CUTTING THROUGH THE HYPE: ARTIFICIAL INTELLIGENCE FOR CLINICAL DECISION SUPPORT IN PSYCHIATRY

A thesis submitted by

Matthew Squires

(BSc, MSc, MTeach(Sec))

For the award of

Doctor of Philosophy

Submitted: June, 2024

ABSTRACT

Mental health conditions are one of the most significant challenges to global society. This thesis explores the integration of artificial intelligence (AI) techniques in psychiatric research, focusing on enhancing the detection, diagnosis, and treatment of depression. While simultaneously exploring the implications of integrating AI into clinical practice. Through the use of empirical experiments the works contained within this thesis demonstrate the potential uses for AI in mental healthcare. These works show AI techniques can reliably predict treatment outcomes to repetitive transcranial magnetic stimulation (rTMS) above the existing state-of-the-art. Combining these methods with explainable AI (XAI) this work identifies candidate biomarkers indicative of treatment response to rTMS. Furthermore, this work shows predictive performance can be improved by using diversity enhanced training data. This thesis includes a novel method for enhancing the diversity of training data. Including experiments which demonstrate the possibility of synthetic data to improve dataset diversity. This thesis also presents a novel method for addressing label bias when detecting suicide risk on social media through semi-supervised deep label smoothing. Empirical experiments show this methods improves classification accuracy by leveraging fuzzy labels and Bayesian techniques. Put together, the research within this thesis highlights the transformative potential of AI in psychiatry, demonstrating the possibility of personalised psychiatry, advocating for innovative data augmentation and regularisation methods to improve model performance. By critically analysing these empirical experiments, this thesis examines the broader implications of AI in psychiatry. It places special emphasis on methods to ensure the ethical and equitable deployment of AI in mental healthcare.

CERTIFICATION OF THESIS

I Matthew Squires declare that the PhD Thesis entitled *Cutting Through the Hype: Artificial Intelligence for Clinical Decision Support in Psychiatry* is not more than 100,000 words in length including quotes and exclusive of tables, figures, appendices, bibliography, references, and footnotes.

This Thesis is the work of Matthew Squires except where otherwise acknowledged, with the majority of the contribution to the papers presented as a Thesis by Publication undertaken by the student. The work is original and has not previously been submitted for any other award, except where acknowledged.

Date: June 2024

Endorsed by:

Xiaohui Tao Principal Supervisor

Professor Raj Gururajan Associate Supervisor

Dr Xujuan Zhou Associate Supervisor

Professor Rajendra Acharya Associate Supervisor

Student and supervisors' signatures of endorsement are held at the University.

STATEMENT OF CONTRIBUTION

This section details the authors contributions to publications produced during the period of candidature:

Paper 1:

Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X., Acharya, U. R., & Li, Y. (2023). Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. *Brain informatics*, 10(1), 10. https://doi.org/10.1186/s40708-023-00188-6

Matthew Squires contributed 80% to this paper. Collectively, Xiaohui Tao, Soman Elangovan, Raj Gururajan, Xujuan Zhou U Rajendra Acharya and Yuefeng Li contributed the remainder.

Paper 2:

Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X., Li, Y., & Acharya, U. R. (2023). Identifying predictive biomarkers for repetitive transcranial magnetic stimulation response in depression patients with explainability. *Computer methods and programs in biomedicine*, 242, 107771. https://doi.org/10.1016/j.cmpb.2023.107771

Matthew Squires contributed 80% to this paper. Collectively, Xiaohui Tao, Soman Elangovan, Raj Gururajan, Xujuan Zhou, Yuefeng Li and U Rajendra Acharya contributed the remainder.

Paper 3:

Squires, M., Tao, X., Elangovan, S., Gururajan, R., Xie. H., Zhou, X., Li, Y., & Acharya, U. R. (2023). DE-CGAN: Boosting rTMS Treatment Prediction with Diversity Enhancing Conditional Generative Adversarial Networks *Under Review*,

https://doi.org/10.48550/arxiv.2404.16913

Matthew Squires contributed 80% to this paper. Collectively, Xiaohui Tao, Soman Elangovan, Raj Gururajan, Hoaran Xie, Xujuan Zhou, Yuefeng Li and U Rajendra Acharya contributed the remainder.

Paper 4:

Squires, M., Tao, X., Elangovan, S., Acharya, U. R., Gururajan, R., Xie, H & Zhou, X. (2024). Enhancing Suicide Risk Detection on Social Media through Semi-Supervised Deep Label Smoothing *Under Review*, https://doi.org/10.48550/arxiv.2405.05795

Matthew Squires contributed 80% to this paper. Collectively, Xiaohui Tao, Soman Elangovan, U Rajendra Acharya, Raj Gururajan, Hoaran Xie and Xujuan Zhou contributed the remainder.

ACKNOWLEDGEMENTS

I would like to express my thanks to those who have supported me throughout this PhD journey. This includes my supervisory team, Xiaohui Tao, Soman Elangovan, Rajendra Acharya, Raj Gururajan and Xujuan Zhou. Particularly, Professor Xiaohui Tao who has dedicated a significant amount of time to my academic development.

Additionally, I would like to acknowledge the support provided by Belmont private hospital including the support from Belmont Private Hospital team members, especially, Ms Mary Williams (CEO), Rachel Stark (Area Manager), Dr Mark Spelman (Psychiatrist), Dr Sean Gills (Psychiatrist), and Dr Tom Moore (Psychiatrist). Without their kind support, this work wouldn't be possible.

I would also like to thank my friends and family for the support and encouragement throughout this PhD. Most significantly, I would like to thank my wife for her understanding and patience during the times of intense focus and stress which has been invaluable.

Finally, I would like to thank all those who have supported me throughout all levels of my academic journey, each of these contributions have enabled me to reach this significant milestone in my research career.

This research has been supported by the Australian Government Research Training Program Scholarship.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
cGAN	Conditional General Adversarial Network
DL	Deep Learning
GAN	General Adversarial Network
GPT	Generative Pre-Trained Transformer
KNN	K- Neareast Neighbours
ML	Machine Learning
MLP	Multi Layer Perceptron
rTMS	Repetitive Transcranial Magnetic Stimulation
RQ	Research Question
SVM	Support Vector Machine
TRD	Treatment Resistant Depression
XAI	Explainable Artificial Intelligence

CHAPTER 1: INTRODUCTION

Mental health conditions significantly burden both society and the individuals who experience them. Globally, mental illness will effect 1 in 5 people (Timmons et al. 2022), with similar statistics observed in Australia (Kasturi et al. 2023). Of these mental health conditions depression places the largest economic burden. Annually, between 1 and 2 million Australians will experience episodes of diagnosable depression (Kasturi et al. 2023). This prevalence makes the investigation of safe and effective treatments a significant societal issue. Recent developments in artificial intelligence (AI) has fuelled interest in investigating the ways AI can be used to support the reduction of this significant disease burden.

AI offers promise as a technology to innovate and disrupt the existing model of mental healthcare. This potential ranges from better targeting treatments to the discovery of previously unidentified biomarkers of mental health conditions. Existing research is highlighting the ability of AI to equal or exceed human performance on a variety of health specific tasks. From the classification of potentially cancerous skin lesions (Furriel et al. 2024) to the analysis of chest x-rays (Miró Catalina et al. 2024). However, significant work is required to bridge the gap between controlled empirical experiments and real world clinical settings. This thesis aims to bridge the gap between research and clinical deployment by exploring the necessary work. It focuses on addressing key issues, including mitigating data bias, incorporating explainable AI (XAI), and handling statistical uncertainty.

1.1 Artificial Intelligence as a disruptor of mental health care

Since the industrial revolution no technological innovation is predicted to be as disruptive as what is expected of AI. From the first Generally Pre-trained Transformer (GPT) (Radford et al. 2018), to AI's breakthrough into the public discourse following the release of ChatGPT (OpenAI 2023). Broadly, "AI is the science and engineering of making intelligent machines, especially intelligent computer programs" (Xu et al. 2023, p.657). In a technical sense, this thesis uses the term AI to encompasses a set of technical tools such as machine learning (ML) and deep learning (DL). Visually this is represented in Figure 1.1.

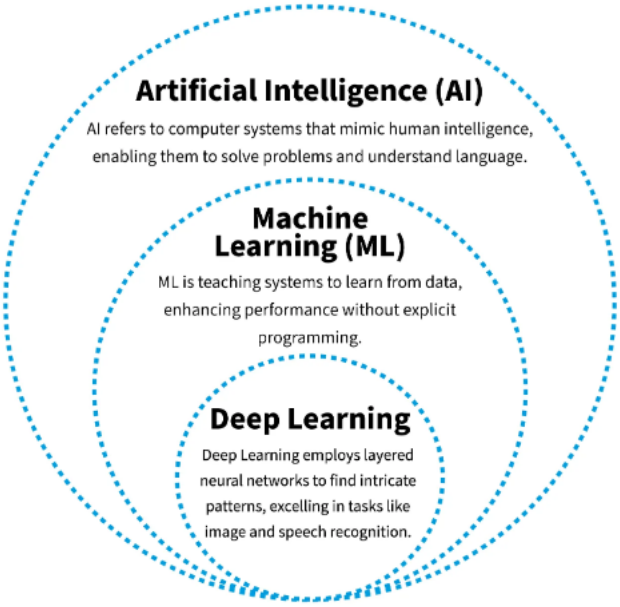

Figure 1.1: Artificial Intelligence and its antecedents. Source: (Interaction Design Foundation - IxDF 2016)

ML then refers to early shallow AI systems such as linear regression, support vector machines (SVM) or k-nearest neighbours (KNN). Most of the recent hype around AI has been driven by DL. DL builds upon the original artificial neural network (ANN) or multilayer perceptron (MLP). In his seminal work, Rosenblatt (1958) is attributed as one of the pioneers of AI through the proposal of the perceptron. The perceptron is one of the foundational features of modern DNN. Modern advances in hardware have allowed for the

layering of perceptrons in such a way it allows for the learning of complex multivariate relationships beyond what existing shallow networks are capable of. These large networks then become highly parameterised with complex foundation models such as ChatGPT-4 (OpenAI 2023) which is believed to include 1.8 trillion paramaters.

The influence and disruption caused by AI is so significant it could be defined as the fourth industrial revolution (Girasa 2020). Disruptive innovation is loosely defined as processes that initially focus on new markets, which are initially inferior to the incumbent, but eventually evolve to meet and exceed the needs of the consumer. (Christensen et al. 2018, Si & Chen 2020). Such technological innovations can challenge existing ways of doing such that in extreme cases the existing methods are no longer valued. Under this definition we can posit a future where AI challenges the existing status quo such that to do health care without the aid of some form of AI becomes unrealistic.

Disruptive innovation is distinct from social disruption where a disruptor, such as a technology, impacts society in such a way that society could not continue without change (van de Poel et al. 2023). Thus we can see at both a business and societal level AI is emerging as an innovation to challenge existing ways of being such that old methods will eventually by outdated and superseded. This contention is echoed in Păvăloaia & Necula (2023), as existing methods are superseded by technologies with entirely new characteristics whose capabilities eventually existing ways of doing. In these preliminary stages of the AI revolution humans have the potential to shape how AI systems are used. This motivates the central theme of this thesis: if AI is to become widespread, it must be used to enhance social good. From a practitioner's perspective, this means improving patient outcomes, mitigating bias, and ensuring that AI benefits all members of society.

AI has captured the attention of both the public and the research community for its expected impact on society. The hype associated with the use of AI, especially in healthcare, is worth investigating. Strange (2024) argues that both positive hype and concern raising about the nature of the use of AI in healthcare are equally unhelpful. Strange (2024) contends these dichotomous positions take away from the nuance required to explore the actual effects of AI in healthcare. Similarly, Xu et al. (2023) argues that high expectations of the opportunities presented by AI are useful for providing momentum for AI funding. However, this leaves AI researchers in the position of maintaining balanced and

realistic expectations. This thesis then moves away from hype to providing a balanced presentation of the strengths, limitations and potential solutions AI in psychiatry could provide. Thus, it is the nuanced discussion of AI applications in healthcare that allows AI to fulfill its potential (Xu et al. 2023).

1.2 Personalised Psychiatry and Trustworthy AI

Personalised medicine utilises the power of individual differences between patients. By using data to quantify patient-to-patient variability, personalised medicine can target treatments based on individual differences (Stefanicka-Wojtas & Kurpas 2023, Suwinski et al. 2019). This model moves away from traditional prescription methods based on population averages. Figure 1.2 demonstrates visually the contrast between the existing healthcare model and the potential of AI-assisted personalised treatment prescription. Vicente et al. (2020) outline their core vision for the "next generation" of healthcare. Under the model, by 2030 they hope healthcare systems are equipped to deliver, "personally tailored, optimised health promotion and disease prevention, diagnosis, and treatment for the benefits of patients" (Vicente et al. 2020, p.2). Fulfilling this outlined goal would truly be disruptive innovation as AI tailored care would make former models of care redundant and improve outcomes for patients.

Despite the ambitious goals of personalised medicine, the progress towards personalised healthcare is not being mirrored in psychiatry:

> "Unlike the high-profile breakthroughs made in personalized medicine, progress in psychiatry, which relies mainly on subjective methods of assessment and firsthand accounts for diagnosis, has lagged behind in delivering personalized treatments. Paradoxically, psychiatry is a field that could benefit greatly from more personalized approaches, owing to the wide heterogeneity of symptoms within individual disorders. Many psychiatric disorders are complex and can be associated with numerous, often thousands of, genetic variants, each, however, with a small effect. Polygenicity and high heterogeneity in psychiatric disorders, combined with environmental and epigenetic effects, suggest the

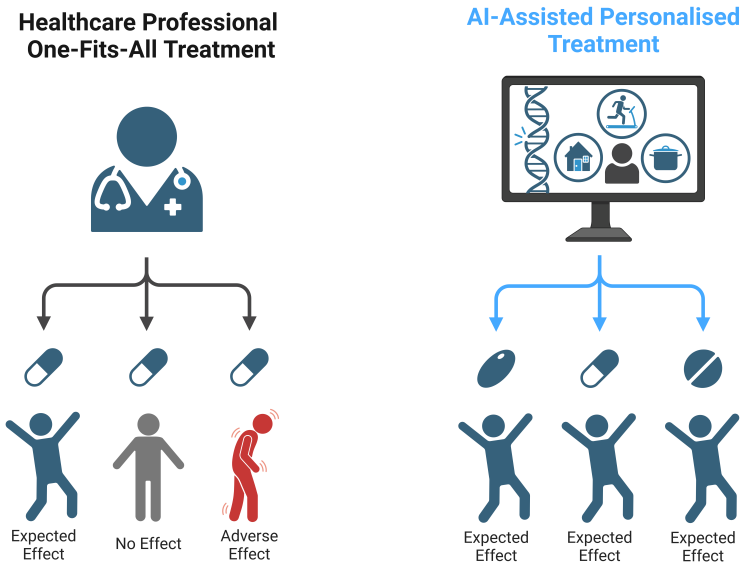

Figure 1.2: A Comparison Between Personalised Medicine and Traditional Healthcare

> need to apply different approaches and lines of action to shaping personalized psychiatry." (excerpt from *'The right treatment for each patient: unlocking the potential of personalized psychiatry'* 2023)

The significant heterogeneity within psychiatric disorders raises another important consideration in the pursuit of personalised psychiatry. For AI to be effectively deployed in this field, the systems must be deemed trustworthy. Diversity, non-discrimination, and fairness are core components of trustworthy AI systems (Díaz-Rodríguez et al. 2023, Cannarsa 2021).

However, small and homogeneous samples may limit the generalisability of results to all members of the population. This is especially true in psychiatry where small sample sizes and lack of large diverse datasets have somewhat limited the application of AI systems. Data augmentation then is one potential tool for improving the trustworthiness of AI systems (Díaz-Rodríguez et al. 2023).

Psychiatry is ideally suited to the hype of AI personalised care. However, further work is required to match the breakthroughs seen in medicine. Additionally, it is imperative AI systems deployed in psychiatry are trustworthy. This clear and obvious gap in the litera-

ture motivates the exploration of AI and its use to personalise psychiatric care. Specifically looking at depression as the mental illness with the greatest burden on Australian society.

Section 1.1 emphasises AI is poised to be a significant disruptor to the healthcare industry. However, a critical and nuanced discussion of both the strengths and limitations of AI in psychiatry is required to ensure the technology meets the needs of the community.

1.3 Depression and rTMS

Depression and suicide are common and expensive mental health issues. Given this prevalence researchers have explored many avenues to aid in the treatment and diagnosis of these issues. Barriers such as reduced access to mental health services (Fitzpatrick et al. 2021), and stigma associated with seeking mental health care (Gaur et al. 2019) are among the factors which prevent people from seeking help.

The global impact of these mental health conditions require the investigation of effective treatments. Frontline treatments to depression vary from pharmacological antidepressants (Thornton et al. 2023), psychological treatments (Malhi et al. 2020) or a combination of the two. However, these frontline treatments vary in their effectiveness. Some patients will benefit greatly from their treatments while others will see little to no improvement in their symptoms.

When Depression is recurring or difficult to treat is referred to as Treatment-resistant depression (Johnston et al. 2019, TRD). Due to its complexity there is no collectively agreed upon definition of TRD. Within the literature most commonly, TRD is defined as depression which does not respond to at least 2 treatments (Hannah et al. 2023). The pursuit of ineffective treatments associated with TRD provides a significant burden to patients. Johnston et al. (2019) showed greater treatment resistance was associated with increased costs and a reduced health related quality of life.

Recent research suggests that what is currently diagnosed as depression under the existing diagnostic model is likely not a single condition, but multiple distinct conditions. The significant heterogeneity offers one explanation for why some patients benefit from

treatment while others do not. Potentially, AI can reimagine existing diagnostic categories, where rTMS is seen as a potential treatment for some subtypes of depression. The challenge for rTMS and other brain stimulation techniques is as follows:

> If brain stimulation is going to become a reliable frontline approach, clinicians will have to determine, among other things, who responds best to which treatments. (Yam 2024, p.4)

rTMS involves electromagnetic stimulation of the brain through coils applied to the patient's scalp (Razza et al. 2018). The applied magnetic field induces an electrical current within the brain which over time alters the underlying structures (George & Taylor 2014). rTMS sessions vary in their intensity and duration. Patients are prescribed treatment times in advance with a variety of treatment times being available to doctors (Fitzgerald et al. 2020).

1.4 Uncertainty quantification and data in Psychiatry

A significant challenge in diagnosing, detecting, and treating mental health conditions is the reliance on self-reporting and clinical judgment. Unlike other areas of medicine, which depend on objective measurements, psychiatry often relies on subjective assessments. Therefore, mechanisms are needed to capture this subjectivity for AI models.

For example, the classification of suicidal behaviours using assessment tools can be difficult (Interian et al. 2017). When it is difficult for human raters to agree with an annotation it is likely difficult for AI models to uncover underlying patterns (Gaur et al. 2019). Handling these inherent uncertainties contained within data is a rapidly expanding field of AI research.

Many people are turning to social media to seek support and share mental health related information (Akhther & Sopory 2022). The use of AI systems to detect depression has seen extensive research. Similarly, Gaur et al. (2019) turned their attention to how text classification can be used to recognize social media users who may be at risk of suicide.

More recently, Guo et al. (2024) reported social media text can be used to identify suicide risk. Suicide is a rapidly growing public health risk (Naseem et al. 2023, Guo et al. 2024)

The automation of these tools using DL methods is a further application of AI in psychiatry. Text classification systems for suicide behaviours could help to connect users sharing their emotions and mental health struggles with health professionals using artificial intelligence models.

As DL systems become more prominent in non-trivial fields such as healthcare, it is important we understand how model decisions are made and how confident we can be in their predictions. Begoli et al. (2019) assert that uncertainty quantification is a necessary next step for the deep learning and artificial intelligence field when systems are being relied upon to make critical medical decisions. Additionally, Ståhl et al. (2020) contend, given deep learning algorithms are not capable of making out-of-domain predictions, it is essential algorithms can express their uncertainty when faced with out-of-sample examples.

Abdar et al. (2021) in their recent survey of uncertainty quantification identify two categories of techniques for equipping models with uncertainty quantification: Bayesian techniques and deep ensembles. Among these techniques the Bayesian technique of Monte Carlo Dropout (MC Dropout). MC Dropout has been applied extensively to research on image segmentation. However, little research has applied uncertainty quantification using MC Dropout to text classification tasks requiring natural language processing (Abdar et al. 2021).

Prominent techniques for equipping models with prediction confidence find their origins in traditional statistics. The core difference between probabilistic DL and traditional DL is the model outputs. Outputs from traditional models are expressed as a single prediction, instead, probabilistic models express their predictions as a probability distribution. Repeated simulations of these distributions, known as Monte Carlo dropout, enable models to express uncertainty. Practically, prediction confidence can be leveraged to enhance the accuracy of models.

.

1.5 Thesis objectives

So far, this chapter has highlighted the significant global impact of mental illness. Specifically, in Australia, depression places the largest burden on society. As AI becomes more prominent and powerful, research is increasingly exploring how AI can aid in the detection, diagnosis, and treatment of depression. This thesis argues that AI represents a disruptive innovation in the healthcare industry. Through research efforts like this thesis, AI is expected to dramatically transform psychiatry, challenge existing diagnostic categories, improve the diagnosis of mental health conditions, and enhance the targeting of treatments.

The purpose of this thesis is to explore the ways in which artificial intelligence can support and enhance the delivery of psychiatric care. It will explore the ways AI can disrupt existing mental health treatment decision making and care This thesis presents novel works which explore the minimum requirements for personalised rTMS delivery, potential biomarkers indicative of treatment response, novel methods for mitigating bias and delivering trustworthy AI.

In doing so, this work explores the following research questions:

1. **RQ1: How can artificial intelligence methods be used to facilitate personalised psychiatry?** Predictive medicine is a rapidly expanding area of research. As part of this work we explore data-driven informatics paradigms to predict treatment outcomes. We seek to identify the minimum data requirements for predicting treatment outcomes for depression treatments.

2. **RQ2: What are the requirements for the use of artificial intelligence as decision support in psychiatry to be effective?** The disruption of AI in psychiatry is certainly hyped in the research community. To meet these use case expectations, it is necessary to investigate and analyse the requirements for the effective deployment of AI based decision support in psychiatry. This analysis includes identifying the required data types, DL methodologies and data for the effective use of AI in psychiatry.

3. **RQ3: How are researchers ensuring artificial intelligence systems in per-**

sonalised psychiatry are trustworthy? In identifying the preconditions for effective decision support systems in psychiatry it is necessary to consider situations in which these requirements are not fulfilled. As such when sufficiently diverse data, or high quality objective data is not available what strategies can be applied to ensure AI systems remain trustworthy?

In addressing these research questions this thesis makes the following contributions:

1. An overview of the current state-of-the-art in AI use in psychiatry, exploring how AI is supporting the detection, diagnosis, and treatment of depression.
2. A detailed comparison of the data required to predict treatment outcomes using AI including the identification of candidate biomarkers indicative of treatment response.
3. A novel framework for data augmentation of depression datasets to enhance the diversity of small datasets with underrepresented values.
4. A methodology for capturing the subjective nature of expert mental health judgements.

1.6 Overview of the Thesis

This thesis examines how AI is poised to disrupt the delivery of mental healthcare, with a focus on its potential to support the treatment, detection, and diagnosis of depression. In addressing the research questions outlined above, this thesis further investigates the implications of a future where the use of AI becomes more widespread. In this sense, if AI is to be used more in clinical settings, significant work must be done to ensure it is fair for all users.

Figure 1.3 provides a detailed overview of the structure of this thesis. The Figure visually represents the relationship between research questions and thesis content to aid the reader. This dissertation is organized as follows:

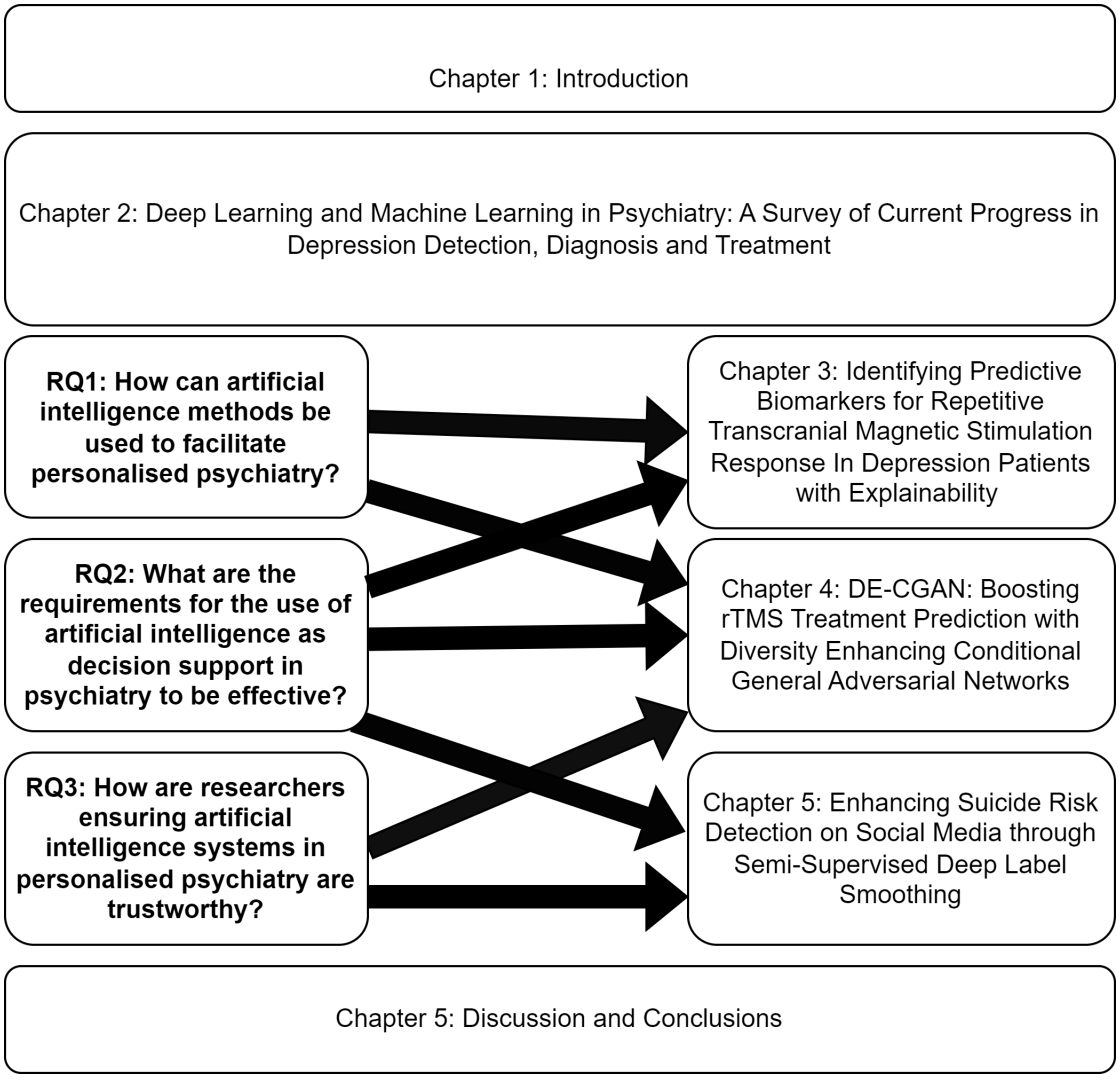

Figure 1.3: Thesis Organisation and Content Framework

Chapter 2: Deep Learning and machine learning in psychiatry: a survey of current progress in depression detection and diagnosis This chapter provides a survey of the existing state-of-the-art methods for the detection and diagnosis of depression. It further explores the ways AI is being integrated into psychiatric care. It highlights the potential of AI-driven methods to support precision psychiatry through the improved targeting of treatments to uncovering new diagnostic categories. This chapter also highlights potential limitations and challenges related to the widespread adoption of AI in psychiatry, including the need for larger and more diverse datasets and the necessity for more robust validation of AI models within the research community.

Chapter 3: Identifying predictive biomarkers for repetitive transcranial magnetic stimulation response in depression patients with explainability In this chapter, the research paper explores the use of DL methods to predict the treatment response to rTMS using fMRI connectivity data. This work explores the data types required to predict treatment outcomes demonstrating neuroimaging data to be superior to demographic data. Furthermore, this research shows DL techniques can reliably predict rTMS treatment outcomes before treatment begins. Additionally, the work incorporates the use of XAI to both identify potential biomarkers indicative of treatment response and make model performance more transparent.

Chapter 4: DE-CGAN: Boosting rTMS Treatment Prediction with Diversity Enhancing Conditional General Adversarial Networks This chapter presents a novel method called Diversity Enhancing Conditional Generative Adversarial Network (DE-CGAN). A novel method for generating synthetic examples of underrepresented features. The model generates a diversity-enhanced dataset, and empirical experiments have shown that it produces more robust predictions compared to models trained on non-diversity-enhanced data.This work highlights the importance of guaranteeing data diversity and ensuring AI models are trained on representative data.

Chapter 5 Enhancing Suicide Risk Detection on Social Media through Semi-Supervised Deep Label Smoothing This chapter investigates the use of AI for improving the detection of suicide risk in social media posts. Introducing a novel methodology of non-uniform label smoothing, these works explore the effects of

leveraging uncertainty between human raters to improve model performance. This work demonstrates how the use of fuzzy labels improve the ability of a DL model to identify social media posts describing suicide risk.

Chapter 6 Discussion and Conclusions This final chapter situates the research presented in this thesis within the broader context of the future applications of AI in psychiatry, highlighting the ongoing work needed to ensure that any implemented AI system is free from bias.

This thesis seeks to contribute to the emerging and expanding interest in the ways AI can be used to support societal good. As AI becomes more prominent in everyday systems, it is important for AI researchers to play a role in ensuring AI meets societal expectations of fairness and ethical behavior. Thus, this thesis details the ways AI and DL can be used to improve patient outcomes, while at every step considering the broader context and ethical responsibility to ensure AI is fair and without discrimination.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This section provides a detailed overview of the use of artificial intelligence techniques for the detection, diagnosis and treatment of depression. This chapter presents the article *Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment.* In this work this thesis explores the current state of the art in the use of AI techniques for the detection, diagnosis and treatment of depression. The findings of this work provides the foundation for papers described in the remaining chapters

2.2 Published paper

Squires *et al. Brain Informatics* (2023) 10:10
https://doi.org/10.1186/s40708-023-00188-6

Brain Informatics

REVIEW Open Access

Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment

Matthew Squires[1*], Xiaohui Tao[1], Soman Elangovan[2], Raj Gururajan[3], Xujuan Zhou[3], U Rajendra Acharya[1] and Yuefeng Li[4]

Abstract

Informatics paradigms for brain and mental health research have seen significant advances in recent years. These developments can largely be attributed to the emergence of new technologies such as machine learning, deep learning, and artificial intelligence. Data-driven methods have the potential to support mental health care by providing more precise and personalised approaches to detection, diagnosis, and treatment of depression. In particular, precision psychiatry is an emerging field that utilises advanced computational techniques to achieve a more individualised approach to mental health care. This survey provides an overview of the ways in which artificial intelligence is currently being used to support precision psychiatry. Advanced algorithms are being used to support all phases of the treatment cycle. These systems have the potential to identify individuals suffering from mental health conditions, allowing them to receive the care they need and tailor treatments to individual patients who are mostly to benefit. Additionally, unsupervised learning techniques are breaking down existing discrete diagnostic categories and highlighting the vast disease heterogeneity observed within depression diagnoses. Artificial intelligence also provides the opportunity to shift towards evidence-based treatment prescription, moving away from existing methods based on group averages. However, our analysis suggests there are several limitations currently inhibiting the progress of data-driven paradigms in care. Significantly, none of the surveyed articles demonstrate empirically improved patient outcomes over existing methods. Furthermore, greater consideration needs to be given to uncertainty quantification, model validation, constructing interdisciplinary teams of researchers, improved access to diverse data and standardised definitions within the field. Empirical validation of computer algorithms via randomised control trials which demonstrate measurable improvement to patient outcomes are the next step in progressing models to clinical implementation.

Keywords Psychiatry, Artificial intelligence, Depression, Deep learning, Neural networks, Treatment response prediction

*Correspondence:
Matthew Squires
Matthew.Squires@usq.edu.au
Full list of author information is available at the end of the article

1 Introduction

Conditions associated with poor mental health place a significant burden on the Australian health care system. Some evidence [1, 2] suggests despite government investment, availability of inpatient mental health services sits below the level of demand. Additionally, demand for mental health services is expected to grow further as the psychological effects of the Coronavirus pandemic are felt by the population [3]. To support increases in demand, modern algorithms have the potential to streamline the diagnosis of mental health conditions and support the improved targeting of treatments utilising a data-driven paradigm.

Advanced computing techniques including machine learning, deep learning and artificial intelligence are well positioned to positively contribute to mental health outcomes of individuals [4]. With these advanced techniques comes the potential for precision medicine. The aim of precision medicine is to tailor treatments to the individual patient as opposed to population averages [5]. More recently, the notion of precision medicine has opened the possibility of personalised mental health care. This personalisation is often referred to as precision psychiatry. Research exploring the ways artificial intelligence, machine learning and big data can be used to support mental health treatment is growing rapidly. Evidence of this growth is demonstrated by Brunn et al. [6] who observed a 250% increase in publications exploring artificial intelligence and psychiatry between 2015 and 2019 on PubMed.

Artificial intelligence will be a part of mental health care in the future. This notion is widely acknowledged by practising psychiatrists [7]. Doraiswamy et al. [7] reported results from a global survey of psychiatrists in which most acknowledge artificial intelligence will impact the future of their profession. However, clinicians vary on the degree of disruption artificial intelligence will have on the field. Few psychiatrists believe artificial intelligence will be able to "provide empathetic care to patients" [7, p. 3]. However, a slim majority believe artificial intelligence will be able to diagnose or predict patient outcomes "better than the average psychiatrist" [7, p. 4]. Whilst opinion differs on the level of artificial intelligence disruption, most clinicians believe that artificial intelligence will never completely replace mental health professionals [8, 9].

While artificial intelligence may never replace the personalised, empathetic care that a psychiatrist can provide, this paper will detail the data-driven informatics approaches positioned to revolutionise the diagnosis, detection and treatment of depression.

Pattern recognition is one of the key strengths of machine and deep learning algorithms. These techniques have shown some promise in identifying generalisable patterns amongst patients suffering mental health conditions. For example, Carrillo et al. [10] demonstrated a Gaussian Naive Bayes classifier using transcribed textual data could successfully categorise healthy controls from patients suffering depression with a $F1$-score of 0.82. Given the observed difficulty in diagnosing mental health conditions, systems with the ability to diagnose depression provide some benefit to Psychiatrists. Compared to other domains of medicine, mental health conditions have no objective markers of disease [11]. This lack of objective marker is one of several key diagnostic challenges in identifying psychopathology [12]. Current diagnostic systems are being questioned due to the significant heterogeneity of symptoms amongst populations diagnosed with the same condition [13]. Unsupervised learning techniques are supporting the identification of distinct subtypes of depression or potentially new diagnosis. Exploring depression heterogeneity, Drysdale et al. [11] used an unsupervised learning technique, hierarchical clustering, to explore functional connectivity amongst patients diagnosed with depression. While the majority of research surveyed in this paper utilises supervised techniques, unsupervised techniques provide researchers with the opportunity to uncover previously unknown relationships. The work by Drysdale et al. [11] uncovered four distinct biotypes of depression based on fMRI scans. Each of these biotypes was shown to respond differently to rTMS treatment. Given each subtype responded differently to treatments it is possible that each subtype represents a unique condition. This work highlights the possibility of artificial intelligence systems to support a transition to new diagnostic taxonomies.

As well as supporting the detection and diagnosis of mental health conditions, modern computing techniques offer the potential to personalise treatment prescription. Currently, clinicians rely on a trial and error approach to find the best antidepressant for a patient [4, 14, 15]. However, groundbreaking research by Chang et al. [16] demonstrates the potential for psychiatrists to evaluate the likely effect of an antidepressant drug before prescribing it. Their work shows using an artificial neural network, the Antidepressant Response Prediction Network, or ARPNet, can reliably predict the effect of an antidepressant prior to treatment. These technologies raise the possibility of treatment tailored to the patient level.

