International Journal on Artificial Intelligence Tools

Mental Health Diagnosis: A Case for Explainable Artificial Intelligence

Grigoris Antoniou, Emmanuel Papadakis, George Baryannis
School of Computing and Engineering, University of Huddersfield
Queensgate, Huddersfield, HD1 3DH, UK
{g.antoniou,e.papadakis,g.bargiannis}@hud.ac.uk

Mental illnesses are becoming increasingly prevalent, in turn leading to an increased interest in exploring artificial intelligence (AI) solutions to **facilitate and** enhance healthcare processes ranging from diagnosis to monitoring and treatment. In contrast to application areas where black box systems may be acceptable, explainability in healthcare applications is essential, especially in the case of diagnosing complex and **sensitive** mental health issues. In this paper, we first summarise recent developments in AI research for mental health, followed by an overview of approaches to explainable AI and their potential benefits in healthcare settings. We then present a recent case study of applying explainable AI for ADHD diagnosis which is used as a basis to identify challenges in realising explainable AI solutions for mental health diagnosis and potential future research directions to address these challenges.

Keywords: Explainable Artificial Intelligence, Mental Health, Healthcare

1. AI for mental health

Mental health illnesses are among the most common health conditions in the United States, where more than 50% of the population will be diagnosed with a mental illness or disorder at some point in their lifetime[1]. In the United Kingdom, mental health problems account for 23% of the burden of disease[2]. According to the Centers for Disease Control and Prevention[3], mental illness is a condition affecting the way people think, feel and behave (or all of the above), which eventually leads to a negative impact on their overall mental health. Such disruptions affect the emotional, psychological and social well-being of an individual, eventually leading to a poor quality of life, degenerated physical health and disease-related disabilities. For instance, depression is considered the primary cause of lost working hours worldwide[4] and is usually intertwined with suicidal behavior or substance misuse, which entail a wide range of social problems.

What is special about identifying and treating mental health probles? Mental disorders frequently occur at early ages. They are heterogeneous, dynamic and multi-causal phenomena[5], which if left untreated, it is likely to turn into chronic problems. On top of that, mental health patients appear to be less respondent and compliant, compared to those suffering from physiological issues, which introduces additional challenges in the management of public mental health. Building diagnostic models requires a multi-faceted inspection of the condition that includes direct observations such as biological, physiological and psychological factors as well as alternative sources of information like social factors and behaviour. In addition, as opposed to somatic conditions, whose diagnosis rely on quantitative measures (i.e. instance laboratory tests), nosological classification and identification of mental disorders relies on self-reported core symptoms originating from questionnaires, clinical interviews and performance tests[6]. Thus, individualised cases that depend on substantial multi-factorial clinical expressions of a disorder or symptomatic overlaps of co-occurring conditions are usually neglected.

The recent surge of individualised medical data has enabled technologies such as Artificial Intelligence (AI) and Machine Learning (ML) to provide clinical decision support with exceptional outcomes. In particular, ML is employed to analyse big and diverse data-sets in order to identify patterns that associate mental disorders with clinical data, biometrics, behaviours and social interactions. As opposed to traditional methods, ML relies on voluminous data that cover numerous and diverse expressions of a disorder. Also, it allows the combination of multi-modal indicators, which is in line with the multi-faceted nature of mental disorders. The generated patterns can be used for the early detection and diagnosis of several disorders[7]. Predicting the likelihood of a patient to have a disorder based on historical diagnoses is one of the most common applications. For instance, prediction of attention deficit hyperactivity disorder (ADHD) based on electroencephalograms[8] or scores of validated self-reported screening questionnaires[9]. Moreover, ML predictive models are also utilised to differentiate conditions with overlapping symptoms such as ADHD and autism (ASD)[10] or Alzheimer patients and normal brain deterioration through aging [11].

ML applications in mental health range from simple models, such as belief networks and decision trees, to complex models like deep neural networks. As expected, complex models tend to outperform simpler ones, however these high-performant models come with the trade of of a black box functionality. During diagnosis, a black box model uses raw data (e.g. images, text) to yield a probability, while obscuring any explanation about its inner mechanics. Although these probabilities are definitely valuable to clinicians, they do not provide any causation in terms of linking individual patient characteristics with the disorder under consideration. This becomes exceptionally challenging in cases of multi-modal based predictions, where models express the correlation of combined clinical data with a disorder, hindering any possible reasoning or discovery of possible interventions. This results into a significant drawback in the domain of mental health, and healthcare in general, which cultivates distrust from medical experts and despite the predictive abilities, it does not improve our understanding on how mental disorders are expressed. Explainable Artificial Intelligence (XAI) approaches aim to address these drawbacks.

