RobIn: A Robust Interpretable Deep Network for Schizophrenia Diagnosis

Daniel Organisciak[a], Hubert P. H. Shum[b], Ephraim Nwoye[c], Wai Lok Woo[a,*]

[a]*Department of Computer and Information Sciences, Northumbria University, Newcastle, UK*
[b]*Department of Computer Science, Durham University, Durham, UK*
[c]*Department of Biomedical Engineering, University of Lagos, Lagos, Nigeria*

Abstract

Schizophrenia is a severe mental health condition that requires a long and complicated diagnostic process. However, early diagnosis is vital to control symptoms. Deep learning has recently become a popular way to analyse and interpret medical data. Past attempts to use deep learning for schizophrenia diagnosis from brain-imaging data have shown promise but suffer from a large training-application gap - it is difficult to apply lab research to the real world. We propose to reduce this training-application gap by focusing on readily accessible data. We collect a data set of psychiatric observations of patients based on DSM-5 criteria. Because similar data is already recorded in all mental health clinics that diagnose schizophrenia using DSM-5, our method could be easily integrated into current processes as a tool to assist clinicians, whilst abiding by formal diagnostic criteria. To facilitate real-world usage of our system, we show that it is interpretable and robust. Understanding how a machine learning tool reaches its diagnosis is essential to allow clinicians to trust that diagnosis. To interpret the framework, we fuse two complementary attention mechanisms, 'squeeze and excitation' and 'self-attention', to determine global attribute importance and attribute interactivity, respectively. The model uses these im-

---

*Corresponding author
    Email addresses:* `daniel.organisciak@northumbria.ac.uk` (Daniel Organisciak), `hubert.shum@durham.ac.uk` (Hubert P. H. Shum), `enwoye@unilag.edu.ng` (Ephraim Nwoye), `wailok.woo@northumbria.ac.uk` (Wai Lok Woo)

portance scores to make decisions. This allows clinicians to understand how a diagnosis was reached, improving trust in the model. Because machine learning models often struggle to generalise to data from different sources, we perform experiments with augmented test data to evaluate the model's applicability to the real world. We find that our model is more robust to perturbations, and should therefore perform better in a clinical setting. It achieves 98% accuracy with 10-fold cross-validation.

1. Introduction

According to the World Health Organisation, schizophrenia can cause a greater level of disability than any other physical or mental illness (World Health Organization, 2011). Diagnosis and treatment must be conducted as early as possible to improve outcomes, so that it does not reach this level of severity (Häfner & Maurer, 2006). However, the symptoms of schizophrenia are similar to those caused by drug use, brain tumours, and other mental health problems like bipolar disorder. This makes schizophrenia notoriously difficult to diagnose, with clinicians going through an exhaustive process to rule out other potential causes. A minimum of six months of observation is required before a schizophrenia diagnosis can be provided, according to DSM-5 criteria (American Psychiatric Association., 2013). Our model predicts the diagnosis after six months from one observation session with 98% accuracy.

Machine learning is increasingly being applied for medical diagnosis in the real world. Identifying problems at an earlier stage helps to manage them before they develop into a serious condition. Employing machine learning for disease diagnosis can also reduce a considerable number of required man-hours: machine learning systems can quickly process information and provide a recommendation. This would allow more patients to receive high-quality early-stage treatment.

To use machine learning to diagnose schizophrenia, we construct a data set that complies with current clinical assessment guidelines. Clinicians currently must assess patients according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (American Psychiatric Association., 2013) criteria A-F. We compose a data set of clinical observations of patients with a mental health condition. The collected attributes in the data set include attributes related to DSM-5 criterion A symptoms and attributes related to the symptoms of other mental health illnesses, to help the model to exclude these illnesses in the same way that psychiatrists must. Moreover, mental health clinics that use DSM-5 criteria will record psychiatric observational data similar to ours. Our model can therefore be applied to data that is already recorded, or after a clinical appointment, with minimal change to current processes required. These data set design choices have been made to demonstrate the ability of our proposed framework to function in the real world. Many machine learning models have shown good performance on laboratory-created data sets, but their real-world impact has been minimal (Tandon & Tandon, 2019b). This paper aims to minimise the gap between research and practice.

Deep Neural Networks (DNNs) have become the most popular machine learning tool in recent years as they automatically learn effective features from the data, rather than rely on human input to design hand-crafted features. Even with expert knowledge, these algorithms were sometimes ineffective, or had unwittingly incorporated the human experts' bias. Therefore, DNNs typically attain stronger performance and generalisation ability compared with traditional machine learning algorithms (Alom et al., 2018).

One major drawback of DNNs is a lack of interpretability. A missed diagnosis could result in death, whereas incorrectly diagnosing a patient with a condition could also result in complications (e.g. due to side-effects of needlessly prescribed medicine). For a clinician to have confidence in a recommendation from a deep learning system, they must be able to understand why the system has come to that recommendation.

We propose a *collaborative attention network* to improve the interpretability
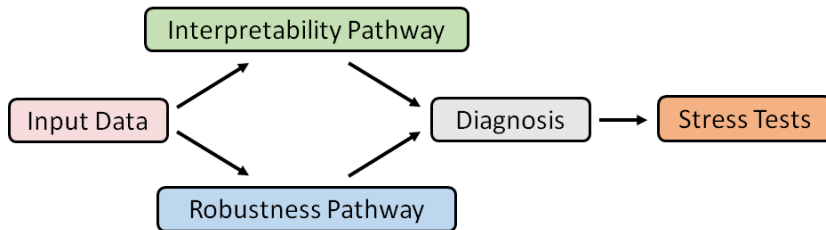
Figure 1: An overview of our proposed method. Data is processed via two pathways: one for robustness and one for interpretability. These pathways are complementary to one another, provide an insight into how the model arrives at a decision, and can generalise to new distributions.

of the system and provide a high quality diagnosis in an efficient manner. It is important to consider *static* and *dynamic attention*. We wish to determine which features the network believes are important in general, and to understand how this changes based on specific values that features take, and combinations of those values. Our mechanism consists of a *squeeze and excitation* (SE) module (Hu et al., 2018) to compute the global feature importance, and a *self-attention* (SA) module (Vaswani et al., 2017) to decipher how features interact individually and with each other.

To deploy a machine learning model in a real-world setting, it must be robust when generalising to data from a different distribution. A typical machine learning experiment will split a data set into a train and a test set. However, despite the test set being unseen during training, its distribution is often independent and identically distributed (i.i.d) to the training set. When a trained model is tested on distributions that are not identical to the training data, it often struggles to generalise to the new distribution (D'Amour et al., 2020).

Squeeze and Excitation networks (Hu et al., 2018) compute very discriminative feature representations by learning which input features at each layer are most important. We posit that this fixed global feature importance allows the network to remain robust to data from new distributions. To show the effectiveness of our method, we design three stress tests at four different strength

levels to alter the distribution of test data. Our stress tests consist of noise addition, data erasing and a combination of the two. We evaluate all compared models across these twelve robustness experiments to determine which would be suitable for deployment in a real-world clinical setting.

An overview of our paper can be found in Figure 1. We provide the following contributions:

1. The collection of psychiatric evaluation data that enables machine learning training for automated diagnosis of schizophrenia.

2. A **Rob**ust, **In**terpretable (RobIn) deep network is developed to model the data to recommend a schizophrenia diagnosis. Our framework outperforms existing machine learning methods such as support vector machines, multi-layer perceptrons, and deep neural networks.

3. RobIn is made interpretable on two levels: the adaptation of channel attention to determine global feature importance, and self-attention for feature interactivity analysis.

4. A total of twelve stress tests are designed to evaluate model performance on a distribution that is not i.i.d. to the training data, to simulate model performance in the real world. Our proposed model outperforms all other methods at remaining robust to perturbations.

We perform experiments on two settings: the standard 90/10 cross-validation protocol that is most commonly used with small data sets, and a 50/50 train/test split with 25 runs as a baseline with which to compare the robustness stress tests. We find that RobIn outperforms other methods on both of these test settings, and is also most robust to distribution shift proposed to test robustness.

The rest of the paper is organised as follows. Related studies are outlined in Section 2. Section 3 provides more details on the data set. Section 4 gives details on the deep learning framework that we propose. Section 5 outlines the attention mechanism helping clinicians to understand the outcomes of the DNN. Section 6 contains our experimental evaluation and stress tests. The paper is concluded in Section 7.

2. Related Work

2.1. Diagnosis of Schizophrenia

Schizophrenia has traditionally been diagnosed via specifications from the Diagnostic and Statistical Manual of Mental Disorders (DSM) or the International Classification of Diseases (ICD), with the most recent versions being DSM-5 (American Psychiatric Association., 2013) and ICD-11 (Khoury et al., 2017) respectively. These traditional specifications recommend a diagnosis based on the symptoms that a patient is exhibiting.