In its earliest form, artificial intelligence aimed to synthetically reproduce human processes [17]. In its infancy, symbolic artificial intelligence was the aim of such research. The goal of symbolic artificial intelligence work was to "carry out a series of logic-like reasoning steps over language like representations" [18, p. 17]. However, symbolic artificial intelligence is no

longer the predominant area of interest for the majority of artificial intelligence researchers. Instead, pattern recognition through the use of artificial neural networks now dominates the field [17]. The seminal work of Rosenblatt [19] provides the first example of the perceptron, the foundation of much of the current work on neural networks. Increasingly, with advances in technology, these networks have become larger leading to the advent of deep learning [20]. The depth, in deep learning refers to the number of hidden layers in an artificial neural network. However, no agreed-upon definition exists to what constitutes a 'deep' neural network [20, 21]. Sheu [22] assert a deep neural network has a minimum of 3 layers, an input layer, a hidden layer and an output layer. However, in general, modern researchers require several hidden layers before declaring a network a deep neural network.

In this paper, we will define artificial intelligence as the broad field of techniques, encompassing all of machine learning, the neural network and deep learning. In turn, machine learning will be used to refer to all non-neural network techniques, regardless of depth. This will include techniques such as linear regression, logistic regression and nearest neighbours. Given the ambiguity in the difference between artificial neural networks and deep learning, the terms will be used somewhat interchangeably. Additionally, to help the reader navigate this paper we have an included a concept map in Fig. 1. This figure provides a high-level representation of the data types and techniques being used to explore the field of depression detection, diagnosis and treatment response prediction.

This paper explores the ways in which modern phenomenons such as machine learning and deep learning are contributing to improvements in the detection, diagnosis and treatment of mental health condition. As such, this article contributes:

- An overview of the current data types and methodologies being used by the research community to progress the detection, diagnosis and treatment response prediction of mental health conditions.
- A survey of the modern computational techniques used for the detection, diagnosis and treatment response prediction of mental health conditions. Including software repositories useful for feature generation.
- A summary of the current methodological and technical limitations facing the field researching precision psychiatry.
- Reflection on the current issues facing the field and possible solutions to guide future research.

Currently, detection systems are the most widely researched areas utilising artificial intelligence to support mental health care. Section 2 provides an overview of the ways modern computational techniques are shaping the detection of mental health conditions. This area of study focuses on the design of systems built using multimodal data, such as audio, video and text data to detect mental health conditions. Section 2.3 provides a summary of the modern systems being used to revolutionise current diagnostic systems, including the vast heterogeneity

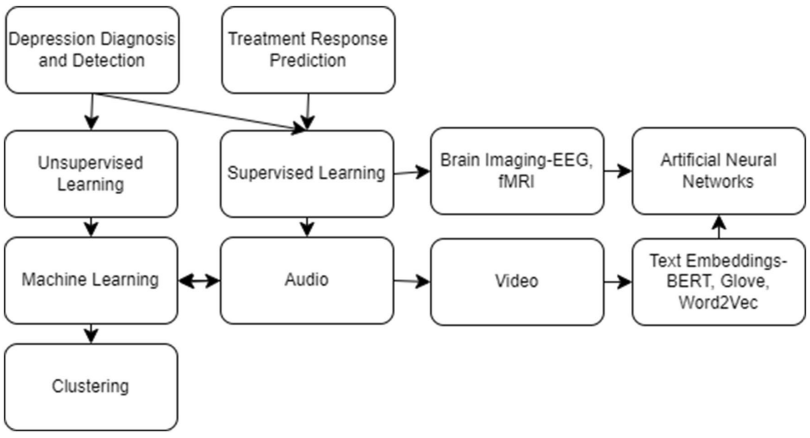

Fig. 1 Content map

within current diagnostic categories. Additionally, Sect. 3 provides an in-depth overview of one of the more recent advances in the literature, treatment response prediction. To date, detection models for mental illness have dominated the literature. More recently, using data to predict how effective a treatment might be has become an exciting area of research with much potential.

2 Informatics paradigms and the diagnosis and detection of depression

Traditionally the study of psychiatry has relied heavily on statistical inference. Inferential statistics are mainly concerned with underlying distributions. Inference "creates a mathematical model of the data generation process to formalize understanding or test a hypothesis about how a system behaves" [23, p. 233]. Where statistical inference focuses on explaining group differences based on a handful of variables. Prediction is instead suited to larger variable sets to make predictions around some target variable. Machine learning is interested in prediction and pattern recognition. Diagnosing a mental health condition requires recognising common patterns associated with a condition to make a prediction at an individual level. More recently, advances in computing processing power have led to the rise of deep learning models.

2.1 Machine learning to support the diagnosis of depression

Depression detection using machine learning has grown quickly, taking advantage of the vast corpus of text generated by social media. The diagnosis of depression from social media data can be understood as a supervised learning task where posts are labelled as depression or not depression. From the literature surveyed two classes of experiments emerge; Research where depression status is confirmed by psychometric test or clinical opinion and research relying on self-report.

When building depression detection systems variables must be preprocessed for model input. Preparing text for machine learning is referred to as Natural Language Processing (NLP). NLP is the process of converting natural language to numerical representations that are computer interpretable. Observed processing techniques within the literature are the LIWC [24], Affective Norms for English Words [25], LabMT [26], Latent Dirichlet Allocation [27], *n*-grams and bag-of-words [28, see Chapter 3]. *N*-grams and bag-of-words are elementary methods to numerically represent text, where bag-of-words is a simple text representation which counts the frequency of each word within a text document [28]. Despite their simplicity, the utility of these methods has been shown on several occasions [29–33]. More recently, audio and visual features have been included with several systems utilising processed audio features [34–36] and others which combine audio and visual information [37, 38].

Text data have become a staple feature of most depression detection systems. In pioneering work, De Choudhury et al. [39] attempted to predict depression in Twitter users. Similarly, Reece et al. [31] sought to use Twitter content to classify depressed users. Both [31, 39] recruited participants via crowdsourcing and validated a depression diagnosis using psychological diagnostic questionnaire. For example, in both [31, 39] participants completed the Center for Epidemiological Studies-Depression (CES-D; [40]) self-report survey. Results from this diagnostic tool were used as the ground truth labels between depressed and non-depressed individuals. In these examples [31, 39] researchers used surveys to attempt to confirm a depression diagnosis, however, some works rely on self reported depression status without survey data. De Choudhury et al. [39] developed one of the earliest depression diagnosis systems in the literature. Motivated by the limitations of self-report questionnaires De Choudhury et al. [39] aimed to construct an objective depression measurement. These early text analysis systems exploring word usage and depression relied on dictionary-based text analysis software. These systems used hard-coded dictionaries of words selected and grouped by their psychometric properties. Primarily used by clinicians these systems sought to explore differences in language use between depressed and non-depressed individuals.

The Linguistic Inquiry and Word Count (LIWC; [24]) was one of the earliest examples of a text analysis software. Before the LIWC, text analysis was generally conducted by human raters, however, this was inefficient, costly, and emotionally draining for judges [41]. Furthermore, raters rarely agreed when evaluating the same piece of writing [41]. Hence, computational solutions provide a faster and more consistent alternative. For depression researchers the LIWC allowed the comparison of language usage between depressed and non-depressed populations. Combining linguistic features, such as the LIWC, with Twitter behavioural data, De Choudhury et al. [39] showed a support vector machines (SVM) classifier could predict a depressive episode up to twelve months in advance. Similarly, in the Japanese context Tsugawa et al. [33] combined linguistic features with users' Twitter information to detect depression on Twitter. Along with analysing the sentiment of posts, Tsugawa et al. [33] show understanding the underlying topics of tweets to be helpful in distinguishing depression status. Combining LDA, a statistical technique used to identify underlying topics within a passage of text [27], with sentiment and twitter data Tsugawa et al. [33] returned an $F1$-score of 0.46. Both [39, 33] these works used

questionnaires to validate depression status. In contrast, Hassan et al. [30] used self-reported depression status to generate a text corpus. Using SVM and multiple linguistic features, Hassan et al. [30] achieved a *F*-score of 0.81 in their depression measurement system. The LabMT and ANEW could be broadly described as classes of sentiment analysers. These dictionaries associate each word with a valence which can be then input into a machine learning classifier. The LabMT word list contains 5000 of the most common words used on popular online platforms such as Twitter [26]. Similarly, The ANEW is a dictionary of words and an associated valence [25]. Furthermore, these tools can be manipulated to a research problem. For example, Shen et al. [42], constructed the Valence, Arousal and Dominance (VAD) tool from the ANEW. Shen et al. [42] assert their VAD tool was useful for explaining human emotions within text documents.

Reece et al. [31] used a random forest classifier to detect depression indicators in a Twitter corpus. Similar to methods described previously, a depression diagnosis was verified using psychological questionnaire. Reporting a *F*1-score of 0.644 Reece et al. [31] assert their work offers strong support for a computational method to detect depression. Similarly, Islam et al. [43] found all LIWC dimensions fed into a KNN showed promise in the detection of depression. Table 1 provides a summary of the classification systems identified under the scope of this survey. However, this table does not include deep learning algorithms or neural networks which are discussed in Sect. 2.2.

Some detection systems base their ground truth labels on the self reported health status of the participant. All of Pirina and Çöltekin [44], Islam et al. [43], Tadesse et al. [32], Shen et al. [42] rely on self-report of depression status. These works used pattern matching to identify depression indicative content, searching for that include sentences like, "I have depression." Depression indicative posts are labelled and used as training data for supervised learning techniques. Unfortunately, when datasets are developed in this manner depression status is never assessed by psychologist or questionnaire. As such, some mislabeled examples must be expected within the dataset [44]. Despite these limitations, large datasets allow researcher to uncover algorithms and feature sets which can be applied to the detection and diagnosis of depression.

The relationship between mental health status and speech is well established [45]. While text features focus on the content of speech, audio features involve the processing of the sound to analyse a variety of measurements. The inclusion of audio features in depression detection systems requires signal processing of the audio for it to be included in classification models. Several open source speech processing repositories exist and are used in the literature including COVAREP [46], openSMILE [47] to aid in feature extraction. Equivalent tools for processing of visual data technologies include measurements such as Facial Action Units (FAU) [37, 38]. Where FAU's "objectively describe facial muscle activations" [48, p. 2].

From Table 1, we see distinct performance difference depending on how depression status was validated. These findings raise concerns around how accurate methods relying on self-report actually are. Existing methods fail to capture this uncertainty inherent within self-reported data. Mental health data is often subjective which makes creating establishing ground truth labels more difficult. Future work should endeavour to adopt emerging data science techniques such as Bayesian Neural Networks (BNN) which are currently being explored to account for inherent data uncertainty.

Table 1 Detection systems and their features

Researcher	Method	Features	Dataset	*F*1-score
McGinnis et al. [35]	Logistic regression and linear SVM	Zero crossing rate, Mel frequency cepstral coefficients and the *Z*-score of the power spectral density	McGinnis et al. [35]	–
Tadesse et al. [32]	SVM	LIWC, LDA and Bigram	Pirina and Çöltekin [44]	0.91
Islam et al. [43]	Coarse KNN	LIWC	Islam et al. [43]	0.71
Reece et al. [31]	Random Forest	LIWC, LabMT, ANEW and Unigram	Reece et al. [31]	0.61
Hassan et al. [30]	SVM	*N*-gram, POS tagger, Sentiment Analyser and Negation	Hassan et al. [30]	0.81
Shen et al. [42]	Multimodal dictionary learning	LIWC, VAD, LDA, word2vec and Twitter behaviour data	Shen et al. [42]	~ 0.85
Deshpande and Rao [29]	Multinominal Naive Bayes	Bag-of-words	Deshpande and Rao [29]	0.83
Tsugawa et al. [33]	SVM	Bag-of-words, LDA, sentiment analysis+user specific information	Tsugawa et al. [33]	0.46
De Choudhury et a.l [39]	SVM	ANEW,LIWC and Twitter behaviour data	De Choudhury et al. [39]	0.68

2.2 Artificial neural networks and deep learning: from hand-crafted features to text embeddings and beyond

To date, the tools described above have shown to be efficacious in the development of depression detection system. For machine learning, feature selection is a vital part of model building. However, the development of these features can be laborious and time consuming [49]. As such, recent approaches have sought to automate the feature selection process. One of the strengths of deep learning algorithms is their ability to learn feature representations without the need for lengthy feature selection process.

More recently, deep learning has been applied to the detection of depression from text, audio and visual features. Similar to the machine learning techniques discussed in Sect. 2.1, deep learning methods are trained using labelled examples to discern patterns between individuals with and without depression. In contrast to traditional machine learning techniques, in general deep learning algorithms do not require hand-crafted features. Advanced deep learning algorithms that use textual data require word embeddings to make text machine readable. These embeddings are vector representations of text documents [28]. Deep learning algorithms use these vector representations to then learn features from the provided data [49]. Neural word embeddings such as Word2Vec [50], Global Vectors for Word Representation [51, GloVE] and more recently transformer based architectures such as Google's Bidirectional Encoder Representation from Transformers [52, BERT] are becoming far more prevalent in depression research for representing text numerically for deep learning models.

To date, little work has applied deep learning to the assessment of psychopathology [53]. There are likely several reasons for the delay in adoption of these techniques. One of which is concerns around the lack of transparency in how deep learning models make their predictions. These concerns have led some [54] to argue against the use of deep learning models for important health-related decisions. Instead preferencing traditional techniques which have greater prediction transparency. Despite concerns about model transparency, deep learning models have been shown to significantly outperform traditional machine learning techniques for the detection of depression. Cong et al. [49] proposed a system which combined XGBoost with an Attentional Bidirectional LSTM (BiLSTM). Their work was tested on the Reddit Self-Reported Depression Dataset (RSDD; [55]). Compared against several systems applied to the same dataset (including an SVM using LIWC features), the authors [49] reported a *F*1-score of 0.60. Despite its performance, previous sections have outlined some issues with self report data (see Sect. 2.1). While the system design may be useful, a dataset trained on a self-reported sample may not be applicable in a clinical setting. Rosa et al. [53] developed a deep learning approach for the recognition of stressed and depressed users. Their work used a dataset constructed using 27,308 labelled Facebook messages. The authors assert their Convolutional Neural Network (CNN) BiLSTM-Recurrent Neural Network (RNN) using SoftMax recorded the best results for recognising depressed users. They [53] reported an *F*1-score of 0.92 with a precision of 0.9 for the recognition of depressed users, significantly outperforming a Random Forest and Naive Bayes. However, it is not clear from their paper how responses were labelled or participants recruited. As highlighted in previous sections how study participants are recruited has a huge impact on model performance.

As such, textual data are commonly used data type for detection of mental health conditions. Building upon the success of text-based systems emerging research is utilising multimodal data to detect depression. The Distress Analysis Interview Corpus (DAIC; [56]) is a database of 621 interviews collected utilising a combination of face to face, teleconference and automated agent interview. The dataset includes text, physiological data (such as electrocardiogram), voice recordings and psychological questionnaire scores. Utilising this dataset, Alhanai et al. [34] combined audio with transcribed transcripts to predict depression categorically using a neural network. Their approach trained two LSTM models separately, one trained on audio features, the other using text features. Each model was trained individually, with their own weights and hyperparameter. The outputs of these two separate models were then concatenated and passed to another LSTM layer. The best performing model reported by Alhanai et al. [34] utilised both text and audio features to report a *F*1-score of 0.77. Highlighting the benefits of combining multiple data types in model performance.

Chen et al. [57] applied a deep learning approach to automate the diagnosis of perinatal depression. Their method used WeChat, a popular social media application, in the design of their system. Participants were recruited from doctors based on their Edinburgh Postnatal Depression Score (EDPS). Their work [57] was built using Long Short Term Memory (LSTM), a type of neural network. In this work the authors assert their findings match the findings of the EDPS in their sample however, little evidence is offered to support this assertion.

Table 2 provides an overview of the surveyed depression detection systems which deploy deep learning models. From this table we see a heavy reliance on text data. Recently, we observe a trend away from hand-crafted

Table 2 Deep learning and neural networks

Researcher	Deep learning architecture	Feature types	Dataset	*F*1-score
Kabir et al. [58]	BERT, DistilBERT	BERT	DEEPTWEET [58]	
Ansari et al. [59]	LSTM with Attention	GLoVE, SenticNet	Reddit, CLPsych 2015, eRisk Dataset	0.77
Wani et al. [60]	CNN, LSTM	Word2Vec, TF-IDF	Wani et al. [60]	0.99
Nemesure et al. [61]	Stacked ensemble	Electronic health records; demographic and medical	Nemesure et al. [61]	–
Zogan et al. [62]	CNN, BiGRU	BERT	Shen et al. [42]	0.91
Wan et al. [63]	Hybrid EEGNet	Resting state EEG	Wan et al. [63]	0.95
Ray et al. [37]	BiLSTM	Audio, text and visual	DIAC [56]	–
Rosa et al. [53]	CNN, BiLSTM and RNN with SoftMax	–	Rosa et al. [53]	0.92
Tadesse et al. [32]	MLP	LIWC, LDA and Bigram	Pirina and Çöltekin [44]	0.91
Tasnim and Stroulia [36]	DNN	Audio	AVEC '17 [64]	0.61
Alhanai et al. [34]	LSTM	Audio and text	DIAC [56]	0.77
Cong et al. [49]	XGBoost and attentional-BiLSTM	–	Yates et al. [55]	0.60
Chen et al. [57]	LSTM	–	Chen et al. [57]	–
Yang et al. [38]	Deep CNN and DNN	Audio and video	AVEC '17 [64]	–

features towards complex neural word embedding models such as those seen in [59, 58, 62]. This mirrors a pattern seen in the data science field in general with powerful text embedding models becoming the current state of the art. Future research should combine interdisciplinary teams to ensure researchers are using the current leading data science techniques. The utility of these deep learning systems for the recognition of depression is quickly growing, however, to date fewer examples exist of systems that model depression treatment effect. While sophisticated deep learning networks are rapidly being utilised in research the lack of transparency of these deep neural networks comes with several limitations for their use in practice. Deep learning systems although promising in their detection are unable to justify or explain why they classify a study participant a certain way. As such, [54] argue so-called 'black box' models should not be used in high stakes fields including healthcare, when a model is not human interpretable.

2.3 Uncovering new diagnostic categories with unsupervised learning and data-driven informatics

Current systems of diagnosis in psychiatry rely on diagnostic labels constructed through research rather than objective measurements of disorder [4]. The problems associated with the diagnosis of mental health conditions are widely acknowledged in the literature. An observed flaw of the diagnosis of mental health conditions is the subjectivity on which it relies. Furthermore, the categorical descriptions of psychopathology ignores heterogeneity of within group variation for specific conditions. For example, Fried and Nesse [65] identified 1030 unique symptom profiles amongst 3703 patients diagnosed with clinical depression as part of the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial. Fried and Nesse [65] go on to conclude "dissatisfaction with the diagnostic criteria of major depressive disorder might be reduced by acknowledging that it is not one coherent condition with a single cause" [65, p. 100].

Categorical diagnosis systems treat conditions as binary entities. Under a categorical approach disease entities or either present or absent [66]. Past research [67, 68] has sought to use neuroimaging to delineate between individuals suffering depression and healthy controls. For example, Yang et al. [68] used fMRI to compare differences in resting state activations, identifying reduced activity in the left dorsolateral prefrontal cortex when compared to prefrontal cortex. More recently, Artificial intelligence has the potential to identify sub groups within disease populations through pattern recognition. This pattern recognition can be referred to as unsupervised learning. In contrast to the supervised tasks surveyed so far, unsupervised algorithms are used to "identify inherent groupings within the unlabeled data" [69, p. 5]. Thus, unsupervised algorithms can be used to identify groupings that transcend existing diagnostic labels [70]. Exemplifying the possibility of new diagnostic criteria, Drysdale et al. [11] utilised hierarchical clustering, a type of unsupervised learning to identify four sub types of depression. Their method, grouped patients based on fMRI connectivity measures. Further exploration showed these sub types could be used to predict treatment response to rTMS. Of note the machine learning classifier was better able to predict treatment response than a model built using symptoms alone [11].

These results offer support for that position that depression may not be one single disease entity but in fact made up of multiple different conditions. More recently, Kuai et al. [71] explored a brain computing approach to construct and evaluate prediction models using different brain states. Kuai et al. [71] argue a brain mapping approach to understanding mental health offers strengths over existing strategies as it allows for hypothesis testing to validate causal results. Future work using brain computing may in fact be used to verify differences in the underlying brain structures of people diagnosed with the same condition.

This section has raised the possibility of either distinct subtypes of depression, or in fact several different underlying conditions distinct from depression. What is significant from the patients perspective is these different depression variants vary in their response to treatment. As such, the use of data to support treatment decisions in mental health has been an area of significant research. As research for personalised medicine has increased so to has work exploring the ways in which psychiatric treatments can be tailored to the individual. One emerging area of interest is the use of machine learning algorithms to predict a patient's response to treatment prior to intervention.

3 Learning systems to predict depression treatment response

Patterns of response to treatments for mental health conditions are often inconsistent. Conventional research aims to find interventions which are successful at the group level [4]. However, as highlighted above, recent research is now uncovering significant heterogeneity of symptoms among patients classified under the same diagnostic label. As such, diagnosis alone are not sufficient to inform treatments [70]. The heterogeneity of categorical diagnostic systems is reflected in the inconsistent response to treatment interventions for patients diagnosed with the same condition. Major depressive disorder provides an example of the difficulties in prescribing treatments and the inconsistency in treatment response and remission rates.

Estimates of remission rates to antidepressant treatments vary from 25 to 33% of patients achieving remission after their first course of treatment [15, 72–74]. However, this does not mean that patients do not go on to achieve remission of their disorder. Some estimates suggest 67% of patients go on to achieve remission after trials of multiple antidepressant treatments [15]. Given this, a preferred method for assigning treatments would be to maximise the likelihood of success. However, currently no standardised way exists of prescribing treatments with clinicians relying on a trial and error approach to find the best [14, 15].

A more desirable option would be to identify likely responders to an intervention prior to treatment. Under this approach, treatments can be targeted to the individual patients who are most likely to derive benefit [4]. This is the aim of precision psychiatry. Precision psychiatry supported by artificial intelligence would allow clinicians to move beyond diagnostic categories and make room for the individual variability of care [70]. Tailoring treatments to the individual has several benefits. If it is possible to predict whether a patient will respond to treatment before commencing the therapeutic intervention. Hence reducing the time spent pursuing likely ineffective treatments. Additionally, time saved reduces both the financial and psychological burden on patients and health care systems [14, 75].

3.1 rTMS response prediction

Repetitive transcranial magnetic stimulation (rTMS) is an evidenced based treatment for depression. However, despite a demonstrated clinical benefit when compared to a control [76] for some patients rTMS is ineffective. Berlim et al. [76] in their meta analysis report a response rate to rTMS treatment of $\approx$ 30% and remission rate of $\approx$ 19%. Similarly, Fitzgerald et al. [77] in their pooled sample review observed a response rate of $\approx$ 46% and remission rate of $\approx$ 30%. According to Koutsouleris et al. [78] the variability of response to rTMS is seen as one of the main barriers to the widespread adaptation of the treatment modality. This section provides an overview of the data science techniques used to delineate rTMS treatment responders from non-responders. Focusing on systems which make predictions on treatment response at the level of individual patients. These treatment response prediction systems employ supervised learning techniques and utilise several types of predictor variables such as neuroimaging (MRI, EEG, fMRI), genetic, phenomenological or a combination of several variable types [79].

The works by Fitzgerald et al. [77] highlights a distinctly bimodal pattern of response to rTMS treatment. This pattern of response is distinguished by patients who respond to the rTMS treatment, and those who see little benefit. Using traditional inferential statistical techniques [77] note no variable alone could delineate between responders and non-responders. This limitation of traditional statistics highlights one strength of artificial intelligence and machine learning approaches. Advanced techniques have the ability to combine and make treatment recommendations based on multiple variables. As such, in situations where one variable alone cannot distinguish between a

responder and non-responder, combinations of variables may have that power. Additionally, these advanced techniques allow for the combination of data from multiple sources. More recently, researchers [11, 14, 75, 78, 80–83] have utilised more sophisticated machine learning techniques to distinguish rTMS responders from non-responders. The works summarised in Table 3, combine physiological measurements such as electroencephalogram (EEG) [14, 75, 80–82] and fMRI [11, 83]. Table 4 provides a brief overview of the common EEG features input into the models described in this survey.

Noting the link between working memory and depression (for example, [87]), Bailey et al. [80] explored the predictive power of working memory related EEG measurements. Models were built combining Montgomery Åsberg Depression Rating Scale [88, MADRS] scores, performance on a working memory test, reaction times and EEG measurements. EEG measurements included connectivity, power, and theta gamma coupling measures. Where connectivity was calculated using weighted Phase Lag Index (wPLI; [89]).

Exploring the relationship between connectivity and rTMS response, Chen et al. [84] investigated the role of connectivity features collected using MRI. In their study, Chen et al. [84] report using functional connectivity maps as features as inputs to their SVM regression analysis. Recently, Hopman et al. [85] deployed a linear SVM using features collected via fMRI, such as connectivity features between the subgenual anterior cingulate cortex, lateral occipital cortex, superior parietal lobule, frontal pole and central opercular cortex. During fivefold cross-validation, the authors present a training accuracy of $\approx$ 97% however, on a held out test set, model performance drops to an average $\approx$ 87% with a 95% confidence interval from 100% to roughly 70% accuracy. Similarly, a SVM model of 30 features the [80] report an $F1$-score of 0.93 and a balanced accuracy of 91%. These metrics were the mean results of a robust internal validation scheme of 200,000 iterations of fivefold cross-validation. Building upon these initial findings [81] explored the utilised linear SVM with resting EEG features collected prior to treatment and after 1 week of treatment to predict rTMS treatment response for depression. Built using 54 features the research utilised 5000 trials of fivefold cross-validation to achieve a balanced prediction accuracy of 86.6%. The 54 features combined measures collected from MADRS questionnaire and quantitative EEG signals Alpha Power, Theta Power, Alpha Connectivity, Theta connectivity, Theta Cordance and Individualised Alpha Peak frequency. Building upon [81, 75] used

Table 3 rTMS treatment response prediction

Author	Condition	Features	Algorithm
Chen et al. [84]	Depression	Resting state MRI	SVM regression
Hopman et al. [85]	Depression	Resting state fMRI	Linear SVM
Bailey et al. [81]	Depression	EEG and MADRS	Linear SVM
Fan et al. [83]	Depression	Resting state fMRI	Hierarchical regression
Hasanzadeh et al. [14]	Depression	EEG	K-NN
Zandvakili et al. [75]	Depression and post-traumatic stress disorder	EEG	Lasso regression and SVM
Bailey et al. [80]	Depression	EEG	Linear SVM
Koutsouleris et al. [78]	Schizophrenia	–	Linear SVM
Drysdale et al. [11]	Depression	fMRI	Hierarchical clustering and SVM
Rostami et al. [86]	Unipolar and bipolar depression	Clinical and demographic	Binary logistic regression
Erguzel et al. [82]	Depression	EEG	Artificial neural network

Table 4 EEG feature summary

Feature	Description
Cordance	The sum of z-transformed absolute and relative power for a frequency band [90]
Coherence	Coherence is a measure of correlation between signals [91, 92]. Contextualised, coherence is operationalised as a measure of functional connectivity between brain regions [75].
Power	A measure of the activity in a frequency band [92]
Theta gamma coupling	Research [93] has shown a relationship between theta gamma coupling and deficits in working memory
Weighted Lag Phase Index (wPLI; [89])	A measure of functional connectivity

machine learning to predict response to rTMS of depression sufferers with comorbid post-traumatic stress disorder (PTSD). However, in contrast to Bailey et al. [81], Zandvakili et al. [75] utilised lasso regression to model treatment prediction. Alpha EEG signal coherence was used to build the lasso prediction model. Coherence is a measure of correlation between signals [91, 92]. Contextualised, coherence is operationalised as a measure of functional connectivity between brain regions [75]. Utilising a regression model the model outputs predicted percentage reductions in scores on the Post-Traumatic Stress Disorder Checklist-5 (PCL-5; [94]) and Inventory of Depressive Symptomatology-Self-Report (IDS-SR; [95, 96]). Reductions of greater than 50% are classified as a clinical response. Continuous predictions of questionnaire score reduction are then converted to classifications. For example, a model that predicts a 60% reduction in IDS-SR for an actual reduction of 65% is the correct. While Zandvakili et al. [75] report an impressive AUC of 0.83 utilising Alpha coherence to predict IDS-SR response and AUC of 0.69 for PCL-5 response classification. These results must be interpreted in the context of high sensitivity (approx. 100%) and low specificity (approx. 50%) suggesting a large number of false positives.

Continuing with the use of pretreatment EEG features [14] sought to predict treatment response to rTMS. Where response was defined as a reduction of Hamilton Rating Scale for Depression (HRSD; [97]) or Beck Depression Inventory (BDI; [98]) by over 50%. Their sample included 46 patients with a balanced sample of responders and non-responders. The model utilised K-NN built on EEG features with the best single feature model built using the Power of beta. This model achieved a classification accuracy of 91.3% when using leave one out cross-validation. The best performing of the multi-feature models included the Power measurements of all bands (Delta, Theta, Alpha, Beta) accuracy remained at the level as the model built using only the power of Beta. However, the model utilising all power features did differ in terms of specificity and sensitivity. Hasanzadeh et al. [14] claim their system built using only pretreatment EEG features offers a better alternative to systems requiring multiple measurements.

To our knowledge [82] provides the only example of a deep learning algorithm for the prediction of rTMS responders. Erguzel et al. [82] explored the possibility of quantitative EEG to predict treatment response using an artificial neural network. The main predictive model utilised Quantitative EEG (QEEG) cordance as the main predictive feature, this is consistent with Bailey et al. [81] who offer some support for the use of cordance as an input feature. Further evidence [99, 100] suggests theta cordance for the discrimination between treatment responders and non-responders. The majority of surveyed papers relying on EEG use hand-crafted features consisting of existing signal processing techniques. However, more recently [63], showed through a novel deep learning CNN, EEG data can be processed directly by a deep learning architecture. This provides an opportunity for future researchers to streamline the data pipeline by inputting EEG data directly into networks.

The literature so far has highlighted the value of rTMS treatment for at a minimum a subset of the population experiencing depression. Additionally, emerging evidence exists to support the use of rTMS for the treatment of schizophrenia [101, 102]. Koutsouleris et al. [78] utilised linear SVM to predict treatment response for schizophrenia to rTMS treatment. Utilising structural MRI they utilised principal component analysis to reduce image features to approximately 25 principal components. According to Koutsouleris et al. [78] response was defined using the positive and negative syndrome scale (PANSS; [103]). In contrast to depression, schizophrenia is characterised by both positive symptoms including hallucinations and delusions as well as negative symptoms such as social withdrawal [104]. As such, response to treatments for schizophrenia is defined as a greater than 20% increase in the positive symptoms sub-scale (PANSS-PS) or greater than 20% increase in the negative symptom sub-scale (PANSS-NS). Hence, response to treatment is classified in terms of response for positive symptoms or negative symptoms. In the active treatment condition a cross validated model produced a balanced accuracy of 85% between responders and non-responders. Consistent with expectation and findings observed by Tian et al. [105] when utilising a leave-one-site-out validation protocol was utilised balanced accuracy dropped to 71%. Koutsouleris et al. [78] provides evidence for machine learning algorithms utility irrespective of condition. With enough data, advanced computing techniques have the potential to support improvements across multiple conditions in psychiatry.

To that end, prediction of responders at the single patient level has become of interest to the research community. The surveyed papers show EEG features to be the most common neuroimaging feature [14, 75, 80–82], with a recent trend towards fMRI and MRI features [83–85]. EEG measurements of interest include connectivity, measured using coherence or wPLI, along with power and cordance. Additional features include depression rating surveys such as MADRS [81]. These observations are consistent with Lee et al. [79] who explored the use of machine learning algorithms to predict treatment outcomes for patients with either depression or bipolar depression. In the current work SVM was the most

widely used algorithm to delineate between treatment responders and non-responders of rTMS treatments. Several studies report exceptional predictive performance (for example, [80]) for their models, however, the studies surveyed rely almost exclusively on cross-validation, an internal validation strategy. Of note [14, 78] included some pseudo-external validation in the form of a leave one group out validation. In their multi-site sample, validation involved holding one site out from training for model evaluation. Interestingly, performance of this model dropped significantly when tested on a site not included in the training set. Future opportunities exist for the streamlining of techniques to preprocess data such as EEG, MRI and fMRI for input into deep learning models. Future work may see networks which automate this preprocessing reducing the need for hand-crafted features.

3.2 Pharmacological intervention response prediction

Currently, robust biomarkers or objective measurements of psychiatric conditions do not exist. However, several studies have identified neuroimaging techniques as "candidates of prognostic biomarkers in major depression disorder" [72, p. 2]. Seminal work by Khodayari-Rostamabad et al. [15] provides an early example of treatment response prediction for antidepressants. Their system utilised pretreatment EEG features combined with a mixed feature analysis [106]-based classifier to predict treatment response prediction. More recently, Jaworska et al. [72] explored the efficacy of several machine learning classifiers for the prediction of treatment response of antidepressants. The work explored, random forests, Adaboost, SVM, classification and regression trees (CART) and the multilayer perceptron (MLP). The best performing model reported by Jaworska et al. [72] was a random forest classifier which combined 117 features from a variety of sources including eLORETA, EEG and clinical features. The model recorded an $F1$-score of 0.901. Despite this impressive performance, models built with large numbers of features are vulnerable to overfitting [107]. Given the problem of overfitting, the more suitable model presented by Jaworska et al. [72] is built using twelve predictive features selected based using extremely randomised trees. This method ranks the predictive power of features using the average impurity score. Of models built using only twelve features, [72] report random forest to have the best prediction performance with an $F1$-score of 0.827 slightly outperforming Adaboost with an $F1$-score of 0.815. Similar to the findings of Drysdale et al. [11], Jaworska et al. [72] assert models built on features incorporating imaging techniques outperformed models built solely on clinical or demographic data. This assertion suggests models neuroimaging techniques to be a more reliable measure of psychiatric health.

While imaging, clinical and demographic features are the predominant features of interest, pioneering works [16, 109, 110] have included genetic features, such as single nucleotide polymorphisms (SNP). Pei et al. [109] collected SNP's via a blood sample where the significance of each allele was determined using logistic regression. The outcome variable of interest was treatment response vs non-response. Continuing with the theme of algorithmic feature set selection, Pei et al. [109] utilised SVM recursive feature elimination. Linear SVM was used in an ensemble approach outperforming single classifiers built using the same predictor variables. This result is consistent with the literature that emphasises the strength of ensemble methods for classification tasks in supervised learning [114]. Similarly, Lin et al. [110] explored the predictive power of SNPs utilising the deep learning algorithm, multilayer feedforward neural networks (MFFN). The work explored the performance capability of the MFFN compared to logistic regression with a feature set of 16 biomarkers and six clinical features to predict both treatment response and remission. For a set of 16 features, the MFFN with up to three hidden layers outperformed logistic regression in both AUC and sensitivity, however, logistic regression achieved slightly better specificity. When the number of features was lowered to six biomarkers, similar to Jaworska et al. [72] performance declined as the number of features dropped. For 6 features, the best AUC score dropped to an AUC of 0.5597 for a single-layer MFFN with the logistic regression achieving higher specificity.

Also utilising a deep learning for the prediction of treatment response, Chang et al. [16] developed a neural network based system, the Antidepressant Response Prediction Network (ARPNet), to predict both the degree of treatment response, as a continuous variable, and whether a patient reaches clinical remission. In contrast to other studies (see [72, 109]), Chang et al. [16] define clinical remission as a greater than 50% reduction in HAM-D score; whereas [110] defined remission as a HDRS score of less than 7. These differences in definitions are significant. As the field strives for clinical use of artificial intelligence systems a standardisation of definitions would be helpful for comparing models. Despite terminology differences, Chang et al. [16] present a robust system to predict response with their model significantly outperforming other widely used classifiers such as linear regression. Similar to Pei et al. [109], Lin et al. [110], ARPnet includes genetic variables and combines this information with neuroimaging biomarkers. The system utilises elastic net feature selection with

hyper parameter tuning conducted using fivefold cross-validation with a test set of 10%. Two features unique to ARPnet is the antidepressent prescription layer of the neural network and the use of ARPnet to predict the degree of treatment response, measured in terms of HAM-D score across time. This novel approach would allow psychiatrists to model the likely response of an antidepressant before prescribing it [16].

While text features were widely used for the detection of depression (see Sect. 2), the use of these features is uncommon in treatment response prediction. Carrillo et al. [10], in a unique method present text analysis as a method for predicting the treatment response to psilocybin. Given, the established relationship between psychological health and language use [115–119], Carrillo et al. [10] first show that a Gaussian Naive Bayes classifier could distinguish between individuals suffering from depression, and healthy controls. Their model was built using features constructed by sentiment analysis collected via interview. Additionally, this Gaussian system able to distinguish responders from non-responders at a level of significance when compared to permutation testing. However, this research is significantly limited by the small sample size of only 17 study participants comprising 7 responders and 10 non-responders.

