The role of trustworthy and reliable AI for multiple sclerosis

Lorin Werthen-Brabants*, Tom Dhaene and Dirk Deschrijver

SUMO Lab, IDLab, INTEC, Ghent University – imec, Ghent, Belgium

This paper investigates the importance of Trustworthy Machine Learning (ML) in the context of Multiple Sclerosis (MS) research and care. Due to the complex and individual nature of MS, the need for reliable and trustworthy ML models is essential. In this paper, key aspects of trustworthy ML, such as out-of-distribution generalization, explainability, uncertainty quantification and calibration are explored, highlighting their significance for healthcare applications. Challenges in integrating these ML tools into clinical workflows are addressed, discussing the difficulties in interpreting AI outputs, data diversity, and the need for comprehensive, quality data. It calls for collaborative efforts among researchers, clinicians, and policymakers to develop ML solutions that are technically sound, clinically relevant, and patient-centric.

1 Introduction

Machine Learning (ML) is increasingly applied to healthcare applications (1). While traditional statistical methods can help with biomarker discovery and recognizing trends and correlations, modern ML techniques such as Deep Learning (DL), are able to uncover complex correlations and provide better results than traditional, simpler techniques (2) due to their universal nature (3). Conversely, as these techniques become more complex, the need for *reliable* and *trustworthy* models increases (4, 5), especially within healthcare. However, building trust does not have a one-size-fits-all solution, resulting in many techniques to be developed to aid decision making.

For an end-user, be it a clinician or a patient, a model that is trustworthy is one that can provide certain guarantees on its predictions, explain its predictions, and provide a notion of uncertainty. For a complex disease such as Multiple Sclerosis (MS), the need for trustworthy models is especially pertinent, as its progression is non-trivially defined, and the decisions made to hinder its progression are important ones. A machine learning system that does not provide adequate reliability metrics, or trustworthy insights, will be less appealing to the end-user when there are high-stakes consequences. In recent years, the need for Trustworthy ML (TML) has also reached mainstream attention with the use of generative AI becoming more prevalent. For example, though Large Language Models have shown impressive results, they may still provide incorrect results, without any notion of uncertainty or trustworthiness (6). This is also known as the "hallucination" effect (7). Complex data and relationships warrant the use of trustworthiness techniques.

In the following Sections, we provide a summary of techniques present in Trustworthy ML (TML) (Section 2), why TML is necessary for MS (Section 3.1), and the associated challenges (Section 3.2).

2 Trustworthy machine learning

2.1 Out-of-distribution generalization

The many ways in which MS progression can occur (different limbs, locations of lesion growth, etc.), makes the disease variable and patient specific. Therefore, training data will rarely contain enough data to cover the full extent of the ways progression can be observed. Furthermore, due to protocols changing regularly and equipment variability, *concept* or *model drift* (8) may pose a real issue when ML models are deployed in the real world. Model drift occurs when new data do not correspond to the data on which a model was trained. As a result, models must be continually adapted so changes in data distributions are captured.

These issues can be tackled by making use of techniques such as domain adaptation (9, 10), a specific case of transfer learning (11), and synthetic data sampling such as SMOTE (12, 13).

The concept of Out-of-Distribution Generalization can be elucidated by considering a concrete example within the MS context. Imagine an ML model trained on data from North American patients. When this model is applied to patients from different geographical regions with distinct genetic and environmental factors, its predictions may falter due to differences in disease manifestation. Domain adaptation techniques can help here by adjusting the model to account for these regional variations. Similarly, synthetic data sampling, like the aforementioned SMOTE technique, can artificially—not necessarily in a representative way—augment the dataset to include underrepresented samples in a given dataset, improving the model's robustness against a wide range of clinical scenarios. However, it must be stressed that data quality is key, and an underrepresented dataset can not fully capture the underlying factors to guarantee good out-of-distribution generalization.

2.2 Explainability and interpretability

A perfectly interpretable AI provides insights into the inner workings and decision process of an AI system. When it comes to the types of ML systems, they can broadly be divided into two categories: white-box models and black-box models.

2.2.1 White-box models

Models that are inherently explainable and interpretable. These are often simpler methods such as linear or logistic regression, the latter of which can be represented as a nomogram (14), a graphical representation of such models that visually convey the weight of different input variables. These models can be fully dissected, so there may be many ways of representing or explaining them.