The Research Domain Criteria (RDoC) (Cuthbert et al., 2014) is a modern approach to understand the underlying neurobiology associated with major mental health issues. Although a proportion of schizophrenia research has shifted from DSM-ICD to RDoC, we note that the RDoC website states "RDoC is not meant to serve as a diagnostic guide, nor is it intended to replace current diagnostic systems." RDoC can be effective supplementary information but the necessary data can be costly to obtain and is therefore less accessible in developing countries, or for patients who must pay to obtain a brain scan but cannot afford to do so. DSM-ICD remains the gold-standard for diagnosis (Barros et al., 2021). For this reason, we design our diagnostic machine learning system using the traditional symptomatic approach, using DSM-5 symptoms.

2.2. Artificial Intelligence for Schizophrenia Diagnosis

DNNs have seen extensive use in the medical domain (Ravì et al., 2017; Zeiser et al., 2021; Rangarajan & Ramachandran, 2021). Clinical data is a very popular data source to apply deep learning (Shickel et al., 2018) for improved diagnosis (Obermeyer & Emanuel, 2016), detection (Lauritsen et al., 2020; da Silva et al., 2021), and risk prediction (Weng et al., 2017). Kelly *et al.* Kelly et al. (2019) identify three challenges for the use of machine learning on clinical data: bias identification, generalisation, and interpretability.

A vast majority of machine learning applications to mental health focus on the analysis of brain scans (Shi et al., 2018; Matsubara et al., 2019; Li et al., 2019;

6

Raza et al., 2019). This is also the case for schizophrenia diagnosis (Oh et al., 2020; Palaniyappan et al., 2019). Qi & Tejedor (2016) apply DNNs to magnetic resonance imaging (MRI) to identify subjects with schizophrenia and schizo-affective disorder. A layer-wise relevance propagation module was proposed by Yan *et al.* (Yan et al., 2017) to interpret a DNN that discriminated between schizophrenia patients with healthy controls. Kalmady et al. (2019) introduce fMRI data that is not corrupted with anti-psychotic medication and develop an ensemble algorithm with multiple parcellations.

Although these results seem promising, the real-world application of machine learning techniques on brain image data for schizophrenia diagnosis remains limited (Barros et al., 2021). There are several problems that are ubiquitous in the domain of machine learning for schizophrenia diagnosis:

- There is a large *training-application gap* (Wainberg et al., 2018), i.e. the trained models accurately describe the source data well, but the models are not shown to be transferable to the real world (Tandon & Tandon, 2019b,a), nor are there explanations of how their models could be used in a clinical setting. In contrast, our model is designed to be directly applicable to data that is routinely collected by schizophrenia diagnosticians. Furthermore, we design stress tests to demonstrate that our model would be robust to data from different sources.

- Data sets consist of only schizophrenic or healthy individuals but do not contain other unhealthy samples with bipolar disorder or drug overdoses (Winterburn et al., 2019). In practice, a model trained on this data is only able to determine if a patient is healthy or unhealthy, but cannot diagnose schizophrenia. Our data contains patients with multiple causes for the observed symptoms ('bipolar affection disorder', 'complex partial seizure', 'drug related disorder', etc.) resulting in a more challenging and realistic data set.

- Brain image scans in schizophrenia data sets often contain patients who have already started taking anti-psychotic drugs. These drugs have a dis-

7

DSM-5 Criteria	Attribute(s)	Label Information
Delusions	'Thought Content'	Includes persecutory and grandiose delusions
Hallucinations	'Thought Perception'	Includes auditory, visual, and tactile hallucination
Disorganised speech	'Speech'	Normal, mute, irrelevant or incoherent
Grossly disorganised /catatonic behaviour	'Mood', 'Attention', 'Concentration'	Good, poor, neutral
Negative symptoms	'Mental State Examination'	Kempt, unkempt, poor eye contact, restless

Table 1: Data Set Attributes Relating to DSM-5 Criteria

tinguishable effect on the brain (Lesh et al., 2015), reducing the reliability of the data sets and the results.

2.3. Interpretable Deep Learning

Deep learning has high requirements before it can be used in real-world medical applications due to the absolute necessities of trust (Nature Biomedical Engineering, 2018) and interpretability (Vellido, 2019; Rebane et al., 2020). *Attention* is commonly used to handle these issues (Cheung & Dahl, 2018; Huk Park et al., 2018). The attention modules highlight important structures within data to assist classification performance. By inspecting these modules, we can better interpret how the neural network comes to a classification decision. There are various types of attention module, including self-attention (Vaswani et al., 2017), squeeze and excitation (Hu et al., 2018) and spatial attention (Wang et al., 2017).

Multiple works within the medical domain have used attention for interpretability. Choi *et al.* (Choi et al., 2016) propose RETAIN, a two-level neural attention model which processes electronic health records to identify important past visits. They then use this information to predict the likelihood of heart failure. Oktay *et al.* (Oktay et al., 2018) build Attention U-net to segment organs on an abdominal computed tomography (CT) scan. This architecture could also be applied for disease diagnosis to segment regions of the image that

the network considers important to arrive at a decision. Paschali *et al.* (Paschali et al., 2019) equip a DNN with a multiple instance learning branch to output a logit heatmap of the neural activations.

Fused attention mechanisms have shown to improve performance for image processing (Organisciak et al., 2020) and natural language processing (Duan et al., 2018). However, we are unaware of attention fusion for medical diagnostic purposes.

3. Data Collection

In this section, we motivate the collected data set, give an overview of its characteristics, and detail how it follows current diagnostic guidelines. We also discuss the selected subjects and the pre-processing performed on the completed data set.

3.1. Data Motivation

Artificial intelligence for schizophrenia diagnosis has received criticism for not being transferable to the real-world (Tandon & Tandon, 2019b). One reason for this is a disconnect between machine learning practitioners and schizophrenia diagnosticians. Machine learning suffers from the ubiquitous practice of competing for the highest scores on benchmark data sets, often giving less regard to real-world applicability.

- The collected data is standardised, based on symptoms from DSM-5 (American Psychiatric Association., 2013). Attributes are taken directly from case files of patients, so the data will be very similar to that which is already collected in mental health clinics. Moreover, this data does not depend on advanced technology (the EEG attribute could easily be discarded), meaning that models developed from this data are more accessible to developing countries with limited resources. Note that our model requires minimal computational resources.

9

- Studies have shown a link between childhood trauma and schizophrenia in both high income (Matheson et al., 2013; Varese et al., 2012) and developing (Mall et al., 2020) countries. Schizophrenia can be particularly dangerous in poorer regions of Africa, where the likelihood of trauma is greater, the availability of treatment is lower, and there is stigma attached to mental illness (Burns, 2012). Most recent works that apply AI to diagnose schizophrenia use images from EEG or fMRI scans (Winterburn et al., 2019). However, these scans are often unavailable in poorer regions and there is no evidence to suggest that models developed in high-income countries can generalise to data from developing countries, either because of hardware differences or genetic differences (Gulsuner et al., 2020).

- The data allows us to predict a clinician's six-month diagnosis from one observational session. This greatly advances the speed that a potential schizophrenia diagnosis could be identified to help inform the clinician's future care of the patient.

- Most research on schizophrenia diagnosis only classify samples with schizophrenia from healthy samples. Mental health clinics rarely face this challenge as patients who do not have schizophrenia will usually have a similar mental health condition. The main difficulty of diagnosis is differentiating schizophrenia from other conditions rather than from healthy samples. Our data set contains samples with a broad variety of health issues to more accurately estimate the difficulty that clinicians face.

3.2. Data Acquisition

The data was collected from a total of 151 subjects, between 2013 and 2018 inclusive, from the Lagos University Teaching Hospital, Lagos, Nigeria. Our data set consists of psychiatrists' direct observations of their patients; consequently, our model gets the same information that psychiatrists have used to determine their diagnosis.

97 samples are positively diagnosed Schizophrenia patients, and the remaining 54 are other patients with different afflictions. It is important to note that the negative samples are patients suffering from other related illnesses; they are not healthy individuals as in most schizophrenia diagnosis data sets used for machine learning research. This makes schizophrenia diagnosis more difficult, but is a significantly more realistic challenge.

The medical records of the patients were obtained from their case files, after obtaining ethical approval from the Lagos University Teaching Hospital Health Ethics Committee. Each subject signed a patient consent form prior to their mental health assessments; the form permits the use of medical information for the purposes of research and education. All records were obtained anonymously in line with ethical approval guidelines, to retain the privacy of the subjects.