2. XAI approaches and their benefits in healthcare

Before delving into particular ways of achieving XAI, it is worth clarifying the relation between explainability and interpretability, which are often used interchangeably in literature [12]. This is, in turn, a side effect of the recent proliferation of ML research which has led to occasions where AI and ML are also used interchangeably, disregarding AI technologies that do not rely on learning from data, i.e. symbolic or knowledge-based ones. We argue that interpretability is a narrower term, referring most often to the interpretation of ML model outputs, or as Biran and Cotton[13] define it, "the degree to which an observer can understand the cause of a decision" of a model. Explainability, on the other hand, is a broader concept that encompasses explanations that may not depend on ML model interpretability, but may rely on modelling expert knowledge and reasoning pathways and may involve psychological, cognitive or philosophical aspects[14].

ML models can either be inherently interpretable or require a post-hoc analysis to be interpreted[15]. The former includes algorithms such as decision trees, regression or Bayesian inference, while the latter includes support vector machines and connectionist approaches based on neural networks. Post-hoc analysis methods are either specific to particular ML algorithms or are applicable regardless of which approach is applied (referred to as model-agnostic interpretability). These include: generating natural language or visual interpretations of model outputs, interpreting particular examples or subsets of a model, or building a surrogate model that produces the same outputs but is more interpretable. The interested reader is referred indicatively to Guidotti et al.[16] and Arrieta et al.[12] for comprehensive reviews of both model-specific and model-agnostic approaches.

Going beyond ML model interpretability, XAI is also achievable through symbolic and knowledge-based approaches which are intrinsically explainable due to their focus on modelling expert knowledge and human reasoning. These include various forms of case-based and rule-based reasoning that may rely on knowledge models such as ontologies or knowledge graphs. More recently, XAI research has led to hybrids that combine knowledge-based and data-driven methods, including well-established neurosymbolic approaches. As Doran et al.[17] argues, "truly explainable AI should integrate reasoning", with interpretability approaches enabling explanations of outcomes but needing to be integrated within a line of reasoning to formulate a comprehensible explanation.

Researchers and practitioners have highlighted the importance of explainability in applications of AI in healthcare and have identified several key benefits such as security and privacy of sensitive medical data, public trust in the use of AI in healthcare and the required skills of healthcare professionals. Adadi and Berrada[18] highlight XAI's potential to support with issues such as security and privacy of sensitive medical data, public trust in the use of AI in healthcare and the required skills of healthcare professionals. Yang et al.[19] emphasises legal aspects, particularly accountability and liability afforded by XAI in healthcare and preventing exploitation of healthcare technologies, especially when there is high risk to life, as well as mitigation of vulnerabilities that can prevent users with malicious intent from exploiting healthcare technologies. Apart from the legal and ethical perspective, Amann et al.[20] explores medical and patient perspectives of XAI in healthcare, such as the ability to resolve disagreement between human experts and AI recommendations, and contribute to keeping patients better informed and reducing likelihood of inaccurate risk perceptions. In the same context, Tonekaboni et al.[21] also adds the ability for clinicians to justify decisions made based on a model's prediction to both patients and colleagues. Finally, the systematic review by Antoniadi et al.[22] summarises desirable and reported benefits of XAI in healthcare, pointing out enhanced decision confidence for clinicians and increased uptake of AI-based clinical decision support systems.

2.1. *Case study: ADHD diagnosis*

Since 2019, our group has been working closely with South-West Yorkshire Partnership NHS Foundation Trust, part of the UK's National Health Service, to develop a world-first solution for adult ADHD diagnosis using AI. Input to our solution is the same clinical data routinely captured by adult ADHD healthcare services, based on National Institute for Health and Care Excellence (NICE) guidelines.

The solution involves a hybrid of two AI-based components. The first is based on a ML model trained from clinical data of past cases. Training of the model was based on 500 past cases, and the two outcomes are yes (has ADHD) and no (does not have ADHD). The other component uses a knowledge-based model using rules captured through extensive interviews of clinical specialists. There are three possible outcomes: yes, no or further assessment needed (complex case requiring specific assessment by senior clinicians). The diagnosis system then adopts a positive or negative diagnostic recommendation when the two components agree; where there is disagreement, the case is referred for further assessment. Based on this approach, an accuracy of 98% was achieved.

