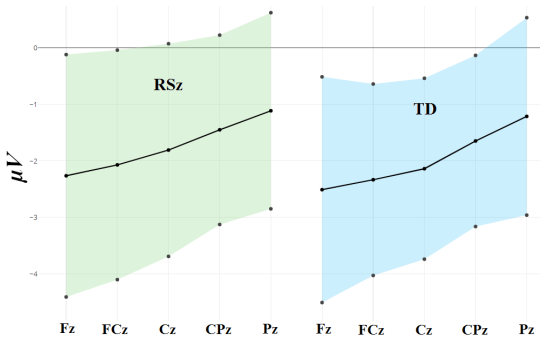

Fig. 4. Group difference in the early negative mean amplitude (mean ± SD) for each midline channel in RSz (green) and TD (blue) children.

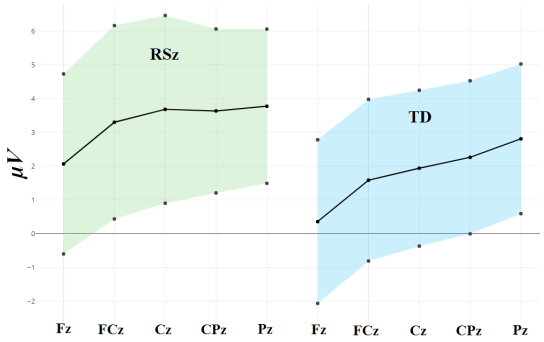

Fig. 5. Group difference in the mid-latency positive mean amplitude (mean ± SD) for each midline channel in RSz (green) and TD (blue) children.

C. Experimental results and model interpretation

1) Traditional machine learning: We first illustrate the difficulty in distinguishing RSz and TD children based on typical features extracted from the brain response to the passive auditory oddball paradigm. In [5], the authors demonstrated that by using peak amplitudes, there exists a statistical difference between RSz and TD children. Considering only assessment A1, we compute the mean amplitude in the respective time windows of interest (early negative and mid-latency positive components) across all participants of each group and discriminate according to each electrode channel. Fig. 4 illustrates the mean and standard deviation of the early negative mean amplitude for each channel for RSz and TD children, and Fig. 5 presents the same for the mid-latency positive mean amplitude. To qualitatively illustrate the limitations of these hand engineering features, we apply PCA [28] and plot the top two components in 2D. Fig. 6 shows the resultant plot. From this figure, the features are not sufficient to discriminate between RSz and TD children because there is not a clear separation between groups. Therefore, we can argue the complex task to differentiate between groups and the necessity to automate this process.

An evaluation of the optimized ML classifiers is presented in Table II. We find that in the assessment A1, all three models

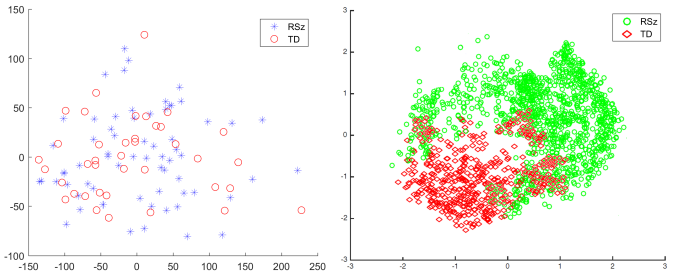

Fig. 6. 2D illustration of the evaluation of extracted features of RSz and TD children through PCA in the assessment phase A1. **Left:** Hand-engineering features using the average values per participant of the early negative and mid-latency positive mean amplitudes. **Right:** Embeddings from the R-CNN model for 5,000 randomly selected trials from the test set.

KNN, SVM and DT overfit to the training data showing poor performance and high variance on the validation data, and achieving average test accuracies of less than 45% (5-CVs). This baseline suggests that the features extracted were insufficient to discriminate between RSz and TD children.

2) Deep learning: An evaluation of all DL models to classify trials in assessment A1 is shown in Table II. The results showed improved performance of the hybrid network 2D-CNN-LSTM in comparison to ML classifiers, and convolutional and recurrent networks. In table III, we list the average results of trial classification using the best model in all assessment phases. This framework achieved an average accuracy across all phases of 89.98% and 69.53% for the validation and test sets, respectively.