The full details of the collected attributes can be found in the appendix. Note that although we collect attributes of 'Age', 'Sex', 'Occupation History', and 'Marital Status', we do not use them in any of our experiments. RobIn is trained on observed features related to mental health, rather than the general characteristics of the individuals; that is, we train with features that represent an underlying cause of schizophrenia.

With any human-labelled data set, there is a small chance of label error. However, if ML methods can consistently infer a diagnosis from input features, the overall data set can be seen as reliable enough to generalise to new data.

3.3. Data Structure

The data set has 31 raw attributes that are used for learning, and a final diagnosis. The attributes collected reflect the criteria used by psychiatrists for Schizophrenia diagnosis, as officially recommended in DSM-5 documentations (American Psychiatric Association., 2013). DSM-5 outlines diagnostic criteria A-F, which must be fulfilled for a patient to be diagnosed with schizophrenia.

Criterion A concerns the characteristic symptoms of the condition, and is where DNNs can be effective. Table 1 presents: the DSM-5 diagnostic criterion
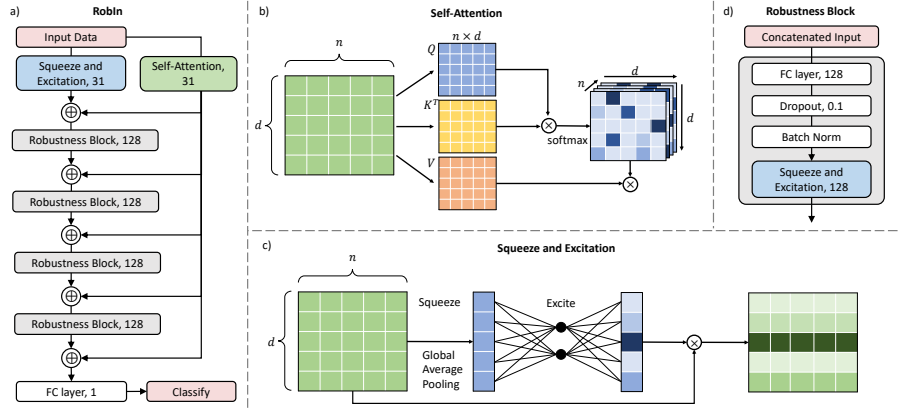
Figure 2: a) An overview of the entire network with output dimensions of each layer represented. Input data goes down the robustness stream and the interpretability stream; b) self-attention mechanism: the input representation is converted into a key, query and value matrix, the cosine distance between each query and each key is found via a matrix multiplication with a higher activation signalling higher alignment between query and key; c) squeeze and excitation: each attribute $i = 1, \ldots, d$ is *squeezed* down to a representative number, a miniature neural network *excites* the squeezed information to evaluate how important each attribute is, then the initial data is multiplied by the importance scores; d) the robustness block we propose in this paper.

A, the attributes selected to reflect criterion A symptoms, and further information on the potential labels of these attributes.

We also collect attributes that commonly occur in other mental health conditions, including depression, bipolar disorder, and schizo-effective disorder. We do this because DSM-5 Criteria D-F are exclusions of substance misuse and other disorders such as such as schizo-affective disorder and autism spectrum disorder.

3.4. Data Pre-processing

Data pre-processing is an essential step in most machine learning pipelines to improve consistency of the algorithm. The collected data set contains a mixture of numeric and text data, which is difficult for a deep learning model to process. Categorical data is converted directly into numeric representations via *label encoding*: for every attribute in the data set, each unique value is assigned a representative number.

Of the selected 31 raw attributes, the majority of samples are missing at least one value. This scenario is common in medical data and is therefore well-studied because, for example, different clinicians record different data. To address this, we set all missing values to -1 before performing label-encoding. This method is easiest to apply to the real world where the majority of samples will also have missing values, rather than having to use imputation methods from external data sources that may not match up to the data being processed.

Experimental results showed that setting missing values to -1 resulted in better performance than attempting single-variate or multi-variate imputation techniques. This is likely because this format explicitly informs the system about data absence, allowing the attention mechanism to learn to assign less emphasis.

4. Deep Neural Networks

Deep Neural Networks (DNN) are biologically inspired structures composed of layers of neurons. Every instance from the input data is passed forward through each DNN layer, until it becomes a one-dimensional feature representation of the initial input. This process is repeated many times to improve the quality of the representation.

Let $f^* : D \longrightarrow \{0,1\}^n$ be the function which correctly maps patient data, $\mathbf{x}_i \in X$, to the true diagnosis, y_i, for all $i = 1, \ldots, n$; i.e.

$$y = f^*(\mathbf{x}). \tag{1}$$

Here, n is the total number of patients, $X \subset \mathbb{R}^{k \times n}$ is collected data with k features of n patients and $\mathbf{x}_i$ refers to the data of patient i.

A DNN learns the function $f : X \longrightarrow \{0,1\}^n$ to get as close to f^* as possible. For $i = 1, \ldots, n$, we obtain

$$\hat{y}_i = f(\mathbf{x}_i; \theta), \tag{2}$$

where $\hat{y}_i$ is the predicted diagnosis of patient i, $\mathbf{x}_i \in X$ is the observed data of patient i, and θ are the weights and biases that the DNN learns. The goal is to minimise deviation from the predicted labels $\hat{y}$ to the true labels y.

The DNN f is comprised of multiple layers $f_1, \ldots, f_L$ given by:

$$\mathbf{h}_1 = f_1(\mathbf{x}; \mathbf{W}_l, \mathbf{b}_l) \quad = \phi(\mathbf{W}_0^\top \mathbf{x} + \mathbf{b}_0), \tag{3}$$

$$\mathbf{h}_l = f_l(\mathbf{h}_{l-1}; \mathbf{W}_l, \mathbf{b}_l) = \phi(\mathbf{W}_l^\top \mathbf{h}_{l-1} + \mathbf{b}_l), \tag{4}$$

where ϕ is the non-linear activation function, $\mathbf{x}$ is the input data, $\mathbf{h}_l$ is the output of the lth hidden layer, $\mathbf{W}_l$ and $\mathbf{b}_l$ are the neuron weights and biases of hidden layer l respectively.

We use binary cross-entropy as the loss function:

$$\mathcal{L}_{\text{BCE}} = -\sum_{x \in X}(y \log(p(x)) + (1 - y) \log(1 - p(x))), \tag{5}$$

where y is the class and p is the probability assigned by RobIn that an observation x belongs to the class $y = 1$.

5. Interpretable Artificial Neural Network

Despite the excellent performance of deep networks, they suffer from not being interpretable. Features at each layer are generated automatically by the network, so it is not clear to humans what these mean. To use machine learning for health applications, it is important to have a degree of understanding of how the algorithm arrives at a conclusion.

We develop a robust, interpretable framework based on *Squeeze and Excitation* and *Self-Attention* to provide insight on the network's decision process. From the input layer, all input features get processed by a robustness pathway,

governed by squeeze and excitation, and an interpretability pathway, indicated via self-attention. After each robustness block, we fuse the output vector with the self-attention, to ensure the interpretable mechanism is influential at all stages of the network. Fusing these attention modules allows the network to discover both *global importance* through squeeze and excitation, and *feature interactivity* through self-attention. We are therefore able to interpret the network on multiple levels. Our overall framework can be found in Figure 2.

5.1. Squeeze and Excitation

A *Squeeze and Excitation* unit (Hu et al., 2018) is a channel-wise attention mechanism that has recently been proposed for convolutional neural networks (CNN). It learns the most informative channels at each layer and assigns them greater weight and reduces the weight of less important channels. We re-design squeeze and excitation for non-sequential data in order to learn the importance of the input channels to help humans understand why a machine has come to a certain conclusion.

First, the information from each attribute is *squeezed* into a representative descriptor via Global Average Pooling. Given the jth attribute $a_j = [a_{1j}, \ldots, a_{bj}]$, where b is the number of samples in the batch, we form the attribute descriptor, c, via:

$$c_j = \text{squeeze}(a) := \frac{1}{b} \sum_{i=1}^{b} a_{ij}. \tag{6}$$

These feature descriptors form a vector $\mathbf{z} = [c_1, \ldots, c_d]$ where d is the dimensionality of each sample.

We then *excite* the network to determine which attributes are most important. $\mathbf{z}$ is passed through a fully connected layer, a ReLU, another fully connected layer and finally a sigmoid activation.

Formally, this excitation is written as:

$$\mathbf{s} = \text{excite}(\mathbf{z}) := \sigma\left(\mathbf{W_2}\delta(\mathbf{W_1}\mathbf{z})\right), \tag{7}$$

where $\mathbf{W_1}, \mathbf{W_2}$ are the parameters of the fully connected layers, δ is the ReLU function and σ is the sigmoid activation function.

5.2. Self-Attention

Self-attention (Vaswani et al., 2017) has transformed the field of natural language processing and is beginning to make similar progress in computer vision (Ramachandran et al., 2019) and graphical data (Veličković et al., 2018). We explore its potential for tabular data.