2.2.2 Black-box models

Models that can not be interpreted easily, and are regarded as a "black box" out of which little or no knowledge can be derived. However, there are techniques that can provide explainability when working with black-box models, such as making use of Shapley values (15, 16) or making use of Deep Learning specific techniques (17) such as Layer-wise Relevance Propagation (18, 19). These are often post-hoc. In practice, these techniques will show a number of features and their importances expressed as a number. This could also be in the form of a heatmap. These feature importances may not always be as readily interpretable and may need training and education to comprehend adequately. Additionally, they do not necessarily *explain* why those features are important.

A classifier that may perform well in its evaluation metrics (sensitivity, specificity, ROC AUC, etc.) may still benefit from explainability methods. In particular, if models were to take into account many multimodal variables, the primary drivers of a given prediction may offer important insight for the user of the machine learning system.

Related to interpretable AI is explainable AI. Rather than being able to fully comprehend the inner workings of a model, an explainable AI model is able to be queried so that a reasonable explanation to the prediction is provided. Explainable AI can be viewed on different levels as well: Global, cohort, and local explainability. Global explainability provides information about the entire population or dataset. Due to the complex nature of the MS disease, valuable insights on a population level are scarce. Cohort explainability gives insight on subsets of the data, which can be more interesting when taking into account certain covariates. In this way, different groups of patients can be identified and correlations within these groups may offer more helpful insights than looking only at a global level. Lastly, local explainability provides insight on the model's output for a single input example. Every patient has a different profile, and therefore local explainability may help acquire insight into the prediction of the model for that specific patient or observation.

2.3 Uncertainty quantification and calibration

2.3.1 Uncertainty quantification

In machine learning models, uncertainty plays a critical, yet understated role in understanding and interpreting predictions. Healthcare specifically can greatly benefit from uncertainty quantification, as it can add a layer of trust between the user and the model (20–22). Two major sources of uncertainty are aleatoric and epistemic uncertainty (23).

2.3.1.1 Aleatoric uncertainty

This type of irreducible uncertainty is inherent in the data itself. It cannot be reduced by adding more data and manifests as the noise within the data. An example of this uncertainty arises when using very few features. For example, a patient's blood pressure is a crucial health metric, but it exhibits natural variability within an individual due to various factors like stress, activity level, time of day, and even the way it is measured.

This uncertainty can be either homoscedastic, when it remains constant for all values (e.g., base noise of a sensor), or

heteroscedastic, when it varies depending on the value of the sample.

2.3.1.2 Epistemic uncertainty.

Epistemic uncertainty arises from the model's limited knowledge. This reducible uncertainty is high when the model has insufficient data to characterize or capture the target variable. Increasing the size of the data set can help reduce epistemic uncertainty. An intuitive example can be demonstrated as follows: Say there are multiple experts for a single disease such as MS. These experts may disagree on a given prognosis, despite all of them being equally trained for such a task. Analogously, in a machine learning model predicting patient outcomes for MS, the model might exhibit high epistemic uncertainty if it has been trained on a limited or non-representative dataset. Just as the disagreement among experts might stem from variations in their individual experiences and interpretations, the model's uncertainty arises from its limited exposure to the diverse manifestations of the disease. By providing the model with more comprehensive data that captures a wider range of patient histories, symptoms, and outcomes, the epistemic uncertainty can be reduced, leading to more consistent and reliable predictions.

Applying uncertainty quantification in MS involves recognizing and managing the inherent unpredictability in patient responses and disease progression. For instance, a model expressing aleatoric uncertainty might show the variability in a patient's symptoms over time, acknowledging that certain aspects of MS progression cannot be predicted with complete precision. Epistemic uncertainty can be illustrated by a model's varying predictions based on different patient subgroups, reflecting limited knowledge about specific MS manifestations. To quantify and capture these uncertainties, techniques like Monte Carlo Dropout (MCD) [24] can be employed, providing a probabilistic understanding of a model's predictions and helping clinicians make informed decisions under uncertainty.

Uncertainty quantification has been applied to lesion detection in MRI images [25–27], often making use of MCD or other methods of obtaining a model that can express uncertainty [28].

2.3.2 Calibration

A well-calibrated machine learning model is one in which the model's predicted probabilities closely match the probabilities observed in the actual data [29]. Mathematically, this is represented as $P(y|\hat{p}(y) = \alpha) = \alpha$. This equation signifies that the probability of an event y occurring, given that the model predicts it with probability α, should ideally be α itself. As a practical example: a model that predicts the probability of 40% disease progression for a patient will ideally be correct 40% of the time of all patients who receive a similar prognosis. For methods such as neural networks, this is not often the case by default, and calibration needs to be improved. Additionally, calibration can also be applied to regressors that output a distribution, rather than a single value. In this case, the confidence interval (such as a 95% confidence interval, for example) can be calibrated to ensure that it matches the observations.