To illustrate the variability of abnormal brain activity across participants in a selected fold, we compare the individual performance for all RSz children included in the test set (13 participants for A1 and A2, and 11 for A3). These results are illustrated via a horizontal boxplot in Fig. 7. It can be observed that the percentage of trials that were correctly classified at individual level for RSz children ranges from 43.2% to 95.2% (median 82.1%) for phase A1, from 50.1% to 91.4% (median 75.2%) for phase A2, and from 58.1% to 81.2% (median 67.1%) for phase A3. This marked intersubject variability (trials across RSz children) reflects challenges related to the heterogeneity and variable frequency of abnormal features within the EEG-based data. Some children exhibit an abnormality unique from all others and so are poorly classified. These patterns limit the ability of the model to generalise across participants, as shown by the difference between validation and test accuracy. It can also be observed in Fig. 7 that identifying trials in assessment A3 is more complex than the other two phases. However in this phase, the classification results at individual level are more consistent, which may indicate that the abnormality is similar among RSz children.

Table III indicates also the participant identification (RSz and TD children) based on the total number of correct trials classified on the test set. We evaluate the performance using the following metrics. *Sensitivity:* percentage of trials from a RSz participant that are correctly classified with the abnormality and are greater than 60%. *Specificity:* percentage of trials from a TD participant that are unrelated to the abnormality and are greater than 60%. In this scenario, the model categorised

TABLE II
TRIAL CLASSIFICATION OF THE LEARNING ALGORITHMS IN A1.

Algorithm	Test Accuracy (%)
KNN	44.50
SVM	43.88
DT	44.75
1D-CNN	45.10
2D-CNN	48.64
LSTM	53.90
1D-CNN-LSTM	55.20
2D-CNN-GRU	69.78
2D-CNN-LSTM	**72.54**

TABLE III
PERFORMANCE FOR CLASSIFYING TRIALS AND PARTICIPANTS FOR EACH ASSESSMENT PHASE WITH THE R-CNN MODEL.

Assessment Phase	Validation Accuracy (%)	Test Accuracy (%)	AUC	Test RSz (Sensitivity (%))	Test TD (Specificity (%))
A1	88.92	72.54	0.77	8/13 (61.53%)	6/8 (75%)
A2	93.59	69.83	0.71	8/13 (61.53%)	7/9 (77.77%)
A3	87.44	67.02	0.65	9/11 (81.81%)	7/8 (87.5%)
Average	**89.98**	**69.80**		**68.29**	**80.09**

Assessment phases. A1: Baseline; A2: ∼24 months; A3: ∼48 months.

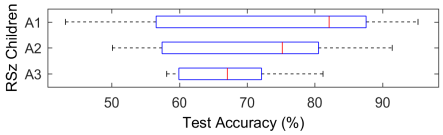

Fig. 7. Performance at the individual level for RSz children from the test set for each assessment phase.

TABLE IV
PERFORMANCE WHEN CLASSIFYING TRIALS AMONG PHASES.

Training Phase	Test Phase	Test Accuracy (%)
A1	A2	78.7
A1	A3	74.1
A2	A1	69.9
A2	A3	70.3
A3	A1	66.2
A3	A2	67.8

the majority of the participants correctly with an average sensitivity of 68.29% and specificity of 80.09% among the three assessment phases.

Finally, Table IV depicts the performance when identifying abnormal patterns that remain among the assessment phases. According to these results, we can confirm the presence of certain discriminative features that are common during the developmental trajectory. Therefore, we demonstrate the robustness of the model to detect vulnerability for schizophrenia across changes in age and putative disease progression.

As an EEG signal is recorded in the form of a multi-channel signal, we can consider this representation an image of one dimension (*i.e.* the window length and number of channels correspond to the width and height of an image, respectively) [29]. Such a representation merits investigation as it may enable the use of common DL benchmark used for image classification. It could be argued that fine-tuning these architectures to classify raw physiological signals may be preferable over training a network from scratch. However, we choose to design and train our own network due to the benefits conferred by a small number of trainable parameters $(1.3M)$. To illustrate this point, we fine-tune a pre-trained ResNet50 by modifying the last two layers (32 and 2 dimensional vectors) for binary classification. We find experimentally that this type of architecture is not suitable for our data due to the large number of training parameters $(> 23M)$ which quickly led to over-fitting. This supports other research that has preferred shallow networks for analysing EEG data [29].