So far this section has explored a variety of data sources used as features for systems that predict treatment response. With the most common physiological feature being EEG. An additional and emerging data type is the use of fMRI neuroimaging [11, 83, 105]. Tian et al. [105] explored resting fMRI features as predictors of escitalopram response in patients suffering depression. The work explored the predictive power of fMRI features across three sites. Using data of 34 patients from Nanjing Brain Hospital across a 7-year period [105] used an SVM classifier to deliver an optimal accuracy of 79.41%. Using permutation test as comparison the authors [105] conclude this result to be significant at the $p < 0.001$ level. Using the minimum redundancy maximum relevancy the authors identified 7–8 features which combined to produce the optimal classifier. Similar to Hasanzadeh et al. [14], Koutsouleris et al. [78], as Tian et al. [105] was a multi-site trial, a leave one group/site out analysis was used as a validation technique. Using one site as the hold out set for more thorough validation which tests model generalisation. For Tian et al. [105] a leave one group out analysis showed performance decrease. This leave one group out protocol achieved accuracy of between 69 and 71% compared to the 79.41% when data were trained and tested at a single site. This performance drop highlights the common limitation of machine learning, model generalisation to unseen data. Similar performance decline is observed by Browning et al. [108] who provide one of few examples of external validation on an independent dataset. Exploring the possibility of baseline Quick Inventory of Depression Severity (QUIDS; [120]) and the face-based emotion recognition task (FERT). Browning et al. [108] observed performance decline from approximately 80% accuracy to 60% accuracy on the independent dataset. Similarly, Chekroud et al. [112] using gradient boosting machines achieved an accuracy score 64.6% during cross-validation compared to an accuracy of 59.6% on an external data a performance drop not in the magnitude of Browning et al. [108]. The difference in relative performance drop could be due to the low accuracy reported in the internal validation stage by Chekroud et al. [112]. Performance comparisons between Browning et al. [108] and Chekroud et al. [112] are further complicated by their different target variables. Browning et al. [108] sought to identify patients who achieved a response to treatment, defined by a greater than 50% reduction in QIDS-SR, in contrast, Chekroud et al. [112] sort to identify clinical remission defined by the QIDS-SR as a final score less than or equal to five.

Several algorithms have been trialled for the prediction of treatment response to pharmacological treatments of depression. A summary of these techniques can be found in Table 5. These algorithms include deep learning techniques such as MFFN [72] and customised neural net-based systems such as those in Chang et al. [16]. Other commonly utilised algorithms include Linear SVM [109, 105], tree-based methods [72, 113] and logistic regression [111].

While the majority of studies discussed in this section report impressive results, they are significantly limited by small samples (see Table 6) and lack of external validation. Commonly, internal validation techniques such as k-fold cross-validation and leave-one-out cross-validation. And others [110, 111] employed repeated cross-validation, the most robust form of internal validation [121]. We observed significant performance drops when data were spread across multiple sites or models tested on independent data. This performance decline highlights the issue of generalisation in machine learning, one of the key barriers to clinical adoption of these techniques [5, 122].

We also note the recent shift towards more sophisticated deep learning techniques, with Tian et al. [105] claiming their MFFN to outperform a logistic regression, [16] reporting their neural net-based system to outperform common strategies such as SVM and random forests. The majority of response prediction studies agreed to a common definition of response as a greater than 50% reduction in score from a psychometric questionnaire used to asses depression severity, with instrument of choice varying across samples. Notably, only Chang et al.

Table 5 Pharmacological treatment response prediction

Author	Features	Algorithm	Validation
Jaworska et al. [72]	EEG and eLORETA	Random forests	Tenfold cross-validation
Browning et al. [108]	Initial QIDS-R and face-based emotional recognition task (FERT)	Linear SVM	External validation on unseen data
Pei et al. [109]	EEG and genetic markers	Linear SVM	Leave-one-out cross-validation
Chang et al. [16]	MRI and genetic markers	Artificial neural network	Holdout set and k-fold cross-validation for hyperparamater tuning
Tian et al. [105]	fMRI	Linear support vector machine	Leave-one-out cross-validation
Carrillo et al. [10]	Speech data	Gaussian Naive Bayes	Sevenfold cross-validation
Lin et al. [110]	Genetic markers	Multilayer feedforward neural network	10 iterations of tenfold cross-validation
Mumtaz et al. [111]	EEG	Logistic regression	100 iterations of tenfold cross-validation
Chekroud et al. [112]	Sociodemographic, questionnaires (such as HAMD), clinical information	Gradient boosting machine	10 iterations of tenfold cross-validation and externally validated on unseen data
Patel et al. [113]	Demographic and neuroimaging	Alternating decision trees	Leave-one-out cross-validation
Khodayari-Rostamabad et al. [15]	Pretreatment EEG	Mixture of factor analysis	100 iterations of leave N out cross-validation

Table 6 Pharmacological treatment response sample summary

Author	Sample size	Definition of response
Jaworska et al. [72]	51	> 50% reduction in MADRS score
Pei et al. [109]	98	> 50% reduction in HDRS 6
Lin et al. [110]	421	–
Chang et al. [16]	121	Remission defined as > 50% reduction in HAM-D
Carrillo et al. [10]	17	> 50% reduction in QIDS
Mumtaz et al. [111]	34	> 50% reduction in BDI-II
Khodayari-Rostamabad et al. [15]	22	> 30% reduction in HAM-D

[16] differed in their definition responder, defining clinical remission as a 50% reduction in HAM-D score.

As artificial intelligence becomes more prevalent in medicine and psychiatry a more standardised framework is required for the testing and validation of deep learning models. Differences in definitions between models make comparison between systems more difficult. As such regulators and the research community should endeavour to standardise definitions; This standardisation would first make the regulation of artificial intelligence systems easier and secondly make communication of model performance more transparent.

4 Discussion: challenges and opportunities

Advances in deep learning, machine learning and natural language processing are slowly being applied to the field of precision psychiatry. This paper serves as a guide for psychiatrists and data science practitioners alike as to the existing state-of-the-art techniques and the open problems which require further work.

Supporting a shift towards precision psychiatry artificial intelligence provides the opportunity for treatment response prediction. Treatment response prediction provides empirical evidence for the likely effect of an intervention. Currently, clinicians rely on trial and error to find the best antidepressant for a patient [4, 14, 15]. As such, treatment response prediction offers a shift from trial and error treatment prescription to evidence-based treatment recommendations supported by data. The surveyed works explore two categories: single patient response prediction for rTMS and pharmacological interventions. These systems utilise any of neuroimaging, demographic and clinical features [79]. Jaworska et al. [72] observed neuroimaging features outperformed clinical and demographic features. This is consistent with Drysdale et al. [11] reports "clinical symptoms alone were not strong predictors of rTMS treatment responsiveness at an individual level" [11, p. 8]. Systems built using neuroimaging techniques consistently demonstrated the ability to delineate between treatment responders and non-responders for both rTMS and drug-based treatments. However, for these systems to be adopted in a clinical setting several limitations must be addressed.

4.1 Challenges and limitations

Through our survey of the literature, we identified some consistent themes for consideration by the research community. The studies reviewed so far report impressive results for the detection, diagnosis and treatment response prediction. Despite impressive results reported

above, none of the works surveyed as yet have been shown to demonstrate improved treatment outcomes for patients. Given the field of personalised psychiatry is not new, with surveyed works spanning a decade. Further collaboration between mental health professionals and data scientists to ensure this research is being converted into improved patient outcomes. This section explores the limitations of existing systems which reduces the possibility of real world application.

4.1.1 Model validation: the need for external validation

Several of the surveyed studies described in previous sections report impressive power for predicting treatment response with several performing above current standards observed in practice. However, several issues exist in moving these research systems to clinical practice. Of the papers reviewed above the most obvious limitation, or barrier to implementation is the issue of model validation.

Of the surveyed articles two studies include multiple sites [78, 105] and two test their models on independent data [108, 112]. Rigorous validation is crucial if machine learning systems are to effectively transition to industry use [122]. The majority of papers cited above use some form of internal validation such as k-fold cross-validation. Widely cited work by Harrell Jr [121] provides a hierarchy of validation techniques used to predict model performance on new data. Using this hierarchy validation techniques range in effectiveness from only reporting the best performing iteration of model performance, to the most powerful validation technique, external validation by an independent research team on new data. Harrell Jr [121] asserts the strongest of internal validation techniques is repeated iterations of k-fold cross-validation. Model validation is of significant importance in the transition of predictive models. Fröhlich et al. [5] notes the path to implementation for predicative artificial intelligence models must include robust internal validation, external validation on independent data and empirical validation as part of a clinical trial.

These views are supported by Browning et al. [108] who contend randomised control trials are necessary to validate model performance to a level that would justify clinical adoption. Of the papers surveyed to date few tested their models on independent data and none included randomised control trials of their systems. With the lack of publicly accessible data for depression, external validation of model performance is challenging. Open datasets would enable researchers to build their models on one dataset and compare performance across samples. This realisation is already being realised by datasets such as ADNI, providing an established research pipeline for the study of Alzheimer's. Providing researchers with datasets for external validation.

4.1.2 Small sample sizes and greater data access

The issue of access to data and sample sizes provides a brief overview of progress in the respective dimensions covered in this review. Data relating to depression detection are widely available compared to data for treatment response prediction. For example, social media text, DIAC [56] and AVEC [64] are widely accessible. Access to data provides computer scientists and researchers the opportunity to compare their systems on the same datasets. In contrast, researchers exploring treatment response prediction at the single patient level are limited by small samples and challenges accessing data. A centralised cloud-based repository of mental health data as proposed by Chen et al. [123] offers one potential solution, however, would be require significant infrastructure to implement.

Treatment response prediction relies more heavily on neuroimaging data. Labelled examples for treatment response prediction are far less available with the surveyed articles relying on small samples. Table 6 provides an overview of the sample sizes used to generate the results discussed in this paper. Consistent with trends identified in Arbabshirani et al. [124], with the exception of [110] the majority of studies surveyed have samples under 150. Arbabshirani et al. [124] assert it is difficult to generalise results from small samples to the broader patient population. Furthermore, it is likely small samples overstate the predictive power of a system [125]. Button et al. [126] assert low statistical power as a result of small sample sizes is a problem of endemic proportions within the field of neuroscience. Combined, with observed publication bias of artificial intelligence systems [125] it is likely the published literature provides only a theoretical upper limit of the current effectiveness of artificial intelligence systems for precision psychiatry. Furthermore, small sample sizes do increase the probability of overfitting [4], leaving researchers to overstate the performance of their model.

For the continued growth of personalised psychiatry research larger datasets become more accessible. The dearth of open datasets is especially true for the study of depression. With the benefits of open data sharing is exemplified by the success garnered from the Alzheimer's Disease Neuroimaging Initiative. Recently, Birkenbihl et al. [122] report the ADNI dataset has now been referenced more than 1300 times. To date there is no equivalent data repository for conditions such as depression. Possible large cloud based solution such as that proposed by Chen et al. [123] may pave the way forward, however, further work is required.

4.2 Future trends and opportunities

The last decade of research has seen rapid advancements in the technologies being used to support mental health care. For the detection and diagnosis of depression we observe a trend away from machine learning algorithms to sophisticated deep learning architectures. Similarly, text classification is moving away from traditional text mining features such as *n*-grams and bag-of-words to more sophisticated transformer-based embeddings such as BERT. However, the transition to deep learning architectures is less evident in treatment response prediction. Despite using quantitative data like EEG, fMRI or MRI, this field is relying on existing technologies such as SVM. Few methods exist where raw neuroimaging data, such as EEG is passed directly to Deep Learning Algorithms. Thus an opportunity exists for the use of deep learning methods to learn feature representations for the treatment response prediction and streamline data preparation.

4.2.1 Causal artificial intelligence

Existing trends in this survey show a move from hypothesis testing, to pattern recognition using artificial intelligence techniques. However, predictive techniques do not establish causality as hypothesis and randomised control trials did. While some confuse pattern recognition for causality, Sgaier et al. [127] asserts "Relying solely on predictive models of AI in areas as diverse as health care, justice, and agriculture risks devastating consequences when correlations are mistaken for causation."

Establishing causation using artificial intelligence would be a significant breakthrough in depression research and precision psychiatry alike. In some medical fields we are starting to see early attempts at establishing causality with the use of deep learning. Wang et al. [128] show their model DeepCausality was able to identify 20 causal factors for identifying drug induced liver disease from electronic health records. Furthermore, advances in brain mapping such as the strategies shown in Kuai et al. [71] may allow for the establishment of causal relationships between changes in brain activity and depression severity

4.2.2 New technologies and automating data pipelines

Recent advances in text embeddings such as BERT, GloVe or Word2Vec are more often being utilised by practitioners to prepare text for depression detection. The use of these transformer-based word embeddings have led to more streamlined data pipelines. Further opportunities exist for data scientists to develop new techniques to process neuroimaging data directly such as the approach proposed by Wan et al. [63]. CNNs are well equipped to handle sequence data and feature work may allow for networks equipped to handle neuroimaging data without prepossessing.

To date, the detection and diagnosis of mental health conditions relies on self-report or clinician-administered questionnaires. Currently, objective biomarkers of psychopathology do not exist [11]. Given this challenge, significant research has explored the possibility of depression detection using text, audio and visual. Currently, evidence [37] suggests the content of speech is the best predictor when compared to audio and visual to delineate between people who are healthy and individuals suffering mental health conditions. Systems designed for depression detection utilise a variety of techniques progressing from elementary machine learning methods to more sophisticated techniques such as deep learning algorithms. Depression detection is the most widely researched area explored within the scope of this survey. This advancement has been driven by the access to significant bodies of text and publicly accessible datasets such as DIAC [56] and AVEC [64].

4.2.3 Uncertainty quantification

As the field strives for clinical implementation of the artificial intelligence systems surveyed further work is required to capture the uncertainty associated with model building. This includes the two types of uncertainty, data uncertainty (aleatoric uncertainty), and epistemic uncertainty, (model uncertainty). The aleatoric uncertainty can be seen in the variations in depression detection system performance depending on how ground truth labels were collected. We noted performance drop off when self-report measures were used as ground truth labels. The use of self-report measures encompasses some inherent uncertainty which existing methods fail to capture. Additionally, if these models are to become prevalent in their use in informing treatment decisions, these models must be able to express their prediction confidence, which currently is not included in model outputs. Bayesian Neural Networks are an emerging technology to encompass both data uncertainty and express prediction confidence. Further to this, more work is required to ensure as models become more complex effort is made to understand the inner workings of these models. Some concerns exist regarding the lack of transparency in how deep learning models make their predictions. These concerns have led some [54] to argue against the use of deep learning models for important health-related decisions. Accurate predictive models which are interpretable are of significant interest to the research community.

5 Conclusions

Much excitement surrounds the potential for artificial intelligence and machine learning to revolutionise psychiatry. This paper provides an overview of the techniques and methodologies available to researchers for the detection, diagnosis and treatment of depression. Whilst every endeavour has been made to ensure the completeness of this survey paper given the speed of progress within the data science community we cannot guarantee all papers within the literature have been included. However, this paper aims to provide an up-to-date assessment of the current position of artificial intelligence's use in the field of psychiatry.

The last decade of research has seen rapid advancements in the technologies being used to support mental health care. For the detection and diagnosis of depression we observe a trend away from machine learning algorithms to sophisticated deep learning architectures. Similarly, text classification is moving away from traditional text mining features such as *n*-grams and bag-of-words to more sophisticated transformer-based embeddings such as BERT. However, the transition to deep learning architectures is less evident in treatment response prediction. Despite using quantitative data like EEG, fMRI or MRI, this field is relying on existing technologies such as SVM. Few methods exist where raw neuroimaging data, such as EEG is passed directly to deep learning algorithms. Thus an opportunity exists for the use of deep learning methods to learn feature representations directly and streamline the treatment response prediction process.

Current limitations of treatment response systems include small sample sizes and model validation. The small samples observed in the treatment response prediction systems described in Sect. 3 make it difficult to generalise findings to the broader population [124]. Additionally, small sample sizes increase the likelihood of model overfitting [4]. Larger, more publicly accessible datasets such as the data pipelines that are well established for the study of Alzheimer's disease (see [122]) would address this issue. Further barriers to the widespread adoption of these systems is the issue of model validation. As noted by Fröhlich et al. [5] the path to implementation for predicative artificial intelligence models includes robust internal validation, external validation and empirical validation as part of a clinical trial. Of the works included within the scope of this review the majority includes only internal validation, falling well below the standard for implementation. To advance the field of personalised psychiatry to the clinic, future work should seek larger datasets and explore empirical validation in the form of randomised control trials. We suggest greater collaboration between healthcare professionals and artificial intelligence researchers may speed up the process of adoption and ensure state-of-the-art techniques are being used to improve health outcomes.

Author contributions
MS contributed with conceptualisation, methodology, data curation, formal analysis, investigation, software, validation and writing—original draft. XT contributed with conceptualisation, methodology, formal analysis, editing and supervision. SE contributed with conceptualisation and supervision. RG contributed with conceptualisation, supervision and administration. XZ contributed with conceptualisation and supervision. URA contributed with methodology, formal analysis, editing and supervision. YL contributed with methodology. All authors read and approved the final manuscript.

Funding
This work is partially funded by The Cannan Institute, Belmont Private Hospital, Brisbane.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable as this is a survey article of existing literature.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba, QLD, Australia. [2]Belmont Private Hospital, QLD, Brisbane, Australia. [3]School of Business, University of Southern Queensland, Springfield, QLD, Australia. [4]School of Computer Science, Queensland University of Technology, Brisbane, QLD, Australia.

Received: 22 October 2022 Accepted: 8 March 2023
Published online: 24 April 2023

References

1. Allison S, Bastiampillai T, O'Reilly R et al (2018) Access block to psychiatric inpatient admission: implications for national mental health service planning. Aust N Z J Psychiatry 52(12):1213–1214. https://doi.org/10.1177/0004867418802901
2. Allison S, Bastiampillai T, Copolov D et al (2019) Psychiatric bed numbers in Australia. Lancet Psychiatry 6(10):e21. https://doi.org/10.1016/s2215-0366(19)30208-1
3. Wind TR, Rijkeboer M, Andersson G et al (2020) The COVID-19 pandemic: the 'black swan' for mental health care and a turning point for e-health. Internet Interv 20(100):317. https://doi.org/10.1016/j.invent.2020.100317
4. Bzdok D, Meyer-Lindenberg A (2018) Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging 3(3):223–230. https://doi.org/10.1016/j.bpsc.2017.11.007
5. Fröhlich H, Balling R, Beerenwinkel N et al (2018) From hype to reality: data science enabling personalized medicine. BMC Med 16(1):1–15. https://doi.org/10.1186/s12916-018-1122-7
6. Brunn M, Diefenbacher A, Courtet P et al (2020) The future is knocking: how artificial intelligence will fundamentally change psychiatry. Acad Psychiatry 44(4):461–466. https://doi.org/10.1007/s40596-020-01243-8
7. Doraiswamy PM, Blease C, Bodner K (2020) Artificial intelligence and the future of psychiatry: insights from a global physician survey. Artif Intell Med 102(101):753. https://doi.org/10.1016/j.artmed.2019.101753
8. Graham S, Depp C, Lee EE et al (2019) Artificial intelligence for mental health and mental illnesses: an overview. Curr Psychiatry Rep 21(11):1–18. https://doi.org/10.1007/s11920-019-1094-0

9. Jiang F, Jiang Y, Zhi H et al (2017) Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2(4):230–243. https://doi.org/10.1136/svn-2017-000101
10. Carrillo F, Sigman M, Slezak DF et al (2018) Natural speech algorithm applied to baseline interview data can predict which patients will respond to psilocybin for treatment-resistant depression. J Affect Disord 230:84–86. https://doi.org/10.1016/j.jad.2018.01.006
11. Drysdale AT, Grosenick L, Downar J et al (2017) Erratum: Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 23(2):264. https://doi.org/10.1038/nm0217-264d
12. Yassin W, Nakatani H, Zhu Y et al (2020) Machine-learning classification using neuroimaging data in schizophrenia, autism, ultra-high risk and first-episode psychosis. Transl Psychiatry 10(1):278. https://doi.org/10.1038/s41398-020-00965-5
13. Allsopp K, Read J, Corcoran R et al (2019) Heterogeneity in psychiatric diagnostic classification. Psychiatry Res 279:15–22. https://doi.org/10.1016/j.psychres.2019.07.005
14. Hasanzadeh F, Mohebbi M, Rostami R (2019) Prediction of rTMS treatment response in major depressive disorder using machine learning techniques and nonlinear features of EEG signal. J Affect Disord 256:132–142. https://doi.org/10.1016/j.jad.2019.05.070
15. Khodayari-Rostamabad A, Reilly JP, Hasey GM et al (2013) A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. Clin Neurophysiol 124(10):1975–1985. https://doi.org/10.1016/j.clinph.2013.04.010
16. Chang B, Choi Y, Jeon M et al (2019) ARPNet: antidepressant response prediction network for major depressive disorder. Genes 10(11):907. https://doi.org/10.3390/genes10110907
17. Dick S (2019) Artificial intelligence. Issue 1. https://doi.org/10.1162/99608f92.92fe150c
18. Garnelo M, Shanahan M (2019) Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. Curr Opin Behav Sci 29:17–23. https://doi.org/10.1016/j.cobeha.2018.12.010
19. Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65(6):386–408. https://doi.org/10.1037/h0042519
20. Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61:85–117. https://doi.org/10.1016/j.neunet.2014.09.003
21. Zhang W, Yang G, Lin Y, et al (2018) On definition of deep learning. In: 2018 world automation congress (WAC). IEEE. https://doi.org/10.23919/wac.2018.8430387
22. Sheu YH (2020) Illuminating the black box: interpreting deep neural network models for psychiatric research. Front Psychiatry 11:551299. https://doi.org/10.3389/fpsyt.2020.551299
23. Bzdok D, Altman N, Krzywinski M (2018) Statistics versus machine learning. Nat Methods 15(4):233–234. https://doi.org/10.1038/nmeth.4642
24. Pennebaker J, Boyd R, Jordan K et al (2015) The development and psychometric properties of LIWC2015. Univeristy of Texas Austin, Austin
25. Bradley MLP (1999) Affective norms for English words (ANEW): instruction manual and affective rating. The Center for Research in Psychophysiology
26. Reagan A (2018) labMTsimple documentation
27. Blei DM, Ng AY, Jordan MI et al (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
28. Beysolow T II (2018) Applied natural language processing with python: implementing machine learning and deep learning algorithms for natural language processing. Apress, Berkeley
29. Deshpande M, Rao V (2017) Depression detection using emotion artificial intelligence. In: 2017 International conference on intelligent sustainable systems (ICISS), pp 858–862. https://doi.org/10.1109/ISS1.2017.8389299
30. Hassan AU, Hussain J, Hussain M, et al (2017) Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. In: 2017 International conference on information and communication technology convergence (ICTC), pp 138–140. https://doi.org/10.1109/ICTC.2017.8190959
31. Reece AG, Reagan AJ, Lix KLM et al (2017) Forecasting the onset and course of mental illness with twitter data. Sci Rep 7(1):13006. https://doi.org/10.1038/s41598-017-12961-9
32. Tadesse MM, Lin H, Xu B et al (2019) Detection of depression-related posts in reddit social media forum. IEEE Access 7:44883–44893. https://doi.org/10.1109/access.2019.2909180
33. Tsugawa S, Kikuchi Y, Kishino F, et al (2015) Recognizing depression from twitter activity. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems—CHI '15. ACM Press. https://doi.org/10.1145/2702123.2702280
34. Alhanai T, Ghassemi M, Glass J (2018) Detecting depression with audio/text sequence modeling of interviews. In: Interspeech 2018. ISCA. https://doi.org/10.21437/interspeech.2018-2522
35. McGinnis EW, Anderau SP, Hruschak J et al (2019) Giving voice to vulnerable children: machine learning analysis of speech detects anxiety and depression in early childhood. IEEE J Biomed Health Inform 23(6):2294–2301. https://doi.org/10.1109/jbhi.2019.2913590
36. Tasnim M, Stroulia E (2019) Detecting depression from voice. Advances in artificial intelligence. Springer International Publishing, Cham, pp 472–478. https://doi.org/10.1007/978-3-030-18305-9_47
37. Ray A, Kumar S, Reddy R, et al (2019) Multi-level attention network using text, audio and video for depression prediction. In: Proceedings of the 9th international on audio/visual emotion challenge and workshop—AVEC '19. ACM Press. https://doi.org/10.1145/3347320.3357697
38. Yang L, Sahli H, Xia X, et al (2017) Hybrid depression classification and estimation from audio video and text information. In: Proceedings of the 7th annual workshop on audio/visual emotion challenge—AVEC '17. ACM Press. https://doi.org/10.1145/3133944.3133950
39. De Choudhury M, Gamon M, Counts S et al (2013) Predicting depression via social media. Proc Int AAAI Conf Web Social Media 7(1):128–137
40. Radloff LS (1977) The CES-D scale. Appl Psychol Meas 1(3):385–401. https://doi.org/10.1177/014662167700100306
41. Tausczik YR, Pennebaker JW (2009) The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol 29(1):24–54. https://doi.org/10.1177/0261927x09351676
42. Shen G, Jia J, Nie L, et al (2017) Depression detection via harvesting social media: a multimodal dictionary learning solution. In: Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI-17, pp 3838–3844. https://doi.org/10.24963/ijcai.2017/536, https://doi.org/10.24963/ijcai.2017/536
43. Islam MR, Kamal ARM, Sultana N, et al (2018) Detecting depression using k-nearest neighbors (KNN) classification technique. In: 2018 international conference on computer, communication, chemical, material and electronic engineering (IC4ME2). IEEE. https://doi.org/10.1109/ic4me2.2018.8465641
44. Pirina I, Çöltekin Ç (2018) Identifying depression on reddit: the effect of training data. In: Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task. Association for Computational Linguistics. https://doi.org/10.18653/v1/w18-5903
45. Cummins N, Sethu V, Epps J et al (2015) Analysis of acoustic space variability in speech affected by depression. Speech Commun 75:27–49. https://doi.org/10.1016/j.specom.2015.09.003
46. Degottex G, Kane J, Drugman T, et al (2014) COVAREP—a collaborative voice analysis repository for speech technologies. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE. https://doi.org/10.1109/icassp.2014.6853739
47. Eyben F, Wöllmer M, Schuller B (2010) Opensmile. In: Proceedings of the international conference on Multimedia—MM '10. ACM Press. https://doi.org/10.1145/1873951.1874246
48. Baltrusaitis T, Zadeh A, Lim YC et al (2018) OpenFace 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE. https://doi.org/10.1109/fg.2018.00019
49. Cong Q, Feng Z, Li F, et al (2018) X-A-BiLSTM: a deep learning approach for depression detection in imbalanced data. In: 2018 IEEE international conference on bioinformatics and biomedicine (BIBM), pp 1624–1627. https://doi.org/10.1109/BIBM.2018.8621230
50. Mikolov T, Sutskever I, Chen K, et al (2013) Distributed representations of words and phrases and their compositionality. https://doi.org/10.48550/ARXIV.1310.4546
51. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Empirical methods in natural language

processing (EMNLP), pp 1532–1543. http://www.aclweb.org/anthology/D14-1162
52. Devlin J, Chang MW, Lee K, et al (2018) Bert: pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/ARXIV.1810.04805
53. Rosa RL, Schwartz GM, Ruggiero WV et al (2019) A knowledge-based recommendation system that includes sentiment analysis and deep learning. IEEE Trans Ind Inform 15(4):2124–2135
54. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x
55. Yates A, Cohan A, Goharian N (2017) Depression and self-harm risk assessment in online forums. CoRR abs/1709.01848. http://arxiv.org/abs/1709.01848, https://arxiv.org/abs/arXiv:1709.01848
56. Gratch J, Artstein R, Lucas GM, et al (2014) The distress analysis interview corpus of human and computer interviews. In: LREC, pp 3123–3128
57. Chen Y, Zhou B, Zhang W, et al (2018) Sentiment analysis based on deep learning and its application in screening for perinatal depression. In: 2018 IEEE third international conference on data science in cyberspace (DSC). IEEE. https://doi.org/10.1109/dsc.2018.00073
58. Kabir M, Ahmed T, Hasan MB et al (2023) DEPTWEET: a typology for social media texts to detect depression severities. Comput Hum Behav 139(107):503. https://doi.org/10.1016/j.chb.2022.107503
59. Ansari L, Ji S, Chen Q et al (2022) Ensemble hybrid learning methods for automated depression detection. IEEE Trans Comput Soc Syst. https://doi.org/10.1109/tcss.2022.3154442
60. Wani MA, ELAffendi MA, Shakil KA et al (2022) Depression screening in humans with AI and deep learning techniques. IEEE Trans Comput Soc Syst. https://doi.org/10.1109/tcss.2022.3200213
61. Nemesure MD, Heinz MV, Huang R et al (2021) Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Sci Rep 11(1):1–9. https://doi.org/10.1038/s41598-021-81368-4
62. Zogan H, Razzak I, Jameel S, et al (2021) DepressionNet: learning multimodalities with user post summarization for depression detection on social media. In: Proceedings of the 44th international ACM SIGIR Conference on Research and Development in Information Retrieval. ACM. https://doi.org/10.1145/3404835.3462938
63. Wan Z, Huang J, Zhang H et al (2020) HybridEEGNet: a convolutional neural network for EEG feature learning and depression discrimination. IEEE Access 8:30332–30342. https://doi.org/10.1109/access.2020.2971656
64. Ringeval F, Pantic M, Schuller B, et al (2017) AVEC 2017. In: Proceedings of the 7th annual workshop on audio/visual emotion challenge—AVEC '17. ACM Press. https://doi.org/10.1145/3133944.3133953
65. Fried EI, Nesse RM (2015) Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR* D study. J Affect Disord 172:96–102. https://doi.org/10.1016/j.jad.2014.10.010
66. Moreland AD, Dumas JE (2008) Categorical and dimensional approaches to the measurement of disruptive behavior in the preschool years: a meta-analysis. Clin Psychol Rev 28(6):1059–1070. https://doi.org/10.1016/j.cpr.2008.03.001
67. Li M, Zhong N, Lu S et al (2016) Cognitive behavioral performance of untreated depressed patients with mild depressive symptoms. PLoS ONE 11(1):e0146356. https://doi.org/10.1371/journal.pone.0146356
68. Yang Y, Zhong N, Imamura K et al (2016) Task and resting-state fMRI reveal altered salience responses to positive stimuli in patients with major depressive disorder. PLoS ONE 11(5):e0155092. https://doi.org/10.1371/journal.pone.0155092
69. Alloghani M, Al-Jumeily D, Mustafina J et al (2019) A systematic review on supervised and unsupervised machine learning algorithms for data science. Unsupervised and semi-supervised learning. Springer International Publishing, Cham, pp 3–21. https://doi.org/10.1007/978-3-030-22475-2_1
70. Bickman L (2020) Improving mental health services: a 50-year journey from randomized experiments to artificial intelligence and precision mental health. Adm Policy Ment Health Ment Health Serv Res 47(5):795–843. https://doi.org/10.1007/s10488-020-01065-8
71. Kuai H, Zhong N, Chen J et al (2021) Multi-source brain computing with systematic fusion for smart health. Inf Fusion 75:150–167. https://doi.org/10.1016/j.inffus.2021.03.009
72. Jaworska N, de la Salle S, Ibrahim MH et al (2019) Leveraging machine learning approaches for predicting antidepressant treatment response using electroencephalography (EEG) and clinical data. Front Psychiatry. https://doi.org/10.3389/fpsyt.2018.00768
73. Pigott HE, Leventhal AM, Alter GS et al (2010) Efficacy and effectiveness of antidepressants: current status of research. Psychother Psychosom 79(5):267–279. https://doi.org/10.1159/000318293
74. Trivedi MH, Rush AJ, Wisniewski SR et al (2006) Evaluation of outcomes with citalopram for depression using measurement-based care in STAR* D: implications for clinical practice. Am J Psychiatry 163(1):28–40. https://doi.org/10.1176/appi.ajp.163.1.28
75. Zandvakili A, Philip NS, Jones SR et al (2019) Use of machine learning in predicting clinical response to transcranial magnetic stimulation in comorbid posttraumatic stress disorder and major depression: a resting state electroencephalography study. J Affect Disord 252:47–54. https://doi.org/10.1016/j.jad.2019.03.077
76. Berlim MT, van den Eynde F, Tovar-Perdomo S et al (2013) Response, remission and drop-out rates following high-frequency repetitive transcranial magnetic stimulation (rTMS) for treating major depression: a systematic review and meta-analysis of randomized, double-blind and sham-controlled trials. Psychol Med 44(2):225–239. https://doi.org/10.1017/s0033291713000512
77. Fitzgerald PB, Hoy KE, Anderson RJ et al (2016) A study of the pattern of response to rTMS treatment in depression. Depress Anxiety 33(8):746–753. https://doi.org/10.1002/da.22503
78. Koutsouleris N, Wobrock T, Guse B et al (2017) Predicting response to repetitive transcranial magnetic stimulation in patients with schizophrenia using structural magnetic resonance imaging: a multisite machine learning analysis. Schizophr Bull 44(5):1021–1034. https://doi.org/10.1093/schbul/sbx114
79. Lee Y, Ragguett RM, Mansur RB et al (2018) Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J Affect Disord 241:519–532. https://doi.org/10.1016/j.jad.2018.08.073
80. Bailey N, Hoy K, Rogasch N et al (2018) Responders to rTMS for depression show increased fronto-midline theta and theta connectivity compared to non-responders. Brain Stimul 11(1):190–203. https://doi.org/10.1016/j.brs.2017.10.015
81. Bailey N, Hoy K, Rogasch N et al (2019) Differentiating responders and non-responders to rTMS treatment for depression after one week using resting EEG connectivity measures. J Affect Disord 242:68–79. https://doi.org/10.1016/j.jad.2018.08.058
82. Erguzel TT, Ozekes S, Gultekin S et al (2015) Neural network based response prediction of rTMS in major depressive disorder using QEEG cordance. Psychiatry Investig 12(1):61. https://doi.org/10.4306/pi.2015.12.1.61
83. Fan J, Tso IF, Maixner DF et al (2019) Segregation of salience network predicts treatment response of depression to repetitive transcranial magnetic stimulation. NeuroImage: Clin 22:101719. https://doi.org/10.1016/j.nicl.2019.101719
84. Chen D, Lei X, Du L et al (2022) Use of machine learning in predicting the efficacy of repetitive transcranial magnetic stimulation on treating depression based on functional and structural thalamo-prefrontal connectivity: a pilot study. J Psychiatr Res 148:88–94. https://doi.org/10.1016/j.jpsychires.2022.01.064
85. Hopman H, Chan S, Chu W et al (2021) Personalized prediction of transcranial magnetic stimulation clinical response in patients with treatment-refractory depression using neuroimaging biomarkers and machine learning. J Affect Disord 290:261–271. https://doi.org/10.1016/j.jad.2021.04.081
86. Rostami R, Kazemi R, Nitsche MA et al (2017) Clinical and demographic predictors of response to rTMS treatment in unipolar and bipolar depressive disorders. Clin Neurophysiol 128(10):1961–1970. https://doi.org/10.1016/j.clinph.2017.07.395