The hybrid approach is key not only for high performance but also for addressing the hard requirement of explainability as posed by clinicians, who would only be willing to adopt a technological solution if they understand the basis of the provided recommendations. On the other hand, the ML model supports adaptation when deploying in new health services, and will lead over time to increased accuracy as more data is collected. Evaluation of our technology in a real clinical setting is underway, funded by the National Institute of Heath Research. Preliminary results have been published[9] and an international patent application was submitted in November 2021.

3. XAI for mental health diagnosis: challenges and directions

Explainability in any diagnosis setting, including mental health, retains patients at the center of the process, ensuring they are included and informed about any decisions related to their health, as well as facilitating fair distribution of resources on a case by case basis[20]. Another essential benefit of explainability, which is tied to mental health, is the aid that intelligent systems may offer to clinicians to discover and better understand the perplexed associations of biology, psychology and behaviour in the expression of mental disorders. While there is a wealth of research in different approaches to provide interpretations and explanations of decisions based on AI models, it is an open question as to how appropriate and applicable these approaches are for the particular case of mental health diagnosis, given that its nature requires going beyond mere statistical significance or simple inference to robust explanations.

Explainability approaches that have been applied in other domains face significant challenges in a mental health diagnosis context. Feature importance based solutions are negatively affected by complex correlation relationships that exist among features carrying a diagnostic capacity such as comorbidities or hidden biological links and which are patient-specific. Another challenge is that several ML interpretability methodologies such as ones based on instances as explanations have not yet been evaluated on complex clinical models (including diagnosis ones)[21]. Also, there is no agreed approach on evaluating produced explanations, increasing the risk of confirmation bias[23]. Inherently explainable (e.g. knowledge-based) methods, on the other hand, face difficulties in scaling up to higher complexity models and may be less accurate than data-driven counterparts.

These challenges lead to several directions that can shape future research on explainable AI for mental health diagnosis. A primary direction involves exploring hybrid approaches, such as the one described in Section 2.1. These have the potential of combining the inherent explainability of knowledge-based approaches that can integrate domain knowledge with the performance afforded by data-driven methodologies, reducing the dependence on data which are often difficult to obtain in mental health cases. In this context, neurosymbolic approaches may hold promise due to the combination of explainable symbolic rules and reasoning with connectionist models[17], providing a natural way of explaining diagnosis through reasoning over model properties and diagnosis criteria. Research on design patterns for hybrid learning and reasoning systems[24] can help identify candidates for yielding XAI diagnosis systems for mental health diagnosis.

Further research can focus on establishing an appropriate set of metrics for evaluating explainability of diagnosis to objectively determine the extent to which diagnosis decisions can be explained to relevant parties (including clinicians, patients and guardians). These metrics should generally be quantifiable, even if they may be based on qualitative factors (e.g. usefulness and satisfaction[12]) and must take into account particular aspects of mental health diagnosis such as stigma and misconceptions around mental health. Moreover, developing XAI approaches in consultation with clinicians and patients and evaluating them in real clinical settings will help increase adoption and applicability levels. Advancing research in any of the aforementioned directions will help further the integration of AI technologies in clinical pathways, facilitating diagnosis and treatment processes in mental health and beyond.