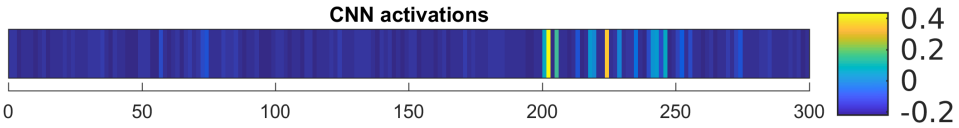

Fig. 8. Overview representation of the extracted activations from the CNN layer of the R-CNN model for random 500 trials correctly classified in the assessment phase A1. All filters from all trials are combined. Colours blue to yellow corresponds to low-high values.

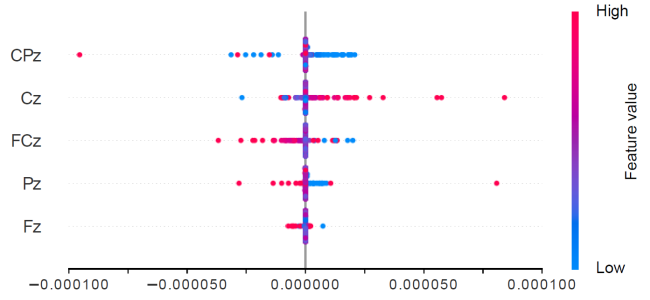

Fig. 9. Impact of each channel on the model output. We show the SHAP value of each channel for every sample of the test set (Red and blue denote high and low feature values respectively). We sort channels by the sum of SHAP value magnitudes over all test samples (*i.e.* first row most important), and use SHAP values to show the distribution of the impacts each channel has on the model output. This reveals that channel CPz is higher in the rank because it has values with most displacement from the decision boundary.

3) Model interpretation: The proposed R-CNN model is able to learn representative features from the raw data automatically. In Fig. 8, we illustrate the heat maps of a feature sample extracted from the 2D CNN layer, *i.e.* a representation of the relevant features from the morphology of the signal. Here, the feature importance varies from blue to yellow, and yellow represents highly discriminative features. Considering an average of 500 randomly selected samples that were correctly classified from the test set, we note that the most important features to identify RSz children are located between 200-225 ms in the post-stimulus window. This information can be corroborated with the group average waveforms depicted in Fig. 2. We also analyse the discriminative nature by randomly sampling 5,000 inputs from the test set in assessment A1 and apply PCA. The output result is shown in Fig. 6. The features are extracted from the last LSTM layer in the R-CNN model. This clearly demonstrates that the resultant sparse vector is able to discriminate a considerable proportion of trials.

We adopt the SHapley Additive exPlanations (SHAP) framework [30], [31] for interpreting the model output. The SHAP values represent what the model would have predicted if the model did not know any information regarding the other features (channels in this scenario). Hence we are capable to provide an overview of the most important channels for the model to identify abnormality. For binary classification, a SHAP value of 0 is taken as the base decision boundary and the negative side of the decision boundary is one class and the positive side is the other (a clear separation of SHAP values is expected only in a perfect model). An overview of discriminative channels according to the SHAP values in evaluating trials from the assessment A1 is illustrated in Fig. 9. These contacts are sorted based on the importance by the sum of SHAP value magnitudes over all samples. Higher features values are shown in red and lower values are in blue. In this scenario, channel CPz is the most important contact on the midline in the passive auditory oddball paradigm for distinguishing RSz children as it provides a higher dispersion of SHAP values from the decision boundary (*i.e* 0.00).

D. Discussion

In the last few years, there has been increasing interest in the potential of traditional machine learning and deep learning techniques in mental illness evaluation [13]. As introduced previously, several works have proposed using ML techniques as a tool for distinguishing between adults with schizophrenia and healthy controls using neuroimaging data. However, there has been considerably less work regarding analysis of electrophysiological recordings of children at risk of schizophrenia. This scenario is more complex because the characteristic brain abnormalities among adults with schizophrenia are more evident than for children putatively at the early stages of psychosis. In children, the abnormalities can be confused with those of other neurological and behavioural problems.