The need for calibration is evident in the lack of information an uncalibrated classifier or regressor provides. Often, as is the case with neural networks, a neural network classifier will collapse to output probabilities close to 100% or 0% consistently, rather than providing accurate probability estimates [29]. As a result, a user of such a system needs to blindly trust the classifier rather than being able to take the confidence of the classifier into account.

3 Discussion

3.1 Why trustworthy ML is necessary for MS research

With the current knowledge of MS and performance of state-of-the-art machine learning models in the field, it stands to reason that there may not be a one-size-fits-all solution to detecting disease progression. Although other types of model (such as image classifiers) may perform very well and can reliably be used in most, if not all, cases, this may not be the case for MS. ML models for this purpose will likely be a tool to aid decision making, rather than a decision maker by itself. To that end, an ML model that just states "yes" or "no" is not sufficient. Rather, more information should be supplied to the user. A trustworthy version of this model will highlight parts of the input that contribute greatly to the prediction, show which global and cohort features are important, and also provide a notion of (un)certainty with the prediction. In this way, the user can:

- Select which predictions to trust and keep, both by using aleatoric and epistemic uncertainty as guides
- Analyze the subgroup in which the prediction fits
- Analyze the specific prediction and the features leading to the prediction

For MS research, the use and adoption of ML will be guided by advances in *trustworthy* ML. MS is a disease marked by its heterogeneity in symptoms, progression, and response to treatment, making reliable analysis of significant importance.

The ability of ML models to process and analyze different types of data—from clinical observations to MRI images—can lead to earlier detection and more precise monitoring of the disease's progression. However, the value of these insights depends on their explainability. Clinicians and patients must be able to understand and trust the model's predictions, necessitating a focus on explainable AI. For example, an ML model might identify subtle changes in brain lesions over time, but this information becomes clinically actionable only when it is presented in an understandable manner. Explainable models can elucidate the factors driving a prediction, thereby enhancing the clinician's ability to make informed treatment decisions.

Moreover, the integration of uncertainty quantification in ML models is particularly relevant for MS. Given the variability in how the disease presents and progresses, models that can express their confidence in predictions are invaluable. They provide clinicians with a more nuanced understanding of each prediction,

facilitating more informed risk-benefit analyses when deciding on treatment plans. A model that indicates a high level of uncertainty in its prediction might prompt further testing or closer monitoring, whereas a prediction made with high confidence could lead to more decisive action.

The importance of trustworthy ML in MS research also extends to patient empowerment. Access to understandable and reliable ML-driven insights can foster better patient-clinician dialogues. When patients understand the basis for predictions about their condition, they are better positioned to make informed decisions about their treatment and lifestyle choices.

3.2 Challenges of trustworthy ML for MS

3.2.1 Integration of ML tools to aid clinical decisions

Integrating ML tools into existing clinical workflows presents another layer of complexity. For these tools to be adopted, they must fit into the highly regulated environment of healthcare. This integration involves designing user interfaces and metrics that are intuitive for clinicians, ensuring that ML predictions are presented in a way that complements decision-making processes rather than complicating them (30). Furthermore, imperfect data pose a problem during the training and prediction stages of an ML model. Data collection can be a laborious task, and in some cases the data cannot be accurately represented due to individual differences in disease expression. This rings especially true in the case of MS.

3.2.2 Usability of uncertainty quantification and explainability techniques

As highlighted previously, UQ and explainability techniques have their merit, as they can highlight potential issues when making use of ML assisted decision systems. However, the end-user may not find much use in the way UQ results are represented in literature. Even explainabilty results have varying degrees of success concerning their usability (31). These techniques could benefit from user studies, as their usability hinges on the representation and, in turn, interpretation by the end-user. For example, rather than providing the clinician and/or patient with a numerical value signifying a "trustworthiness" score or certainty otherwise, larger trust could be gained by comparing the patient with other patients that have similar disease trajectories. This opacity can hinder trust and acceptance, especially in a high-stakes field like healthcare where understanding the "why" behind a diagnosis or prognosis is as crucial as the outcome itself (31).