Self-attention learns three new representations for each input feature: a key, query, and value. Give an input feature, i: the value vector is a hidden representation of that feature consisting of d attributes; the associated key vector indexes the value vector; and the query vector learns to find keys that are relevant to the value vector.

Given a query, q_i, we calculate attention score a_{ij} of all keys, k_j, using the cosine similarity:

$$a_{ij} = q_i \cdot k_j = ||q_i|| \times ||k_j|| \cos(\theta). \tag{8}$$

Noting that the dimensionality of all queries and keys remains constant, the more similar q_i and k_j are, the larger $q_i \cdot k_j$ will be. A softmax function is applied to polarise the attention weights for each query. Finally, the value matrix, V, is then multiplied by these attention scores. This is demonstrated in Figure 2 b). The entire process can be neatly summarised as

$$\text{Attention}(K, Q, V) = \text{softmax}(K^\top Q)V, \tag{9}$$

with $K = \mathbf{W}_k x$, $Q = \mathbf{W}_q x$ and $V = \mathbf{W}_v x$, where $\mathbf{W}_k, \mathbf{W}_q$, and $\mathbf{W}_v$ are learned weights matrices for keys, queries and values, respectively.

To interpret the model, self-attention provides more information than squeeze and excitation. Rather than showing just the raw importance of each feature, self-attention shows us the importance of every feature j to each feature i. For example, squeeze and excitation may suggest that a feature i is more important than feature j, however it may be the case that j is important if i takes on a particular value (or is missing). Self-attention can handle this more complex circumstance.

16

5.3. Robust Interpretable Network

The architecture is designed with three principles in mind: performance, interpretability, and robustness. We marry the two attention mechanisms described, with squeeze and excitation contributing to performance and robustness, and self-attention mostly benefiting performance and interpretability.

In particular, squeeze and excitation has been widely shown to provide more informative features when used at each successive layer. We follow this standard design for the main branch of our network. Self-attention on the other hand is often used sparingly due to its computational cost being $\mathcal{O}(d^2)$ where d is the number of features. This occurs due to the matrix multiplication in Equation 9. Another issue is that we wish to apply self-attention directly to the data to be able to interpret the final result; however, self-attention applied in early layers often performs an averaging effect in early layers, and only starts to highlight important feature interactions at later layers (Ramsauer et al., 2020). This problem is exacerbated when multiple self-attention modules are used throughout the network.

To solve these problems, we include one self-attention mechanism in our framework, and concatenate its output with the feature representation obtained from each layer. This means that the self-attention mechanism acts as a direct path from the input data set to the output classification. This means that the weights of the self-attention mechanism are partially backpropagated directly from the classification layer, rather than the indirect path that is usually seen. This is important for interpretability because the attention heatmaps have to be powerful enough to contribute to the final classification, otherwise performance would heavily suffer. We also find that this self-attention module returns more descriptive, polarised attention heatmaps compared with a traditional implementation, so clinicians can attain more insight from studying them. Each epoch requires 0.012 seconds to train. Total training time lasted 23 seconds. Our full RobIn architecture is shown in Figure 2.

17

Method	Acc.	F1-score	AuC	Precision	Sensitivity	Specificity
MLP	91.33 ± 3.63	92.84 ± 3.03	91.11 ± 3.85	89.83 ± 5.27	96.87 ± 2.60	85.36 ± 7.59
DNN	92.00 ± 2.70	93.19 ± 2.46	92.49 ± 2.85	93.19 ± 4.04	93.67 ± 2.86	91.31 ± 4.94
SVM	94.67 ± 3.14	95.55 ± 2.71	95.62 ± 2.52	99.00 ± 1.62	92.90 ± 5.00	98.33 ± 2.71
Tree	96.00 ± 3.31	96.84 ± 2.65	97.15 ± 2.38	$\mathbf{100.00} \pm 0.00$	94.29 ± 4.77	$\mathbf{100.00} \pm 0.00$
SENN	96.67 ± 2.91	97.39 ± 2.25	97.22 ± 2.51	98.89 ± 1.80	96.11 ± 3.31	98.33 ± 2.71
SANN	96.67 ± 3.33	97.12 ± 2.92	96.51 ± 3.32	96.64 ± 2.79	97.78 ± 3.61	95.24 ± 7.30
RobIn	$\mathbf{98.00} \pm 2.30$	$\mathbf{98.56} \pm 1.62$	$\mathbf{98.33} \pm 3.33$	99.00 ± 1.62	$\mathbf{98.33} \pm 2.71$	98.33 ± 5.00

Table 2: Comparison with Baseline Machine Learning Techniques for the Diagnosis of Schizophrenia with 90/10 Cross Validation

6. Evaluation

6.1. Evaluation Protocol

We consider two settings to test machine learning models on our schizophrenia data set:

1. 10-fold cross-validation,

2. A 50/50 train/test split repeated 25 times, to test the ability of all models with less training data, and to have more test data to perturb with the designed stress tests to more reliably evaluate the performance of all models under a different distribution.

We performed hyperparameter tuning on all models with manual search. In all cases, we report the strongest performance that we could obtain. We also report confidence intervals at the 95% significance level. For fairness, all experiments are performed with the same random seed. The implementation of all frameworks was performed in PyTorch on an NVIDIA Geforce GTX 2080ti. The full code will be released upon acceptance.

We compare three attention models that we propose: our **Rob**ust **In**terpretable Deep Network, **RobIn**, a squeeze and excitation network with four-layers, **SENN**, and a self-attention network containing four consecutive self-attention layers (i.e. no fully-connected layers), **SANN**. These models are compared with standard approaches: a decision tree, **tree**, a support vector machine, **SVM**, a

Method	Acc.	F1-score	AuC	Precision	Sensitivity	Specificity
SVM	83.57 ± 1.47	86.67 ± 1.28	83.38 ± 1.44	89.27 ± 1.76	84.78 ± 2.46	81.97 ± 2.96
MLP	84.43 ± 1.80	88.24 ± 1.35	82.14 ± 2.18	85.04 ± 2.49	$\mathbf{92.22} \pm 1.37$	72.07 ± 4.46
DNN	84.69 ± 2.30	88.37 ± 2.30	82.53 ± 1.85	85.34 ± 2.30	92.12 ± 1.51	72.93 ± 3.90
Tree	85.49 ± 1.85	88.15 ± 1.68	85.33 ± 1.92	90.29 ± 2.17	86.71 ± 2.62	$\mathbf{83.95} \pm 3.51$
SENN	85.97 ± 1.69	$\mathbf{89.68} \pm 2.32$	85.45 ± 1.99	89.68 ± 2.32	88.47 ± 1.88	82.43 ± 3.94
SANN	86.40 ± 1.82	89.36 ± 1.47	85.30 ± 2.35	88.66 ± 1.47	90.47 ± 1.60	80.14 ± 4.02
RobIn	$\mathbf{86.45} \pm 1.46$	89.19 ± 1.19	$\mathbf{86.08} \pm 1.75$	$\mathbf{90.51} \pm 2.17$	88.42 ± 1.83	83.73 ± 3.77

Table 3: Comparison with Baseline Machine Learning Techniques for the Diagnosis of Schizophrenia on a 50/50 Train/Test Split (25 runs)

Reference	Data Type	Samples	Classification Type	Accuracy
Li et al. (2020)	fMRI	183	SZ v HC	80.5
Niu et al. (2019)	fMRI	82	SZ v HC	90.8
Palaniyappan et al. (2019)	fMRI	57	SZ v BD	76.3
Salvador et al. (2019)	fMRI+sMRI	211	SZ v HC	84.0
Oh et al. (2020)	sMRI	926	SZ v HC	88.6
Phang et al. (2020)	EEG	84	SZ v HC	91.7
Calhas et al. (2020)	EEG	84	SZ v HC	95.0
RobIn	Clinical	151	SZ v OD	98.0

Table 4: Comparison with State-of-the-Art Methods using Brain Imaging Data

SZ = Schizophrenia, BD = Bipolar Disorder, HC = Healthy Controls, OD = Other Diseases

neural network with one layer, **MLP**, and a deep neural network consisting of four layers, **DNN**. Note that we can consider the comparison between RobIn, SENN, DNN, and MLP an ablation study where we remove the self-attention mechanism, the squeeze and excitation layers and the additional hidden layers respectively.