We have focused on providing a baseline for identifying RSz children by exploring traditional ML and DL techniques and presenting the model interpretation of what the models are learning for further clinical analysis. Multi-pipeline studies have been proposed as a useful way to disentangle what aspects work best to analyse EEG recordings from a passive auditory oddball paradigm. Contrary to expectations, all traditional ML approaches tested performed poorly. Using hand-crafted features such as the early negative and mid-latency positive mean amplitudes is not sufficient to obtain good segregation between trials from RSz and TD children. Therefore, it is important to be aware of the challenges and limitations when applying traditional ML to psychosis. On the other hand, we have demonstrated that by using the raw data from five selected channels of the midline with a traditional R-CNN model, we can learn important discriminative features to classify trials and participants with a vulnerability for schizophrenia. The features extracted from the trained model illustrate the importance of deep learning for salient information retrieval to support clinical research evaluations. This information can be used in the interpretation of the location and periodicity of possible discriminative features during the brain response to auditory changes. Having a clear understanding of the data expected for this paradigm will help to corroborate what the models are learning.

The identification of the most important channel in the model decision can have a significant impact on the feasibility of the auditory recording sessions. Recording sessions to capture the physiological data in children can be challenging (*e.g.* some children might find the procedure irritating). Hence using only a few channels that discriminate effectively alleviates the time needed to apply the electrodes, which facilitate

each recording task.

The proposed methodology is a critical step in providing an automated tool for early intervention before illness onset. However, the model's prediction is still far from reaching the ideal performance. We note that this task is highly complex because of the increased heterogeneity of participants and abnormal brain activity at different ages. For this reason, an interesting direction for future research is the adoption of architectures that incorporate attention networks and external memory modules capable of mapping relationships across participants [32].

IV. Conclusion

This paper presents the first application of traditional machine learning and deep learning techniques for analysing electrophysiological recordings during a passive auditory oddball paradigm in children with vulnerability for schizophrenia. We evaluate the capability of deep learning models and show that they allow for the detection of children at the early stages of psychosis when the effects of evolving illness are subtle. We applied several neural network architectures to learn invariant markers from raw EEG recordings of a cognitive task. In particular, an R-CNN model reached the best identification performance in distinguishing abnormal brain activity, which confirms the utility of a brain activity response as a biomarker. We have illustrated the discriminative nature of the learned embedding, the location of discriminative features in the selected post-stimulus window and the channel importance to discriminate early stages of psychosis.

References

[1] K. R. Laurens, S. Hodgins, G. L. Mould, S. A. West, P. L. Schoenberg, R. M. Murray, and E. A. Taylor, "Error-related processing dysfunction in children aged 9 to 12 years presenting putative antecedents of schizophrenia," *Biological Psychiatry*, vol. 67, no. 3, pp. 238–245, 2010.

[2] K. R. Laurens and A. E. Cullen, "Toward earlier identification and preventative intervention in schizophrenia: evidence from the london child health and development study," *Social psychiatry and psychiatric epidemiology*, vol. 51, no. 4, pp. 475–491, 2016.

[3] P. Fusar-Poli, S. Borgwardt, A. Bechdolf, J. Addington, A. Riecher-Rössler, F. Schultze-Lutter, M. Keshavan, S. Wood, S. Ruhrmann, L. J. Seidman *et al.*, "The psychosis high-risk state: a comprehensive state-of-the-art review," *JAMA psychiatry*, vol. 70, no. 1, pp. 107–120, 2013.

[4] M. A. Niznikiewicz, "Neurobiological approaches to the study of clinical and genetic high risk for developing psychosis," *Psychiatry research*, 2019.

[5] J. M. Bruggemann, H. V. Stockill, R. K. Lenroot, and K. R. Laurens, "Mismatch negativity (MMN) and sensory auditory processing in children aged 9–12 years presenting with putative antecedents of schizophrenia," *International journal of psychophysiology*, vol. 89, no. 3, pp. 374–380, 2013.

[6] R. J. Atkinson, P. T. Michie, and U. Schall, "Duration mismatch negativity and p3a in first-episode psychosis and individuals at ultra-high risk of psychosis," *Biological psychiatry*, vol. 71, no. 2, pp. 98–104, 2012.

[7] I. C. Gould, A. M. Shepherd, K. R. Laurens, M. J. Cairns, V. J. Carr, and M. J. Green, "Multivariate neuroanatomical classification of cognitive subtypes in schizophrenia: a support vector machine learning approach," *NeuroImage: Clinical*, vol. 6, pp. 229–236, 2014.

