6

Moving Toward Explainable Decisions of Artificial Intelligence Models for the Prediction of Functional Outcomes of Ischemic Stroke Patients

Esra Zihni[1] • Bryony L. McGarry[1,2] • John D. Kelleher[1,3]

[1]PRECISE4Q, Predictive Modelling in Stroke, Technological University Dublin, Dublin, Ireland; [2]School of Psychological Science, University of Bristol, Bristol, UK; [3]ADAPT Research Centre, Technological University Dublin, Dublin, Ireland

Author for correspondence: Esra Zihni, PRECISE4Q, Predictive Modelling in Stroke, Technological University Dublin, Dublin, Ireland. E-mail: esra.zihni@tudublin.ie

Cite this chapter as: Zihni E, McGarry BL, Kelleher JD. Moving Toward Explainable Decisions of Artificial Intelligence Models for the Prediction of Functional Outcomes of Ischemic Stroke Patients. In Linwood SL, editor. *Digital Health*. Brisbane (AU): Exon Publications. Online first 2021 Nov 26.

Doi: https://doi.org/10.36255/exon-publications-digital-health-explainable-decisions

Abstract: Artificial intelligence has the potential to assist clinical decision-making for the treatment of ischemic stroke. However, the decision processes encoded within complex artificial intelligence models, such as neural networks, are notoriously difficult to interpret and validate. The importance of explaining model decisions has resulted in the emergence of explainable artificial intelligence, which aims to understand the inner workings of artificial intelligence models. Here, we give examples of studies that apply artificial intelligence models to predict functional outcomes of ischemic stroke patients, evaluate existing models' predictive power, and discuss the challenges that limit their adaptation to the clinic. Furthermore, we identify the studies that explain which model features are

In: Linwood SL, editor. *Digital Health*. Exon Publications, Brisbane, Australia.
ISBN: 978-0-6453320-1-8. Doi: https://doi.org/10.36255/exon-publications-digital-health

Copyright: The Authors.

License: This open access article is licenced under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) https://creativecommons.org/licenses/by-nc/4.0/

essential in predicting functional outcomes. We discuss how these explanations can help mitigate concerns around the trustworthiness of artificial intelligence systems developed for the acute stroke setting. We conclude that explainable artificial intelligence is a must for the reliable deployment of artificial intelligence models in acute stroke care.

Keywords: explainable artificial intelligence; ischemic stroke; neural networks; functional outcome prediction

INTRODUCTION

Artificial intelligence (AI) focuses on developing computational systems that emulate human expertise on complex tasks. There have been many developments in AI, particularly related to the field of deep learning (DL), resulting in modern AI systems that could rival human performance on complex decision-making tasks on large amounts of data. Studies suggest that the performance of AI on a range of medical diagnostic tasks is equivalent to healthcare professionals (1). Consequently, there is much interest in the potential of AI in the development of clinical decision support systems (CDSS) (2). The purpose of CDSSs is not to replace the clinician but to increase the efficiency and efficacy of medical diagnosis and treatment. CDSSs will be particularly beneficial in the care of ischemic stroke patients due to the heterogeneity, complexity, and time-critical nature of the condition and the wealth of physiological information available from neuroimaging. AI algorithms can provide decision support for ischemic stroke in various clinical tasks, such as diagnosis, onset time estimation, and prognosis (3–5). However, AI systems can be met with resistance from health care professionals because even if the system is accurate, it is not obvious how it arrives at its decisions; this is the so-called "black box problem" (6).

The main goal of this chapter is to provide an insight into the potential of AI for developing CDSSs to be used in the hyperacute stroke setting. This way, we wish to make AI-based methods more accessible and understandable to those less familiar with AI. To this end, we first discuss the potential benefits of medical AI for developing CDSSs for ischemic stroke. We focus specifically on predicting functional outcomes, referring to the patient's overall mobility and level of functioning in day-to-day life (7). To contextualize the studies discussed in this chapter, we provide an overview of machine learning (ML) and DL terminology. We then give examples of studies that have used AI methods to predict functional outcomes using clinical variables and neuroimaging, followed by a discussion of the potential barriers to implementing AI systems in the clinical setting. As one of the avenues to overcome the barriers, we introduce the concept of explainable AI (xAI) and refer to studies that use explainability to give insight into the inner mechanisms of functional outcome prediction models. We place particular attention on explanation methods that identify the important features that drive a prediction and refrain from providing an in-depth review. Lastly, we evaluate the findings of these studies to understand what the explanations mean in terms of future directions. We conclude the chapter by discussing the necessity of explainable decision in CDSSs developed for acute stroke care.

THE BENEFITS OF MEDICAL-AI FOR ISCHEMIC STROKE

A major benefit of using AI for developing CDSSs is facilitating a patient-specific, adaptive (rather than reactive) approach to treatment decisions. Currently, ischemic stroke patients are treated following population-based guidelines, which recommend intravenous recombinant tissue plasminogen activator (IV-rtPA) or endovascular therapy (EVT) with or without rtPA for large vessel occlusions, but within specific time windows (8, 9). However, although these treatments are safe and effective for most that meet the eligibility criteria, due to the heterogeneity of the disease, there are some patients that meet eligibility criteria that could be harmed by treatment or may not benefit either way and other patients who are not eligible (e.g., unknown onset time) that could still have benefitted. Therefore, predicting the functional outcome of patients based on information available on admission, given a specific treatment option, will help clinicians identify, on a case-by-case basis and in a contextually informed manner, patients who will (or will not) benefit from different forms of treatment.