We report the metrics: 'Accuracy', 'Precision', 'Sensitivity', 'Specificity', 'F1
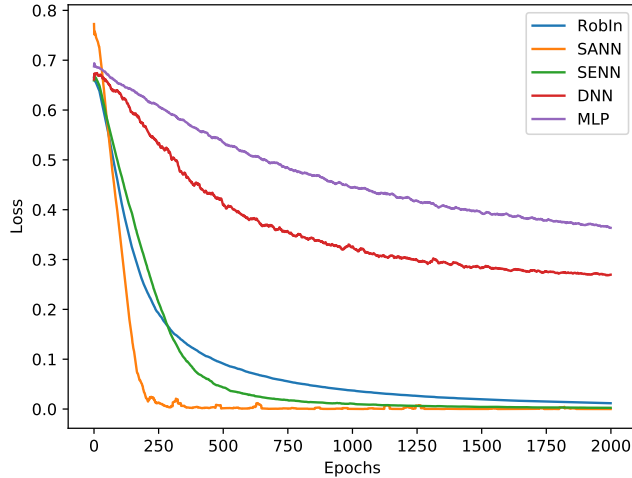
Figure 3: Loss graphs of compared models.

Score' and 'AuC', where

$$\text{Accuracy} = \tfrac{\text{TP+TN}}{\text{TP+TN+FP+FN}}, \tag{10}$$

$$\text{Precision} = \tfrac{\text{TP}}{\text{TP+FP}}, \tag{11}$$

$$\text{Sensitivity} = \tfrac{\text{TP}}{\text{TP+FN}}, \tag{12}$$

$$\text{Specificity} = \tfrac{\text{TN}}{\text{TN+FP}}, \tag{13}$$

are calculated via true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). From these measures we obtain the F1 Score and AuC:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}. \tag{14}$$

The AuC is the area under a receiver operating characteristic (ROC) curve, plot on the unit square with axes sensitivity and (1-specificity). An AuC of 0.5 indicates random guessing.

6.2. Model Performance

The effectiveness of the compared models on the 10-fold cross-validation setting is evaluated in Table 2. RobIn attains the best results across the most
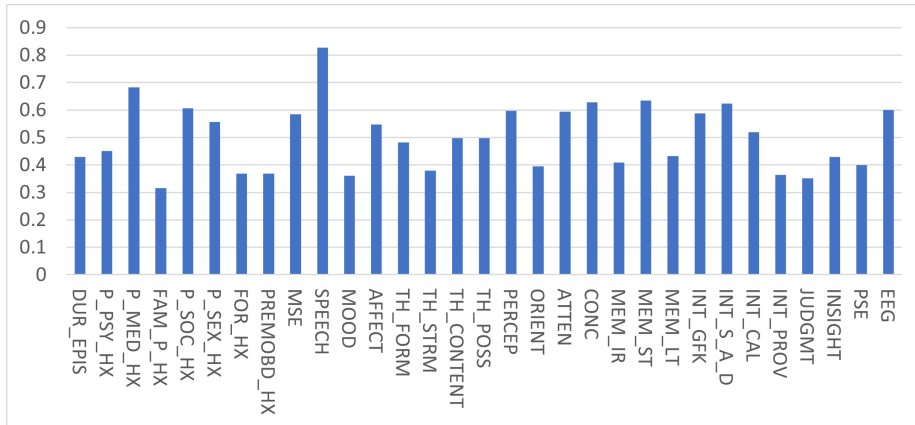
20

Figure 4: Global Feature Importance of RobIn: as expected, no features are entirely discarded, but certain features such as speech, past medical history and concentration are thought to be important.
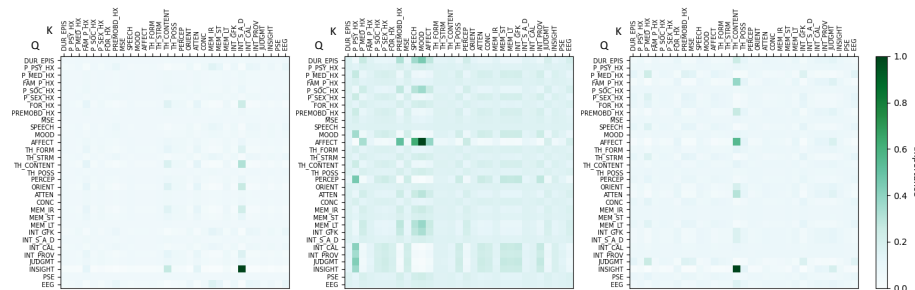

Figure 5: Heatmaps generated by the self-attention module. Darker squares indicate higher importance assigned by the queries (rows) to the keys (columns).

descriptive measures: accuracy, F1-score and AuC. RobIn outperforms MLP and DNN to a statistically significant level on all three measures, and SVM on Accuracy and F1-score. Given that the decision tree also obtains a good performance, it is very difficult to get a statistically significant result. Nonetheless, we get a statistically significant improvement on F1-score at a 95% significance level, and get p-values of 0.086 and 0.121 for accuracy and AuC, respectively.

The effectiveness of the compared models over 25 runs on the 50/50 split is evaluated in Table 3. These results are shown as a pre-cursor to the robustness tests in Section 6.5. With a much smaller training set, we expect the traditional models to outperform the deep models. However, with the same random seed, RobIn outperforms all other models on Accuracy and AuC. SENN achieves a higher F1 score than RobIn; however, it does so with a very high standard deviation. Although our results are not statistically significantly better than the decision tree, Section 6.5 shows that the scores from the decision tree and SANN are unreliable because they over-predict a positive diagnosis because the data set has more positive samples than negative. This means that the performance of these methods has been artificially inflated. Although they are capable of predicting from an i.i.d. test set, they cannot be trusted in a clinical setting.

We also provide loss curves of compared models in Figure 3. MLP and DNN are not able to converge to a low loss, indicating that they underfit the training data. On the other hand, SANN reduces towards zero very sharply, but then appears unstable. This indicates that training data is being memorised. RobIn and SENN appear to converge in an appropriate fashion, with RobIn having a shallower slope after 300 epochs, indicating it is least likely to be overfitting.

6.3. Comparison with Brain-Imaging Research

To place our work into context of the ongoing research of using artificial intelligence to diagnose schizophrenia, in Table 4, we compare with a selection of state-of-the-art works that were conducted using brain imaging data. For a comprehensive study, see (Cortes-Briones et al., 2021).

Most data sets are relatively small, with the second highest number of samples being 211. The exception is (Oh et al., 2020), who combine different publicly available data sets. To release the full potential of modern machine learning algorithms, large amounts of data are required. However, brain imaging data is prohibitive to obtain. We believe that out work establishes a proof-of-concept for the use of clinical data to diagnose schizophrenia, which should be easier to obtain and process, allowing for significantly larger data sets to be constructed in the future.

In terms of performance, RobIn applied to clinical data exceeds the performance of all image classifiers. Most brain-imaging methods are only required to discriminate between patients with schizophrenia and healthy controls, whereas we have a more difficult test setting as our control group consists of patients with other mental illnesses. This demonstrates the potential of involving clinical data within diagnostic systems, either alone or in combination with brain-imaging techniques.

6.4. Interpretability

6.4.1. Global Attribute Importance

An example of the global attribute importance is shown in Figure 4. At this stage of the network, the squeeze and excitation mechanism needs to consider all attributes, and it therefore does not give any feature a particularly low score. However, we can see that the model determines speech ability to be very important in finding a diagnosis, while other attributes like past medical history, concentration, short-term memory, and EEG scan are also deemed important.

These scores remain identical for test data that is considered. This allows our system to remain robust to different stress tests that we conduct, which indicates that this global attribute importance is essential for the model to generalise to real-world data.
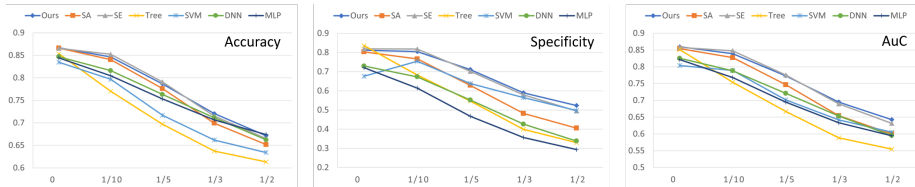
23

Figure 6: Robustness comparisons with the addition of noise, $X \sim \mathcal{N}(\mu = 0, \sigma^2)$, where the x-axis is σ^2.
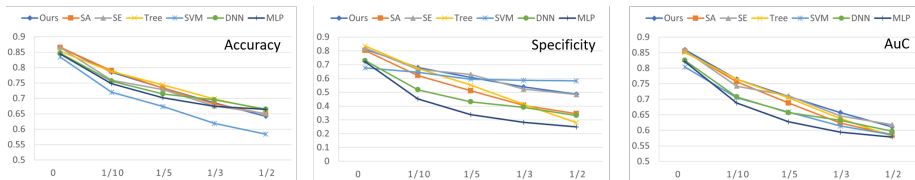

Figure 7: Robustness comparisons with the removal of data points where the x-axis signifies the fraction of values that were removed from the test data.