References

1. R. C. Kessler, G. P. Amminger, S. Aguilar-Gaxiola, J. Alonso, S. Lee and T. B. Üstün, Age of onset of mental disorders: a review of recent literature, *Current Opinion in Psychiatry* **20**(4) (2007) 359–364.
2. Department of Health, No health without mental health: A cross-Government mental health outcomes strategy for people of all ages, tech. rep., HM Government (2011).
3. Centers for Disease Prevention and Control, Mental health https://www.cdc.gov/mentalhealth/index.htm.
4. M. Marcus, M. T. Yasami, M. van Ommeren, D. Chisholm and S. Saxena, Depression: A Global Public Health Concern, tech. rep., WHO Department of Mental Health and Substance Abuse (2012).
5. V. Roessner, J. Rothe, G. Kohls, G. Schomerus, S. Ehrlich and C. Beste, Taming the chaos?! Using eXplainable Artificial Intelligence (XAI) to tackle the complexity in mental health research, *Eur Child Adolesc Psychiatry* **30** (2021) 1143–1145.
6. K. Katahira and Y. Yamashita, A theoretical framework for evaluating psychiatric research strategies, *Computational Psychiatry* **1** (2017) 184–207.
7. A. B. Shatte, D. M. Hutchinson and S. J. Teague, Machine learning in mental health: a scoping review of methods and applications, *Psychol Med* **49**(9) (2019) 1426–1448.
8. A. E. Alchalabi, S. Shirmohammadi, A. N. Eddin and M. Elsharnouby, FOCUS: detecting ADHD patients by an EEG-based serious game, *IEEE Transactions on Instrumentation and Measurement* **67**(7) (2018) 1512–1520.
9. I. Tachmazidis, T. Chen, M. Adamou and G. Antoniou, A hybrid AI approach for supporting clinical diagnosis of attention deficit hyperactivity disorder (ADHD) in adults, *Health Information Science and Systems* **9**(1) (2021) 1–8.
10. M. Duda, N. Haber, J. Daniels and D. Wall, Crowdsourced validation of a machine-learning classification system for autism and ADHD, *Transl Psychiatry* **7**(5) (2017).
11. N. T. Doan, A. Engvig, K. Zaske, K. Persson, M. J. Lund, T. Kaufmann, A. Cordova-Palomera, D. Alnæs, T. Moberget, A. Brækhus, M. L. Barca, J. E. Nordvik, K. Engedal, I. Agartz, G. Selbæk, O. A. Andreassen and L. T. Westlye, Distinguishing early and late brain aging from the alzheimer's disease spectrum: consistent morphological patterns across independent samples, *NeuroImage* **158** (2017) 282–295.
12. A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila and F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion* **58** (2020) 82–115.
13. O. Biran and C. V. Cotton, Explanation and justification in machine learning: A survey, in *IJCAI-17 workshop on explainable AI (XAI)* (IJCAI, 2017).
14. A. Adadi and M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), *IEEE Access* **6** (2018) 52138–52160.
15. W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl and B. Yu, Definitions, methods, and applications in interpretable machine learning, *Proceedings of the National Academy of Sciences* **116**(44) (2019) 22071–22080.
16. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti and D. Pedreschi, A Survey of Methods for Explaining Black Box Models, *ACM Comput. Surv.* **51**(5) (2018).
17. D. Doran, S. Schulz and T. R. Besold, What Does Explainable AI Really Mean?

A New Conceptualization of Perspectives, in *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with AI*IA 2017*, ed. T. R. B. amd Oliver Kutz **2071**, (CEUR, 2017).
18. A. Adadi and M. Berrada, Explainable AI for Healthcare: From Black Box to Interpretable Models, in *Embedded Systems and Artificial Intelligence*, eds. V. Bhateja, S. C. Satapathy and H. Satori *Advances in Intelligent Systems and Computing: Proceedings of ESAI 2019, Fez, Morocco* **1076**, (Springer, Singapore, 2020).
19. G. Yang, Q. Ye and J. Xia, Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond, *Information Fusion* **77** (2022) 29–52.
20. J. Amann, A. Blasimme, E. Vayena, D. Frey and V. I. Madai, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, *BMC Med Inform Decis Mak* **20**(310) (2020).
21. S. Tonekaboni, S. Joshi, M. D. McCradden and A. Goldenberg, What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use, in *Proc. MLHC 2019*, eds. F. Doshi-Velez, J. Fackler, K. Jung, D. C. Kale, R. Ranganath, B. C. Wallace and J. Wiens *Proceedings of Machine Learning Research* **106**, (PMLR, 2019), pp. 359–380.
22. A. M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker and C. Mooney, Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review, *Appl Sci* **11**(11) (2021).
23. M. Ghassemi, L. Oakden-Rayner and A. L. Beam, The false hope of current approaches to explainable artificial intelligence in health care, *The Lancet Digital Health* **3**(11) (2021) e745–e750.
24. F. van Harmelen and A. ten Teije, A Boxology of Design Patterns for Hybrid Learning and Reasoning Systems, *Journal of Web Engineering* **18**(1-3) (2019) 97–124.

Acknowledgments

This report is independent research partially funded by the National Institute for Health Research Artificial Intelligence in Health and Care Award (An Artificial Intelligence Algorithm for Diagnosing Attention Deficit Hyperactivity Disorder (ADHD) in Adults, AI_AWARD01612), and NHSX. The views expressed in this publication are those of the author(s) and not necessarily those of the National Institute for Health Research, NHSX or the Department of Health and Social Care.