3.2.3 Out-of-distribution data, diverse data, available data

Data diversity and availability are critical factors that significantly influence the development and performance of ML models in MS research. MS is a disease with a highly variable clinical course and a wide range of symptoms that differ from patient to patient. This heterogeneity necessitates a rich and diverse dataset that captures the broad spectrum of the disease. After all, deep learning techniques are prone to overfitting, and may have performance below acceptable levels as a result (21, 32). Initiatives such as MSBase (33, 34) attempt to address the issue of out-of-distribution performance by providing multi-center data. The amount of data by itself may give the end-user a reason to trust a model, given enough diversity. Data quality is another concern, with issues such as missing values, inconsistent data entry, and the need for standardization across different data sources complicating the development of reliable ML models. Introducing diversity by including measurements that stray away from purely medical imaging or clinical data may also provide a new avenue of research, potentially discovering novel biomarkers. Future work should focus on developing models that can adapt to individual patient variations and incorporating emerging data types such as Motor Evoked Potentials (35, 36) into ML models.

4 Conclusion

This paper underscores the importance of trustworthiness in Machine Learning (ML) applications for Multiple Sclerosis (MS). Key aspects such as explainability, uncertainty quantification and calibration, and out-of-distribution generalization have been explored. Additionally, the challenges in integrating ML into clinical workflows and the hurdles posed by data diversity and availability have been discussed.

The authors urge the research community and healthcare providers to prioritize the development and implementation of trustworthy ML solutions for MS (and healthcare in general). There is an urgent need to foster partnerships between computer scientists, neurologists, and patients. This collaboration will ensure the development of ML solutions that are not only technically sound but also clinically relevant and patient-centric. Making comprehensive, high-quality data sets accessible while respecting privacy concerns is crucial. Initiatives should focus on standardizing data collection and sharing practices to aid in the development of more effective ML models. ML tools must be integrated into clinical workflows in a way that is intuitive and enhances decision-making processes. This involves designing user-friendly interfaces and ensuring that clinicians are adequately trained to use these tools effectively.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

LW-B: Conceptualization, Investigation, Writing – original draft, Writing – review & editing. TD: Funding acquisition,

Supervision, Writing – review & editing. DD: Funding acquisition, Supervision, Writing – review & editing.

Funding

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. (2017) 2:230–43. doi: 10.1136/svn-2017-000101

2. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with alphafold. *Nature*. (2021) 596:583–9. doi: 10.1038/s41586-021-03819-2

3. Cybenko G. Approximation by superpositions of a sigmoidal function. *Math Control Signals Syst*. (1989) 2:303–14. doi: 10.1007/BF02551274

4. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clin Infect Dis*. (2018) 66:149–53. doi: 10.1093/cid/cix731

5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. (2023) 29:1930–40. doi: 10.1038/s41591-023-02448-8

6. Sejnowski TJ. Large language models and the reverse turing test. *Neural Comput*. (2023) 35:309–42. doi: 10.1162/neco_a_01563

7. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. Data from: Chatgpt and other large language models are double-edged swords (2023)

8. Tsymbal A. *The Problem of Concept Drift: Definitions and Related Work*. Technical report (TCD-CS-2004-15). Dublin: Computer Science Department, Trinity College Dublin (2004). Vol. 106. p. 58.

9. Farahani A, Voghoei S, Rasheed K, Arabnia HR. A brief review of domain adaptation. In: *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020* (2021). p. 877–94.

10. Valverde S, Salem M, Cabezas M, Pareto D, Vilanova JC, Ramió-Torrentà L, et al. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage Clin*. (2019) 21:101638. doi: 10.1016/j.nicl.2018.101638

11. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data*. (2016) 3:1–40. doi: 10.1186/s40537-016-0043-6

12. Branco D, Martino B, Esposito A, Tedeschi G, Bonavita S, Lavorgna L. Machine learning techniques for prediction of multiple sclerosis progression. *Soft Comput*. (2022) 26:12041–55. doi: 10.1007/s00500-022-07503-z

13. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. *J Artif Intell Res*. (2002) 16:321–57. doi: 10.1613/jair.953

14. Kattan MW, Marasco J. What is a real nomogram? In: *Seminars in Oncology*. Elsevier (2010). Vol. 37. p. 23–6.

15. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inform Process Syst*. (2017) 30:4768–77.