6.4.2. Attribute Interactivity

In Figure 5, we see examples of the attention activations of the self-attention mechanism. These heatmaps visualise the features that are deemed important to arrive at a diagnosis. In particular, we can observe that a model often finds that the insight of a patient is a useful query.

Note that because of the design of RobIn, more polarised activations are obtained because we connect the first layer with the final layer. In self-attention frameworks, the first layer activations are usually averaged while final layer activations are more polarised. However, this is not beneficial for interpreting feature importance because by the time polarised heatmaps are obtained, features have already undergone several layers of abstraction.

6.5. Stress Tests

DNNs are difficult to implement in the real world because, even though they perform well on unseen samples from the same data set, they struggle to generalise to unseen samples from a different source (D'Amour et al., 2020).
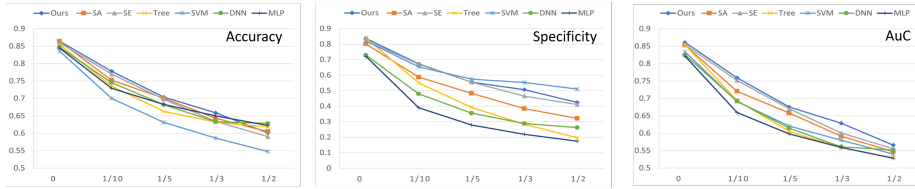
Figure 8: Robustness comparisons with the addition of noise and the removal of data points. The x-axis signifies the variance of the normal curve from which the noise was sampled and then added to the test data, and the fraction of values that were removed from the test data.

This is because the test set is independent and identically distributed (iid) to training set, but real-world data rarely has an identical distribution to small-sample collected data sets. We design three stress tests to randomly alter the distribution of the test set at varying strengths to evaluate the robustness of all models:

1. **Noise**: We sample noise $X \sim \mathcal{N}(\mu = 0, \sigma^2)$ and add this noise distribution to the test data to simulate different ways that a clinician might record a patient observation

2. **Data Erasing**: We randomly eliminate a fraction $1/m$ of the test data to simulate different ways that missing values could impact performance

3. **Noise + Data Erasing**: We combine 1) and 2) to provide a challenging, realistic test of model robustness

All models are evaluated on the original 50/50 train/test split with the same random seed, with the test set altered via the aforementioned perturbations. For each stress-test, we evaluate $\sigma^2, 1/m$ at values $\{1/10, 1/5, 1/3, 1/2\}$. Note that all perturbations are applied after the initial training data has been normalised between 0 and 1.

The performances under stress tests are presented in Figures 6 - 8. We found that accuracy and F1-score were poor indicators of model performance under these stress-tests because the data set is skewed with more positive samples than

negative ones. As data became more noisy, models would often over-predict a positive diagnosis to obtain the highest accuracy. This can be seen from the specificity graphs, where the decision tree and the neural networks without squeeze and excitation modules all drop below 0.35. This indicates that models have learnt to make decisions based on the data set characteristics, rather than the features provided. We report the accuracy and specificity to present this phenomenon, but base our robustness evaluation on the AuC - the only reliable evaluation metric because it is a function of specificity.

The figures show that the decision tree and SANN are not deployable because they show a tendency to over-predict a positive diagnosis. Those models are not suitable for real-world deployment because, counter to our data set, the proportion of people who actually have schizophrenia is very low.

Across all three stress tests, RobIn and SENN suffer least from perturbations of the test data, indicating that they are most suitable for deployment in the real world. This gives credence to our claim that the squeeze and excitation mechanism improves model robustness.

6.5.1. Additional Noise

As seen in Figure 6, when noise $X \sim \mathcal{N}(\mu = 0, 1/2)$ is added, the decision tree does barely better than random guessing with an AuC of 0.555, whereas RobIn maintains an AuC of 0.643 which is very impressive given the severity of the perturbation. Moreover, when $X \sim \mathcal{N}(\mu = 0, 1/5)$ perturbs the data, the AuC of RobIn falls only by 0.088, in contrast to the 0.186 drop by the decision tree. In the realistic scenario when $\sigma^2 = 1/10$, all methods using attention see very little performance loss, whereas again the decision tree performs worst with a loss of 0.099, more than RobIn suffers with σ^2 is twice as large.

6.5.2. Removed Values

In Figure 7, the removal of values appears to be a much more difficult task, as shown by the sharp drop in AuC even when only 1/10 values are removed. RobIn and SENN maintain the highest AuCs throughout the stress levels. All models
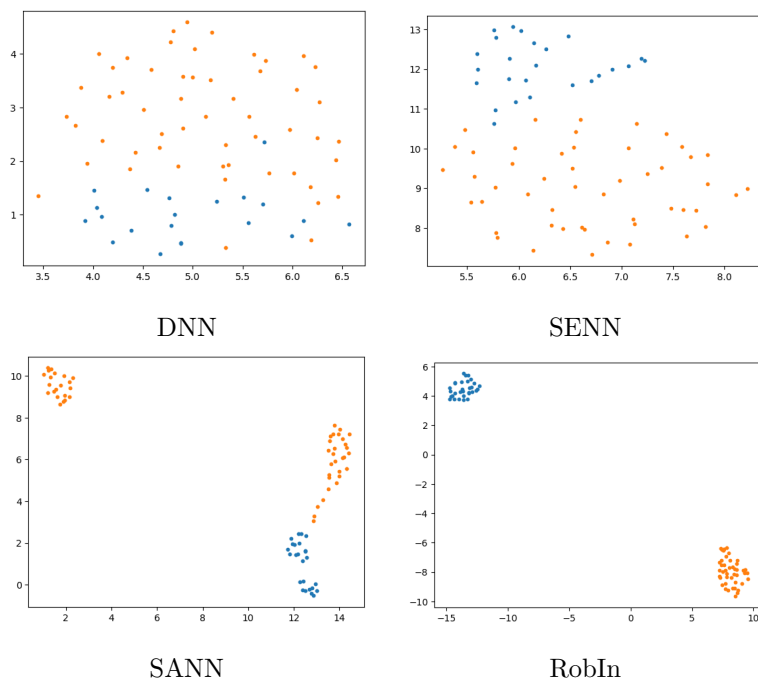
Figure 9: UMAP results across tested neural network models. Blue dots are negative samples, orange dots are positive samples.

struggle with half of the values removed and are close to random guessing. At the 1/3 stress level, RobIn has 0.011 greater AuC than the next best, SENN. It is worth noting that the specificity of the SVM remains fairly constant as more values are removed; however, it has by far the lowest accuracy at all stress levels and only consistently outperforms MLP on AuC, despite the high specificity. It over-predicts a negative diagnosis, resulting in a high number of true negatives and a low number of false positives.

6.5.3. Combination

The combination of the previous two stress tests in Figure 8 is the most realistic test, where noise simulates difference of opinion between clinicians and the data point removal simulates values that are often missing. RobIn outperforms all non-attention based methods on AuC at $1/10, 1/5$, and $1/3$ stress levels,

with a 95% confidence interval. Furthermore, in the 1/3 level, RobIn attains a 0.629 ± 0.0257 AuC, a statistically significant improvement on the next best method, SENN, which attains 0.601 AuC. RobIn even maintains the highest accuracy until the 1/3 level, despite the propensity of other methods to cheat. It is clear that RobIn is the most robust method and most suitable for real-world deployment.

6.5.4. Discussion of Stress Tests

The weakness of traditional models at handling these stress tests demonstrates the necessity of conducting evaluation to test robustness of models. Machine learning models often overfit on the data distribution they are trained on, and therefore struggle to generalise to unseen distributions, particularly when training data is small. Conducting these stress tests has highlighted significant flaws in models such as MLPs, SVMs and decision trees, despite the fact they appear to work well on the uncorrupted data.

Furthermore, stress tests appear to be a strong method to evaluate a model's ability to reason with imbalanced data. As data perturbation gets stronger, models that are less robust will classify samples into the majority class, whereas more robust models will continue to reason with the corrupted data. This indicates that more robust models have a more sophisticated ability to retrieve correlations between subsets of features and outputs, so are not overly reliant on individual features that can become corrupted.

6.6. Feature Visualisation

In Figure 9, we use UMAP (McInnes et al., 2020) to visualise the 2-dimensional manifold of feature representations of different neural networks.

In the top left, the output of the DNN is disorganised and there is a lot of crossover between the two classes. SENN has a clearer boundary between classes, but there is only a small distance between the two classes; many samples being nearer to samples from the opposite class than their own. In the bottom left, SANN does a better job of clustering samples but there are a subset of

positive samples very close to the negative samples. This shows that the model is unsure about a vast percentage of positive samples.