87. Joormann J, Gotlib IH (2008) Updating the contents of working memory in depression: interference from irrelevant negative material. J Abnormal Psychol 117(1):182–192. https://doi.org/10.1037/0021-843x.117.1.182
88. Montgomery SA, Åsberg M (1979) A new depression scale designed to be sensitive to change. Br J Psychiatry 134(4):382–389. https://doi.org/10.1192/bjp.134.4.382
89. Hardmeier M, Hatz F, Bousleiman H et al (2014) Reproducibility of functional connectivity and graph measures based on the phase lag index (PLI) and weighted phase lag index (wPLI) derived from high resolution EEG. PLoS ONE 9(10):e108648. https://doi.org/10.1371/journal.pone.0108648
90. Tas C, Cebi M, Tan O et al (2015) EEG power, cordance and coherence differences between unipolar and bipolar depression. J Affect Disord 172:184–190. https://doi.org/10.1016/j.jad.2014.10.001
91. Mohanty R, Sethares WA, Nair VA et al (2020) Rethinking measures of functional connectivity via feature extraction. Sci Rep 10(1):1298. https://doi.org/10.1038/s41598-020-57915-w
92. Xiao R, Shida-Tokeshi J, Vanderbilt DL et al (2018) Electroencephalography power and coherence changes with age and motor skill development across the first half year of life. PLoS ONE 13(1):e0190276. https://doi.org/10.1371/journal.pone.0190276
93. Goodman MS, Kumar S, Zomorrodi R et al (2018) Theta-gamma coupling and working memory in Alzheimer's dementia and mild cognitive impairment. Front Aging Neurosci 10:101. https://doi.org/10.3389/fnagi.2018.00101
94. Blevins CA, Weathers FW, Davis MT et al (2015) The posttraumatic stress disorder checklist for DSM-5 (PCL-5): development and initial psychometric evaluation. J Trauma Stress 28(6):489–498. https://doi.org/10.1002/jts.22059
95. Rush AJ, Gullion CM, Basco MR et al (1996) The inventory of depressive symptomatology (IDS): psychometric properties. Psychol Med 26(3):477–486. https://doi.org/10.1017/s0033291700035558
96. Rush AJ, Carmody T, Reimitz PE (2000) The inventory of depressive symptomatology (IDS): clinician (IDS-C) and self-report (IDS-SR) ratings of depressive symptoms. Int J Methods Psychiatr Res 9(2):45–59
97. Hamilton M (1960) A rating scale for depression. J Neurol Neurosurg Psychiatry 23(1):56–62. https://doi.org/10.1136/jnnp.23.1.56
98. Beck AT (1961) An inventory for measuring depression. Arch Gen Psychiatry 4(6):561–571. https://doi.org/10.1001/archpsyc.1961.01710120031004
99. Bares M, Brunovsky M, Novak T et al (2014) QEEG theta cordance in the prediction of treatment outcome to prefrontal repetitive transcranial magnetic stimulation or venlafaxine ER in patients with major depressive disorder. Clin EEG Neurosci 46(2):73–80. https://doi.org/10.1177/1550059413520442
100. Hunter AM, Nghiem TX, Cook IA et al (2017) Change in quantitative EEG theta cordance as a potential predictor of repetitive transcranial magnetic stimulation clinical outcome in major depressive disorder. Clin EEG Neurosci 49(5):306–315. https://doi.org/10.1177/1550059417746212
101. Kennedy NI, Lee WH, Frangou S (2018) Efficacy of non-invasive brain stimulation on the symptom dimensions of schizophrenia: a meta-analysis of randomized controlled trials. Eur Psychiatry 49:69–77. https://doi.org/10.1016/j.eurpsy.2017.12.025
102. Shi C, Yu X, Cheung EF et al (2014) Revisiting the therapeutic effect of rTMS on negative symptoms in schizophrenia: a meta-analysis. Psychiatry Res 215(3):505–513. https://doi.org/10.1016/j.psychres.2013.12.019
103. Kay SR, Fiszbein A, Opler LA (1987) The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull 13(2):261–276. https://doi.org/10.1093/schbul/13.2.261
104. Picchioni MM, Murray RM (2007) Schizophrenia. BMJ 335(7610):91–95. https://doi.org/10.1136/bmj.39227.616447.be
105. Tian S, Sun Y, Shao J et al (2019) Predicting escitalopram monotherapy response in depression: the role of anterior cingulate cortex. Hum Brain Mapp 41(5):1249–1260. https://doi.org/10.1002/hbm.24872
106. Ghahramani Z, Hinton GE, et al (1996) The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, University of Toronto
107. Ransohoff DF (2004) Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 4(4):309–314. https://doi.org/10.1038/nrc1322
108. Browning M, Kingslake J, Dourish CT et al (2019) Predicting treatment response to antidepressant medication using early changes in emotional processing. Eur Neuropsychopharmacol 29(1):66–75. https://doi.org/10.1016/j.euroneuro.2018.11.1102
109. Pei C, Sun Y, Zhu J et al (2019) Ensemble learning for early-response prediction of antidepressant treatment in major depressive disorder. J Magn Reson Imaging 52(1):161–171. https://doi.org/10.1002/jmri.27029
110. Lin E, Kuo PH, Liu YL et al (2018) A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. Front Psychiatry. https://doi.org/10.3389/fpsyt.2018.00290
111. Mumtaz W, Xia L, Yasin MAM et al (2017) A wavelet-based technique to predict treatment outcome for major depressive disorder. PLoS ONE 12(2):e0171409. https://doi.org/10.1371/journal.pone.0171409
112. Chekroud AM, Zotti RJ, Shehzad Z et al (2016) Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry 3(3):243–250. https://doi.org/10.1016/s2215-0366(15)00471-x
113. Patel MJ, Andreescu C, Price JC et al (2015) Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. Int J Geriatr Psychiatry 30(10):1056–1067. https://doi.org/10.1002/gps.4262
114. Yang Y (2017) Ensemble learning. Temporal data mining via unsupervised ensemble learning. Elsevier, Amsterdam, pp 35–56. https://doi.org/10.1016/b978-0-12-811654-8.00004-x
115. Al-Mosaiwi M, Johnstone T (2018) In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clin Psychol Sci 6(4):529–542. https://doi.org/10.1177/2167702617747074
116. Edwards T, Holtzman NS (2017) A meta-analysis of correlations between depression and first person singular pronoun use. J Res Personal 68:63–68. https://doi.org/10.1016/j.jrp.2017.02.005
117. Rude S, Gortner EM, Pennebaker J (2004) Language use of depressed and depression-vulnerable college students. Cognit Emot 18(8):1121–1133. https://doi.org/10.1080/02699930441000030
118. Stirman SW, Pennebaker JW (2001) Word use in the poetry of suicidal and nonsuicidal poets. Psychosom Med 63(4):517–522. https://doi.org/10.1097/00006842-200107000-00001
119. Ziemer KS, Korkmaz G (2017) Using text to predict psychological and physical health: a comparison of human raters and computerized text analysis. Comput Hum Behav 76:122–127. https://doi.org/10.1016/j.chb.2017.06.038
120. Rush A, Trivedi MH, Ibrahim HM et al (2003) The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-c), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 54(5):573–583. https://doi.org/10.1016/s0006-3223(02)01866-8
121. Harrell FE Jr (2015) Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer, Berlin
122. Birkenbihl C, Emon MA et al (2020) Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia—lessons for translation into clinical practice. EPMA J 11(3):367–376. https://doi.org/10.1007/s13167-020-00216-z
123. Chen J, Wang N, Deng Y et al (2020) Wisdom as a service for mental health care. IEEE Trans Cloud Comput 8(2):539–552. https://doi.org/10.1109/tcc.2015.2464820
124. Arbabshirani MR, Plis S, Sui J et al (2017) Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. NeuroImage 145:137–165. https://doi.org/10.1016/j.neuroimage.2016.02.079
125. Widge AS, Bilge MT, Montana R et al (2019) Electroencephalographic biomarkers for treatment response prediction in major depressive illness: a meta-analysis. Am J Psychiatry 176(1):44–56. https://doi.org/10.1176/appi.ajp.2018.17121358
126. Button KS, Ioannidis JPA, Mokrysz C et al (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14(5):365–376. https://doi.org/10.1038/nrn3475
127. Sgaier SK, Huang V, Charles G (2020) The case for causal AI. Stanf Soc Innov Rev 18:50–55. https://doi.org/10.48558/KT81-SN73
128. Wang X, Xu X, Tong W et al (2022) DeepCausality: a general AI-powered causal inference framework for free text: a case study of LiverTox. Front Artif Intell. https://doi.org/10.3389/frai.2022.999289

Publisher's Note

2.3 Links and implications

This work identified the potential opportunity for the use of DL for treatment response prediction. While some work has explored the us of DL for evaluating and predicting response to frontline depression treatments, such as antidepressants. A dearth of literature exists applying DL for predicting rTMS treatment outcomes, existing work in rTMS has focused largely on shallow ML techniques. Furthermore, this work highlights the issue of small datasets which is discussed further in Chapter 3.3 which emphasises the importance of mitigating data bias to improve AI fairness.

For the preceding chapters, however, this chapter has identified one of the primary challenges in psychiatry a lack of objective biomarkers for detecting and diagnosing depression. XAI techniques then provide a useful tool for understanding the influence of certain biomarkers on response to treatment and hence their impact on depression.

CHAPTER 3: PAPER 2 - IDENTIFYING PREDICATIVE BIOMARKERS FOR REPETITIVE TRANSCRANIAL MAGNETIC STIMULATION RESPONSE IN DEPRESSION PATIENTS WITH EXPLAINABILITY

3.1 Introduction

This chapter presents the work *Identifying predictive biomarkers for repetitive transcranial magnetic stimulation response in depression with explainability.* This work extends on the findings of the previous chapter and in consultation with our industry partners at Belmont Private Hospital compares the strength of a variety of demographic, psychological and physiological variables for predicting the outcome of rTMS treatment. This work serves as a proof of concept for the use of DL for the prediction of rTMS treatment response in the first of its kind work. Furthermore, this chapter shows the utility of XAI techniques for identifying predictive biomarkers which may be of use to clinicians and researchers alike

3.2 Published paper

Computer Methods and Programs in Biomedicine 242 (2023) 107771

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Identifying predictive biomarkers for repetitive transcranial magnetic stimulation response in depression patients with explainability

Matthew Squires [a,*], Xiaohui Tao [a,*], Soman Elangovan [b], Raj Gururajan [c], Xujuan Zhou [c], Yuefeng Li [d], U. Rajendra Acharya [e]

[a] School of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba, Australia
[b] Belmont Private Hospital, Brisbane, Australia
[c] School of Business, University of Southern Queensland, Springfield, Australia
[d] School of Computer Science, Queensland University of Technology, Brisbane, Australia
[e] School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, Australia

ARTICLE INFO

Keywords:
Repetitive transcranial magnetic stimulation
Deep learning
Explainable AI
Depression

ABSTRACT

Repetitive Transcranial Magnetic Stimulation (rTMS) is an evidence-based treatment for depression. However, the patterns of response to this treatment modality are inconsistent. Whilst many people see a significant reduction in the severity of their depression following rTMS treatment, some patients do not. To support and improve patient outcomes, recent work is exploring the possibility of using Machine Learning to predict rTMS treatment outcomes. Our proposed model is the first to combine functional magnetic resonance imaging (fMRI) connectivity with deep learning techniques to predict treatment outcomes before treatment starts. Furthermore, with the use of Explainable AI (XAI) techniques, we identify potential biomarkers that may discriminate between rTMS responders and non-responders. Our experiments utilize 200 runs of repeated bootstrap sampling on two rTMS datasets. We compare performances between our proposed feedforward deep neural network against existing methods, and compare the average accuracy, balanced accuracy and F1-score on a held-out test set. The results of these experiments show that our model outperforms existing methods with an average accuracy of 0.9423, balanced accuracy of 0.9423, and F1-score of 0.9461 in a sample of 61 patients. We found that functional connectivity measures between the Subgenual Anterior Cingulate Cortex and Centeral Opercular Cortex are a key determinant of rTMS treatment response. This knowledge provides psychiatrists with further information to explore the potential mechanisms of responses to rTMS treatment. Our developed prototype is ready to be deployed across large datasets in multiple centres and different countries.

1. Introduction

Depression is a highly prevalent and debilitating mental illness [1]. As such, finding effective and efficient treatments for depression is a high priority. Repetitive Transcranial Magnetic Stimulation (rTMS) is an evidence-based treatment for depression [2–4]. rTMS involves electromagnetic stimulation of the brain that aims to alter its underlying structures to improve a patient's symptoms [7]. However, the patterns of response to this treatment are inconsistent [5]. Evidence [6,9,5] suggests that the distribution of response to rTMS is bimodal. For some patients, rTMS treatment will lead to a significant reduction in depression severity. However, others see minimal improvement in their depression rating scale scores post-treatment. Given this disparity, current work is investigating the potential of using artificial intelligence (AI) to predict treatment outcomes and personalize mental healthcare.

To date, existing research has sought to predict response to rTMS treatment using machine learning (ML) algorithms. These systems aim to delineate between responders and non-responders in rTMS treatment. Thus, the problem can be defined as a supervised binary classification task. Existing works [21,22,11–13,15–17,10,9,18] have applied a variety of algorithms to predict the response to rTMS treatment. Methods include linear support vector machines [22,11,12], linear regression [13]

* Corresponding authors.
E-mail addresses: matthew.squires@usq.edu.au (M. Squires), xiaohui.tao@usq.edu.au (X. Tao), soman.elangovan@healthecare.com.au (S. Elangovan), Raj.Gururajan@usq.edu.au (R. Gururajan), xujuan.zhou@usq.edu.au (X. Zhou), y2.li@qut.edu.au (Y. Li), Rajendra.Acharya@usq.edu.au (U.R. Acharya).

https://doi.org/10.1016/j.cmpb.2023.107771
Received 8 June 2023; Received in revised form 12 August 2023; Accepted 19 August 2023
Available online 25 August 2023

and k-nearest neighbours [15]. These surveyed methods vary from those that rely on features collected after treatment has begun, to emerging methods that utilize pre-treatment measures only.

Predicting the treatment outcome before it begins is the goal of personalized mental healthcare. However, such examples are in the minority. For example, only Hopman et al. [11] and Hasanzadeh et al. [15] predicted treatment outcomes before starting treatment. In their work, Hasanzadeh et al. [15] utilized pre-treatment EEG features to accurately predict rTMS treatment outcomes in roughly 90% of cases. Hopman et al. [11] instead used pre-treatment functional magnetic resonance imaging (fMRI) to predict treatment outcomes. These results suggest that there is scope for the use of more advanced techniques, such as deep neural networks (DNN), to predict the outcome of rTMS in patients [7]. This observation is echoed by [12], who assert that future work could explore the efficacy of deep learning (DL) algorithms for predicting the treatment response. For example, the linear support vector machine (SVM) used by Hopman et al. [11] performed excellently during cross-validation, however, Hopman et al. [11] reported sharp declines in predictive performance on a held-out test set. Therefore, an opportunity exists to explore more sophisticated algorithms, such as DNNs, which are known - under the right settings - to generalize well on unseen data. As such, our work seeks to address the following research questions:

1. Can a deep neural network improve upon existing methods for predicting treatment response on a held-out test set?
2. Which features most predict treatment response?
3. In what circumstances is the proposed network vulnerable to misclassification?
4. Are there any commonalities in misclassification errors that can be communicated to the end user to improve clinical utility?

To address these research questions, we compare empirically existing shallow ML methods against our proposed DNN. Furthermore, with the aim of increasing the value of work for end users and in collaboration with domain experts, we utilize explainable artificial intelligence (XAI) techniques to identify the features that are most predictive of treatment response. In addressing this research question, we aim to identify candidate biomarkers indicative of treatment response. Additionally, to support the potential implementation of our model, we present model knowledge, which is an extension of the 'model facts labels' presented by [20]. Model knowledge is our process of rigorously evaluating model performance, including potential limitations. By gathering model knowledge, we can enhance the clinical utility by explicitly declaring circumstances such as when the model performs well or is vulnerable to prediction errors, which in turn promotes trust in end users. Lack of trust in AI models is seen as a key barrier to its implementation in healthcare [19]. Through addressing these research questions, our work makes the following contributions:

- A robustly validated regularised deep feedforward neural network that predicts the treatment outcome of rTMS before treatment commences.
- A robust analysis of rTMS treatment response patterns through the use of two datasets, namely, the differences in the predictive power of self-reported psychometric data against fMRI connectivity measures.
- The use of XAI techniques, including SHAPLEY values, to identify candidate biomarkers indicative of response to rTMS treatment.

These findings help to provide confidence to clinicians in our model while also uncovering new knowledge for depression researchers.

The current work proposes a DNN model for predicting treatment response to rTMS. We use multi-modal data to predict treatment outcomes and explore XAI techniques to add support and robustness to our model. As such, our work aims to produce the first DNN model to model rTMS treatment outcomes that includes explainability. The paper is structured as follows. The following section reviews existing strategies for treatment response prediction used to evaluate rTMS. Given the dearth of literature exploring DL architectures in rTMS, we include an exploration of DL systems applied in other medical contexts. Additionally, we survey some methods used to produce interpretable AI systems. In Section 2, we present the research problem, a summary of the dataset, formally define the research problem and introduce the notation. Section 2 also introduces our proposed model, the baseline models, and model hyperparameters. Details on the experiment design, performance measuring schemes and experiment results are included in Section 3. Finally, comments about our findings and proposed future directions from this research topic are included in Section 4.

1.1. Related work

Previous studies [21,22,11–13,15–17,10,9,18] have applied a variety of techniques to predict treatment outcomes to rTMS. To date, ML algorithms have performed well on this binary classification task. A summary of the current literature is shown in Table 1.

This Table shows that several feature modalities, including electroencephalogram (EEG), fMRI, psychological, and demographic features, have been used to predict treatment response to rTMS. Existing work has relied on shallow machine learning methods like linear SVM (LSVM) and k-nearest neighbours algorithms (KNN). Recently, Shadabi et al. [22] became the first to apply DL methods to rTMS response prediction when they explored the ability of a convolutional neural network (CNN) to predict treatment response to rTMS using EEG features. CNNs are well suited to the temporal data collected by EEG, with the authors reporting a 97.1% average accuracy after 10 fold cross validation.

Previously, existing research has relied on shallow ML methods. For example, Hopman et al. [11] deployed a LSVM using features collected via fMRI. They used connectivity features between the subgenual anterior cingulate cortex and lateral occipital cortex, superior parietal lobule, frontal pole and central opercular cortex. During five-fold cross validation, the authors presented a training accuracy of $\approx 97\%$ however, on a unseen test set, model performance dropped to an average of $\approx 87\%$, with a 95% confidence interval from 100% to roughly 70% accuracy. Further implementations of a LSVM include Bailey et al. [12], who built a LSVM classifier composed of 54 features. These features consisted of a combination of mood and EEG measurements collected at baseline and after one week of treatment. In addition to measurements collected at these two-time points, features were extracted for the change between week 1 and baseline. Each feature was standardized, which is a common technique in ML. Testing of the final classifier was validated against 5000 runs of five-fold validation. For this LSVM, Bailey et al. [12] reported a mean balanced accuracy of 86.60%. As part of their conclusions, they felt that the efficacy of existing algorithms for the prediction of treatment response could be improved [12].

DL algorithms are capable of modelling complex relationships and yield high classification performances. Moving from ML to DL to predict rTMS treatment outcomes is the potential next step in mental healthcare. The strength of DL architectures is the ability to model complex multi-variable relationships with improved accuracy [26]. Hence, the opportunity exists for DL methods to be applied to rTMS modelling using fMRI connectivity features. However, the challenge in applying DL methods to critical domains such as mental health care is the distrust toward DL methods due to their lack of interpretability [19]. XAI is a field of AI research that focuses on the inner workings of complex DL models. DL models are more powerful for identifying relationships than more interpretable shallow methods. However, there is a trade-off between performance and interpretability. XAI techniques aim to eliminate that trade-off by increasing the interpretability of DL models. Furthermore, Hopman et al. [11] observed that their LSVM failed to generalise well to unseen data. By contrast, DNNs are known to gener-

Table 1
rTMS depression treatment response prediction.

Author	Modality	Features	Algorithm	Performance	Validation
Ebrahimzadeh et al. [21]	EEG	EEG beta power, Correlation Dimension (CD), Permutation entropy (PE), Fractal dimension (FD), Lempel-Ziv Complexity (LZC), Power spectral density, Frontal and prefrontal cordance	SVM*	94.31% average accuracy after cross validation	10 fold cross validation
Shahabi et al. [22]	EEG	Continuous Wavelet Transform	CNN	97.1% average accuracy after cross validation	ten-fold cross validation
Hopman et al. [11]	fMRI	Connectivity features: subgenual anterior cingulate cortex, lateral occipital cortex, superior parietal lobule, frontal pole and central opercular cortex	LSVM	87% accuracy on a held out test set	five-fold cross validation
Bailey et al. [12]	EEG and Mood	Alpha power, theta power, alpha connectivity, theta connectivity, theta cordance, individualised alpha peak frequency (iAPF) and MADRS	LSVM	Mean balanced accuracy of 86.60%	5000 runs of five-fold cross validation
Fan et al. [13]	fMRI	Network Segregation of the Salience Network*	Regression	Coefficient of determination of 0.27	NA
Hasanzadeh et al. [15]	EEG	Power of beta*	K-NN	Accuracy of 91.3%	Leave-one-out cross validation
Bailey et al. [17]	Mood, Behaviour and EEG	Alpha power, theta power, gamma power, alpha connectivity, theta connectivity, gamma connectivity, theta gamma coupling, MADRS, working memory and reaction time	LSVM	F1 score = 0.93	200000 runs of five fold cross validation
Drysdale et al. [9]	fMRI	Connectivity features	Hierarchical Clustering and SVM	Balanced accuracy of 90.39%	Leave-one-out cross validation

* Best performing model.

alise well to unseen data, which potentially addresses the performance decline found in studies such as [11]. Additionally, the inclusion of explainability can improve trust in end users and take advantage of DNN's improved performance over existing methods.

At present, there is a dearth of literature exploring XAI and rTMS treatment. A recent review by [19] argued for the importance of including XAI through methodologies like SHAP values to enhance trust in DL methods. Thus, our work is motivated to enhance trust in DL methods from psychiatrists through both improving performance in rTMS response prediction, and including XAI in our approach.

2. Materials and methods

This section outlines the problem statement, the datasets used and defines the research problem to be explored. Here, we provide some background information on DNNs and their development, before we present our model, which uses quantitative data to evaluate the treatment effects of rTMS. Additionally, this section includes details about our strategies for reducing overfitting and internally validating our model.

2.1. Problem statement

The effectiveness of rTMS for the treatment of depression is now well-established [27]. Significant evidence shows rTMS to be a safe and effective intervention for treatment-resistant depression [2–4]. Despite this effectiveness, some patients will see no significant improvement in their depression severity following rTMS treatment [5]. To address these inconsistent response patterns, we are exploring ways to better target rTMS treatment toward patients who are likely to see the most benefit. In order to personalize care, AI can be deployed to support psychiatrists [28]. The aim of our work is to explore the potential of a DNN architecture to predict response to rTMS treatment and identify any potential biomarkers indicative of treatment response.

2.2. Research design

The current work aims to test the efficacy of a DNN to predict the treatment outcome of rTMS. Utilising empirical experiments, we seek to investigate whether DL offers any improvement over existing methods. As part of this work, we identify the features that provide the most information for treatment response in the hope of identifying the key biomarkers. Additionally, our experiments compare self reported measures or fMRI connectivity features for predicting treatment response. By utilising XAI techniques, we present new knowledge that can aid clinicians in prescribing treatments.

2.3. Datasets

To address whether a DNN can provide robust predictions compared to existing methods, we utilise two datasets. The first dataset was used in published work by Hopman et al. [11] and made publicly available in [29]. The data includes several fMRI features, along with a patient's treatment outcomes. A summary of the mean connectivity measures across response type can be seen in Table 2. Further detailed summary statistics of this dataset, including associated ethics approval, can be found in [29].

The second dataset is new data collected from a large private hospital in Australia that specializes in the delivery of rTMS care. A summary of the relevant psychological variables collected in this data is shown in Table 3. This Table shows the mean survey scores between groups. The psychological health information in this Table was collected using

Table 2
Mean connectivity measurements by group in Dataset 1: Hopman [29].

	Responders	Non-responders
N (n = 61)	33	28
Frontal Pole Connectivity	0.0252	−0.1026
Occipital Cortex Connectivity	0.0667	−0.0467
Superior Parietal Lobule Connectivity	0.0802	−0.0450
Centeral Opercular Cortex Connectivity	0.1142	0.0266
Left Lateral Occipital Cortex Connectivity	0.0607	−0.0440
Right Lateral Occipital Cortex Connectivity	0.0386	−0.0476

Table 3
Mean DASS measurements by groups in Dataset 2: data collected from Belmont hospital.

	Responders	Non-responders
N (n = 133)	83	50
Depression Baseline	29.4819	27.4000
Anxiety Baseline	17.0843	16.3600
Stress Baseline	25.4578	23.4800
Depression after 10 sessions	17.1566	25.9800
Anxiety after 10 sessions	11.4698	13.7200
Stress after 10 sessions	15.3253	20.4400

the Depression, Anxiety and Stress Subscale [DASS 30]. DASS is a self-report survey measuring three dimensions of mental health: depression, anxiety and stress. Patients are required to complete a baseline survey prior to treatment, then in the rTMS program, they complete the DASS survey 3 times during treatment. An additional survey is completed after 10 sessions of rTMS, and a final measurement about patients is collected following treatment.

In the current work, the DASS-21 was used, a short form of the 42 item DASS. Each dimension of mental health in the DASS-21 has a maximum score of 42. For the depression dimension, a score of greater than 21 is deemed severe depression [23]. Participants in this study consented to DASS data being used for the study of rTMS treatment. Ethics approval was obtained from the University's Human Research Ethics Committee to use and analyse collected data.

2.4. Problem definition

The current work seeks a function that optimizes the classification of patients as responders or non-responders to rTMS treatment. In addition, this function provides a model of the treatment effect of rTMS dependent on either psychological or neuroimaging based variables.

Formally, let X be a dataset containing Patients P and class label Y, where:

$$X = \{(p_1, y_1), (p_2, y_2), (p_3, y_3) \dots (p_n, y_n)\} \tag{1}$$

Each patient p has a set of connectivity measures C such that

$$C = \{FP, OC, SPL, COC, lOCL, lOCR\}$$
$$y_i = \begin{cases} 1 & \text{if } \frac{\Delta d}{d_0} \leq -0.5 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Patients are assigned a class label according to the function in Equation (2). Any patient who experiences a greater than 50% reduction in depression severity, $\Delta d \leq -0.5$, is classified as a responder and assigned the class label $y_i = 0$. Conversely, patients who see a less than 50% reduction, $\Delta d > -0.5$, are classed as non responders and receive the label $y_i = 0$. The target variable for this binary classification task is y. As such, we seek a classifier

$$h[p] = y$$

which minimizes the prediction error between classes. See Table 4.

Table 4
Symbol descriptions.

Symbol	Description
X	Dataset
Y,*y*	Set of class labels and a patient's class label
p	A patient
d_0, Δd	Depression severity at baseline, change in depression severity following treatment
FP	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Frontal Pole
OC	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Occipital Cortex
SPL	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Superior Parietal Lobule
COC	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Centeral Opercular Cortex
lOCL	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Left Lateral Occipital Cortex
lOCR	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Right Lateral Occipital Cortex

2.5. Deep neural networks

DL is a subfield of ML that builds upon existing neural network architectures by increasing the number of hidden layers of a network [24]. This increased depth allows for modelling of increasingly complex nonlinear functions [25]. The complexity makes it possible for models to learn complex representations of existing data, which may not be observable using traditional inferential statistics or standard ML techniques.

The basis for the artificial neural network (ANN) is found in the seminal work of Rosenblatt et al. [31]. Where initially a single perceptron defined a linear decision boundary between a binary set of classes, the multilayer perceptron (MLP) adds the concept of a hidden layer. The hidden layer involves multiple perceptrons, with each perceptron sharing an edge with each node in the hidden layer. This increase in model complexity increases the model's predictive power beyond linear functions to learning complex nonlinear decision boundaries between classes [33]. The MLP is a feedforward neural network applied to classification and regression [32]. An MLP with several hidden layers is referred to as a DNN [8]

A crucial aspect of the performance of the MLP is the training of the network. Training refers to the model weights being tuned so that predicted outputs match the expected outputs or ground truth values of the data [33]. The process of tuning these weights or parameters can be referred to as learning. Rarely can a model match all examples with their ground truth labels, therefore, we need a function to monitor the performance of the model during training. Training aims to minimize a loss function to obtain the weights so that the difference between the expected and predicted outcomes is minimized [32].

2.6. Regularisation

Modern solutions have enabled the fitting of increasingly complex functions to data. However, the added complexity of networks with several hidden layers increases the risk of overfitting. That is, where functions simply memorise datasets. Regularisation encompasses a class of tools used to reduce the risk of model overfitting. Common strategies for reducing the risk of overfitting include: early stopping, weight regularisation and dropout [24].

Early stopping involves monitoring a metric during training and ending training when the selected value stops improving [34]. In addition to an unseen test dataset, we use a validation set during model training in our project. Validation loss is monitored throughout training, and for each trained model as part of our bootstrap resampling, patience was set to 100. We set a minimum improvement in validation loss of 0.05 as being required to continue model training.

Table 5
Model hyperparameters.

Hidden Layers	4
Layer Width	10
Activation Function	reLU
Loss Function	Binary Crossentopy
Regularisation Layers	4
Test set size	20%
Epochs	2000 or until early stopping criteria met

Dropout involves turning a proportion of parameter weights down to zero. Conceptually, we can perceive this as 'dropping' edges between nodes. Srivastava et al. [35] first proposed dropout as a regularisation strategy to add noise to a neural network. The introduction of noise through 'dropping' connections between neurons forces the network being trained on the data to identify the true nature of the signal within the data. In turn, this reduces an overparametrised network's ability to memorize the dataset. The benefit in identifying the true signal from the data means a greater potential for identifying meaningful patterns within the data. As such, each layer of our trained model includes a dropout probability of 0.3.

The final hyperparameters of our model are shown in Table 5. These final hyperparameters were selected after an iterative model building process. Through several cycles of experiments, we monitored how changes in model hyperparameters impacted model performance. Through continual refinement and the aim of creating a model robust to overfitting, we settled on the final model hyperparameters. These selected values achieve the goal of a model that generalizes well to unseen data when compared against existing methods.

2.7. Experimental design

This section provides an overview of the empirical experiments used to explore the research questions outlined in Section 1. The current work presents two experiment arms. In the first arm, we test our proposed DNN on data collected by [11]. This data includes fMRI connectivity features from 61 patients suffering from depression treated by rTMS. In this experiment arm, we assess the ability of our model to discriminate rTMS responders from non-responders neuroimaging features.

Our second set of experiments utilises a privately collected dataset from Belmont Private Hospital, Brisbane, Australia. This data includes the records of 133 patients who undertook rTMS treatment. However, in contrast to the first experiment, the features of the second experiment arm include only features collected through a self-reported questionnaire.

Rigorous validation of ML and DL algorithms is essential to ensure the robustness of reported results. The validation of AI systems for healthcare is an important step in the transition to clinical practice [36]. Harrel [37] asserts that the strongest form of internal validation is repeated bootstrap resamples, with analysis of the target variable repeated for each resample. This process ensures that a relationship between input variables and target variables exists, thus increasing the robustness of the results.

These experiments are designed to compare the performance of self reported psychometric measures of depression severity against quantitative fMRI measures, in predicting treatment outcomes to rTMS. Through these experiments, we aim to identify candidate biomarkers that explain the patterns of response to rTMS treatment.

2.8. Baseline models

Our baseline models include a LSVM, as proposed by [21,11,12], and a KNN classifier [21]. Additionally, we include XGBoost and random forests as baseline models, which are widely used in healthcare with explainability [19]. The hyperparameters of all baseline models were optimized using grid search. In our final experiments, we used the best hyperparameter set found during grid search to compare against our proposed DNN. Full baseline model hyperparameters are listed in Appendix A, Table A.11.

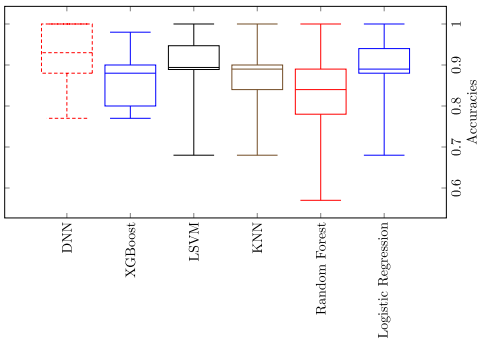

Fig. 1. Box plot showing the accuracies obtained using various algorithms.

For our second dataset, we explore the potential of DNNs to model changes in depression severity from psychometric questionnaires. For this experiment, we utilize the only model that includes the same features. This baseline model is provided by Feffer et al. [6].

As outlined in Section 1.1, Feffer et al. used early symptom improvement to predict treatment response to rTMS. In their study, they proposed that patients with $< 20\%$ reduction in depression severity after 2 weeks of treatment (10 sessions) are unlikely to respond to treatment. As such, inline with the model proposed by Feffer et al., patients with $\leq 20\%$ improvement in symptoms after 10 sessions are classed as non-responders, and the remaining cases are defined as responders. The Feffer et al. [6] model reported high sensitivity but low specificity.

2.9. Performance measuring schemes

Performance metrics are required to evaluate models and make comparisons between them. These metrics differ slightly depending on the nature of the outcome variable. Common metrics used for the evaluation of classification models in psychiatry include F1 score and accuracy, as used in Chang et al. [38]. Additionally, in line with Bailey et al. [12], we have included balanced accuracy to assess performance in both the positive and negative cases.

3. Results

We present the results of our two experiment arms (Experiments 1 and 2) in the two sub-sections below.

3.1. Experiment 1: fMRI connectivity measures to predict rTMS treatment outcomes

Recent work by Hopman et al. [11] proposed a LSVM for the early prediction of treatment response to rTMS. Their works combined fMRI features with a Linear SVM to predict treatment outcomes. Our experiments compare the performance of several baseline models against our proposed DNN architecture over 200 repeated bootstrap samples. The distribution of test set accuracy for each algorithm is shown in Fig. 1.

From this diagram we see while most algorithms have an upper limit of correctly predicting all cases in the test set. The DNN finds this optimal solution more frequently across all trials. Followed by the logistic regression, and the LSVM. With the observed LSVM performance closely mirroring the performance that [11] obtained on the same dataset.

Table 6
Summary of average model performance matrices obtained from 200 bootstrap resamples.

Model	Accuracy	F1 score	Balanced accuracy
DNN	0.9423 (0.0605)	0.9461 (0.0561)	0.9423 (0.0618)
XGBoost	0.7813 (0.0823)	0.7916 (0.0794)	0.7804 (0.0830)
LSVM	0.9107 (0.0588)	0.9127 (0.0584)	0.9115 (0.0587)
KNN	0.8913 (0.06842)	0.8942 (0.06753)	0.8918 (0.0686)
Random Forests	0.8416 (0.0822)	0.8506 (0.0790)	0.8404 (0.0830)
Logistic Regression	0.9047 (0.0636)	0.9071 (0.0647)	0.9052 (0.0635)

Table 7
Summary of average model performance in Experiment 2.

Model type	Feature Set	F1-score	Accuracy	Balanced accuracy
Feffer et al. [6]	Domain Knowledge	0.857	0.815	0.791
DNN	DASS Scores	0.772	0.630	0.500

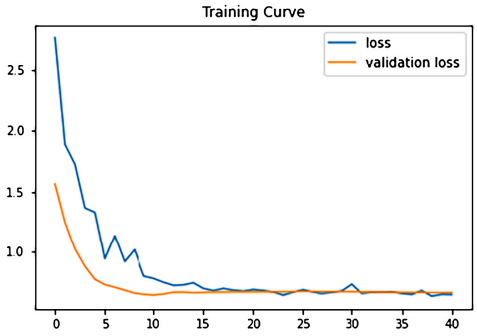

Fig. 2. A graph of training loss against validation loss.

Table 6 shows the summary of results obtained using our model compared against our baseline models. Reported results are the average of 200 bootstrap resamples with standard deviations included in parentheses. Again, our proposed DNN outperformed all baseline models across the reported metrics.

In addition to the described summary metrics obtained using 200 samples, the LSVM identified the optimal solution for correctly classifying the unseen test set 21 times compared to 64 times using the DNN. To demonstrate the robustness of the DNN model we have also included training curves. One way to ensure the robustness of DNN performance is to monitor training loss compared to validation loss. The significant divergence between training and validation loss, when validation loss deteriorates significantly compared to training loss, indicates that the model is overfitting. The training curve from 1 of the 200 trained networks is shown in Fig. 2. The figure shows no significant divergence between losses. During our testing when regularisation was removed, the model was prone to overfitting. This was demonstrated by a significant divergence between validation loss and training loss.

It may be noted from the performance of both the LSVM and DNN that a signal exists between variables and patterns of response. This motivated us to further explore which variables are most significant for correctly predicting treatment response. Based on the results of our experiments, a DNN with the hyperparameters described in Table 5 outperforms the existing baseline models across all metrics.

3.2. Experiment 2: self-reported DASS scores to predict final rTMS treatment outcome

Extending our current work, we investigate the potential for self-reported measures to predict rTMS treatment outcomes. Existing work has shown early changes in symptom severity to be a reasonable predictor of rTMS treatment outcome. Extending upon this work, we explore whether a DNN can identify the relationship between self-reported depression severity and treatment response.

The results shown in Table 7 highlight that when using DASS scores, the preferred method to predict treatment response is domain knowledge as described in [6]. These results highlight that the fMRI connectivity features are superior to self reported DASS scores. Surprisingly, the DNN was unable to pick up on the relationship between early symptom improvement and final treatment outcome.