16. Basu S, Munafo A, Ben-Amor AF, Roy S, Girard P, Terranova N. Predicting disease activity in patients with multiple sclerosis: an explainable machine-learning approach in the mavenclad trials. *CPT Pharmacom Syst Pharmacol*. (2022) 11:843–53. doi: 10.1002/psp4.12796

17. Chakraborty S, Tomsett R, Raghavendra R, Harborne D, Alzantot M, Cerutti F, et al. Interpretability of deep learning models: a survey of results. In: *2017 IEEE Smartworld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*. IEEE (2017). p. 1–6.

18. Creagh AP, Lipsmeier F, Lindemann M, Vos MD. Interpretable deep learning for the remote characterisation of ambulation in multiple sclerosis using smartphones. *Sci Rep*. (2021) 11:14301. doi: 10.1038/s41598-021-92776-x

19. Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR. Layer-wise relevance propagation: an overview. *Explainable AI*. (2019) 11700:193–209. doi: 10.1007/978-3-030-28954-6_10

20. Seoni S, Jahmunah V, Salvi M, Barua PD, Molinari F, Acharya UR. Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013–2023). *Comput Biol Med*. (2023) 165:107441. doi: 10.1016/j.compbiomed.2023.107441

21. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. *Nat Mach Intell*. (2019) 1:20–3. doi: 10.1038/s42256-018-0004-1

22. Lambert B, Forbes F, Doyle S, Dehaene H, Dojat M. Trustworthy clinical ai solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. *Artif Intell Med*. (2024) 150:102830. doi: 10.1016/j.artmed.2024.102830

23. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn*. (2021) 110:457–506. doi: 10.1007/s10994-021-05946-3

24. Gal Y, Ghahramani Z. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: *International Conference on Machine Learning*. PMLR (2016). p. 1050–9.

25. Nair T, Precup D, Arnold DL, Arbel T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Med Image Anal*. (2020) 59:101557. doi: 10.1016/j.media.2019.101557

26. Molchanova N, Raina V, Malinin A, La Rosa F, Muller H, Gales M, et al. Novel structural-scale uncertainty measures and error retention curves: application to multiple sclerosis. In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE (2023). p. 1–5.

27. Tousignant A, Lemaître P, Precup D, Arnold DL, Arbel T. Prediction of disease progression in multiple sclerosis patients using deep learning analysis of MRI data. In: *International Conference on Medical Imaging with Deep Learning*. PMLR (2019). p. 483–92.

28. Lambert B, Forbes F, Doyle S, Tucholka A, Dojat M. Fast uncertainty quantification for deep learning-based MR brain segmentation. In: *EGC 2022-Conference Francophone Pour l'Extraction et la Gestion des Connaissances*. (2022). p. 1–12.

29. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: *International Conference on Machine Learning*. PMLR (2017). p. 1321–30.

30. Dabbs ADV, Myers BA, Mc Curry KR, Dunbar-Jacob J, Hawkins RP, Begey A, et al. User-centered design and interactive health technologies for patients. *Comput Inform Nurs*. (2009) 27:175. doi: 10.1097/NCN.0b013e31819f7c7c

31. Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J. Explainable, trustworthy, and ethical machine learning for healthcare: a survey. *Comput Biol Med.* (2022) 149:106043. doi: 10.1016/j.compbiomed.2022.106043

32. Mårtensson G, Ferreira D, Granberg T, Cavallin L, Oppedal K, Padovani A, et al. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *Med Image Anal.* (2020) 66:101714. doi: 10.1016/j.media.2020.101714

33. Butzkueven H, Chapman J, Cristiano E, Grand'Maison F, Hoffmann M, Izquierdo G, et al. Msbase: an international, online registry and platform for collaborative outcomes research in multiple sclerosis. *Mult Scler J.* (2006) 12:769–74. doi: 10.1177/1352458506070775

34. De Brouwer E, Becker T, Werthen-Brabants L, Dewulf P, Iliadis D, Dekeyser C, et al. Machine-learning-based prediction of disability progression in multiple sclerosis: an observational, international, multi-center study. *PLoS Digit Health.* (2024) 3: e0000533. doi: 10.1371/journal.pdig.0000533

35. Rossini PM, Rossi S. Clinical applications of motor evoked potentials. *Electroencephalogr Clin Neurophysiol.* (1998) 106:180–94. doi: 10.1016/S0013-4694 (97)00097-7

36. Yperman J, Popescu V, Van Wijmeersch B, Becker T, Peeters LM. Motor evoked potentials for multiple sclerosis, a multiyear follow-up dataset. *Sci Data.* (2022) 9:207. doi: 10.1038/s41597-022-01335-0