On the other hand, RobIn does the best job of categorically separating positive and negative samples, with a large distance between the two clusters that are formed. This highly organised output manifold demonstrates the ability of RobIn to categorise data, while also demonstrating the benefits of combining the two attention mechanisms. Squeeze and excitation forms a strong decision boundary whilst self-attention contributes to clustering similar data.

7. Conclusion and Discussions

Applications of machine learning to schizophrenia diagnosis have struggled to bridge the gap between theory and application. To combat this, we have collected a data set that complies with current clinical guidelines outlined by DSM-5. This data will be of a similar format to records that are already kept by clinicians for patient monitoring.

We have developed a robust, interpretable network that outperforms other machine learning methods for the diagnosis of schizophrenia. Our network contains a squeeze and excitation mechanism that works on a global scale to give an overview of feature importance and a self-attention mechanism that works on a local level to assess feature interactivity. We have also comprehensively tested all methods with stress tests to evaluate their potential to generalise to new data sources. Our proposed method is the most robust of all compared models on all three stress tests. The evidence suggests that this is because of the static weights learnt by the squeeze and excitation mechanism. Therefore, we conclude that self-attention contributes to model interpretability and squeeze and excitation contributes to model robustness, whilst both contribute to overall performance.

For future work, we wish to collect a larger data set of clinical data from several different test sites to further evaluate the performance of RobIn, so that RobIn or similar frameworks could be used in real-world clinical decision

making. There are two streams of work that we are particularly interested in: a) *combining clinical and brain-imaging features* - we have demonstrated that deep learning can use clinical data to diagnose schizophrenia, we want to fuse clinical data with associated image data to allow the model to have all of the available information when making a decision, and to use cross-attention to link clinical features with highlighted areas on the fMRI scan; b) *attentional machine feedback for clinicians* - we wish to analyse high-attention features that RobIn finds when training on a larger data set to identify particularly influential features. Current works that utilise brain-imaging aim to identify regions of importance to help improve clinician's ability of diagnosing from brain images themselves. In a similar way, we wish to identify behavioural features that clinicians should pay special focus to when interacting with a client for the first time.

Acknowledgement

References

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Van Esesn, B. C., Awwal, A. A. S., & Asari, V. K. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, .

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). doi:`10.1176/appi.books.9780890425596. dsm05`.

Barros, C., Silva, C. A., & Pinheiro, A. P. (2021). Advanced eeg-based learning approaches to predict schizophrenia: Promises and pitfalls. *Artificial Intelligence in Medicine*, *114*, 102039. URL: `https://www.sciencedirect.`

com/science/article/pii/S0933365721000324. doi:`https://doi.org/10.1016/j.artmed.2021.102039`.

Burns, J. K. (2012). The social determinants of schizophrenia: an african journey in social epidemiology. *Public Health Reviews*, *34*, 8.

Calhas, D., Romero, E., & Henriques, R. (2020). On the use of pairwise distance learning for brain signal classification with limited observations. *Artificial Intelligence in Medicine*, *105*, 101852. URL: `https://www.sciencedirect.com/science/article/pii/S0933365719311261`. doi:`https://doi.org/10.1016/j.artmed.2020.101852`.

Cheung, B. L. P., & Dahl, D. (2018). Deep learning from electronic medical records using attention-based cross-modal convolutional neural networks. In *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)* (pp. 222–225). doi:`10.1109/BHI.2018.8333409`.

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., & Stewart, W. (2016). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 3504–3512). Curran Associates, Inc.

Cortes-Briones, J. A., Tapia-Rivas, N. I., D'Souza, D. C., & Estevez, P. A. (2021). Going deep into schizophrenia with artificial intelligence. *Schizophrenia Research*, .

Cuthbert, B. N. et al. (2014). The RDoC framework: continuing commentary. *World Psychiatry*, *13*, 196.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D. et al. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, .

Duan, C., Cui, L., Chen, X., Wei, F., Zhu, C., & Zhao, T. (2018). Attention-fused deep matching network for natural language inference. In *IJCAI* (pp. 4033–4040).

Gulsuner, S., Stein, D., Susser, E., Sibeko, G., Pretorius, A., Walsh, T., Majara, L., Mndini, M., Mqulwana, S., Ntola, O. et al. (2020). Genetics of schizophrenia in the south african xhosa. *Science*, *367*, 569–573.

Häfner, H., & Maurer, K. (2006). Early detection of schizophrenia: current evidence and future perspectives. *World psychiatry : official journal of the World Psychiatric Association (WPA)*, *5*, 130–138. 17139339[pmid].

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huk Park, D., Anne Hendricks, L., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., & Rohrbach, M. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8779–8788).

Kalmady, S. V., Greiner, R., Agrawal, R., Shivakumar, V., Narayanaswamy, J. C., Brown, M. R. G., Greenshaw, A. J., Dursun, S. M., & Venkatasubramanian, G. (2019). Towards artificial intelligence in mental health by improving schizophrenia prediction with multiple brain parcellation ensemble-learning. *npj Schizophrenia*, *5*, 2. doi:`10.1038/s41537-018-0070-8`.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, *17*, 195. doi:`10.1186/s12916-019-1426-2`.

Khoury, B., Kogan, C., & Daouk, S. (2017). International classification of diseases 11th edition (icd-11). In V. Zeigler-Hill, & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 1–6). Cham: Springer International Publishing. doi:`10.1007/978-3-319-28099-8_904-1`.

Lauritsen, S. M., Kalør, M. E., Kongsgaard, E. L., Lauritsen, K. M., Jørgensen, M. J., Lange, J., & Thiesson, B. (2020). Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artificial Intelligence in Medicine*, *104*, 101820. URL: `https://www.sciencedirect.com/science/article/pii/S0933365719303173`. doi:`https://doi.org/10.1016/j.artmed.2020.101820`.

Lesh, T. A., Tanase, C., Geib, B. R., Niendam, T. A., Yoon, J. H., Minzenberg, M. J., Ragland, J. D., Solomon, M., & Carter, C. S. (2015). A multimodal analysis of antipsychotic effects on brain structure and function in first-episode schizophrenia. *JAMA psychiatry*, *72*, 226–234.

Li, G., Han, D., Wang, C., Hu, W., Calhoun, V. D., & Wang, Y.-P. (2020). Application of deep canonically correlated sparse autoencoder for the classification of schizophrenia. *Computer Methods and Programs in Biomedicine*, *183*, 105073. URL: `https://www.sciencedirect.com/science/article/pii/S0169260719307655`. doi:`https://doi.org/10.1016/j.cmpb.2019.105073`.

Li, X., Zhang, X., Zhu, J., Mao, W., Sun, S., Wang, Z., Xia, C., & Hu, B. (2019). Depression recognition using machine learning methods with different feature generation strategies. *Artificial Intelligence in Medicine*, *99*, 101696. URL: `https://www.sciencedirect.com/science/article/pii/S0933365719300296`. doi:`https://doi.org/10.1016/j.artmed.2019.07.004`.

Mall, S., Platt, J. M., Temmingh, H., Musenge, E., Campbell, M., Susser, E., & Stein, D. J. (2020). The relationship between childhood trauma and schizophrenia in the genomics of schizophrenia in the xhosa people (sax) study in south africa. *Psychological medicine*, *50*, 1570.

Matheson, S., Shepherd, A. M., Pinchbeck, R., Laurens, K., & Carr, V. J. (2013). Childhood adversity in schizophrenia: a systematic meta-analysis. *Psychological medicine*, *43*, 225.

Matsubara, T., Tashiro, T., & Uehara, K. (2019). Deep neural generative model of functional mri images for psychiatric disorder diagnosis. *IEEE Transactions on Biomedical Engineering*, *66*, 2768–2779. doi:`10.1109/TBME.2019.2895663`.

McInnes, L., Healy, J., & Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction. `arXiv:1802.03426`.

Nature Biomedical Engineering (2018). Towards trustable machine learning. *Nature Biomedical Engineering*, *2*, 709–710. doi:`10.1038/s41551-018-0315-x`.

Niu, Y.-W., Lin, Q.-H., Qiu, Y., Kuang, L.-D., & Calhoun, V. D. (2019). Sample augmentation for classification of schizophrenia patients and healthy controls using ica of fmri data and convolutional neural networks. In *2019 Tenth International Conference on Intelligent Control and Information Processing (ICICIP)* (pp. 297–302). doi:`10.1109/ICICIP47338.2019.9012169`.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future - big data, machine learning, and clinical medicine. *The New England journal of medicine*, *375*, 1216–1219. doi:`10.1056/NEJMp1606181`. 27682033[pmid].

Oh, J., Oh, B.-L., Lee, K.-U., Chae, J.-H., & Yun, K. (2020). Identifying schizophrenia using structural mri with a deep learning algorithm. *Frontiers in Psychiatry*, *11*, 16. doi:`10.3389/fpsyt.2020.00016`.

Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Y Hammerla, N., Kainz, B. et al. (2018). Attention u-net: Learning where to look for the pancreas. *arXiv e-prints*, (pp. arXiv–1804).

Organisciak, D., Ho, E. S. L., & Shum, H. P. H. (2020). Makeup style transfer on low-quality images with weighted multi-scale attention. In *Proceedings of the 2020 International Conference on Pattern Recognition* ICPR '20.

Palaniyappan, L., Deshpande, G., Lanka, P., Rangaprakash, D., Iwabuchi, S., Francis, S., & Liddle, P. F. (2019). Effective connectivity within a triple network brain system discriminates schizophrenia spectrum disorders from psychotic bipolar disorder at the single-subject level. *Schizophrenia research*, *214*, 24–33.

Paschali, M., Naeem, M. F., Simson, W., Steiger, K., Mollenhauer, M., & Navab, N. (2019). Deep learning under the microscope: Improving the interpretability of medical imaging neural networks. *arXiv preprint arXiv:1904.03127*, .

Phang, C.-R., Noman, F., Hussain, H., Ting, C.-M., & Ombao, H. (2020). A multi-domain connectome convolutional neural network for identifying schizophrenia from eeg connectivity patterns. *IEEE Journal of Biomedical and Health Informatics*, *24*, 1333–1343. doi:`10.1109/JBHI.2019.2941222`.

Qi, J., & Tejedor, J. (2016). Deep multi-view representation learning for multi-modal features of the schizophrenia and schizo-affective disorder. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 952–956). doi:`10.1109/ICASSP.2016.7471816`.

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., & Shlens, J. (2019). Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems* (pp. 68–80). Curran Associates, Inc. volume 32.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K. et al. (2020). Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, .

Rangarajan, A. K., & Ramachandran, H. K. (2021). A preliminary analysis of ai based smartphone application for diagnosis of covid-19 using chest x-ray images. *Expert Systems with Applications*, (p. 115401).

Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. (2017). Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, *21*, 4–21. doi:`10.1109/JBHI.2016.2636665`.

Raza, M., Awais, M., Ellahi, W., Aslam, N., Nguyen, H. X., & Le-Minh, H. (2019). Diagnosis and monitoring of alzheimer's patients using classical and deep learning techniques. *Expert Systems with Applications*, *136*, 353–364.

Rebane, J., Samsten, I., & Papapetrou, P. (2020). Exploiting complex medical data with interpretable deep learning for adverse drug event prediction. *Artificial Intelligence in Medicine*, *109*, 101942.

Salvador, R., Canales-Rodríguez, E., Guerrero-Pedraza, A., Sarró, S., Tordesillas-Gutiérrez, D., Maristany, T., Crespo-Facorro, B., McKenna, P., & Pomarol-Clotet, E. (2019). Multimodal integration of brain images for mri-based diagnosis in schizophrenia. *Frontiers in Neuroscience*, *13*. URL: `https://www.frontiersin.org/article/10.3389/fnins.2019.01203`. doi:`10.3389/fnins.2019.01203`.

Shi, J., Zheng, X., Li, Y., Zhang, Q., & Ying, S. (2018). Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of alzheimer's disease. *IEEE Journal of Biomedical and Health Informatics*, *22*, 173–183. doi:`10.1109/JBHI.2017.2655720`.

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*, 1589–1604. doi:`10.1109/JBHI.2017.2767063`.

da Silva, D. B., Schmidt, D., da Costa, C. A., da Rosa Righi, R., & Eskofier, B. (2021). Deepsigns: A predictive model based on deep learning for the early detection of patient health deterioration. *Expert Systems with Applications*, *165*, 113905.

Tandon, N., & Tandon, R. (2019a). Machine learning in psychiatry-standards and guidelines. *Asian journal of psychiatry*, *44*, A1.

Tandon, N., & Tandon, R. (2019b). Using machine learning to explain the heterogeneity of schizophrenia. realizing the promise and avoiding the hype. *Schizophrenia Research*, *214*, 70–75.

Varese, F., Smeets, F., Drukker, M., Lieverse, R., Lataster, T., Viechtbauer, W., Read, J., Van Os, J., & Bentall, R. P. (2012). Childhood adversities increase the risk of psychosis: a meta-analysis of patient-control, prospective-and cross-sectional cohort studies. *Schizophrenia bulletin*, *38*, 661–671.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* NIPS'17 (pp. 6000–6010). USA.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. *International Conference on Learning Representations*, . Accepted as poster.

Vellido, A. (2019). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, (pp. 1–15).

Wainberg, M., Merico, D., Delong, A., & Frey, B. J. (2018). Deep learning in biomedicine. *Nature biotechnology*, *36*, 829–838.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., & Tang, X. (2017). Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3156–3164).

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, *12*, e0174944.

Winterburn, J. L., Voineskos, A. N., Devenyi, G. A., Plitman, E., de la Fuente-Sandoval, C., Bhagwat, N., Graff-Guerrero, A., Knight, J., & Chakravarty, M. M. (2019). Can we accurately classify schizophrenia patients from healthy controls using magnetic resonance imaging and machine learning? A multi-method and multi-dataset study. *Schizophrenia Research*, *214*, 3–10. doi:`10.1016/j.schres.2017.11.038`.

World Health Organization (2011). WHO methods and data sources for global burden of disease estimates 2000-2011, .

Yan, W., Plis, S., Calhoun, V. D., Liu, S., Jiang, R., Jiang, T., & Sui, J. (2017). Discriminating schizophrenia from normal controls using resting state functional network connectivity: A deep neural network and layer-wise relevance propagation method. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6). doi:`10.1109/MLSP.2017.8168179`.

Zeiser, F. A., da Costa, C. A., de Oliveira Ramos, G., Bohn, H. C., Santos, I., & Roehe, A. V. (2021). Deepbatch: A hybrid deep learning model for interpretable diagnosis of breast cancer in whole-slide images. *Expert Systems with Applications*, *185*, 115586.

FEATURE	DESCRIPTION	VALUES
AGE	Age of patient	Age e.g. 23, 34
SEX	Sex of patient	Male, female
OCCUP_HX	Occupation history	Unemployed, occupation
MAR_STA	Marital status	Married, single, divorced, widow
DUR_EPIS	Episode duration (length of time the patient has suffered the symptoms)	Time in months
P_PSY_HX	Past psychiatric history	No, yes (e.g. rape, mental illness, etc.)
P_MED_HX	Past medical history	No, disease suffered in past (e.g. diabetes)
FAM_P_HX	Family psychiatric history	Yes, no
P_SOC_HX	Past social history	Yes, no
P_SEX_HX	Past sexual history	Normal, experience (e.g. masturbate, gonorrhea, etc.)
FOR_HX	Forensic history	Yes, no
PREMOB_HX	Pre-morbid history	Normal, introvert, extrovert, melancholic
MSE	Mental state examination	Kempt, unkempt, poor eye contact, restless
SPEECH	Speech status	Normal, reduced volume, mute, slurred, decreased tone, irrelevant, incoherent
MOOD	Mood of the patient at the time of report	Euphoric, neutral, happy, relaxed, fine/ok, worried, sad, irritable
AFFECT	Affect of the patient at time of report	Depressed, reactive, blunt, restricted, congruent, abnormal
TH_FORM	Thought form at time of report	Logical, abnormal
TH_STRM	Thought stream at time of report	Reduced, normal, increased
TH_CONTENT	Thought content at time of report	Persecutory delusion, auditory hallucination, normal, obsession, grandiose delusion, disorder
TH_POSSESSION	Thought possession at time of report	Impaired, normal
PERCEP	Perception at time of report	No, auditory hallucination, visual hallucination, tactile hallucination, olfactory hallucination, preoccupation
ORIENT	Time, place and position orientation at time of report	Oriented in TPP, no
ATTEN	Attention status at time of report	Rousable, poor
CONC	Concentration status at time of report	Good, reduced, poor
MEM_IR	Immediate recall status at time report	Good, fair, poor
MEM_ST	Short-term memory status at time report	Good, fair, poor
MEM_LT	Long-term memory status at time report	Good, fair, poor
INT_GFK	Intelligence test of general fund of knowledge	Good, fair, poor
INT_S_A_D	Intelligence test of similarity and difference	Good, fair, poor
INT_CAL	Intelligence test of arithmetic	Good, fair, poor
INT_PROV	Intelligence test of proverbs	Good, fair, poor
JUDGMT	Intellectual judgement status at time of report	Good, poor
INSIGHT	Insight status at time of report	Good, partial ,poor
PSE	Physical state examination status	Good, normal, pale
EEG	Electroencephalogram (to exclude brain tumour possibility status)	Normal, altered