3.3. Explainable AI (XAI) approaches to identify potential biomarkers indicative of response to treatment

This paper has emphasized the importance of understanding model performance. This position is echoed by Tjoa and Guan [41], who asserted that when DNNs and AI models are applied to non-trivial tasks, improving model understanding is imperative. Methods for assessing feature importance vary from global to local explanations. Global methods explore feature importance from a global scope [42]. In contrast, local methods provide an explanation as to which variables are contributing to the prediction of an individual case within the dataset.

We consider two methods for ranking feature importance: a global and a local method. A global post-hoc method that is commonly used to interpret AI methods is permutation feature importance (PFI) [25]. A PFI score involves the shuffling of one variable within the testing set before the data containing the shuffled feature is input into the trained model [25]. Similar to ablation, the more significant the decline in the model's performance metrics, the greater the relative importance to the model. This process is then repeated throughout all variables in the dataset. One limitation of this approach is that any correlation between a shuffled feature and an unshuffled feature may lead to underestimating the importance of a feature [25]. This issue is similar to the issue of colinearity in simple linear regression.

Utilizing the PFI score, Table 8 shows the relative performance declines in performances associated with each variable. This Table highlights that shuffling of COC leads to the most significant performance decline in model performance. This is measured by the change in test set accuracy by iteratively shuffling each variable. For clarity, the relative performance change attributed to each feature is shown visually in Fig. 3.

Table 8
Summary of feature importance scores.

Feature	Test set accuracy	Percentage drop in accuracy
FP	92.3077	−7.6923
OC	76.9230	−23.0769
SPL	61.5384	−38.4616
COC	53.8462	−46.1538
LOCL	92.3077	−7.6923
LOCR	92.3077	−7.6923

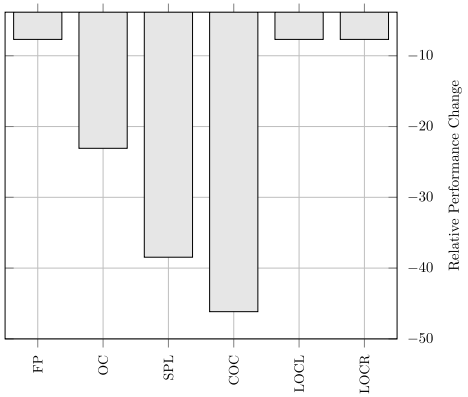

Fig. 3. Relative change in the performance due to various features.

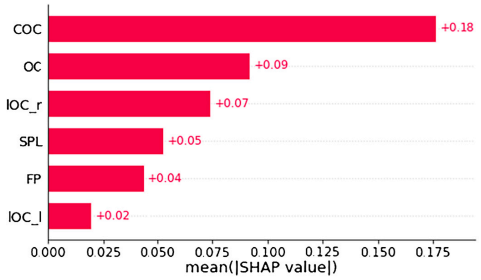

Fig. 4. Average SHAP values on Training Set.

3.3.1. Shapley (SHAP) values

One local method used for assessing feature importance is the SHAP value [43]. Inspired by the seminal work of Shapley [44], Lundberg and Lee [43] introduced the SHAP value. The Shapley value is a game theoretic approach to measure a player's contribution to an end goal in an n-player cooperative game. SHAP values then provide a local explanation for the contribution of each feature to a final output.

Using the SHAP values calculated in Python's SHAP package offers support for computing PFI score results. Fig. 4 shows that COC contributes significantly to model predictions, followed by OC. These findings mirror the results of the PFI score except for SPL, which is ranked much lower by SHAP value when compared to the results in Fig. 3.

One strength of local approaches to XAI is the ability to investigate SHAP values for individual cases. We can use this to instill greater trust from clinicians in our model to support its use.

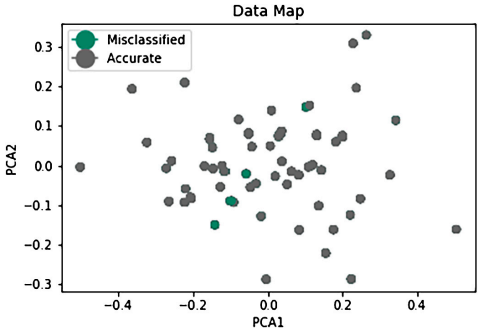

Fig. 5. Plot of Principal Component 2 (PCA2) versus Principal Component 1 (PCA1).

3.4. Model limitations

To the best of our knowledge, we are the first group to use a DNN with fMRI connectivity features for the classification of rTMS response. By using connectivity measures collected before treatment, our network reliably predicts the final treatment outcome. Motivated by Sendak et al. [20], we have included a detailed overview of our model, including potential limitations and the relative contribution of each feature to the model's overall performance. These contributions are an essential step to support the transition from research to clinical use.

Given the high accuracy reported in Section 2.7, it is useful to give some attention to any misclassified examples. By aiming to understand their occurrences, this investigation provides end users - in our case, psychiatrists - with a complete understanding of the model's behaviour.

Amershi et al. [39] present several guidelines for human-AI interaction. These guidelines emphasise the importance of setting clear expectations for the quality and capability of AI systems. Additionally, [39] highlight the importance of making the user aware of situations when an AI system may make mistakes. Formalizing this process, Sendak et al. [20] present model facts, a systematic approach to documenting a ML model designed for clinicians, including advice on interpreting model outputs and warnings. As noted by Sendak et al. [20], warnings regarding the use of an AI model are rarely discussed in the literature. Model limitations must be acknowledged if proposed models are to have an impact in clinical practice.

Exploring the performance of our model, we investigated similarities between commonly misclassified cases. Swayamdipta et al. [40] present a novel methodology for recognizing areas of uncertainty within a large corpus of text. Existing works focus on identifying mislabelled examples within the training data. In contrast, our work focuses on identifying portions of the data that are mislabelled in the test set. Using this novel adaptation, we identified a portion in the lower left-hand quadrant of Fig. 5 that is vulnerable to misclassification. The clustering of these misclassified examples motivated us to explore the hypothesis that these points may share commonalities.

3.4.1. Extreme values are vulnerable to being misclassified

Given the clustering of misclassified examples in Fig. 5, we hypothesise that these values may share some similarities. Identifying these commonalities is an important step in communicating the potential limitations of our model to clinicians. Analysis of the proposed model indicated that after 200 runs, our model accurately predicts treatment response on a held-out test set $\approx 92\%$ of the time. This leaves roughly 8% of cases being misclassified. A variable-wise comparison of distributions by t-test is shown in Table 9. It can be noted from the Table that,

Table 9
Comparison of variable distributions in correctly labelled and mislabelled cases.

Variable	t	p-value
FP	−0.8274	0.4113
OC	1.9119	0.0607
SPL	−0.1517	0.8800
COC	0.5420	0.5899
lOCL	1.2261	0.2250
lOCR	1.2907	0.2018

Table 10
Exploring the differences in OC between examples correctly labelled and mislabelled examples.

		95% Confidence interval	
Group	Mean	Lower bound	Upper bound
Correct	0.0297	−0.0044	0.06372
Misclassified	−0.0313	−0.0749	0.01220

although marginally short of the level of significance, there is some difference between groups in the OC variable.

Through further analysis of group differences as shown in Table 10, we can see the differences between values that were accurately classified and misclassified. Misclassified values had, on average, lower connectivity measures in the OC variable compared to the correctly classified values.

Rerunning an analysis of our model with the removal of the OC variable shows a drop in performance over 200 bootstrap resamples. Highlighting OC is valuable in discerning between classes, however, it does have some observed failure cases. These are important considerations, given each rejected positive case is a patient who may be denied access to treatment when they may actually benefit from it, or conversely, a patient who commits time to receive treatment and sees no benefit. As such, we include the limitation or warning of misclassifications in our model between the 95% confidence interval of −0.0749 to 0.0122. These model limitations can be communicated to end users.

4. Discussions

The current work demonstrates that a feedforward DNN model can accurately predict the treatment outcome of rTMS before treatment. With rigorous internal validation, our work shows a DNN using fMRI connectivity features outperforms existing baseline methods. In our experiments, the performance of prominent ML algorithms like XGBoost and random forests was disappointing. It may be noted that tree-based algorithms like XGBoost have under-performed when the number of samples is less than 500 [14]. Furthermore, the baseline LSVM reproduces the findings of Hopman et al. [11], offering additional support for the use of fMRI features and their ability to predict rTMS treatment outcomes. These findings further emphasize the potential of fMRI connectivity measures as biomarkers for response to rTMS treatment. Furthermore, the current work reiterates that demographic and psychometric variables alone are insufficient to identify patterns of response to rTMS treatment. Even when using sophisticated algorithms, psychometric variables could not improve on the existing rule-based methods proposed by Feffer et al. [6]. Using these psychometric variables, a DL model was unable to identify the association between early change in depression severity and treatment outcome, similar to Feffer et al. [6].

Our work utilizes high levels of internal validation to ensure robust results in an important setting: the psychiatric care of those suffering from depression. Along with this validation, we demonstrated the significant impact of regularisation on model performance to reduce the risks of overfitting. These initial findings will become increasingly significant as larger rTMS datasets become available to further explore the potential of verifying these results against independent datasets. Our results highlight the benefits of using DNNs with several hidden layers compared against shallow ML methods in modelling complex relationships. The proposed architecture outperforms other shallow methods in terms of F1 score, balanced accuracy, and accuracy. This superior predictive performance may be due to the ability of DL algorithms to model complex multi-variable relationships. Shallow methods, such as traditional linear algorithms, are unable to recognize these complicated relationships. In practice, the interplay between treatment, psychiatrists and patient variables is more complex than can be modelled using linear models. The proposed model here consists of 1191 parameters, highlighting the complexity of the model compared to shallow methods.

One thing to note is that the current work has been developed using a limited number of complete records. In the future, we plan to impute the missing records to increase the size of the data. Furthermore, participants for whom data is incomplete may have left the study due to a lack of improvement in their psychological health, leaving only patients who benefited. The risk, then, is that the remaining sample is not truly representative of the true population of patients receiving rTMS treatment. Also, there is a possibility that the model may be overfitting to the current distribution of patients. Further work involving data collected from multiple centres could help to improve the robustness of this model. While the current work is completed using data where classes are relatively balanced, it is not known how the current method would perform if training data was imbalanced. Future work could attempt to incorporate methods that are robust to uneven class distributions of the target variable.

5. Conclusions

In this work, we have proposed a novel DL architecture to predict the outcome of rTMS treatment using fMRI connectivity features. To the best of our knowledge, we are the first to apply both a DNN and combine a DNN with XAI to rTMS response prediction using fMRI connectivity features. Through empirical experiments, we showed that a DNN using fMRI connectivity measures outperforms existing state-of-the-art algorithms. In our repeated bootstrap simulations, we demonstrate our model finds the optimal solution in an unseen test set more frequently than other methods. The demonstrated robustness of this model moves the field closer to clinical implementation over existing shallow methods. Furthermore, our work demonstrates neuroimaging variables are superior to psychometric variables in predicting treatment response to rTMS. Additionally, using XAI techniques, our work shows functional connectivity measures between the Subgenual Anterior Cingulate Cortex and Centeral Opercular Cortex to be a key determinant for rTMS treatment response. These findings are validated using both SHAP values and relative feature importance. The current work improves upon existing methods by including XAI and predicting treatment outcomes before the start of treatment. However, the main limitation of this work is that only a small dataset has been used to develop and test the model. In the future, we plan to use larger datasets from various centres and ethnicities to improve the accuracy of our work.

Statement of ethical approval

Ethical approval for this project was granted by the Universities Human Research Ethics Committee (H21REA026).

Funding

This work is partially funded by The Cannan Institute, Belmont Private Hospital, Brisbane. The authors declare no competing interests.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We gratefully acknowledge the support from Belmont Private Hospital team members, especially, Ms Mary Williams (CEO), Rachel Stark (Area Manager), Dr Mark Spelman (Psychiatrist), Dr Sean Gills (Psychiatrist), and Dr Tom Moore (Psychiatrist). Without their kind support, this work wouldn't be possible.

Appendix A

A.1. Baseline model hyperparameters

Table A.11
Baseline models and their hyperparameters.

Model	Hyperparameter	Value
Linear SVM	C	100
	Gamma	1
	Kernel	Linear
KNN	Distance Metric	Manhattan
	k	4
	Weight	Uniform
XGBoost	colsample_bytree	0.7
	learning_rate	0.01
	max_depth	3
	n_estimators	200
	subsample	1
Random Forest	bootstrap	True
	max_depth	10
	max_features	auto
	min_samples_leaf	1
	min_samples_split	2
	n_estimators	50
Logistic Regression	C	5.4287
	Penalty	L1

References

[1] D. Schofield, M. Cunich, R. Shrestha, R. Tanton, L. Veerman, S. Kelly, M. Passey, Indirect costs of depression and other mental and behavioural disorders for Australia from 2015 to 2030, BJPsych Open 5 (3) (May 2019).

[2] P.B. Fitzgerald, K.E. Hoy, J. Reynolds, A. Singh, R. Gunewardene, C. Slack, S. Ibrahim, Z.J. Daskalakis, A pragmatic randomized controlled trial exploring the relationship between pulse number and response to repetitive transcranial magnetic stimulation treatment in depression, Brain Stimul. 13 (1) (Jan. 2020) 145–152.

[3] C.A. Conelea, N.S. Philip, A.G. Yip, J.L. Barnes, M.J. Niedzwiecki, B.D. Greenberg, A.R. Tyrka, L.L. Carpenter, Transcranial magnetic stimulation for treatment-resistant depression: naturalistic treatment outcomes for younger versus older patients, J. Affect. Disord. 217 (Aug. 2017) 42–47.

[4] C.L. Hovington, A. McGirr, M. Lepage, M.T. Berlim, Repetitive transcranial magnetic stimulation (rTMS) for treating major depression and schizophrenia: a systematic review of recent meta-analyses, Ann. Med. 45 (4) (May 2013) 308–321.

[5] P.B. Fitzgerald, K.E. Hoy, R.J. Anderson, Z.J. Daskalakis, A study of the pattern of response to rTMS treatment in depression, Depress. Anxiety 33 (8) (Apr. 2016) 746–753.

[6] K. Feffer, H.H. Lee, F. Mansouri, P. Giacobbe, F. Vila-Rodriguez, S.H. Kennedy, Z.J. Daskalakis, D.M. Blumberger, J. Downar, Early symptom improvement at 10 sessions as a predictor of rTMS treatment outcome in major depression, Brain Stimul. 11 (1) (Jan. 2018) 181–189.

[7] M. Squires, X. Tao, S. Elangovan, R. Gururajan, X. Zhou, U.R. Acharya, A Novel Genetic Algorithm Based System for the Scheduling of Medical Treatments, Expert Systems with Applications, vol. 195, Elsevier BV, Jun. 2022, p. 116464.

[8] O.A. Montesinos López, A. Montesinos López, J. Crossa, Fundamentals of artificial neural networks and deep learning, in: Multivariate Statistical Machine Learning Methods for Genomic Prediction, Springer International Publishing, 2022, pp. 379–425.

[9] A.T. Drysdale, L. Grosenick, J. Downar, K. Dunlop, F. Mansouri, Y. Meng, R.N. Fetcho, B. Zebley, D.J. Oathes, A. Etkin, A.F. Schatzberg, K. Sudheimer, J. Keller, H.S. Mayberg, F.M. Gunning, G.S. Alexopoulos, M.D. Fox, A. Pascual-Leone, H.U. Voss, B. Casey, M.J. Dubin, C. Liston, Erratum: resting-state connectivity biomarkers define neurophysiological subtypes of depression, Nat. Med. 23 (2) (Feb. 2017) 264.

[10] N. Koutsouleris, T. Wobrock, B. Guse, B. Langguth, M. Landgrebe, P. Eichhammer, E. Frank, J. Cordes, W. Wölwer, F. Musso, G. Winterer, W. Gaebel, G. Hajak, C. Ohmann, P.E. Verde, M. Rietschel, R. Ahmed, W.G. Honer, D. Dwyer, F. Ghaseminejad, P. Dechent, B. Malchow, P.M. Kreuzer, T.B. Poeppl, T. Schneider-Axmann, P. Falkai, A. Hasan, Predicting response to repetitive transcranial magnetic stimulation in patients with schizophrenia using structural magnetic resonance imaging: a multisite machine learning analysis, Schizophr. Bull. 44 (5) (Aug. 2017) 1021–1034.

[11] H. Hopman, S. Chan, W. Chu, H. Lu, C.-Y. Tse, S. Chau, L. Lam, A. Mak, S. Neggers, Personalized prediction of transcranial magnetic stimulation clinical response in patients with treatment-refractory depression using neuroimaging biomarkers and machine learning, J. Affect. Disord. 290 (Jul. 2021) 261–271.

[12] N. Bailey, K. Hoy, N. Rogasch, R. Thomson, S. McQueen, D. Elliot, C. Sullivan, B. Fulcher, Z. Daskalakis, P. Fitzgerald, Differentiating responders and non-responders to rTMS treatment for depression after one week using resting EEG connectivity measures, J. Affect. Disord. 242 (Jan. 2019) 68–79.

[13] J. Fan, I.F. Tso, D.F. Maixner, T. Abagis, L. Hernandez-Garcia, S.F. Taylor, Segregation of salience network predicts treatment response of depression to repetitive transcranial magnetic stimulation, NeuroImage Clin. 22 (2019) 101719.

[14] M. Zou, W-G. Jiang, Q-H. Qin, Y-C. Liu, M-L. Li, Optimized XGBoost model with small dataset for predicting relative density of Ti-6Al-4V parts manufactured by selective laser melting, Materials, MDPI AG 15 (15) (Aug. 2022) 5298, https://doi.org/10.3390/ma15155298.

[15] F. Hasanzadeh, M. Mohebbi, R. Rostami, Prediction of rTMS treatment response in major depressive disorder using machine learning techniques and nonlinear features of EEG signal, J. Affect. Disord. 256 (Sep. 2019) 132–142.

[16] A. Zandvakili, N.S. Philip, S.R. Jones, A.R. Tyrka, B.D. Greenberg, L.L. Carpenter, Use of machine learning in predicting clinical response to transcranial magnetic stimulation in comorbid posttraumatic stress disorder and major depression: a resting state electroencephalography study, J. Affect. Disord. 252 (Jun. 2019) 47–54.

[17] N. Bailey, K. Hoy, N. Rogasch, R. Thomson, S. McQueen, D. Elliot, C. Sullivan, B. Fulcher, Z. Daskalakis, P. Fitzgerald, Responders to rTMS for depression show increased fronto-midline theta and theta connectivity compared to non-responders, Brain Stimul. 11 (1) (Jan. 2018) 190–203.

[18] T.T. Erguzel, S. Ozekes, S. Gultekin, N. Tarhan, G.H. Sayar, A. Bayram, Neural network based response prediction of rTMS in major depressive disorder using QEEG cordance, Psychiatry Investig. 12 (1) (2015) 61.

[19] H.W. Loh, C.P. Ooi, S. Seoni, P.D. Barua, F. Molinari, U.R. Acharya, Application of Explainable Artificial Intelligence for Healthcare: A Systematic Review of the Last Decade (2011–2022), Computer Methods and Programs in Biomedicine, vol. 226, Elsevier BV, Nov. 2022, p. 107161.

[20] M.P. Sendak, M. Gao, N. Brajer, S. Balu, Presenting machine learning model information to clinical end users with model facts labels, npj Digit. Med. 3 (1) (Mar. 2020).

[21] E. Ebrahimzadeh, F. Fayaz, L. Rajabion, M. Seraji, F. Aflaki, A. Hammoud, Z. Taghizadeh, M. Asgarinejad, H. Soltanian-Zadeh, Machine Learning Approaches and Non-linear Processing of Extracted Components in Frontal Region to Predict rTMS Treatment Response in Major Depressive Disorder, Frontiers in Systems Neuroscience, vol. 17, Frontiers Media SA, Mar. 2023.

[22] M.S. Shahabi, A. Shalbaf, R. Rostami, R. Kazemi, A convolutional recurrent neural network with attention for response prediction to repetitive transcranial magnetic stimulation in major depressive disorder, Sci. Rep. 13 (1) (Jun. 2023), https://doi.org/10.1038/s41598-023-35545-2, Springer Science and Business Media LLC.

[23] I.N. Beaufort, G.H. De Weert-Van Oene, V.A.J. Buwalda, J.R.J. De Leeuw, A.E. Goudriaan, The depression, anxiety and stress scale (DASS-21) as a screener for depression in substance use disorder inpatients: a pilot study, Eur. Addict. Res. 23 (5) (2017) 260–268, https://doi.org/10.1159/000485182, S. Karger AG.

[24] F. Chollet, Deep Learning with Python, second edition, Manning Publ., Dec. 2021, [Online]. Available: https://www.ebook.de/de/product/40499536/francois_chollet_deep_learning_with_python_second_edition.html.

[25] Y. han Sheu, Illuminating the black box: interpreting deep neural network models for psychiatric research, Front. Psychiatry 11 (Oct. 2020).

[26] S. Itani, M. Rossignol, At the crossroads between psychiatry and machine learning: insights into paradigms and challenges for clinical applicability, Front. Psychiatry 11 (Sep. 2020).

[27] P.B. Fitzgerald, M.S. George, S. Pridmore, The evidence is in: repetitive transcranial magnetic stimulation is an effective, safe and well-tolerated treatment for patients with major depressive disorder, Aust. N.Z. J. Psychiatry (Aug. 2021) 000486742110430.

[28] P.M. Doraiswamy, C. Blease, K. Bodner, Artificial intelligence and the future of psychiatry: insights from a global physician survey, Artif. Intell. Med. 102 (Jan. 2020) 101753.

[29] H. Hopman, S. Chan, W. Chu, H. Lu, C-Y. Tse, S. Chau, L. Lam, A. Mak, S. Neggers, Personalized prediction of repetitive transcranial magnetic stimulation clinical response in medication-refractory depression data, Data Brief 37 (Aug. 2021) 107264, https://doi.org/10.1016/j.dib.2021.107264, Elsevier BVc.

[30] S. Lovibond, P.F. Lovibond, Manual for the Depression Anxiety Stress Scales, 2nd ed., Psychology Foundation, Sydney, 1995.

[31] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (6) (1958) 386–408.

[32] H. Ramchoun, M. Amine, J. Idrissi, Y. Ghanou, M. Ettaouil, Multilayer perceptron: architecture optimization and training, Int. J. Interact. Multimed. Artif. Intell. 4 (1) (2016) 26.
[33] E. Aldana-Bobadila, A. Kuri-Morales, I. Lopez-Arevalo, A.B. Rios-Alvarado, An unsupervised learning approach for multilayer perceptron networks, Soft Comput. 23 (21) (Nov. 2018) 001.
[34] X. Ying, An overview of overfitting and its solutions, J. Phys. Conf. Ser. 1168 (Feb. 2019) 022022.
[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (Jan. 2014) 1929–1958.
[36] C. Birkenbihl, M.A. Emon, H. Vrooman, S. Westwood, S. Lovestone, M. Hofmann-Apitius, H. F., Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia - lessons for translation into clinical practice, EPMA J. 11 (3) (Jun. 2020) 367–376.
[37] F.E. Harrell Jr, Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, Springer, 2015.
[38] B. Chang, Y. Choi, M. Jeon, J. Lee, K.-M. Han, A. Kim, B.-J. Ham, J. Kang, ARPNet: antidepressant response prediction network for major depressive disorder, Genes 10 (11) (Nov. 2019) 907.
[39] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P.N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, Guidelines for human-AI interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, May 2019.
[40] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N.A. Smith, Y. Choi, Dataset cartography: mapping and diagnosing datasets with training dynamics, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 9275–9293, Online. https://doi.org/10.18653/v1/2020.emnlp-main.746.
[41] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): toward medical XAI, IEEE Trans. Neural Netw. Learn. Syst. 32 (11) (Nov. 2021) 4793–4813.
[42] L. Gianfagna, A.D. Cecco, Explainable AI with Python, Springer International Publishing, 2021.
[43] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4768–4777.
[44] L.S. Shapley, A value for n-person games, in: Contributions to the Theory of Games (AM-28), Volume II, Princeton University Press, Dec. 1953, pp. 307–318.

3.3 Links and implications

Based on the results of the literature review, this paper is the first of its kind to show DL can reliably predict the outcome of rTMS treatment using fMRI connectivity features. The work distinguishes fMRI connectivity features to be a better predictor of rTMS outcome than psychological and demographic features alone. Furthermore, this work shows connectivity between the Subgenual Anterior Cingulate Cortex and the Centeral Opercular Cortex to be the most influential predictor in the DL model.

This work makes two broader observations on the current state of AI in psychiatry. Firstly, this work shows that imaging techniques are superior to easier to collect variables for predicting treatment response. However, this work identifies a subset of patients which the DL model routinely misclassified. Looking through the hype of AI in psychiatry it is important these limitations are acknowledged and investigated such that AI models are not biased to any population group. As such, methods to address the miscllasification of underrepresented values should be explored.

CHAPTER 4: PAPER 3 - DE-CGAN: BOOSTING RTMS TREATMENT PREDICTION WITH DIVERSITY ENHANCING CONDITIONAL GENERATIVE ADVERSARIAL NETWORKS

4.1 Introduction

Previously, Chapter 1.6 identified the limited access to patient data as a significant challenge to deploying AI in psychiatry. DL models perform at their best when trained on large and diverse datasets. However, access to diverse data has so far been a challenge for researchers. This problem is not insignificant. Diversity and fairness are core components of creating trustworthy AI systems. Models trained on homogeneous datasets may in turn perform poorly on more diverse test sets decreasing trust in the system.

Data augmentation then provides a strategy for overcoming these limitations. If we can use AI to generate high quality synthetic examples perhaps we can artificially increase the size of existing datasets. Furthermore, the use of data augmentation methods could be used to increase the diversity of datasets with underrepresented populations. In this chapter, a novel data augmentation technique is presented which proposes the use of DL methods to increase the diversity of rTMS training data.

4.2 Published paper

DE-CGAN: Boosting rTMS Treatment Prediction with Diversity Enhancing Conditional Generative Adversarial Networks

Matthew Squires[1*], Xiaohui Tao[1], Soman Elangovan[2], Raj Gururajan[3], Haoran Xie[4], Xujuan Zhou[3], Yuefeng Li[5], U Rajendra Acharya[1]

[1*]School of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba, Australia.
[2]Belmont Private Hospital, Brisbane, Australia.
[3]School of Business, University of Southern Queensland, Springfield, Australia.
[4]Department of Computing and Decision Sciences, Lingnan University, Hong Kong SAR, China.
[5]School of Computer Science, Queensland University of Technology, Brisbane, Australia.

*Corresponding author(s). E-mail(s): matthew.squires@usq.edu.au;
Contributing authors: xiaohui.tao@usq.edu.au;
soman.elangovan@healthecare.com.au; Raj.Gururajan@usq.edu.au;
hrxie@ln.edu.hk; xujuan.zhou@usq.edu.au; y2.li@qut.edu.au;
rajendra.acharya@usq.edu.au;

Abstract

Repetitive Transcranial Magnetic Stimulation (rTMS) is a well-supported, evidence-based treatment for depression. However, patterns of response to this treatment are inconsistent. Emerging evidence suggests that artificial intelligence can predict rTMS treatment outcomes for most patients using fMRI connectivity features. While these models can reliably predict treatment outcomes for many patients for some underrepresented fMRI connectivity measures DNN models are unable to reliably predict treatment outcomes. As such we propose a novel method, Diversity Enhancing Conditional General Adversarial Network (DE-CGAN) for oversampling these underrepresented examples. DE-CGAN creates

synthetic examples in difficult-to-classify regions by first identifying these data points and then creating conditioned synthetic examples to enhance data diversity. Through empirical experiments we show that a classification model trained using a diversity enhanced training set outperforms traditional data augmentation techniques and existing benchmark results. This work shows that increasing the diversity of a training dataset can improve classification model performance. Furthermore, this work provides evidence for the utility of synthetic patients providing larger more robust datasets for both AI researchers and psychiatrists to explore variable relationships.

Keywords: Data Augmentation, AI Fairness, Mental Health, Data and Bias

1 Introduction

Depression is a highly prevalent and debilitating mental illness [1]. For some, treatment will lead to a reduction in their depression severity. However, approximately one-third of patients will see a minimal reduction in the severity of their depression [2, 3]. For many patients suffering from TRD, rTMS provides some relief. However, other patients will see little improvement in their depression severity. Given the variance in observed response rates, research is exploring the use of artificial intelligence (AI) techniques to predict treatment response to rTMS [4–15]. These works aim to delineate between responders and non-responders to treatment before or early in the treatment cycle. Thus reducing the financial and psychological toll that ineffective treatments place on patients [10, 11].

Personalising psychiatric treatments through the use of AI is a rapidly expanding area of research. Despite this interest, progress towards implementation has been slowed by challenges of model generalisation and access to appropriately sized data. Koppe et al. [16] discuss some of the challenges of implementing deep learning (DL) in psychiatry. DL algorithms require large, diverse datasets to produce their best results, however, these large samples are generally not available in psychiatry [16]. When models are built on large, diverse and representative training sets the better the model will generalise to unobserved data [17]. As such methods to address the lack of quality psychiatric data are required as a method to improve model generalisation.

Enhancing fairness in AI is an important topic of research. Potential challenges to fairness of AI include data bias [18]. When limited examples of a phenomenon exists within the training set, classifiers trained on these datasets tend to generalize poorly [19]. Furthermore, the class imbalance problem is one of the most significant challenges in data mining [20]. Oversampling of the minority class in a dataset is one established strategy to balance datasets with few examples of the minority class [21, 22]. However, these methods fail to address oversampling the feature space with distinct values which could belong to either class. Thus we seek a method which can both enhance the diversity of training dataset and create synthetic examples conditioned on class label. Conditional Generative Adversarial Networks (CGANs) offer an alternative to traditional generative networks by conditioning synthetic examples on additional

information, such as class label. This paper proposes a novel methodology, Diversity Enhancing Conditional General Adversarial Network (DE-CGAN) for boosting the diversity of underrepresented samples conditioned on class labels. This proposed method seeks to boost the diversity of sparse regions of the feature space with synthetic examples belonging to either class, distinct from the problem of oversampling the minority class.

Recently, we showed a Deep Neural Network (DNN) could reliably predict rTMS treatment outcomes using fMRI connectivity features collected before treatment [4]. However, this research also identified certain connectivity values, particularly values of functional connectivity between the Subgenual Anterior Cingulate Cortex and Occipital Cortex which a DNN consistently mislabeled. In an effort to address these misclassifications, we explore the use of generative networks to oversample examples of these difficult-to-classify patients and increase the diversity of the available data.

Synthetic data is artificially created data [23]. Broadly, two classes of methods exist for the generation of synthetic data: process-driven and data-driven [24]. Process-driven methods "derive synthetic data from computational or mathematical models of an underlying physical process" [24, p.2]. However, process-driven methods cannot be implemented when a system cannot be modeled. As such, data-driven methods have surged in popularity in both research and public discourse. These data-driven methods use an underlying data distribution to create synthetic examples, which ideally, come from the same distribution [23, 25]. The benefit of such methods is that generated data is assumed to contain the same characteristics as the original data but protects the privacy of the original data. This protection is especially important for data that contains sensitive information, such as medical data. However, as Jain et al. [26] noted, that generated examples come from the same probability distribution is the best case. More likely, however, that generated examples are less diverse than the original data. As such, we explore methods for evaluating the quality of synthetic (X, Y) pairs to increase the diversity of rTMS datasets to improve classification performance. The current work aims to evaluate the use of synthetic data points to enhance the diversity of rTMS training data. We do this exploring the following research question.

1. How does the use of synthetic rTMS patients impact classification model performance on a real test set?
2. How does oversampling with synthetic patients of underrepresented fMRI connectivity features alter model performance on a held-out test set of real patients?

In answering these questions the current paper makes the following contributions:

- A novel framework we are calling Diversity Enhancing Conditional General Adversarial Network (DE-CGAN) for oversampling underrepresented fMRI connectivity features with synthetic values and their class labels.
- A diversity-enhanced rTMS dataset for the study of rTMS patterns of response to treatment.
- Empirical experiments showing a diversity-enhanced training set improves model performance on a held-out test set of real examples.

This paper is structured as follows. Section 2 provides a review of the current state-of-the-art techniques, for generating synthetic data. Including those used specifically for generating medical data. Section 3 the details of the constructed generative network are outlined. Section 4 provides the experiment design and evaluation metrics with experiment results presented in Section 5. Finally, Section 6 provides a commentary on experiment results.

2 Related Work

The class imbalance problem is one of the most significant challenges in data mining [20]. Class imbalance describes a situation where a target class is underrepresented in the training data compared to other classes [27, 28]. When training sets lack diversity across classes performance can decline on the underrepresented examples [22]. To address the class imbalance issue a variety of sampling and data augmentation techniques have been proposed [29].

Methods of data augmentation vary from traditional oversampling methods to more recent advances such as the use of AI to create synthetic examples. Synthetic Minority Oversampling TEchnique [30, SMOTE] is one of the first oversampling methods to address the issue of class imbalance. SMOTE proposes using synthetic examples by oversampling the minority class. Synthetic examples are created by taking the distance between minority samples and creating a synthetic example in this region. However, some limitations of SMOTE have been identified, particularly when oversampling in regions surrounded by majority class samples which can introduce unnecessary noise [31, 32].

Alternatively, generative networks and more recently, deep generative networks Wang et al. [33, see], have been used across a variety of sectors to generate synthetic examples. For example, in finance [34], health [35], and remote sensing [36]. In healthcare, generative networks have largely been used to synthetically augment neuroimaging [35] and electronic health records [37]. For a detailed overview of GANs and their applications see [38].

While SMOTE works to create linear combinations of existing minority samples modern AI techniques generate synthetic examples while capturing relationships between variables. Recently, [39] proposed a correlation-capturing Generative Adversarial Network (CorGAN) to synthesize electronic health records. Their method combines the use of a 1-dimensional Convolutional GAN with a convolutional autoencoder to discretize continuous values and produce the desired output. Convolutional autoencoders have been proposed as an alternative to the standard autoencoder [40]. In their work, Torfi and Fox [39] showed CorGan to outperform baseline models in creation of medical records and synthetic EEG data. Utilising these same generative techniques researchers have shifted from computer vision to the generation of synthetic medical information. For example, Choi et al. [41] showed generative adversarial networks (GANs) were able to create realistic electronic health records. The architecture, named medGAN, produced predominantly realistic-looking electronic health records as judged by a doctor of medicine. More recently, Torfi et al. [42] proposed convolution GANs for the generation of synthetic medical information. Their work utilised this

architecture to construct synthetic medical data. Goncalves et al. [24] explored several data-driven techniques for the generation of synthetic electronic health records. Their work evaluated the extent to which expected variable relationships were maintained using Surveillance, Epidemiology and End Results (SEER) a widely known cancer dataset. In contrast to Choi et al. [41] and Torfi et al. [42], Goncalves et al. [24] question the efficacy of GANs for creating realistic synthetic data. When comparing several techniques "the generative adversarial network-based model MC-MedGAN failed to generate data with similar statistical characteristics to the real dataset" [24, p.30]. These promising but contrasting findings motivate the exploration of GANs to model synthetic psychiatric data.

Emerging evidence suggests data augmentation methods, such as generative networks, could be used to enhance the diversity of health datasets. For example, Behal et al. [43] proposed Minority Class Rebalancing through Augmentation by Generative modeling (MCRAGE). The work proposes a conditional denoising diffusion probabilistic model to generate synthetic examples. MCRAGE aims to balance datasets of electronic health records based on demographic features such as race, gender and age. Their work shows a classifier trained on a synthetic dataset created by a diffusion improves classifier performance when compared to the original dataset.

DL is positioned to significantly disrupt healthcare. However, progress towards the implementation of AI has been slowed by the limited availability of diverse datasets. One possible strategy to overcome these barriers is the use of synthetic data [25]. Synthetic data may have the potential to create more diverse datasets [44]. To date the use of GANs in healthcare has focused largely on the generation of synthetic images [35]. This motivates our work to explore the use of generative techniques to augment psychiatric data. Existing solutions address imbalances in the class label y. However, these methods fail to target under represented regions in feature spaces $[x_1, x_2, x_3, \cdots, x_n]$. Behal et al. [43] provides one example of balancing imbalanced data based using feature space variables. This motivates our work to balance feature space variables in psychiatry and evaluate the impact this has on a classifier trained on a synthetic dataset

3 Methods

3.1 Problem Statement

Existing research has shown fMRI connectivity features can reliably predict the outcome of rTMS treatment using a DNN [4]. As part of this existing work we identified repeatedly mislabelled examples share common characteristics. To address this issue, the current work proposes using synthetically generated patients to enhance the diversity of a training set in these difficult to classify regions. Previous work has explored algorithmic methods to oversample minority classes in imbalanced datasets. In contrast, this work seeks to generate under represented examples and their class labels to improve model performance. The current work presents DE-CGAN a novel method for enhancing the diversity of rTMS training sets. DE-CGAN seeks to balance datasets by oversampling difficult to classify regions of connectivity between the Subgenual Anterior Cingulate Cortex and Occipital Cortex.