[8] J. Huang, Q. Zhu, X. Hao, X. Shi, S. Gao, X. Xu, and D. Zhang, "Identifying resting-state multifrequency biomarkers via tree-guided group sparse learning for schizophrenia classification," *IEEE journal of biomedical and health informatics*, vol. 23, no. 1, pp. 342–350, 2019.

[9] G. Fond, E. Bulzacka, M. Boucekine, F. Schürhoff, F. Berna, O. Godin, B. Aouizerate, D. Capdevielle, I. Chereau, T. D'Amato *et al.*, "Machine learning for predicting psychotic relapse at 2 years in schizophrenia in the national face-sz cohort," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 92, pp. 8–18, 2019.

[10] S. Vieira, Q.-y. Gong, W. H. Pinaya, C. Scarpazza, S. Tognin, B. Crespo-Facorro, D. Tordesillas-Gutierrez, V. Ortiz-García, E. Setien-Suero, F. E. Scheepers *et al.*, "Using machine learning and structural neuroimaging to detect first episode psychosis: Reconsidering the evidence," *Schizophrenia bulletin*, 2019.

[11] L. Squarcina, U. Castellani, M. Bellani, C. Perlini, A. Lasalvia, N. Dusi, C. Bonetto, D. Cristofalo, S. Tosato, G. Rambaldelli *et al.*, "Classification of first-episode psychosis in a large cohort of patients using support vector machine and multiple kernel learning techniques," *NeuroImage*, vol. 145, pp. 238–245, 2017.

[12] A. Craik, Y. He, and J. L. P. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *Journal of neural engineering*, 2019.

[13] A. B. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: a scoping review of methods and applications," *Psychological medicine*, pp. 1–23, 2019.

[14] J. Dakka, P. Bashivan, M. Gheiratmand, I. Rish, S. Jha, and R. Greiner, "Learning neural markers of schizophrenia disorder using recurrent neural networks," in *NIPSW, Machine Learning for Health*, 2017.

[15] S. L. Oh, J. Vicnesh, E. J. Ciaccio, R. Yuvaraj, and U. R. Acharya, "Deep convolutional neural network model for automated diagnosis of schizophrenia using eeg signals," *Applied Sciences*, vol. 9, no. 14, p. 2870, 2019.

[16] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015, pp. 2625–2634.

[17] K. R. Laurens, S. Hodgins, B. Maughan, R. M. Murray, M. L. Rutter, and E. A. Taylor, "Community screening for psychotic-like experiences and other putative antecedents of schizophrenia in children aged 9–12 years," *Schizophrenia research*, vol. 90, no. 1-3, pp. 130–146, 2007.

[18] J. R. Murphy, C. Rawdon, I. Kelleher, D. Twomey, P. S. Markey, M. Cannon, and R. A. Roche, "Reduced duration mismatch negativity in adolescents with psychotic symptoms: further evidence for mismatch negativity as a possible biomarker for vulnerability to psychosis," *BMC psychiatry*, vol. 13, no. 1, p. 45, 2013.

[19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[22] D. Ahmedt-Aristizabal, S. Denman, K. Nguyen, S. Sridharan, S. Dionisio, and C. Fookes, "Understanding patients' behavior: Vision-based analysis of seizure disorders," *IEEE journal of biomedical and health informatics*, vol. 23, no. 6, pp. 2583–2591, 2019.

[23] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," *arXiv preprint arXiv:1511.06448*, 2015.

[24] M. Golmohammadi, S. Ziyabari, V. Shah, E. Von Weltin, C. Campbell, I. Obeid, and J. Picone, "Gated recurrent networks for seizure detection," in *SPMB*, 2017, pp. 1–5.

[25] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, vol. 14, no. 2, 1995, pp. 1137–1145.

[26] M. U. Guide, "The mathworks," *Inc., Natick, MA*, vol. 5, no. 333, p. 4, 1998.

[27] F. Chollet, "Keras," 2015.

[28] I. Jolliffe, *Principal component analysis*. Springer, 2011.

[29] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[30] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NIPS*, 2017, pp. 4765–4774.

[31] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim *et al.*, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomedical Engineering*, vol. 2, no. 10, p. 749, 2018.

[32] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Task specific visual saliency prediction with memory augmented conditional generative adversarial networks," in *WACV*, 2018, pp. 1539–1548.