3.2 Problem Definition

The current work aims to show our proposed framework DE-CGAN can boost the diversity of our rTMS patient dataset with synthetic patients. Through enhancing the diversity of the training set we seek to improve the performance of a classification model which predicts rTMS patient outcomes using fMRI functional connectivity measures.

To test this empirically we evaluate the impact of augmenting our training data with varying proportions of synthetic patients. We use this hybrid dataset to train a new classifier and test its performance, and hence the quality of our synthetically generated examples on a test set of real examples.

For our experiments, we define real patients and their associated fMRI connectivity features as follows:

$$D_{real} = \{(x_i, y_i)\}_{i=1}^{N}$$

and

$$x_i = [FP, OC, SPL, COC, lOCL, lOCR]$$

DE-CGAN generates synthetic patients and their class labels:

$$D_{synth} = \{(x'_j, y'_j)\}_{j=1}^{M}$$

and

$$x'_j = [FP', OC', SPL', COC', lOCL', lOCR']$$

Combing the datasets let the resultant dataset be defined as:

$$D_{hybrid} = D_{real} \cup D_{synth}$$

Where α defines the proportion of synthetic patients within the training data set, such that:

$$D_{hybrid}(\alpha) = \frac{D_{synth}}{D_{real}}$$

3.3 Conceptual Model

This section describes the conceptual model of the framework deployed to address the problem described above. Based on the results of our literature review this is the first model deployed to generate synthetic rTMS patients with the aim of enhancing the diversity of our dataset. Through this oversampling we aim to improve predictive model performance to improve the generalisability of our model to unseen data.

Figure 1 describes the conceptual framework that underpins our work. Previously. in Squires et al. [4] we showed that a DNN trained on fMRI connectivity features could reliable predict rTMS treatment outcomes before treatment begins. As part of this work we identified a portion of the dataset which was regularly mislabelled. Using these results, we use these mislabeled examples to train a conditional general adversarial network (CGAN) to oversample the mislabelled examples with class labels. These works are distinct to existing works which oversample the minority class. By

Table 1 Symbol Descriptions

Symbol	Description
D_{real}	Original Dataset from [7]
D_{synth}	Synthetic examples created by DE-CGAN
D_{hybrid}	Training dataset denoted by $D_{real} \cup D_{synth}$
(x, y)	An original patient's features with class label
(x', y')	Synthetic patient and class label
α	DE-CGAN hyperparamater, sets the proportionality ofD_{synth} to D_{real}
FP	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Frontal Pole
OC	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Occipital Cortex
SPL	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Superior Parietal Lobule
COC	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Centeral Opercular Cortex
lOCL	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Left Lateral Occipital Cortex
lOCR	Functional connectivity measure between Subgenual Anterior Cingulate Cortex and Right Lateral Occipital Cortex

augmenting and extending the dataset with synthetic patients each with their own class label.

To evaluate our models performance we leverage our previously described results to evaluate the impacts of including synthetic examples in model training data. The details of which are described in Section 4.3.

3.4 Data Augmentation

Data Augmentation is a class of regularization techniques aimed at improving model performance by making modifications to the training data [17]. One data augmentation technique involves the use of generative models to generate synthetic examples within from the same distribution as the training data [16]. Mumuni and Mumuni [17] refers to these as 'data synthesis methods' methods which create new synthetic training examples generally through methods such as GANs or VAEs.

To date data augmentation research has largely focused on image [45] and signal-based tasks [46]. In the medical space these data augmentation techniques have largely been applied to medical imaging [47, 48]. For example, Frid-Adar et al. [49] showed that synthetically augmenting a dataset of computed tomography (CT) images improves

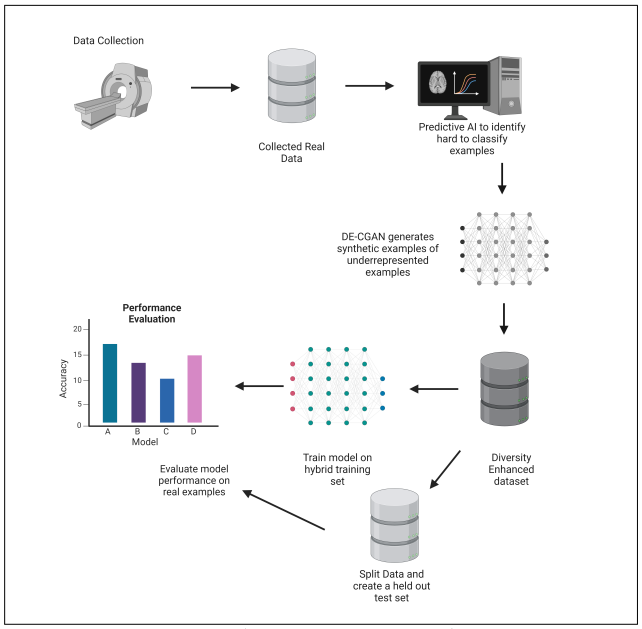

Fig. 1 DE-CGAN conceptual model (Created with BioRender.com)

model performance. While to date, the majority of research has focused on health-related data augmentation, however, to our knowledge, little work has explored data augmentation in psychiatry.

3.4.1 GAN

The classical GAN was first proposed by Goodfellow et al. [50], which harnesses game theory to generate synthetic examples based on the training data set. The game takes place between two networks, the generator, and the discriminator. In classical architecture, the role of the generator is to create realistic examples derived from the training data set. The goal of the generator is to deceive the discriminator into mislabeling fake examples as real. In contrast, the goal of the discriminator is to accurately discern fake examples. Through training, the game reaches a state of Nash equilibrium where neither the discriminator nor the generator is able to further advance their position. Nash [51] in the seminal work showed all 2-person zero-sum games have such an equilibrium point. In theory, the generated examples should mirror the probability distribution of the training data on which the network was trained. However, many

in the research community question this assertion [26, 52, 53]. The first generative network was described by Goodfellow et al. [50]. The theoretical proof provided by Goodfellow et al. [50] assumes a non-parameterized network of infinite capacity. However, in practice network size cannot be infinite. This bounding of network size has led some [26, 52, 53] to question whether the generated data is representative of the data on which the model was trained. Arora et al. [52] asserts even for popular GAN variants mode collapse remains a significant issue. The manifestation of the reduction in the diversity of generated examples has significant implications. Recently, Jain et al. [26] demonstrated the inherent distributional decay and biases of several popular GAN variants. In novel works, Jain et al. [26] used generative networks to create images of engineering staff members. Generated images were then assessed by human raters along with race and gender. Jain et al. [26] reports synthetic images of engineering professors were more likely to have lighter skin tones and masculine features when compared to the original distribution of actual images. These works add tangible examples of differences between synthetic data and the original data distribution.

3.4.2 CGAN

Extending the original GAN the CGAN was first proposed by [54]. The CGAN involves training the original GAN on some additional information, such as a class label [54]. The strength of the CGAN over the original GAN is by conditioning on class labels outputs will also have class labels. For example, Sun et al. [55] show a novel cGAN architecture to generate labeled facial expression images. Mert [56] showed the use of synthetic datasets developed using CGANs improved classifier performance on several medical datasets. Compared to other data augmentation frameworks the CGAN is prefered for this problem as it allows for the conditioning synthetic examples on class labels. Thus capturing important relationships between feature variables and class label.

3.5 DE-CGAN Model Architecture and Hyperparamaters

This Section provides a detailed overview of our proposed DE-CGANs model architecture. Algorithm 1 describes the workflow for our DE-CGAN framework. As part of this method we identify frequently mislabelled examples when evaluating a DNN trained and evaluated on real examples. We then use these frequently mislabelled examples as training data for a CGAN. This produces synthetic examples in difficult to classify regions of the training dataset. The output of our DE-CGAN is a diversity enhanced training dataset for evaluation.

The hyperparameters of both the generator and discriminator are shown in Table 2 and Table 3 respectively. The architecture of these networks is motivated by the architecture in [29] who adopted similar hyperparameters in designing a CGAN to generate synthtetic breast cancer data

Algorithm 1 Diversity Enhancing CGAN Workflow

Input: D_{real}
Output: D_{hybrid}
Parameters: α, DNN, CGAN,
Train Deep Neural Network with D_{real}
for each input x in D_{real} **do**
    Predict label y using the Deep Neural Network
end for
Pass the labeled data to **Mislabelled**
function MISLABELLED(x, y)
    Identify Mislabelled cases
    Initialize Conditional GAN with Mislabelled examples
end function
Conditional GAN Details:
  Generator G**:** Receives a noise vector z and label y, generates synthetic data samples $G(z, y) = x^*$
  Discriminator D**:** Receives both real data samples $(x|y, y)$ and synthetic data samples $(x^*|y, y)$, outputs a probability (via sigmoid) indicating the likelihood of the sample being real
Train Conditional GAN (Generator G and Discriminator D) iteratively
for a specified number of iterations or until convergence **do**
    Train Discriminator D to distinguish between real and generated data
    Train Generator G to create synthetic data that fools Discriminator D
end for
Generate α synthetic cases using CGAN Generator D_{synth}
Split D_{real} into a training set D_{train} and a test set D_{test}
Transform $D_{train} \cup D_{synth} = D_{hybrid}$

Table 2 Generator Hyperparameters

Hidden Layers	2
Layer Width	128, 64
Activation Function	Leaky reLU
Latent Dimensions	100

4 Experiment Overview

4.1 Experiment Design

This section outlines the experiments used to evaluate the quality of datasets augmented by our DE-CGAN. The current work aims to show that augmenting training data with synthetically generated underrepresented examples improves the performance of a binary classification model. This work seeks to show that augmented training data is superior to the original training data.

Table 3 Discriminator Hyperparameters

Hidden Layers	2
Layer Width	64, 32
Activation Function	Leaky reLU
Loss Function	Binary Cross-entopy
Optimizer	Adam
Learning Rate	0.0002

As part of our experiments, we evaluate the performance of varying proportions of augmented data on classifier performance. These experiments compare the performance of a classifier trained using data augmented by a traditional GAN, our DE-CGAN and the original dataset. We test these training datasets empirically using 200 runs of Monte Carlo simulation. During each simulation, a DNN is trained using an augmented dataset with a validation set of 20% used during each training iteration. Early stopping is used to terminate the training process once model performance stabilises. The performance of each model is tested on a held-out test set of real examples. Further details of this methodology are described in Section 4.3.

4.2 Dataset

The original data used in this work is publicly available in Hopman et al. [7, 57]. The data includes several fMRI features, along with a patient's rTMS treatment outcomes. Further detailed summary statistics of this dataset, including associated ethics approval, can be found in [57]. The input variables used from this data are detailed in 1.

4.3 Evaluation metrics

This section further outlines the methods and metrics used to evaluate the quality of augmented training sets:

- **Classification** Evaluating generative networks is an open problem [58]. Esteban et al. [59] proposed a novel framework for evaluating the quality of synthetic examples. Train Synthetic Test Real (TSTR) involves training a classifier on synthetic examples and evaluating its performance on real examples. If the distribution of the synthetic data matches the real data then we would expect the performance of a classifier trained using synthetic examples to perform similarly to that of a classifier trained on the real dataset. This method has several benefits over training on real sets and evaluating on synthetic data. Given the limitations of many generative methods is mode collapse, that is, when synthetic examples become less diverse than the original training data. Hence, the performance of a model on trained on real data and tested on synthetic data may overestimate the quality of these synthetic examples.
- **Hypothesis Testing** We propose the use of two-tail proportion test to evaluate any differences between the proportion of optimal solutions found using training sets augmented with varied proportions of synthetic data.

Table 4 Model Hyperparameters

Hidden Layers	4
Layer Width	10
Activation Function	reLU
Loss Function	Binary Crossentopy
Regularisation Layers	4
Test set size	20%
Epochs	2000 or until early stopping criteria met

4.3.1 Classification Model and Evaluation

Using the TSTR framework evaluation metrics are required to test the classifier performance on the held-out test set. Given the aim of this work is to evaluate the impact of augmenting training data with synthetic examples it is important the classifier remains constant across examples. The model hyperparameters are shown in Tab 4. In each simulation the network is retrained after shuffling the augmented data set.

To evaluate performance we use commonly used deep learning metrics: accuracy, described in Equation 1, balanced accuracy, described in Equation 2 and f1-score in Equation 3. Both balanced accuracy and f1 score are selected as they consider classification performance across classes.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{N}} \tag{1}$$

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right) \tag{2}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

4.3.2 Hypothesis Testing

Further to the classification metrics described above we use hypothesis testing to compare the amount of optimal solutions found using varying proportions of augmented data. The proportion test, described in Equation 4, can evaluate the extent to which differences in proportions between the amount optimal solutions are found

$$z = \frac{(P_{gan} - P_{dist})^2}{SD(P_{gan})} \tag{4}$$

Where

$$SD(P_{gan}) = \sqrt{\frac{P_{gan} \cdot Q_{gan}}{n}}$$

5 Results

This section describes the results of our empirical experiments to evaluate the effectiveness of our proposed model DE-CGAN, and the synthetic patients generated by our model.

As described above the performance of our proposed method was evaluated using 200 monte carlo simulation. The average performance and standard deviation of these models is described in Table 5. These results show a hybrid dataset of 10% synthetic patients from DE-CGAN to be the best performing on a held out test set. Followed by a hybrid dataset of 5% synthetic patients. These models including synthetic examples created by DE-CGAN Followed by our benchmark model presented in Squires et al. [4] with the original dataset described in [7] and no synthetic examples.

Table 5 Mean model performance metrics and standard deviations

Model	Accuracy	F1-Score	Balanced Accuracy
DE-CGAN ($\alpha = 0.05$)	0.9280 (0.0664)	0.9310 (0.6666)	0.9286 (0.0661)
DE-CGAN ($\alpha = 0.10$)	**0.9360 (0.0662)**	**0.9414 (0.0589)**	**0.9355 (0.0684)**
DE-CGAN ($\alpha = 0.15$)	0.9073 (0.0817)	0.9085 (0.0908)	0.9085 (0.0802)
DE-CGAN ($\alpha = 0.20$)	0.9112 (0.0787)	0.9124 (0.0806)	0.9130 (0.0775)
Squires et al. [4]	0.9227 (0.0696)	0.9276 (0.0647)	0.9224 (0.0708)
CGAN ($\alpha = 0.05$)	0.9246 (0.0617)	0.9280 (0.0610)	0.9248 (0.0613)
CGAN ($\alpha = 0.10$)	0.9134 (0.0767)	0.9189 (0.0735)	0.9128 (0.0773)
CGAN ($\alpha = 0.15$)	0.9023 (0.0858)	0.9100 (0.0813)	0.9007 (0.0874)

Visually the distribution of accuracies is displayed in Figure 2. This figure visually shows changes in accuracies across models through the empirical experiments with outliers. Visually we see similarities between the DE-CGAN ($\alpha = 0.05$) and DE-CGAN ($\alpha = 0.10$) distributions with our benchmark model Squires et al. [4] and the baseline model CGAN ($\alpha = 0.05$).

In further analysis of our results, we report the frequency of optimal solutions obtained by each algorithm. These results are shown in Figure 3. From these results, we see both DE-CGAN ($\alpha = 0.05$) and DE-CGAN ($\alpha = 0.10$) are the only models to obtain the optimal solution more frequently than the benchmark model in Squires et al. [4].

To formally compare the proportion of optimal solutions found by each model we use the proportion test to compare the frequencies against our benchmark model. From Table 6 we see the only model which varies significantly is DE-CGAN ($\alpha = 0.10$) which shows using our DE-CGAN to create a training set made up of 10% synthetic cases produces performance above that from using the original dataset without synthetic data.

6 Discussion

The current work shows the generation of synthetic psychiatric patients and their class labels improves the performance of a DNN classification model. These findings emphasize the importance of diverse datasets in deep learning and psychiatric research. This work introduces our novel framework DE-CGAN a novel method for the oversampling

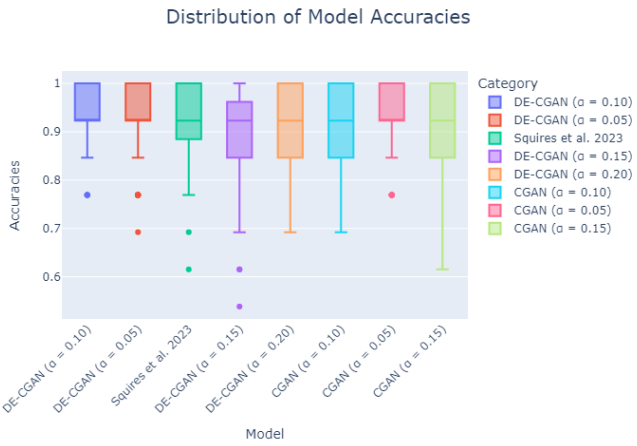

Fig. 2 Box plot showing the distribution of accuracies obtained using various algorithms.

Table 6 Proportion Test Results

Model Name	Z	P	Significant
DE-CGAN ($\alpha = 0.05$)	-0.5327	0.59612	$p > 0.05$
DE-CGAN ($\alpha = 0.10$)	-2.1779	0.02926	$p < 0.05^*$
DE-CGAN ($\alpha = 0.15$)	1.4438	0.1499	$p > 0.05$
DE-CGAN ($\alpha = 0.20$)	0.4344	0.6672	$p > 0.05$
CGAN ($\alpha = 0.05$)	0.5443	0.5892	$p > 0.05$
CGAN ($\alpha = 0.10$)	0	1	$p > 0.05$
CGAN ($\alpha = 0.15$)	1.4438	0.1499	$p > 0.05$

of difficult to classify fMRI connectivity features and synthetic class labels. These findings have significant implications in the field of psychiatry a field where the use of AI has been impacted by small sample sizes [16].

Previous work has introduced methods for oversampling minority classes in classification tasks with imbalanced class labels. In the field of data mining this is referred to as the class imbalance problem [20]. The class imbalance problem occurs in classification tasks where examples for some classes are underrepresented when compared to the majority classes. In contrast, our problem deals with the under-representation of fMRI connectivity measures at certain bandwidths. Our work shows that identifying examples prone to misclassification and generating synthetic examples paired with a synthetic class label in the distribution of difficult to classify regions can lead to improvements in model performance.

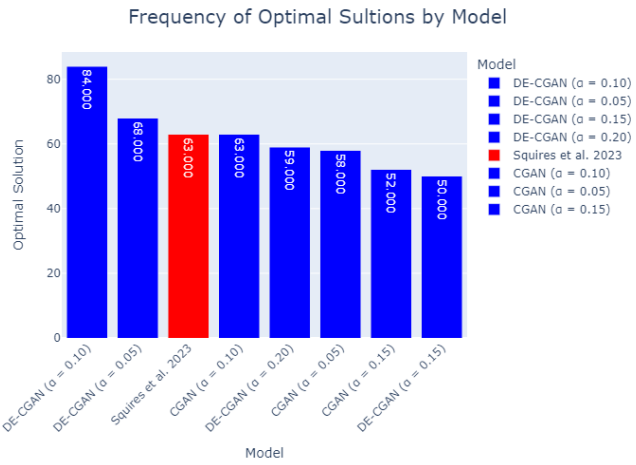

Fig. 3 Bar chart showing the frequency of optimal solutions by algorithm.

Artificial intelligence systems are being deployed rapidly in areas of non-trivial importance. Increasingly, researchers are exploring ways to deploy artificial intelligence systems to personalise medical care. Deep learning algorithms rely heavily on the data on which they are trained. However, problematically, the training data on which these algorithms rely is often not representative with insufficient examples of certain features. Tasci et al. [60] assert under-representation of demographic, biological and outcome variables in data all contribute to algorithmic bias.

Model performance is generally dependent on the data on which it is trained. As the artificial intelligence field pushes towards deployment of predictive models high-quality, representative data is essential Chen et al. [25]. The importance of high quality data is twofold. Data that is truly representative of the general population is likely to improve the generalisation of models. An observed limitation of deep learning models. Additionally, data bias reduces the generalisability of artificial intelligence to unseen populations. Each of these issues is central to the fairness and equity of data science and its implementation in the field.

The current work shows a training dataset of 10% DE-CGAN examples achieves the optimal solution on a holdout test data set of real cases significantly more frequently than a model trained on real examples only. These works provide evidence for the generation of synthetic fMRI connectivity features and their class labels to over-sample datasets with difficult-to-classify patients due to under representation.

Furthermore, these findings emphasize that synthetically generated patients have the potential to supplement small datasets to improve model performance in small

sample sizes. Additionally, these synthetic examples allow psychiatrists further data points to investigate relationships between variables and their impact on mental health.

While this work highlights the potential of synthetic records to increase the diversity of small datasets, this work also demonstrates as shows as the proportion of synthetic examples in the training data increases classification performance deteriorates. We propose two potential reasons for this performance decline: Firstly as the aim of DE-CGAN is to increase the proportion of underrepresented cases, as the proportion of these underrepresented cases makes up a greater proportion of the dataset the representation of the data population skews outside the true population such that the classification model is unable to classify the majority of examples. Secondly, it is possible that due to identified issues with GANs such as mode collapse, as the proportion of synthetic examples increases the synthetic data becomes less diverse.

The general belief of generative networks is the synthetic data created comes from the same probability distribution as the training data from which they are created. However, as Jain et al. [26] noted, that generated examples come from the same probability distribution is the best case. More likely, however, is generated examples are less diverse than the original data. As such, future work should explore explore further methods for generating diverse datasets for the purpose of sharing sufficiently diverse data between research organisations.

7 Conclusion

This paper presents Diversity Enhancing Conditional Generative Adversarial Network, DE-CGAN, a novel framework for oversampling of underrepresented fMRI connectivity features and their class labels. Deep learning models require large and diverse datasets to perform optimally. When diverse datasets are not available model performance can start to deteriorate and is unlikely to generalise well to unseen data. In psychiatry, large datasets are difficult to source due to privacy and legal obligations. DE-GAN provides an option for balancing and extending datasets in psychiatry.

The current work demonstrates increasing the diversity of a training dataset of fMRI connectivity features with synthetic examples improves performance on a held-out test set of real examples. This work provides evidence for the viability of synthetically generated patients to increase the size and diversity of datasets which provides psychiatrists with more data to explore relationships between connectivity features and treatment outcomes.

Future work should explore the potential of larger synthetic datasets for the study of rTMS response prediction. These synthetic datasets should maintain the characteristics of the original data to allow for sharing between research groups where legal obligations may prevent the sharing of actual data.

References

[1] Schofield, D., Cunich, M., Shrestha, R., Tanton, R., Veerman, L., Kelly, S., Passey, M.: Indirect costs of depression and other mental and behavioural disorders for

australia from 2015 to 2030. BJPsych Open **5**(3) (2019) https://doi.org/10.1192/bjo.2019.26

[2] Sforzini, L., Worrell, C., Kose, M., Anderson, I.M., Aouizerate, B., Arolt, V., Bauer, M., Baune, B.T., Blier, P., Cleare, A.J., Cowen, P.J., Dinan, T.G., Fagiolini, A., Ferrier, I.N., Hegerl, U., Krystal, A.D., Leboyer, M., McAllister-Williams, R.H., McIntyre, R.S., Meyer-Lindenberg, A., Miller, A.H., Nemeroff, C.B., Normann, C., Nutt, D., Pallanti, S., Pani, L., Penninx, B.W.J.H., Schatzberg, A.F., Shelton, R.C., Yatham, L.N., Young, A.H., Zahn, R., Aislaitner, G., Butlen-Ducuing, F., Fletcher, C., Haberkamp, M., Laughren, T., Mäntylä, F.-L., Schruers, K., Thomson, A., Arteaga-Henríquez, G., Benedetti, F., Cash-Gibson, L., Chae, W.R., Smedt, H.D., Gold, S.M., Hoogendijk, W.J.G., Mondragón, V.J., Maron, E., Martynowicz, J., Melloni, E., Otte, C., Perez-Fuentes, G., Poletti, S., Schmidt, M.E., Ketterij, E., Woo, K., Flossbach, Y., Ramos-Quiroga, J.A., Savitz, A.J., Pariante, C.M.: A delphi-method-based consensus guideline for definition of treatment-resistant depression for clinical trials. Molecular Psychiatry **27**(3), 1286–1299 (2021) https://doi.org/10.1038/s41380-021-01381-x

[3] Ionescu, D.F., Rosenbaum, J.F., Alpert, J.E.: Pharmacological approaches to the challenge of treatment-resistant depression. Dialogues in Clinical Neuroscience **17**(2), 111–126 (2015) https://doi.org/10.31887/dcns.2015.17.2/dionescu

[4] Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X., Li, Y., Acharya, U.R.: Identifying predictive biomarkers for repetitive transcranial magnetic stimulation response in depression patients with explainability. Computer Methods and Programs in Biomedicine **242**, 107771 (2023) https://doi.org/10.1016/j.cmpb.2023.107771

[5] Adamson, M., Hadipour, A.L., Uyulan, C., Erguzel, T., Cerezci, O., Kazemi, R., Phillips, A., Seenivasan, S., Shah, S., Tarhan, N.: Sex differences in rTMS treatment response: A deep learning-based EEG investigation. Brain and Behavior (2022) https://doi.org/10.1002/brb3.2696

[6] Chen, D., Lei, X., Du, L., Long, Z.: Use of machine learning in predicting the efficacy of repetitive transcranial magnetic stimulation on treating depression based on functional and structural thalamo-prefrontal connectivity: A pilot study. Journal of Psychiatric Research **148**, 88–94 (2022) https://doi.org/10.1016/j.jpsychires.2022.01.064

[7] Hopman, H.J., Chan, S.M.S., Chu, W.C.W., Lu, H., Tse, C.-Y., Chau, S.W.H., Lam, L.C.W., Mak, A.D.P., Neggers, S.F.W.: Personalized prediction of transcranial magnetic stimulation clinical response in patients with treatment-refractory depression using neuroimaging biomarkers and machine learning. Journal of Affective Disorders **290**, 261–271 (2021) https://doi.org/10.1016/j.jad.2021.04.081

[8] Bailey, N., Hoy, K., Rogasch, N., Thomson, R., McQueen, S., Elliot, D., Sullivan, C., Fulcher, B., Daskalakis, Z., Fitzgerald, P.: Differentiating responders and non-responders to rTMS treatment for depression after one week using resting EEG connectivity measures. Journal of Affective Disorders **242**, 68–79 (2019) https://doi.org/10.1016/j.jad.2018.08.058

[9] Fan, J., Tso, I.F., Maixner, D.F., Abagis, T., Hernandez-Garcia, L., Taylor, S.F.: Segregation of salience network predicts treatment response of depression to repetitive transcranial magnetic stimulation. NeuroImage: Clinical **22**, 101719 (2019) https://doi.org/10.1016/j.nicl.2019.101719

[10] Hasanzadeh, F., Mohebbi, M., Rostami, R.: Prediction of rTMS treatment response in major depressive disorder using machine learning techniques and nonlinear features of EEG signal. Journal of Affective Disorders **256**, 132–142 (2019) https://doi.org/10.1016/j.jad.2019.05.070

[11] Zandvakili, A., Philip, N.S., Jones, S.R., Tyrka, A.R., Greenberg, B.D., Carpenter, L.L.: Use of machine learning in predicting clinical response to transcranial magnetic stimulation in comorbid posttraumatic stress disorder and major depression: A resting state electroencephalography study. Journal of Affective Disorders **252**, 47–54 (2019) https://doi.org/10.1016/j.jad.2019.03.077

[12] Bailey, N.W., Hoy, K.E., Rogasch, N.C., Thomson, R.H., McQueen, S., Elliot, D., Sullivan, C.M., Fulcher, B.D., Daskalakis, Z.J., Fitzgerald, P.B.: Responders to rTMS for depression show increased fronto-midline theta and theta connectivity compared to non-responders. Brain Stimulation **11**(1), 190–203 (2018) https://doi.org/10.1016/j.brs.2017.10.015

[13] Koutsouleris, N., Wobrock, T., Guse, B., Langguth, B., Landgrebe, M., Eichhammer, P., Frank, E., Cordes, J., Wölwer, W., Musso, F., Winterer, G., Gaebel, W., Hajak, G., Ohmann, C., Verde, P.E., Rietschel, M., Ahmed, R., Honer, W.G., Dwyer, D., Ghaseminejad, F., Dechent, P., Malchow, B., Kreuzer, P.M., Poeppl, T.B., Schneider-Axmann, T., Falkai, P., Hasan, A.: Predicting response to repetitive transcranial magnetic stimulation in patients with schizophrenia using structural magnetic resonance imaging: A multisite machine learning analysis. Schizophrenia Bulletin **44**(5), 1021–1034 (2017) https://doi.org/10.1093/schbul/sbx114

[14] Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R.N., Zebley, B., Oathes, D.J., Etkin, A., Schatzberg, A.F., Sudheimer, K., Keller, J., Mayberg, H.S., Gunning, F.M., Alexopoulos, G.S., Fox, M.D., Pascual-Leone, A., Voss, H.U., Casey, B., Dubin, M.J., Liston, C.: Erratum: Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nature Medicine **23**(2), 264–264 (2017) https://doi.org/10.1038/nm0217-264d

[15] Erguzel, T.T., Ozekes, S., Gultekin, S., Tarhan, N., Sayar, G.H., Bayram, A.:

Neural network based response prediction of rTMS in major depressive disorder using QEEG cordance. Psychiatry Investigation **12**(1), 61 (2015) https://doi.org/10.4306/pi.2015.12.1.61

[16] Koppe, G., Meyer-Lindenberg, A., Durstewitz, D.: Deep learning for small and big data in psychiatry. Neuropsychopharmacology **46**(1), 176–190 (2020) https://doi.org/10.1038/s41386-020-0767-z

[17] Mumuni, A., Mumuni, F.: Data augmentation: A comprehensive survey of modern approaches. Array **16**, 100258 (2022) https://doi.org/10.1016/j.array.2022.100258

[18] Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., Naganawa, S.: Fairness of artificial intelligence in healthcare: review and recommendations. Japanese Journal of Radiology **42**(1), 3–15 (2023) https://doi.org/10.1007/s11604-023-01474-3

[19] Dunlap, L., Umino, A., Zhang, H., Yang, J., Gonzalez, J.E., Darrell, T.: Diversify Your Vision Datasets with Automatic Diffusion-Based Augmentation. arXiv (2023). https://doi.org/10.48550/ARXIV.2305.16289

[20] Li, T., Wang, Y., Liu, L., Chen, L., Chen, C.L.P.: Subspace-based minority oversampling for imbalance classification. Information Sciences **621**, 371–388 (2023) https://doi.org/10.1016/j.ins.2022.11.108

[21] Farahany, Z., Wu, J., Sajjadul Islam, K.M., Madiraju, P.: Oversampling techniques for predicting covid-19 patient length of stay. In: 2022 IEEE International Conference on Big Data (Big Data), pp. 5253–5262 (2022). https://doi.org/10.1109/BigData55660.2022.10020253

[22] Turlapati, V.P.K., Prusty, M.R.: Outlier-smote: A refined oversampling technique for improved detection of covid-19. Intelligence-Based Medicine **3-4**, 100023 (2020) https://doi.org/10.1016/j.ibmed.2020.100023

[23] Lucini, F.: The real deal about synthetic data. MIT Sloan Management Review **63**(1), 1–4 (2021). Copyright - Copyright Massachusetts Institute of Technology, Cambridge, MA Fall 2021; Last updated - 2021-11-30

[24] Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., Sales, A.P.: Generation and evaluation of synthetic patient data. BMC Medical Research Methodology **20**(1) (2020) https://doi.org/10.1186/s12874-020-00977-1

[25] Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F.K., Mahmood, F.: Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering **5**(6), 493–497 (2021) https://doi.org/10.1038/s41551-021-00751-8

[26] Jain, N., Olmo, A., Sengupta, S., Manikonda, L., Kambhampati, S.: Imperfect imaganation: Implications of gans exacerbating biases on facial data augmentation and snapchat face lenses. Artificial Intelligence **304**, 103652 (2022) https://doi.org/10.1016/j.artint.2021.103652

[27] Sharma, H., Gosain, A.: Oversampling methods to handle the class imbalance problem: A review. In: Patel, K.K., Santosh, K.C., Patel, A., Ghosh, A. (eds.) Soft Computing and Its Engineering Applications, pp. 96–110. Springer, Cham (2023)

[28] Fajardo, V.A., Findlay, D., Jaiswal, C., Yin, X., Houmanfar, R., Xie, H., Liang, J., She, X., Emerson, D.B.: On oversampling imbalanced data with deep conditional generative modelsf. Expert Systems with Applications **169**, 114463 (2021) https://doi.org/10.1016/j.eswa.2020.114463

[29] Strelcenia, E., Prakoonwit, S.: Improving cancer detection classification performance using gans in breast cancer data. IEEE Access **11**, 71594–71615 (2023) https://doi.org/10.1109/ACCESS.2023.3291336

[30] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research (2011) https://doi.org/10.48550/ARXIV.1106.1813

[31] Sowjanya, A.M., Mrudula, O.: Effective treatment of imbalanced datasets in health care using modified smote coupled with stacked deep learning algorithms. Applied Nanoscience (2022) https://doi.org/10.1007/s13204-021-02063-4

[32] Yang, W., Pan, C., Zhang, Y.: An oversampling method for imbalanced data based on spatial distribution of minority samples sd-kmsmote. Scientific Reports **12**(1) (2022) https://doi.org/10.1038/s41598-022-21046-1

[33] Wang, Q., Luo, L., Xie, H., Rao, Y., Lau, R.Y.K., Zhang, D.: A deep data augmentation framework based on generative adversarial networks. Multimedia Tools and Applications **81**(29), 42871–42887 (2022) https://doi.org/10.1007/s11042-022-13476-w

[34] Sun, H., Deng, Z., Chen, H., Parkes, D.C.: Decision-Aware Conditional GANs for Time Series Data. arXiv (2020). https://doi.org/10.48550/ARXIV.2009.12682

[35] Arora, A., Arora, A.: Generative adversarial networks and synthetic patient data: current challenges and future perspectives. Future Healthcare Journal **9**(2), 190–193 (2022) https://doi.org/10.7861/fhj.2022-0013

[36] Alkhalifah, T., Wang, H., Ovcharenko, O.: Mlreal: Bridging the gap between training on synthetic data and real data applications in machine learning. Artificial Intelligence in Geosciences **3**, 101–114 (2022) https://doi.org/10.1016/j.aiig.2022.09.002

[37] Li, J., Cairns, B.J., Li, J., Zhu, T.: Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. npj Digital Medicine **6**(1) (2023) https://doi.org/10.1038/s41746-023-00834-7

[38] Dash, A., Ye, J., Wang, G.: A review of generative adversarial networks (gans) and its applications in a wide variety of disciplines: From medical to remote sensing. IEEE Access, 1–1 (2023) https://doi.org/10.1109/access.2023.3346273

[39] Torfi, A., Fox, E.A.: CorGAN: Correlation-Capturing Convolutional Generative Adversarial Networks for Generating Synthetic Healthcare Records. arXiv (2020). https://doi.org/10.48550/ARXIV.2001.09346

[40] Zhang, X.: Application of discrete event simulation in health care: a systematic review. BMC Health Services Research **18**(1) (2018) https://doi.org/10.1186/s12913-018-3456-4

[41] Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W.F., Sun, J.: Generating multi-label discrete patient records using generative adversarial networks. In: Doshi-Velez, F., Fackler, J., Kale, D., Ranganath, R., Wallace, B., Wiens, J. (eds.) Proceedings of the 2nd Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 68, pp. 286–305. PMLR, ??? (2017). https://proceedings.mlr.press/v68/choi17a.html

[42] Torfi, A., Fox, E.A., Reddy, C.K.: Differentially private synthetic medical data generation using convolutional GANs. Information Sciences **586**, 485–500 (2022) https://doi.org/10.1016/j.ins.2021.12.018

[43] Behal, K., Chen, J., Fikes, C., Xiao, S.: MCRAGE: Synthetic Healthcare Data for Fairness. arXiv (2023). https://doi.org/10.48550/ARXIV.2310.18430

[44] Savage, N.: Synthetic data could be better than real data. Nature (London) (2023)

[45] Naveed, H., Anwar, S., Hayat, M., Javed, K., Mian, A.: Survey: Image Mixing and Deleting for Data Augmentation. arXiv (2021). https://doi.org/10.48550/ARXIV.2106.07085

[46] Lashgari, E., Liang, D., Maoz, U.: Data augmentation for deep-learning-based electroencephalography. Journal of Neuroscience Methods **346**, 108885 (2020) https://doi.org/10.1016/j.jneumeth.2020.108885

[47] Garcea, F., Serra, A., Lamberti, F., Morra, L.: Data augmentation for medical imaging: A systematic literature review. Computers in Biology and Medicine **152**, 106391 (2023) https://doi.org/10.1016/j.compbiomed.2022.106391

[48] Goceri, E.: Medical image data augmentation: techniques, comparisons and interpretations. Artificial Intelligence Review **56**(11), 12561–12605 (2023) https:

//doi.org/10.1007/s10462-023-10453-z

[49] Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 289–293. IEEE, ??? (2018)

[50] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks. arXiv (2014). https://doi.org/10.48550/ARXIV.1406.2661

[51] Nash, J.: Non-cooperative games. The Annals of Mathematics **54**(2), 286 (1951) https://doi.org/10.2307/1969529

[52] Arora, S., Risteski, A., Zhang, Y.: Do gans learn the distribution? some theory and empirics. In: ICLR (2018)

[53] Arora, S., Zhang, Y.: Do GANs actually learn the distribution? An empirical study. arXiv (2017). https://doi.org/10.48550/ARXIV.1706.08224

[54] Mirza, M., Osindero, S.: Conditional Generative Adversarial Nets. arXiv (2014). https://doi.org/10.48550/ARXIV.1411.1784

[55] Sun, Z., Zhang, H., Bai, J., Liu, M., Hu, Z.: A discriminatively deep fusion approach with improved conditional gan (im-cgan) for facial expression recognition. Pattern Recognition **135**, 109157 (2023) https://doi.org/10.1016/j.patcog.2022.109157

[56] Mert, A.: Enhanced dataset synthesis using conditional generative adversarial networks. Biomedical Engineering Letters **13**(1), 41–48 (2022) https://doi.org/10.1007/s13534-022-00251-x

[57] Hopman, H., Chan, S., Chu, W., Lu, H., Tse, C.-Y., Chau, S., Lam, L., Mak, A., Neggers, S.: Personalized prediction of repetitive transcranial magnetic stimulation clinical response in medication-refractory depression data. Data in Brief **37**, 107264 (2021) https://doi.org/10.1016/j.dib.2021.107264

[58] Borji, A.: Pros and cons of GAN evaluation measures: New developments. Computer Vision and Image Understanding **215**, 103329 (2022) https://doi.org/10.1016/j.cviu.2021.103329

[59] Esteban, C., Hyland, S.L., Rätsch, G.: Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. arXiv (2017). https://doi.org/10.48550/ARXIV.1706.02633

[60] Tasci, E., Zhuge, Y., Camphausen, K., Krauze, A.V.: Bias and class imbalance in oncologic data—towards inclusive and transferrable AI in large scale oncology

data sets. Cancers **14**(12), 2897 (2022) https://doi.org/10.3390/cancers14122897

4.3 Links and implications

This thesis has demonstrated DL can reliably predict the treatment outcome to rTMS using fMRI connectivity features. However, Chapter 2.3 demonstrated this technology does not perform as well for a subset of patients. If AI is to be deployed safely and fairly in psychiatry then we must strive that AI systems are designed to mitigate bias. Thus working towards the stated aim of personalised psychiatry, "the right treatment for **every** patient" (right treatment for each patient:unlocking the potential of personalized psychiatry' 2023). This chapter demonstrates techniques like data augmentation can be used to mitigate the bias of underrepresented examples. By synthetically boosting the diversity of training datsets we cam improve performance.

CHAPTER 5: PAPER 4 - ENHANCING SUICIDE RISK DETECTION ON SOCIAL MEDIA THROUGH SEMI-SUPERVISED DEEP LABEL SMOOTHING

5.1 Introduction

Chapter 1.6 highlighted one of the most significant challenges of diagnosis in psychiatry is the lack of objective marker of disease. Even amongst experts there can be disagreement on some diagnosis. The challenge associated with this is that DL models expect a level of certainty. However, when experts disagree then to use only binary classification scheme does not accurately represent the true nature of a system. This chapter presents a novel take on representing this uncertainty using label smoothing. Utilising this technique, this chapter seeks to more accurately represent ground truth labels in fuzzy systems by transposing hard binary classification labels to smoothed labels.

5.2 Published paper

Enhancing Suicide Risk Detection on Social Media through Semi-Supervised Deep Label Smoothing

Matthew Squires[a], Xiaohui Tao[a], Soman Elangovan[b], U Rajendra Acharya[a], Raj Gururajan[c], Haoran Xie[d], Xujuan Zhou[c]

[a]*School of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba, Australia,*
[b]*Belmont Private Hospital, Brisbane, Australia,*
[c]*School of Business, University of Southern Queensland, Springfield, Australia,*
[d]*Department of Computing and Decision Sciences, Lingnan University, Hong Kong SAR, China,*

Abstract

Suicide is a prominent issue in society. Unfortunately, many people at risk for suicide do not receive the support required. Barriers to people receiving support include social stigma and lack of access to mental health care. With the popularity of social media, people have turned to online forums, such as Reddit to express their feelings and seek support. This provides the opportunity to support people with the aid of artificial intelligence. Social media posts can be classified, using text classification, to help connect people with professional help. Text classification of social media posts for the detection of mental health distress is a large and ever-expanding field. However, these systems fail to account for the inherent uncertainty in classifying mental health conditions. Unlike other areas of healthcare, mental health conditions have no objective measurements of disease often relying on expert opinion. Thus when formulating deep learning problems involving mental health, using hard, binary labels does not accurately represent the true nature of the data. In these settings, where human experts may disagree, fuzzy or soft labels may be more appropriate. The current work introduces a novel label smoothing method which we use to capture any uncertainty within the data. We test our approach on a five-label multi-class classification problem. We show, our semi-supervised deep label smoothing method improves classification accuracy above the existing state of the art. Where existing research reports an accuracy of 43% on the Reddit C-SSRS dataset, using empirical experiments to evaluate our novel label smoothing method, we im-

Preprint submitted to Artificial Intelligence in Medicine *May 9, 2024*

prove upon this existing benchmark to 52%. These improvements in model performance have the potential to better support those experiencing mental distress. Future work should explore the use of probabilistic methods in both natural language processing and quantifying contributions of both epistemic and aleatoric uncertainty in noisy datasets.

Keywords: Label Smoothing, Mental Health, Probabilistic Deep Learning, Uncertainty Quantification

1. Introduction

Depression and Suicide are significant issues in society. Given this, researchers have explored many avenues to aid in the treatment and diagnosis of these mental health conditions. Barriers such as reduced access to mental health services [1], and stigma associated with seeking mental health care [2] are among the factors which prevent people from seeking help. As such, many people are turning to social media to seek support and share mental health related information [3]. The use of Artificial Intelligence (AI) systems to detect depression has seen extensive research. Building upon existing work, Gaur et al. [2] investigated the use of text classification to recognise social media users who may be a suicide risk. Text classification systems for suicide behaviours could help to connect users sharing their emotions and mental health struggles with health professionals using AI models.

A difficulty of mental health research, in contrast to other fields of healthcare, mental health conditions have no objective markers of disease [4]. This lack of objective markers is one of several key challenges in identifying psychopathology [5]. Furthermore, human raters can find classifying suicidal behaviours using assessment tools to be difficult [6]. When it is difficult for human raters to agree on a ground truth label it is likely difficult for AI to uncover underlying patterns [2]. In turn, this presents a difficulty for deep learning models which have traditionally relied on binary labels. Fuzzy logic, however, allows for an alternative view. When borders between groups are unclear fuzzy variables "facilitate gradual transitions between states and, consequently, possess a natural capability to express and deal with observation and measurement uncertainties" [7, p.4]. The inadequacy of binary, ore one hot encoded ground truth labels has also been explored in the field of text emotion classification [8]. Where fuzzy emotions can be used to capture text which may convey multiple emotions, where a binary mapping fails to

capture an accurate ground truth.

Uncertainty is ever present in the mental health field due to the reliance on self-reporting and observation. When uncertainty is present in the labelling of ground truth, it seems unreasonable to use traditional hard labels where $y \in 0, 1$. In these settings binary variables as ground truth labels do not represent the true nature of a system. Label smoothing is a technique which involves subtracting a small value from the true class, and distributing the subtracted value evenly across each remaining class. Label smoothing is uniform, that is the distribution remains constant across all remaining labels, as below:

$$y = \begin{cases} 1 - \alpha & \text{if } y_i = 0 \\ \alpha/(k-1) & \text{otherwise} \end{cases} \tag{1}$$

These updated labels are referred to as soft labels. Initially proposed as a regularisation technique to help prevent model overfitting some research investigates the use of soft labels to improve model performance. Recently, Zhang et al. [9] showed non-uniform label smoothing to improve model performance on benchmark datasets such as CIFAR-100 and ImageNet. Our work extends the use of non-uniform distribution label smoothing to text classification and introduces a novel strategy using Bayesian techniques to generate the smoothed labels. These smoothed labels are likely more representative of the fuzzy nature of classifications in the mental health space. As such, this paper seeks to explore issues of uncertainty central to the use of AI in mental health care. Formalised in the following research questions:

- How can the uncertainty in ground truth labels be expressed in settings where expert opinion may be divided?
- How does incorporating uncertainty into labels impact model performance?
- Can the underlying truth label distribution be found?

To explore these research questions we utilise the Reddit C-SSRS dataset first presented by [2]. The data includes posts by 500 Reddit users assessed by experts using the Columbia Suicide Severity Rating Scale (C-SSRS). According to Gaur et al., [2] user posts were labelled by four practising clinical psychiatrists with pairwise annotator agreement varying between $\approx 80\%$ and

$\approx$ 60%, however, one-hot encoded variables fail to capture the uncertainty between raters.

Our work is motivated by the idea that it is difficult for deep learning models to identify underlying relationships in data when the labels do not represent the true nature of the system. Thus if human mental health care experts do disagree on how to classify a post then this uncertainty must be expressed. As such, we propose a novel label smoothing method which builds upon existing work to more accurately capture the uncertainty of ground truth labels. Through our experiments exploring the detection of mental distress of social media posts we make the following contributions:

- A novel fuzzy semi supervised deep learning method for label smoothing to represent uncertainties in ground truth labels
- A state of the art text classification model, achieving accuracy surpassing existing models tested on the same dataset using fuzzy labels;
- An exploration of the use of fuzzy class membership for the use of text classification

This paper is structured as follows. Section 2 provides an overview of existing methods used for uncertainty quantification. These techniques, predominantly used for image segmentation, and the dearth of literature in mental health care motivates this work. This section identifies the gap in the literature and the opportunity for the use of uncertainty estimation in text classification. Section 3 provides an overview of the methods, techniques and data set used to present our novel uncertainty estimation techniques incorporated with text classification. In the remaining sections, Section 5 provides a summary of the performance of the current work compared against the baseline model. Finally, Section 6 and 7 provide concluding remarks explaining the models behaviour and offering future directions for the field.

2. Related Work

The quantification of uncertainty when using deep learning in healthcare is expanding rapidly. This is due to the acknowledgement that if deep learning is to be used in critical settings, such as healthcare, uncertainty quantification must be further developed [10]. Probabilistic deep learning methods, such as Bayesian Neural Networks, Deep Ensembles and Monte-Carlo (MC)

dropout are common techniques proposed to explore these uncertainties. To date, the use of stochastic methods in healthcare focuses on medical imaging and image processing (see [11]). For example, Bayesian techniques were applied to the detection of oral cancer from intraoral images, the diagnosis of COVID-19 from X-rays and the classification of brain lesions from MRI images.

In their survey, Abdar et al. [12] identify aleatoric and epistemic uncertainty as the two main categories of uncertainty. Epistemic uncertainty refers to uncertainty resulting from a model's lack of understanding. Hüllermeir and Waegeman [13] define epistemic uncertainty as "the uncertainty caused by a lack of knowledge." In the practical sense for AI or deep learning models Hüllermeir and Waegeman [13] assert epistemic uncertainty "refers to the ignorance of the agent or decision maker." In this context, the "agent or decision maker" could refer to an AI agent or deep learning model. The understanding that epistemic uncertainty refers to the lack of knowledge by a model, has resulted in epistemic uncertainty to be more commonly referred to as model uncertainty [14]. Model uncertainty is taken to occur when a model is exposed to an example which lies outside of the distribution on which it was trained. Given this, the logical solution to handle epistemic uncertainty is to provide more data [14]. While this may be possible for active learners in fields with large and ever-expanding data sets, in some cases, such as psychology and psychiatry large data sets are not always accessible, making the expression of epistemic uncertainty hugely important.

In contrast, aleatoric uncertainty is defined as "noise inherent in the data distribution" [15]. For example, annotation ambiguity. Where annotation ambiguity refers to uncertainty regarding labelling examples in supervised learning tasks. Put another way, Hüllermeir and Waegeman [13] contend aleatoric uncertainty refers to the randomness inherent to data collection. For this reason, aleatoric uncertainty is also known as data uncertainty. The notion of inherent randomness in data collection is of particular significance in the field of psychiatry and psychology. For many conditions within the mental health space, few or no objective biomarkers of disease exist. As such, the discipline relies heavily on subjective measurement to quantify disease. For example, expert psychiatrists Gaur et al. [2] labelled posts on Reddit according to a five-label suicide severity rating scale. Four practising clinical psychiatrists annotated each post according to the five-label classification scheme. Results reported in [2] show that pairwise annotator agreement on the annotation of posts varied from 79& to 65%. This irreducible random-

ness due to the subjective nature of mental health conditions is an important consideration when modelling mental health conditions. This disagreement between annotators motivates our work, to explore the utility of fuzzy labels and label smoothing as a more accurate representation of the true data distribution.

To date, few methods have been explored to represent the subjectiveness of human raters in generating labels for mental health conditions. Data cleaning methods such as label correction were trialled in [16], which showed using clustering to correct potentially noisy labels improved model performance. Given the success of label correction, our work seeks to explore the possibility of soft labels to represent uncertainty within the labels. These works move beyond label correction, to label smoothing.

Techniques for capturing this uncertainty can be divided into Bayesian techniques and deep ensembles. Song et al. [17] applied a Bayesian Deep Neural Network to the classification of oral cancer from a large image dataset. The work utilised Monte Carlo Dropout a common Bayesian technique to express prediction uncertainty. Dropout is a common regularisation technique used to reduce the chances of model overfitting. The technique involves 'dropping' a percentage of nodes from a hidden layer [18]. This dropping is achieved by turning the parameter weights of all edges to a given node to zero, to in essence eliminate the impact of that node on the network. The dropout rate indicates the percentage of nodes from a layer which should be dropped. For example, in Song et al. [17] the dropout rate was set to 0.5 or 50% of nodes. Seminal work by Gal and Ghahramani [19] showed combining Monte Carlo simulation with dropout could be used to quantify the uncertainty in model predictions. As hidden layer nodes are dropped randomly, repeatedly simulating the same network with dropout has the effect of in essence training different networks and ultimately leading to more robust networks. The output predictions of each network can then be averaged [19] and the variance of these predictions is defined as the prediction uncertainty [17]. Song et al. [17] show, the performance of their Convolution Neural Network (CNN) improves when the uncertainty threshold is varied. As such, the aim of uncertainty quantification is that images with high levels of uncertainty can be referred to human experts for further evaluation.

Additionally, Gour and Jain [20] provide a second example of the combination of MC dropout with a CNN. Gour and Jain [20] utilise a CNN for the diagnosis of COVID-19 from x-ray images. The work utilises a dropout rate of 0.425 with a pre-trained CNN. Furthermore, [21] utilised MC Dropout for

the classification of brain lesions from MRI scans. Wu et al. [21] trained a Deep Convolutional Neural Network (DCNN) with MC Dropout in a teacher-student framework. The teacher-student framework involves the training of two networks, the first model, the teacher is initially trained on the dataset. The second model, the student uses the predictions made from the teacher model with the initial data to make final predictions [22]. Although not Bayesian, Gjestang et al. [22] provide an example of the use of a teacher-student framework for the classification of gastrointestinal images. In Gjestang et al. [22], the teacher model is first trained on unlabeled data to generate pseudo labels. The student is then trained using the teacher-generated labels, the original data set and the original labels with the aim of minimizing the loss function against the target labels. In a novel approach by Wu et al. [21] the teacher is a Bayesian Deep Network. The student then receives as inputs the Bayesian probabilities output by the Teacher. However, Wu et al. [21] do not provide details on the dropout rate of their network.

A recent review by Abdullah et al. [11] provides a comprehensive overview of the current state of Bayesian deep learning techniques within the field. There exists limited research which incorporates uncertainty quantification and model confidence into natural language processing and text classification problems. Abdullah et al. assert [11] they "are not aware of any published work on medical Natural Language Processing that has used Bayesian deep learning (p.36522, [11]). This lack of published works exploring probabilistic methods, uncertainty quantification and text classification provides the gap which further motivates this work. Additionally, significant calls exist for the addition of uncertainty estimation to deep learning models [10]. Given the lack of research exploring uncertainty estimation and text classification and the acknowledged need for deep learning models to express their prediction confidence. Our work seeks to address the needs of the research community by exploring both model and label uncertainty to produce better-performing models, ultimately leading to improved mental health outcomes.

3. Method

3.1. Conceptual Model

This section provides an overview of our novel semi-supervised label smoothing method and the experiments used to evaluate its efficiency when compared to existing methods. As part of the current work, we explore both aleatoric and epistemic uncertainty associated with classifying social media

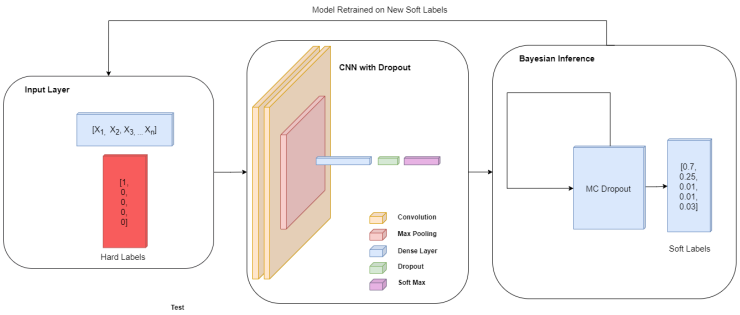

Figure 1: Conceptual Model

posts and their suicide risk. For the purpose of this work, aleatoric uncertainty can be thought of as the uncertainty associated with the ground truth labels.

Our first experiment explores the label uncertainty which is generated by the subjective nature of the diagnosis of mental health conditions. Given this subjectiveness, human raters can at times disagree on the assessment of the same social media post. To capture this uncertainty we propose fuzzy smoothed labels, generated by our novel label smoothing approach using Bayesian techniques. Additionally, we investigate the impact of prediction confidence on

This section details the mechanisms of our approach and experiments utilised to measure the efficacy of our model.

3.2. Problem Definition

The current work can be formally defined as a text classification problem. We utilise a dataset, $\mathcal{D}$ which consists of posts from n Reddit users, such that $\mathcal{X} = \{x_1, x_2, x_3 \cdots, x_n\}$. Users' posts are labelled according to a five-label suicide classification scheme. Such that a post from a user, x_n belongs to a single class $\mathcal{K} = \{k_1, k_2, k_3 ... k_5\}$.

Our work explores the impact of using smoothed labels such that an example can belong partially to several classes. For example, the target label of a post x_n could partially belong to multiple classes as demonstrated in equation 1. In the current work, we introduce a novel method of obtaining

the probability of class membership and explore the impact of uniform and non-uniform label smoothing methods against existing methods on the same dataset.

Where our aim is the classifier function which minimises the cost function C during model testing on a hold-out-test-set. Where C is categorical cross-entropy, the multiclass case of the widely used binary cross-entropy, is defined as:

$$\mathcal{C} = -\sum_{k=1}^{k} y_k \cdot log(\hat{y_k}) \tag{2}$$

3.3. Conceptual Model

The curated data set provided by Guar et al. [2] contains posts from 500 Reddit users labelled according to a five-label suicide severity risk. To perform text classification, natural language must be made machine-readable.

Before text can be used in machine learning tasks it must be prepared for input into deep learning models. All associated preprocessing and model building for this project was constructed in Python with the aid of the software package Keras [23], a popular deep learning library. Preparing natural language text for input to deep learning models requires word embeddings. Text embeddings are the process of converting data to numerical representations [24]. A variety of word embedding techniques are available to machine learning researchers to convert Reddit posts to input vectors. Following tokenisation, the input layer takes as input sequential arrays of length 5,041 which is passed to the word embeddings layer. The word embeddings layer learns the most effective vector representation of the model inputs to perform the text classification task.

Figure 1 details the proposed model architecture. Using this architecture. The word embedding vector is then passed to the first convolutional layer for model training. The convolutional layer is passed to a second convolutional layer which is then down sampled via a pooling layer. The technical benefits of combining convolutional and max pooling layers for text classification are discussed in Section 3.3.1.

Following the max pooling later, the output is flattened before being passed to the final fully connected layer. The fully connected layer utilises the softmax activation function with five output nodes. In our proposed approach Bayesian Inference is used to obtain updated smoothed labels which are then

used to retrain the network. Our experiments explore the effectiveness of this approach to represent the uncertainty faced when diagnosing mental health conditions when human experts disagree on a classification. The aim of this work is to capture these disagreements in the data to build a more effective model.

3.3.1. Convolutional Neural Networks

An extension of the multilayer perceptron is the convolutional neural network (CNN) and deep convolutional neural network. Convolutional Neural networks are the workhorse of pattern recognition for images [25]. The input vector of a multilayer perceptron is typically a flat one-dimensional array or vector. In contrast, the CNN is equipped to take multi-dimensional vectors as inputs. This allows CNNs to take as input a three-dimensional array associated with the red, green and blue components of a two-dimensional image.

The core components of the CNN which differ from the Multilayer Perceptron (MLP) are the convolutional and max pooling layers. The convolutional layer is known to "learn local patterns—in the case of images, patterns found in small 2D windows of the inputs" [26]. This concept of spacial dimensionality is not possible for flattened inputs to MLPs which instead are constrained to pattern recognition across the entire input vector. The size of the area the convolutional layer looks to for these local patterns is referred to as the kernel [25].

In addition to excelling in computer vision tasks, CNNs have demonstrated high levels of performance in text classification [27]. As with images, this success is attributed to the ability of convolutional layers to learn local patterns within data. In the task of pattern recognition in text, convolutions are able to capture patterns within sequential text. Goldberg [28] asserts the combination of one-dimension convolution and pooling operates as an n-gram detector. With [29] going further, suggesting pooling acts similarly to a feature extraction layer, with only class discriminative n-grams passed through the pooling layer. A pooling layer can be thought of as downsampling the size of the original input [25], this ensures relevant information is captured for text classification.

3.3.2. Fuzzy Label Smoothing

Label smoothing is a regularisation technique that has been shown to improve model performance. The smoothing of hard labels involves removing

a small value from the true case and distributing that value across all classes. A formal definition of uniform label smoothing as described by Shen et al. [30] is given below:

$$y_i = \begin{cases} 1 - \alpha & \text{if } y_i = 0 \\ \alpha/(k-1) & \text{otherwise} \end{cases} \tag{3}$$

In contrast to existing methods which either use a uniform distribution or denoise data through label correction, our fuzzy label smoothing method allows for partial multi-class membership. Figure 1, details our architecture. We use the initial hard labels in the first training batch before repeatedly simulations using MC Dropout. These simulations produce updated non-uniform smoothed labels. These smoothed labels follow a non-uniform distribution which looks to simulate the uncertainty of ground truth labels. More accurately representing the disagreement in human raters to more accurately represent the underlying distribution.

3.4. Bayesian Inference and MC Dropout

Deep learning models have demonstrated human-like or exceeded human performance on many tasks. The concept of the artificial neural network is not new. The foundations for the artificial neuron found their origins in the 1950s. The building block of the artificial network, the perceptron was founded by Rosenblatt et al. [31], and their seminal work on the perceptron. Advances in computing power in the 2010s provided the opportunity for deep neural networks [32]. Advances in computing power, have allowed for the chaining of perceptrons together in multiple layers providing the 'depth' of the deep neural network. Building upon traditional neural networks, Bayesian models have been shown to capture prediction confidence. One popular Bayesian technique is MC dropout, a technique first described in [19].

The following section emphasises the differences between the functionality of a traditional artificial neural network and one which utilises the MC Dropout method. Consider the traditional supervised learning task:

$$\hat{y} = f(x') \tag{4}$$

In this task, we seek a function f which minimises the cost function C.

$$\arg\min C[f] \tag{5}$$

Thus a simple feed-forward neural network, with sigmoid activation θ can be denoted as:

$$f(x_i) = \theta(x_i \cdot w + b) \tag{6}$$

As we look to model more complex relationships we can increase the width and depth of the network. Width refers to the number of nodes in a given layer whereas depth refers to the number of hidden layers. Hence deep learning refers to networks with several hidden layers. With these additional hidden layers, the weight term above is replaced with a set of weights W connecting each input node with the hidden layer and a further set of weights connecting each node of the hidden layer with the output node. This set of weights is a weight constellation [32].

Before exploring the use of Monte Carlo Dropout as a Bayesian technique we first provide familiarisation with Bayesian Neural Networks and Bayes theorem. Bayes theorem is shown in Equation 7.

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} \tag{7}$$

The regularisation technique dropout essentially involves turning a proportion of parameter weights down to zero. Visually we can represent this by dropping edges between nodes. When Monte Carlo dropout is used, multiple passes through the network are simulated iteratively reducing the weights of a proportion of nodes from a hidden layer down to zero. When dropout is applied during testing each pass through the network will result in a different weight constellation W. Thus for a simulation of $T = 100$, we get a set of 100 weights and 100 output predictions. Equation 8 can be interpreted as the probability of y, given the input value x_i from the data set $\mathcal{D}$ is equal to the average of the probabilities for the x_ith example from each weight constellation, where $\mathcal{W} = (W_1, W_2, W_3 \dots W_t)$.

$$P(y|x_i, \mathcal{D}) = \frac{1}{T} \sum_{T}^{t=1} P(y|x_i, w_t) \tag{8}$$

Algorithmically we can describe the MC Dropout process as a Monte Carlo simulation through T passes over a test sample. The final output of the Monte Carlo simulation is a probability distribution of membership to each class where the maximum is defined as the final output.

Hence, for a multiclass problem with k, classes, we can compute the probability of class membership for each weight matrix w_t. Such that. The probability of class membership, p_K is given by Equation: 9.

$$p_k = \frac{1}{T}\sum_{t=1}^{T} p_{k_t} \tag{9}$$

4. Experiments

4.1. Experiment Settings

4.1.1. Label Smoothing

Our computational experiments explore the effects of hard labels, traditional uniform label smoothing and our novel fuzzy label smoothing methods on model performance. To evaluate these methods data was divided into training and test set. In line with the proportions used in [2] 20% of examples were used as a hold-out-test-set. For each experiment, the same convolutional model was used with only the label type modified. The experiment conditions were the original hard labels, uniform label smoothing and finally our novel approach. The results reported are the performance on the test set.

4.2. Baseline Models

The baseline model used for the current work is the model presented in [2]. Their work provides the existing state-of-the-art performance on the current dataset. The dataset used for this study is sourced from Gaur et al. [2]. Gaur et al. [2] provide important work on the assessment of suicide risk. The results of which are described later in Section 5. Their work provides a novel five-label classification scheme of suicide risk. The five-label scheme classifies Reddit posts from prominent mental health subreddits as either Suicidal Ideation (ID), Suicidal Behaviour (SB), Actual Attempt (AT), Suicide Indicator (IN) or Supportive (SU). Descriptive statistics for the frequency of class membership are described in Table 1.

4.3. Metrics

Common metrics used for binary classification tasks include Accuracy, Precision, Recall and F1-score. However, the current problem is a multi-class problem where the number of classes, k, exceeds 2 (ie. $k > 2$) as in

Table 1: Frequency of Class Membership

Label	Frequency
Ideation	171
Behaviour	77
Attempt	45
Indicator	99
Supportive	108

the binary sense (where $k = 2$). Although this paper explores a multi-class problem, common binary classification metrics can be used by incorporating some modifications.

For multi-class classification, the accuracy metric remains the same as in binary classification. Correctly labeled examples are divided by the total number of examples in the test set. However, the precision and recall metrics vary slightly from those used for binary classification.

From Grandini et al. [33]:

$$Precision_k = \frac{TP_k}{TP_k + FP_k}$$

$$Recall_k = \frac{TP_k}{TP_k + FN_k}$$

Where TP, is the number of correctly labeled examples and FP is the number of false positives. This Precision metric can be calculated by class or can be summarised using either micro or macro average.

$$MacroAveragePrecision = \frac{\sum_{k=1}^{k} Precision_k]}{k}$$

or

$$MicroAveragePrecision = \frac{\sum_{k=1}^{k} TP_k}{\sum_{k}^{k} Column_k}$$

The challenge of computing summary metrics for multi-class problems is an open problem in data science. Traditionally, a single summary metric, such as F1-score is preferred to capture mode performance. However, the F1-score using Micro averages fails to capture differences in class sizes [33]. Takahashi et al. [34] assert "the inherit drawback of multi-class F1 scores [is]

that these scores do not summarize the data appropriately when a large variability exists between classes. "[34, p.4966]. The current problem includes large variability in classes. As such, along with average Precision and Recall, we include balanced accuracy. Weighted balanced accuracy is the only metric which captures variability in class sizes. Classes are weighted proportional to their class frequency [33].

$$WeightedBalancedAccuracy = \frac{\sum_{k=1}^{k} Recall_k \cdot \frac{1}{w_k}}{k \cdot w}$$

Where w_k is the assigned weight for each class, k and weights are proportional to the frequency of each class within the sample.

5. Results

Our experiments investigate the effect of different labelling methods on model performance. Table 2 provides an overview of the results of our experiments based on the labelling method used. Initially, we see the results reported by Gaur et al. using their approach. We note, that our proposed architecture trained using hard labels slightly outperforms the results reported by Gaur et al. in accuracy, weighted balanced accuracy, macro average precision and macro average recall.

Table 2: Classification Accuracy by Labelling Method

Method	Accuracy	Weighted Balanced Accuracy	Macro Average Precision	Macro Average Recall
Gaur et al. [2]	0.4312	0.2567	0.2903	0.2734
Hard Labels	0.4451	0.3036	0.3337	0.3036
Label Smoothing $\alpha = 0.1$	0.4699	0.3698	**0.5284**	0.3698
Label Smoothing $\alpha = 0.05$	0.4783	0.4226	0.4364	0.4266
Deep Bayesian Label Smoothing	**0.5233**	0.4923	**0.4721**	**0.4777**

Label smoothing when $\alpha = 0.1$ improves on both Gaur et al. and the original hard labels. This alpha value records the best macro average precision of all tested methods. Uniform label smoothing with $\alpha = 0.05$ produces slight improvements in accuracy and weighted balanced accuracy over the previous alpha value however, we see a slight decline in macro average precision.

Utilising our novel deep Bayesian label smoothing method achieved the best accuracy (52.33%) and the best weighted balanced accuracy (49.23%), by a clear margin with the second best weighted balanced accuracy at 42.26% demonstrating the ability of the non-uniform label smoothing approach to more accurately predict across all classes within the data

In summary, Table 2 shows steady improvements in classification accuracy depending on the labelling method used. Uniform soft labels report similar levels of classification accuracy. Whereas, our proposed, non-uniform fuzzy label smoothing method, improves upon existing methods to produce a greater weighted balanced average, and recall over existing methods.

6. Discussion

The overarching aim of this paper was to explore the effect of various labelling methods to predict mental health distress. The difficulty in developing models on systems which rely heavily on subjective measurements, such as mental health conditions, is binary labels to do not accurately capture disagreement between experts. To capture this uncertainty we present Deep Bayesian Label Smoothing, a new method for softening target labels of deep neural networks to improve prediction accuracy. Given many mental health conditions do not have objective markers of disease. This novel approach is designed to incorporate uncertainty into ground truth labels. Which it is hoped will in turn more accurately represent the true nature of a system, thus leading to improvements in model performance.

Our experiments show, that using soft labels generated using non-uniform label smoothing leads to improved performance on a held-out test set. Existing works on this Reddit C-SSRS curated data set was presented by [2]. Interestingly, the existing state-of-the-art work overwhelmingly predicts the predominant class of the dataset. That is, 92% of predictions of the model presented by [2] on the test set are made on the most frequent class of the data. The baseline model exceedingly predicts suicidal ideation, and posts displaying supportive behaviours. This model behaviour is exemplified by low values in precision and recall for the remaining classes. These output metrics

suggest the model is failing to uncover the underlying function representation, and is instead making predictions probabilistically. In contrast, our proposed model demonstrates more consistent performance across all classes. This consistent performance suggests some underlying patterns leading to inputs being classified in a certain way do exist. Past research has suggested CNNs may act as n-gram detectors. Goldberg [28] contend that convolutional networks that combine convolutions with pooling layers are useful "when we we expect to find strong local clues regarding class membership, but these clues can appear in different places in the input" [28, p.3]. In the sense of the current problem, it is possible there exist class discriminant n-grams identified by the CNN model.

A review of model performance highlights clear variations in performance across classes. The baseline model records low precision and recall for the suicide attempts, suicide behaviours and suicide indicators classes. The entropy-filtered models presented in the current works significantly outperform the baseline model, however, still have no true positives for the suicide attempt class. Our reported results suggest there is no class which was consistently classified as a suicide attempt. Thus the in-text relationships required to detect suicide attempts were not learnt by the CNN model. The inability to detect posts labeled as suicide attempts is one limitation of this model. Demonstrating further improvement of suicide risk classification models is required to excede human performance. Given easy-to-classify posts such as *"Its been a year for me since I survived my Suicide attempt"* should clearly be classified as posts describing a suicide attempt. This post is easily classified when inspected however, it is clear models to date have been unable to detect this class of post. To advance this current work, the use of more sophisticated text representation techniques which incorporate context may improve model accuracy. Examples of text representation techniques which represent contexts such as Glove or BERT [35]. It appears it is difficult for the current work to interpret the context, that is if a suicide attempt is referring to another user's post or indicative of a user's own suicide dataset.

The network proposed in this work utilises an embedding layer, two convolutional layers and a max pooling layer. Adopting the convolution-pooling architecture of which the benefits are described in depth in [28]. It is possible deploying deeper convolutional networks, which incorporate more hidden layers may benefit performance. However, deeper networks, these deeper networks become computationally expensive to deploy. Typically, probabilistic methods require large amounts of computational resources. As such,

it is possible very deep convolutional networks with probabilistic methods may become too computationally large to compute on a single machine in reasonable computation time and hence difficult to deploy.

Additionally, as stated in Section 2, the authors of [2] state that each post had varying levels of agreement on each post. However, [2] only presents group-level annotator agreement. The level of inter-annotator agreement at a post level, however, is not included in the dataset. Further details regarding the levels of post uncertainty, hence, understanding the true distribution of aleatoric uncertainty would likely benefit the model's performance. Given this, we assume the distribution of agreement, or aleatoric noise within the data set to not be constant. Future work may explore further probabilistic techniques which account for the varying levels of heteroscedastic aleatoric uncertainty throughout the distribution.

7. Conclusion

The current paper presents Deep Bayesian Label Smoothing, a novel method for generating soft labels for measurements with high levels of uncertainty. We show through empirical experiments our proposed method improves model performance when compared to hard labels and uniform label smoothing. Additionally, the current work provides a text classification system for the assessment of suicide behaviours based on a five-label classification scheme. The incorporation of Bayesian uncertainty techniques into the proposed works greatly improves upon the existing state-of-the-art model on the same dataset. Furthermore, our model provides one of the first adaptions of MC Dropout on a medical text classification task with the majority of work to date focusing on computer vision and image segmentation tasks. Future work may benefit by incorporating word representation models equipped to understand language context such as GloVe or BERT to better classify hard-to-detect classes, such as whether a post is indicative of a suicide attempt behaviour. This is especially true in areas where annotator ambiguity is high.

Acknowledgment

This work is partially supported by a grant from Cannon Institute and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. LU HLCA/E/301/23). We gratefully

acknowledge the support from Belmont Private Hospital team members, especially Ms Mary Williams (CEO), Rachel Stark (Area Manager), Dr Mark Spelman (Psychiatrist), Dr Sean Gills (Psychiatrist), and Dr Tom Moore (Psychiatrist). Without their kind support, this work wouldn't be possible.

References

[1] S. J. Fitzpatrick, T. Handley, N. Powell, D. Read, K. J. Inder, D. Perkins, B. K. Brew, Suicide in rural australia: A retrospective study of mental health problems, health-seeking and service utilisation, PLOS ONE 16 (7) (2021) e0245271. doi:10.1371/journal.pone.0245271.

[2] M. Gaur, A. Alambo, J. P. Sain, U. Kursuncu, K. Thirunarayan, R. Kavuluru, A. Sheth, R. Welton, J. Pathak, Knowledge-aware assessment of severity of suicide risk for early intervention, in: The World Wide Web Conference on - WWW '19, ACM Press, 2019. doi:10.1145/3308558.3313698.

[3] N. Akhther, P. Sopory, Seeking and sharing mental health information on social media during COVID-19: Role of depression and anxiety, peer support, and health benefits, Journal of Technology in Behavioral Science (jan 2022). doi:10.1007/s41347-021-00239-x.

[4] A. T. Drysdale, L. Grosenick, J. Downar, K. Dunlop, F. Mansouri, Y. Meng, R. N. Fetcho, B. Zebley, D. J. Oathes, A. Etkin, A. F. Schatzberg, K. Sudheimer, J. Keller, H. S. Mayberg, F. M. Gunning, G. S. Alexopoulos, M. D. Fox, A. Pascual-Leone, H. U. Voss, B. Casey, M. J. Dubin, C. Liston, Erratum: Resting-state connectivity biomarkers define neurophysiological subtypes of depression, Nature Medicine 23 (2) (2017) 264–264. doi:10.1038/nm0217-264d.

[5] W. Yassin, H. Nakatani, Y. Zhu, M. Kojima, K. Owada, H. Kuwabara, W. Gonoi, Y. Aoki, H. Takao, T. Natsubori, N. Iwashiro, K. Kasai, Y. Kano, O. Abe, H. Yamasue, S. Koike, Machine-learning classification using neuroimaging data in schizophrenia, autism, ultra-high risk and first-episode psychosis, Translational Psychiatry 10 (1) (aug 2020). doi:10.1038/s41398-020-00965-5.

[6] A. Interian, M. Chesin, A. Kline, R. Miller, L. S. Hill, M. Latorre, A. Shcherbakov, A. King, B. Stanley, Use of the columbia-suicide severity rating scale (c-SSRS) to classify suicidal behaviors, Archives of Suicide Research 22 (2) (2017) 278–294. doi:10.1080/13811118.2017.1334610.

[7] D. Dubois, H. Prade, An introduction to fuzzy systems, Clinica Chimica Acta 270 (1) (1998) 3–29. doi:10.1016/s0009-8981(97)00232-5.

[8] Z. Li, X. Li, H. Xie, F. L. Wang, M. Leng, Q. Li, X. Tao, A novel dropout mechanism with label extension schema toward text emotion classification, Information Processing & Management 60 (2) (2023) 103173. doi:10.1016/j.ipm.2022.103173.

[9] C.-B. Zhang, P.-T. Jiang, Q. Hou, Y. Wei, Q. Han, Z. Li, M.-M. Cheng, Delving deep into label smoothing, IEEE Transactions on Image Processing 30 (2021) 5984–5996. doi:10.1109/tip.2021.3089942.

[10] E. Begoli, T. Bhattacharya, D. Kusnezov, The need for uncertainty quantification in machine-assisted medical decision making, Nature Machine Intelligence 1 (1) (2019) 20–23. doi:10.1038/s42256-018-0004-1.

[11] A. A. Abdullah, M. M. Hassan, Y. T. Mustafa, A review on bayesian deep learning in healthcare: Applications and challenges, IEEE Access 10 (2022) 36538–36562. doi:10.1109/ACCESS.2022.3163384.

[12] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, S. Nahavandi, A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information Fusion 76 (2021) 243–297. doi:10.1016/j.inffus.2021.05.008.

[13] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, Machine Learning 110 (3) (2021) 457–506. doi:10.1007/s10994-021-05946-3.

[14] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 5580–5590.

[15] J. Liu, J. Zhang, N. Barnes, Modeling aleatoric uncertainty for camouflaged object detection, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2022. doi:10.1109/wacv51458.2022.00267.

[16] A. Haque, V. Reddi, T. Giallanza, Deep learning for suicide and depression identification with unsupervised label correction (2021). doi:10.48550/ARXIV.2102.09427.

[17] B. Song, S. Sunny, S. Li, K. Gurushanth, P. Mendonca, N. Mukhia, S. Patrick, S. Gurudath, S. Raghavan, I. Tsusennaro, S. T. Leivon, T. Kolur, V. Shetty, V. R. Bushan, R. Ramesh, T. Peterson, V. Pillai, P. Wilder-Smith, A. Sigamani, A. Suresh, moni Abraham Kuriakose, P. Birur, R. Liang, Bayesian deep learning for reliable oral cancer image classification, Biomedical Optics Express 12 (10) (2021) 6422. doi:10.1364/boe.432365.

[18] J. Caldeira, B. Nord, Deeply uncertain: Comparing methods of uncertainty quantification in deep learning algorithms, ArXiv abs/2004.10710 (2020).

[19] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning (Jun. 2015). arXiv:1506.02142.

[20] M. Gour, S. Jain, Uncertainty-aware convolutional neural network for COVID-19 x-ray images classification, Computers in Biology and Medicine 140 (2022) 105047. doi:10.1016/j.compbiomed.2021.105047.

[21] J. Wu, X. Liu, Y. Liao, Difficulty-aware brain lesion segmentation from MRI scans, Neural Processing Letters 54 (3) (2022) 1961–1975. doi:10.1007/s11063-021-10714-4.

[22] H. L. Gjestang, S. A. Hicks, V. Thambawita, P. Halvorsen, M. A. Riegler, A self-learning teacher-student framework for gastrointestinal image classification, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), 2021, pp. 539–544. doi:10.1109/CBMS52027.2021.00087.

[23] F. Chollet, et al., Keras, https://keras.io (2015).

[24] S. Raaijmakers, Deep Learning for Natural Language Processing, Manning Publications, 2022.
URL https://www.ebook.de/de/product/36217598/stephan_raaijmakers_deep_learning_for_natural_language_processing.html

[25] K. O'Shea, R. Nash, An introduction to convolutional neural networks (Nov. 2015). arXiv:1511.08458.

[26] F. Chollet, Deep Learning with Python, Second Edition, MANNING PUBN, 2021.
URL https://www.ebook.de/de/product/40499536/francois_chollet_deep_learning_with_python_second_edition.html

[27] M. Huang, H. Xie, Y. Rao, Y. Liu, L. K. M. Poon, F. L. Wang, Lexicon-based sentiment convolutional neural networks for online review analysis, IEEE Transactions on Affective Computing 13 (3) (2022) 1337–1348. doi:10.1109/taffc.2020.2997769.

[28] Y. Goldberg, A primer on neural network models for natural language processing (Oct. 2015). arXiv:1510.00726.

[29] A. Jacovi, O. S. Shalom, Y. Goldberg, Understanding convolutional neural networks for text classification (Sep. 2018). arXiv:1809.08037.

[30] Z. Shen, Z. Liu, D. Xu, Z. Chen, K.-T. Cheng, M. Savvides, Is label smoothing truly incompatible with knowledge distillation: An empirical study (2021). doi:10.48550/ARXIV.2104.00676.

[31] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain., Psychological Review 65 (6) (1958) 386–408. doi:10.1037/h0042519.

[32] O. Duerr, B. Sick, E. Murina, Probabilistic Deep Learning: With Python, Keras and Tensorflow Probability, MANNING PUBN, 2020.
URL https://www.ebook.de/de/product/37932395/oliver_duerr_beate_sick_elvis_murina_probabilistic_deep_learning_with_python_keras_and_tensorflow_probability.html

[33] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview (Aug. 2020). arXiv:2008.05756.

[34] K. Takahashi, K. Yamamoto, A. Kuchiba, T. Koyama, Confidence interval for micro-averaged f1 and macro-averaged f1 scores, Applied Intelligence 52 (5) (2021) 4961–4972. doi:10.1007/s10489-021-02635-5.

[35] U. Naseem, I. Razzak, S. K. Khan, M. Prasad, A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models, ACM Transactions on Asian and Low-Resource Language Information Processing 20 (5) (2021) 1–35. doi:10.1145/3434237.

5.3 Links and implications

This work presents a novel method for representing uncertainty in fuzzy systems. Originally this work was intended as a stepping stone to applying NLP techniques to EHR, however, the time to digitize the records of the industry partner exceded the time of this thesis. This work then provides an example of a system which could be utilised to detect psychological distress from text. Future work may seek to utilise these kind of detection systems from patient notes with existing rTMS treatment response prediction systems to evaluate the ability for clinical notes to contribute to the prediction of treatment response.

CHAPTER 6: DISCUSSIONS AND CONCLUSIONS

This thesis sets out to explore the ways in which AI has the potential to disrupt traditional mental health care. In doing so, this thesis provides evidence for the ways AI can be used to improve outcomes for people suffering mental health distress. Depression is the mental health condition which places the largest burden on Australia's health care system (Kasturi et al. 2023). Given the prevalence of depression in Australia this thesis has largely focused on how AI is poised to disrupt the treatment, diagnosis and detection of depression.

To achieve these stated aims this thesis addressed the following research questions stated in Chapter :

1. **RQ1: How can artificial intelligence be used to disrupt existing models of mental health treatment decision making and care?**
2. **RQ2: How can artificial intelligence methods be used to facilitate personalised psychiatry ?**
3. **RQ3: What are the implications and considerations required for the use of artificial intelligence as decision support in psychiatry?**

Through empirical experiments this thesis has sought to address these research questions. The results detailed across this research have shown, for the first time, AI can reliably predict which patients will respond to rTMS treatment. These groundbreaking results offer the potential for a future where patients are screened using neuroimaging prior to rTMS treatment to assess their suitability. Furthermore, the research contained within

this thesis has demonstrated that increasing the diversity of training data with synthetic examples can enhance the fairness of AI models. It also presents a novel method for handling the uncertainty of ground truth labels associated with subjective mental health decisions. As a result of these works, this thesis has attempted to dissect hype from reality, providing a critical overview of the current state-of-the-art of AI and its applications to psychiatric practice. Additionally, this thesis has explored the limitations that need to be addressed before the widespread ethical adoption of AI for improved patient outcomes can become a reality.

The empirical experiments detailed throughout this thesis made use of a variety of datasets collected across multiple countries. In the Australian context, data collected at Belmont Private Hospital, Brisbane was used across multiple studies. Additionally, open source data from Hopman et al. (2021) provided additional resources from which to complete these experiments. This data along with the research questions outlined above allow this thesis to makes the following contributions:

- **An overview of the current state of the art AI use in Psychiatry by exploring the ways in which AI is supporting the detection, diagnosis and treatment of depression**

  This thesis contributes a survey which outlines the many ways AI is directly impacting the detection, diagnosis and treatment of depression. AI-driven tools are shown to detect depression through analysing speech, text and facial expressions. Unsupervised AI methods are being shown to identify new disease categories and some methods are being used to better target treatments

- **A detailed comparison of the data required to predict treatment outcomes using AI, including the identification of candidate biomarkers indicative of treatment response**

  Throughout this thesis the data requirements to support the personalisation of psychiatry are explored. This thesis shows for treatment response prediction of rTMS neuroimaging data is superior to demographic and self reported psychological data. By combining DL with fMRI connectivity features the outcomes to rTMS treatment can be reliably predicted for most patients. By incorporating XAI this

work identifies candidate biomarkers indicative of treatment response

- **A novel framework for data augmentation of depression datasets to enhance the diversity of small datasets with underrepresented values**

  In exploring the implications of AI use in psychiatry it is important to consider algorithms' fairness. This includes ensuring methods are trained on large and diverse training sets. However, these large datasets are often unavailable. For these situations, this thesis demonstrates through the use of a novel data augmentation framework, DE-CGAN, synthetic examples can be used to increase the diversity of training datasets. These experiments demonstrate that increasing the diversity of training data can enhance model performance across all examples, underscoring the importance of representative training data.

- **A methodology for capturing the subjective nature of expert mental health judgements** One of the most significant challenges in mental healthcare is the lack of objective disease markers. AI is gradually uncovering potential candidate biomarkers. For example, research contained in this thesis shows functional connectivity between the Subgenual Anterior Cingulate Cortex and Central Opercular Cortex as a key determinant of treatment response to rTMS treatment in depression patients. However, in many circumstances quantitative data is unavailable. In these circumstances, patients may be assessed based on self reported depression severity or as judged by human expert. Despite this, the complexity of mental health conditions means human experts can sometimes disagree. To improve the performance of DL models trained in these contexts this thesis presents a novel method of non-uniform label smoothing to capture this uncertainty.

As we discuss the broader context of the research works contained within this thesis. We recall, Chapter 1.6 provides an overview of the current state of the art of AI use in psychiatry. This chapter details opportunities for the use of DL to predict treatment outcomes to rTMS. Furthermore, the chapter explores the challenges of data access in psychiatry, given the strict legal and privacy concerns regarding data in psychiatry. Additionally, the survey article contends one of the primary challenges in psychiatry that AI is well positioned to address is the lack of objective biomarkers of disease in depression. DL methods are well suited to the task of identifying distinct patient groups where these

distinct groups may respond differently to the same treatment, and hence the concept of personalised psychiatry.

6.1 Treatment Response Prediction

In Chapter 2.3 this thesis has demonstrated that under certain circumstances DL models can predict treatment outcomes to rTMS. In *Identifying predictive biomarkers for repetitive transcranial magnetic stimulation response in depression patients with explainability* it is shown that self reported and demographic data is insufficient to predict treatment outcomes above our baseline model. Where the selected baseline model is a simple rule-based algorithm which uses changes in early treatment depression severity. Therefore, in these works, we demonstrate, consistent with existing research, that an imaging technique is necessary to realise the potential of AI-powered personalised psychiatry.

This thesis has shown that DL models can reliably predict the outcome of rTMS treatment prior to its commencement using fMRI connectivity features. This finding paves the way for a future where treatment prescriptions can be informed by AI algorithms. Reducing the burden placed on patients of exploring ineffective treatments. As it becomes inevitable AI will be part of the future of practice in psychiatry (see Doraiswamy et al. 2020) further work is required to ensure results of empirical experiments are replicated in clinical settings. Additionally, more work is required to understand how these algorithms will co-exist with practitioners.

Squires et al. (2023) illustrates both the exciting potential of AI in psychiatry and the current shortcomings, highlighting the future work needed for AI to become an integral part of everyday clinical practice. For these types of technologies to be become widely adopted models must be trained on robust and diverse datasets (Furriel et al. 2024). The absence of these datasets is something explored in (Squires, Tao, Elangovan, Gururajan, Xie, Zhou, Li & Acharya 2024). Furthermore, external validation of DL models is required to ensure model performance across all circumstances. While little regulation exists in this space, as AI becomes more relevant it is essential these regulations become more widespread. At this stage the AI community has worked towards self regulation of best

practice (Miró Catalina et al. 2024). Similarly, further work is required to better understand how AI and practitioners will work together, including processes of responsibility. Ueda et al. (2023) contend clear processes of responsibility must be developed for AI to be deployed in healthcare.

> "Physicians should be responsible for verifying AI-generated diagnoses and integrating them into the clinical decision-making process. This may involve critically evaluating the AI outputs, considering them along with other relevant clinical information, and making informed decisions regarding patient care. Conversely, AI developers have a responsibility to ensure the accuracy, reliability, and fairness of their algorithms. This includes addressing biases and continuously improving the algorithms based on feedback from the clinical community." (Ueda et al. 2023, p.7)

Reducing and mitigating any bias is a consistent concern expressed throughout this thesis and the literature more broadly. As regulation has failed to keep pace with developments in the field at present the field is forced to regulate itself. External validation of models has become the gold standard of testing the efficacy of AI algorithms (Miró Catalina et al. 2024). In lieu of this, AI professionals must include detailed information regarding how a model is trained and the quality of the training data to allow clinicians to make the most informed decisions. Sendak et al. (2020) propose "model facts" as an approach for documenting model limitations. This methodology includes the documentation of any ML/DL model including advice on interpreting model outputs and warnings. If AI is to fulfill in its promise then AI-human interaction must be at the forefront of model design. This framework is demonstrated in (Squires et al. 2023) where it is documented the model performs poorly on some regions of the feature space.

6.2 Data Augmentation for robust and fair AI in Psychiatry

Ueda et al. (2023) emphasise fairness as one of the core principals in AI ethics, they assert data bias can occur when data is not representative. In Chapter 3.3 this thesis presents *DE-CGAN: Boosting rTMS Treatment Prediction with Diversity Enhancing Conditional*

Generative Adversarial Networks a novel framework for enhancing the diversity of training datasets. This process involves generating conditional examples of regions of the training dataset which have been previously difficult to classify.

Squires, Tao, Elangovan, Gururajan, Xie, Zhou, Li & Acharya (2024) shows improving the diversity of DL training datasets boosts model performance. These works show utilising a novel data augmentation methodology which first identifies under-represented samples before generating synthetic examples in the latent space improves model accuracy across all samples.

In many cases this access to diverse multi-site data is difficult or not possible (Yang et al. 2022). The sensitive nature of collected data in many instances prevents the sharing of data across multiple sites. Future works which focus on multiple sites and multiple imaging machines are required to move closer to the deployment of developed models. This thesis then raises two questions: *if we can reliably predict the outcome of treatment who is responsible when the AI makes mistakes?* and *what level of model performance is deemed necessary for these technologies to be deployed in a clinical setting?*. Beyond the use of generative AI to boost the diversity of the data feature space work is needed to address other barriers to fairness of AI in psychiatry.

Large and diverse training datasets are one way to enhance the probability of model generalisation. However, further considerations are required to ensure AI models that are designed to improve patient outcomes do not perpetuate or exacerbate existing social biases and inequalities (Timmons et al. 2022).

> "The proliferation of available data and technological methods for extracting insights from them might give the impression that such technologies are unbiased. However, just as humans' past experiences and personal values influence their decision-making in biased ways, algorithms built by people on data collected by people are also subject to bias." (Timmons et al. 2022, p.1064)

For example, Şahin et al. (2023) showed in their surveyed articles general model predictions were biased against individuals with lower educational attainment. Şahin et al.

(2023) contend any future work should include fairness assessments. A call which is echoed in Timmons et al. (2022). Altong with the ethical considerations required as AI becomes more prevalent in healthcare and psychiatry are the legal and societal implications. Naik et al. (2022) argues further work is required to answer the question *who is responsible for AI decision making?* Where doctors are directly responsible for their clinical decisions, at present data scientists are not legally accountable for AIs actions.

This section makes clear ensuring diverse training datasets is one area of reducing AI bias. However, the implications of the greater use of AI are much more widespread. Further work is required to ensure AI algorithms are designed to consider a wider range of ethical considerations. In the future, AI researchers may be encouraged to include fairness metrics as standard practice in research outputs.

6.3 Inter-rater uncertainty and ground truth labels

Timmons et al. (2022) contends label-bias is a bias that occurs when labels used to train AI algorithms are incomplete. Label-bias could be due to mistakes, differences in experience level or variations in how different clinicians interpret the data.

Any inter-rater disagreement has the potential to both reduce model performance and contribute to any bias. Squires, Tao, Elangovan, Acharya, Gururajan, Xie & Zhou (2024) shows training a model using a novel, non-uniform label smoothing technique to account for label uncertainty improves model performance on the original binary labels. This is consistent with other research which has demonstrated other methods of handling inter-expert variability, such as label fusion have improved model calibration (Lemay et al. 2022).

6.4 Limitations and Assumptions

This thesis has outlined the ways AI can potentially support improved outcomes for people suffering mental health conditions. The research within this thesis has demonstrated above state-of-the-art performance on a number of DL tasks related to the treatment and

detection of mental health conditions.

In completing this research it it assumed beyond the standard data cleaning and prepossessing data science techniques the data is correct. Any issues in data collection beyond what is expected in a standard data science project may impact the findings encompassed within this thesis. Additionally, we assume self-reported data is accurate to the extent that self-reported data can be. This research was completed against the context of the COVID-19 pandemic which adds an additional extraneous variable that may have impacted empirical experiments in an unknown way. However, this thesis assumes the data used is representative of the broader population and the research collected can be generalised to the broader population.

The results reported in this thesis are based on empirical DL experiments, which are inherently stochastic. Every effort is made to ensure their correctness, such as employing robust validation measures like extensive Monte Carlo simulations. Despite this, some variability in results during reproduction could be expected due to the probabilistic nature of AI experiments. Nonetheless, the use of robust internal validation aims to mitigate this limitation.

At the time of publishing, (Squires et al. 2023) was the first work to utilise DL to predict treatment outcomes to rTMS before treatment began. The standing as one of the only works to utilise DL methods to predict treatment outcomes was recently confirmed in Jin et al. (2024). In their review, Jin et al. (2024) identified only shallow learning methods used in existing works. Squires et al. (2023) demonstrates, to our knowledge, for the first time DL methods can predict treatment outcomes to rTMS using thorough internal validation. The strength of adopting a DL approach compared to existing methods is the known superiority of DL methods to generalise to unseen data and to identify complex non-linear relationships.

The implications of AI algorithms demonstrating performance equivalent to human clinicians (Furriel et al. 2024) is the need to ensure such performance is consistent across contexts. This requires diverse datasets. Squires et al. (2023) showed for the best outcomes quantitative imaging data is superior to qualitative, demographic or self report data. Hence, if AI is to become widely utilised in a clinical setting pre-treatment neu-

roimaging will need to become more widespread. Chen et al. (2023) note the trend towards imaging to support the diagnosis of mental health conditions. It is conceivable to envision a future where combining AI with imaging supersedes existing qualitative measures, making imaging the predominant method for diagnosing mental health conditions. However, greater access to neuroimaging is an open challenge that will require interdisciplinary intervention from a variety of stakeholders.

As an extension to the lack of access to neuroimaging in the treatment pathway. The lack of large and diverse datasets greatly inhibits research by both AI and mental health researchers alike. Greater collaboration between research entities may facilitate the requisite datasets. However, in the mean time, the use of generative AI to generate more robust datasets of synthetic data is an important area of research. Recently, Savage (2023) argues synthetic data may be superior to standard data. This trend is being observed across disciplines as some experts expect the availability of data to slow rapidly which may also slow AI research progress (Villalobos et al. 2022).

6.5 Conclusions

The research contained in this thesis sought to explore the ways AI can contribute to improved mental healthcare. As part of this academic journey, this thesis has demonstrated a novel DNN framework combined with XAI to predict rTMS treatment outcomes above the existing state-of-the-art, while also identifying potential candidate biomarkers indicative of treatment response. Additionally, this thesis has shown that boosting the diversity of training datasets with synthetic examples can improve model performance.

Furthermore, this research sought to explore the essential implications and considerations of future AI use in clinical psychiatry. The exploration of these significant considerations has motivated important research into methods for reducing AI bias in decision support systems. This thesis presents methods for reducing label bias, such as the novel non-uniform label smoothing method proposed here, which improves the ability of a DNN to identify those at risk of suicide.

From the body of work constructed in this thesis, it is clear that for AI to fulfill its most

optimistic potential in psychiatry, much further work is required. This thesis has demonstrated much of the exciting hype and promise being shown throughout the research community. However, additional work that explores its potential use cases and evaluates any limitations is essential. This thesis shows that, in the best-case scenario, AI/DL can be used to personalise psychiatric care. However, further work is required to mitigate potential biases and address the ways these systems can be integrated into current workflows. Furthermore, this thesis demonstrates one potential strategy for addressing data bias with diversity-enhancing data augmentation. While these works demonstrate potential opportunities through rigorous internal validation, AI in psychiatry will have difficulty progressing beyond research until such performance can be validated on external datasets in studies across multiple sites.

6.6 Future Work

Large multi-site testing should be one of the highest priorities for deploying treatment response prediction. Until models are validated across multiple sites, their high performance will remain limited to contained datasets, restricting their broader application and effectiveness. Without large multi-site experiments, even work incorporating the use of regularisation techniques and best practice internal validation, will leave questions about AI's ability to generalise to unseen clinical data unanswered. This will hinder AI from fulfilling the most optimistic expectations of personalized psychiatry.

Similarly, further work around the interactions between clinicians and AI-informed decision support systems is a critical step in deploying the systems discussed throughout this thesis. Equipping AI with uncertainty measures and prediction confidence is one way to improve the interactions between decision support systems and clinicians.

REFERENCES

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V. & Nahavandi, S. (2021), 'A review of uncertainty quantification in deep learning: Techniques, applications and challenges', *Information Fusion* **76**, 243–297.

Akhther, N. & Sopory, P. (2022), 'Seeking and sharing mental health information on social media during COVID-19: Role of depression and anxiety, peer support, and health benefits', *Journal of Technology in Behavioral Science* .

Begoli, E., Bhattacharya, T. & Kusnezov, D. (2019), 'The need for uncertainty quantification in machine-assisted medical decision making', *Nature Machine Intelligence* **1**(1), 20–23.

Cannarsa, M. (2021), 'Ethics guidelines for trustworthy ai', *The Cambridge handbook of lawyering in the digital age* pp. 283–97.

Chen, Q., Zhong, Y., Jin, C., Zhou, R., Dou, X., Yu, C., Wang, J., Xu, H., Tian, M. & Zhang, H. (2023), 'Nuclear psychiatric imaging: the trend of precise diagnosis for mental disorders', *European Journal of Nuclear Medicine and Molecular Imaging* **51**(4), 1002–1006.

Christensen, C. M., McDonald, R., Altman, E. J. & Palmer, J. E. (2018), 'Disruptive innovation: An intellectual history and directions for future research', *Journal of Management Studies* **55**(7), 1043–1078.

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E. & Herrera, F. (2023), 'Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation', *Information Fusion* **99**, 101896.

Fitzgerald, P. B., Hoy, K. E., Reynolds, J., Singh, A., Gunewardene, R., Slack, C., Ibrahim, S. & Daskalakis, Z. J. (2020), 'A pragmatic randomized controlled trial exploring the relationship between pulse number and response to repetitive transcranial magnetic stimulation treatment in depression', *Brain Stimulation* **13**(1), 145–152.

Fitzpatrick, S. J., Handley, T., Powell, N., Read, D., Inder, K. J., Perkins, D. & Brew, B. K. (2021), 'Suicide in rural australia: A retrospective study of mental health problems, health-seeking and service utilisation', *PLOS ONE* **16**(7), e0245271.

Furriel, B. C. R. S., Oliveira, B. D., Prôa, R., Paiva, J. Q., Loureiro, R. M., Calixto, W. P., Reis, M. R. C. & Giavina-Bianchi, M. (2024), 'Artificial intelligence for skin cancer detection and classification for clinical environment: a systematic review', *Frontiers in Medicine* **10**.

Gaur, M., Alambo, A., Sain, J. P., Kursuncu, U., Thirunarayan, K., Kavuluru, R., Sheth, A., Welton, R. & Pathak, J. (2019), Knowledge-aware assessment of severity of suicide risk for early intervention, *in* 'The World Wide Web Conference on - WWW '19', ACM Press.

George, M. S. & Taylor, J. J. (2014), Theoretical basis for transcranial magnetic stimulation, *in* 'A Clinical Guide to Transcranial Magnetic Stimulation', Oxford University Press.

Girasa, R. (2020), *AI as a Disruptive Technology*, Springer International Publishing, pp. 3–21.

Guo, J.-W., Kimmel, J. & Linder, L. A. (2024), 'Text analysis of suicide risk in adolescents and young adults', *Journal of the American Psychiatric Nurses Association* **30**(1), 169–173. PMID: 35130778.
URL: *https://doi.org/10.1177/10783903221077292*

Hannah, L. A., Walsh, C. M., Jopling, L., Perez, J., Cardinal, R. N. & Cameron, R. A. (2023), 'Economic evaluation of interventions for treatment-resistant depression: A systematic review', *Frontiers in Psychiatry* **14**.

Hopman, H., Chan, S., Chu, W., Lu, H., Tse, C.-Y., Chau, S., Lam, L., Mak, A. & Neggers, S. (2021), 'Personalized prediction of transcranial magnetic stimulation

clinical response in patients with treatment-refractory depression using neuroimaging biomarkers and machine learning', *Journal of Affective Disorders* **290**, 261–271.

Interaction Design Foundation - IxDF (2016), 'What is artificial intelligence (ai)?', `https://www.interaction-design.org/literature/topics/ai`. Retrieved April 1, 2024.

Interian, A., Chesin, M., Kline, A., Miller, R., Hill, L. S., Latorre, M., Shcherbakov, A., King, A. & Stanley, B. (2017), 'Use of the columbia-suicide severity rating scale (c-SSRS) to classify suicidal behaviors', *Archives of Suicide Research* **22**(2), 278–294.

Jin, M. X., Qin, P. P., Xia, A. W. L., Kan, R. L. D., Zhang, B. B. B., Tang, A. H. P., Li, A. S. M., Lin, T. T. Z., Giron, C. G., Pei, J. J. & Kranz, G. S. (2024), 'Neurophysiological and neuroimaging markers of repetitive transcranial magnetic stimulation treatment response in major depressive disorder: A systematic review and meta-analysis of predictive modeling studies', *Neuroscience Biobehavioral Reviews* **162**, 105695.

Johnston, K. M., Powell, L. C., Anderson, I. M., Szabo, S. & Cline, S. (2019), 'The burden of treatment-resistant depression: A systematic review of the economic and quality of life literature', *Journal of Affective Disorders* **242**, 195–210.

Kasturi, S., Oguoma, V. M., Grant, J. B., Niyonsenga, T. & Mohanty, I. (2023), 'Prevalence rates of depression and anxiety among young rural and urban australians: A systematic review and meta-analysis', *International Journal of Environmental Research and Public Health* **20**(1), 800.

Lemay, A., Gros, C., Karthik, E. N. & Cohen-Adad, J. (2022), 'Label fusion and training methods for reliable representation of inter-rater uncertainty'.

Malhi, G. S., Bell, E., Bassett, D., Boyce, P., Bryant, R., Hazell, P., Hopwood, M., Lyndon, B., Mulder, R., Porter, R., Singh, A. B. & Murray, G. (2020), 'The 2020 royal australian and new zealand college of psychiatrists clinical practice guidelines for mood disorders', *Australian amp; New Zealand Journal of Psychiatry* **55**(1), 7–117.

Miró Catalina, Q., Vidal-Alaball, J., Fuster-Casanovas, A., Escalé-Besa, A., Ruiz Comellas, A. & Solé-Casals, J. (2024), 'Real-world testing of an artificial intelligence al-

gorithm for the analysis of chest x-rays in primary care settings', *Scientific Reports* **14**(1).

Naik, N., Hameed, B. M. Z., Shetty, D. K., Swain, D., Shah, M., Paul, R., Aggarwal, K., Ibrahim, S., Patil, V., Smriti, K., Shetty, S., Rai, B. P., Chlosta, P. & Somani, B. K. (2022), 'Legal and ethical consideration in artificial intelligence in healthcare: Who takes responsibility?', *Frontiers in Surgery* **9**.

Naseem, U., Thapa, S., Zhang, Q., Hu, L., Rashid, J. & Nasim, M. (2023), 'Incorporating historical information by disentangling hidden representations for mental health surveillance on social media', *Social Network Analysis and Mining* **14**(1).

OpenAI (2023), 'Gpt-4 technical report'.

Păvăloaia, V.-D. & Necula, S.-C. (2023), 'Artificial intelligence as a disruptive technology—a systematic literature review', *Electronics* **12**(5), 1102.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. et al. (2018), 'Improving language understanding by generative pre-training'.

Razza, L. B., Moffa, A. H., Moreno, M. L., Carvalho, A. F., Padberg, F., Fregni, F. & Brunoni, A. R. (2018), 'A systematic review and meta-analysis on placebo response to repetitive transcranial magnetic stimulation for depression trials', *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **81**, 105–113.

right treatment for each patient:unlocking the potential of personalized psychiatry', T. (2023), *Nature Mental Health* **1**(9), 607–608.

Rosenblatt, F. (1958), 'The perceptron: A probabilistic model for information storage and organization in the brain.', *Psychological Review* **65**(6), 386–408.

Savage, N. (2023), 'Synthetic data could be better than real data', *Nature (London)* .

Sendak, M. P., Gao, M., Brajer, N. & Balu, S. (2020), 'Presenting machine learning model information to clinical end users with model facts labels', *npj Digital Medicine* **3**(1).

Si, S. & Chen, H. (2020), 'A literature review of disruptive innovation: What it is, how it works and where it goes', *Journal of Engineering and Technology Management* **56**, 101568.

Squires, M., Tao, X., Elangovan, S., Acharya, U. R., Gururajan, R., Xie, H. & Zhou, X. (2024), 'Enhancing suicide risk detection on social media through semi-supervised deep label smoothing'.

Squires, M., Tao, X., Elangovan, S., Gururajan, R., Xie, H., Zhou, X., Li, Y. & Acharya, U. R. (2024), 'De-cgan: Boosting rtms treatment prediction with diversity enhancing conditional generative adversarial networks'.

Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X., Li, Y. & Acharya, U. R. (2023), 'Identifying predictive biomarkers for repetitive transcranial magnetic stimulation response in depression patients with explainability', *Computer Methods and Programs in Biomedicine* **242**, 107771.

Ståhl, N., Falkman, G., Karlsson, A. & Mathiason, G. (2020), Evaluation of uncertainty quantification in deep learning, *in* 'Information Processing and Management of Uncertainty in Knowledge-Based Systems', Springer International Publishing, pp. 556–568.

Stefanicka-Wojtas, D. & Kurpas, D. (2023), 'Personalised medicine—implementation to the healthcare system in europe (focus group discussions)', *Journal of Personalized Medicine* **13**(3), 380.

Strange, M. (2024), 'Three different types of ai hype in healthcare', *AI and Ethics* .

Suwinski, P., Ong, C., Ling, M. H. T., Poh, Y. M., Khan, A. M. & Ong, H. S. (2019), 'Advancing personalized medicine through the application of whole exome sequencing and big data analytics', *Frontiers in Genetics* **10**.

Thornton, N. L. R., Black, W., Bognar, A., Dagge, D., Gitau, T., Hua, B., Joks, G., King, J., Lord, A., Scott, E. M., Callander, J. S., Ting, S. & Liu, D. (2023), 'Establishing an esketamine clinic in australia: Practical recommendations and clinical guidance from an expert panel', *Asia-Pacific Psychiatry* **15**(4).

Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., Comer, J. S., Brewer, L. C., Frazier, S. L. & Chaspari, T. (2022), 'A call to action on assessing and mitigating bias in artificial intelligence applications for mental health', *Perspectives on Psychological Science* **18**(5), 1062–1096.

Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T. & Naganawa, S. (2023), 'Fairness of artificial intelligence in healthcare: review and recommendations', *Japanese Journal of Radiology* **42**(1), 3–15.

van de Poel, I., Hopster, J., Löhr, G., Ziliotti, E., Buijsman, S. & Brey, P. (2023), *1: Introduction*, Open Book Publishers, pp. 11–32.

Vicente, A. M., Ballensiefen, W. & Jönsson, J.-I. (2020), 'How personalised medicine will transform healthcare by 2030: the icpermed vision', *Journal of Translational Medicine* **18**(1).

Villalobos, P., Sevilla, J., Heim, L., Besiroglu, T., Hobbhahn, M. & Ho, A. (2022), 'Will we run out of data? an analysis of the limits of scaling datasets in machine learning'.

Xu, Z., Biswas, B., Li, L. & Amzal, B. (2023), 'Ai/ml in precision medicine: A look beyond the hype', *Therapeutic Innovation amp; Regulatory Science* **57**(5), 957–962.

Yam, P. (2024), 'Brain stimulation poised to move from last resort to frontline treatment', *Proceedings of the National Academy of Sciences* **121**(7).

Yang, J., Soltan, A. A. S. & Clifton, D. A. (2022), 'Machine learning generalizability across healthcare settings: insights from multi-site covid-19 screening', *npj Digital Medicine* **5**(1).

Şahin, D., Kambeitz-Ilankovic, L., Wood, S., Dwyer, D., Upthegrove, R., Salokangas, R., Borgwardt, S., Brambilla, P., Meisenzahl, E., Ruhrmann, S., Schultze-Lutter, F., Lencer, R., Bertolino, A., Pantelis, C., Koutsouleris, N. & Kambeitz, J. (2023), 'Algorithmic fairness in precision psychiatry: analysis of prediction models in individuals at clinical high risk for psychosis', *The British Journal of Psychiatry* **224**(2), 55–65.