Minimizing the time-to-treatment is critical in treating ischemic stroke due to the fast-evolving tissue death (10). CDSSs will help minimize this time, especially in complex and ambiguous cases, by providing formal support for the experienced clinician's treatment decisions. Knowing which treatment a patient is likely to benefit from will enable more efficient coordination of clinical resources for acute treatment, such as preparing patients for transfer to a comprehensive stroke center for EVT. CDSS will also mean less experienced physicians can make treatment decisions for acute stroke patients, which will benefit hospitals that do not have the expertise and technologies of primary stroke centers.

ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, AND DEEP LEARNING

ML is a subfield of AI that involves developing computer programs that extract patterns from data. In the context of traditional ML, a dataset is a tabular representation of information where each column is a feature (e.g., age, sex, blood pressure), and each row contains the features' values describing an instance, i.e., a patient profile. The standard task in ML is to create a computer program that can estimate the missing value for one feature (known as the target feature) based on the values of the other features. For example, ML can create a computer program to estimate blood pressure based on age, sex, and other features. The estimated value for the target feature is known as a prediction, and a program that can map from a set of feature values to an estimate of a missing feature value is known as a predictive model.

An ML algorithm is a computer program that takes a dataset as input and returns a predictive model. Running an ML algorithm on a dataset to create a model is known as training the model. Once a model has been trained, it can be used to estimate the value of the target feature for new instances (patients) that were not in the original dataset. ML algorithms require structured or tabular data as input features, which take the form of categories, numbers, or values, such as patient demographics (e.g., age, sex), comorbidities (e.g., diabetes, hypertension) and radiological variables such as the ASPECTS score (11), presence of a lesion (yes/no) and clot location. A limitation of ML methods is that getting the data in a structured format usually requires a domain expert, such as a clinician or neuroradiologist, to identify these important features.

DL is a subtype of ML that loosely mimics the human brain's neural pathways. What is distinctive about DL systems is that they can automatically learn what features in the data that are most important to generate accurate predictions. This ability removes the need for a human expert to hand-design the features considered by a model and is important because it enables models to process data that is very difficult for humans to design features from, for example, to define what voxel patterns are most useful to predict a lesion. In more technical terms, this means that DL systems can perform high-level abstractions from structured and unstructured data without pre-processing (12). In this context, unstructured data refers to data that is not easily represented in a tabular manner, such as text, audio, and imaging. DL methods can therefore process data considerably more complex than traditional ML algorithms, such as 2D slices or 3D volumes from neuroimaging such as computerized tomography (CT) or magnetic resonance imaging (MRI). This ability is particularly advantageous in the context of acute stroke because different imaging methods reveal different aspects of stroke pathophysiology, and so there is a large amount of information that can be obtained from scans obtained in the early hours of ischemia (13, 14). The fact that DL algorithms can simultaneously extract useful features from complex unstructured data and learn complex mappings from sets of inputs to an output means that for stroke treatment, these systems can learn to process and use both medical imaging data and structured information (such as clinical variables) to inform the model's decision. This way, they can potentially return more accurate decisions than a system restricted to only processing structured data.

The predictive performance of models is usually reported by testing them on validation data that is not part of model training using various performance measures. Accuracy, a popular performance measure that gives the percentage of true predictions amongst all predictions, has demonstrated over-optimistic results, especially when using imbalanced datasets (15). Thus, the field has shifted towards using alternative scores such as Area under the Receiver Operator Characteristics curve (AUC). The Receiver Operator Characteristics (ROC) curve shows the trade-off between sensitivity and specificity at different decision thresholds. In functional outcome prediction, the AUC can be interpreted as the estimated probability that a randomly selected patient who experienced an unfavorable outcome had a higher risk score than a patient who had experienced a favorable outcome. AUC is the standard measure reported across all ML and DL based studies in predicting functional outcomes.

ARTIFICIAL INTELLIGENCE-BASED MODELS FOR FUNCTIONAL STROKE OUTCOME PREDICTION

In this section, we focus specifically on studies that use modified Ranking Scale (mRS) (Table 1), (16) , as the main functional outcome measure, because it is a

TABLE 1 The modified Rankin Scale (mRS)

Grade	Symptoms
0	None
1	No significant disability despite symptoms: able to carry out all usual duties and activities
2	Slight disability: unable to carry out all previous activities, but able to look after own affairs without assistance
3	Moderate disability: requiring some help, but able to walk without assistance
4	Moderately severe disability: unable to walk without assistance, unable to attend to needs without assistance
5	Severe disability: bed-ridden, incontinent, and requiring constant nursing care and attention
6	Dead*

Patients are graded on the scale of 0–6. *The initial mRS was 0–5 and the 6th grade was added later. For outcome prediction in clinical trials, the mRS is usually dichotomized where good functional outcome is a score 0 – 2 and poor functional outcome 3–6. However other trials have analyzed the mRS ordinally. For full details on the mRS, see (16).

common form of communication across all those involved in the patient's care pathway (7), has good reproducibility (16), is the most prevalent functional outcome measure in contemporary stroke trials (17) and is therefore frequently used as the primary outcome measure in AI prediction studies (3, 4). However, it is worth noting that the discussion points of this chapter are relevant to most AI-based prediction studies in stroke.

Before the introduction of AI, 90-days mRS was predicted using rule-based scores based on clinical information available on admission, such as the patient's age, National Institute of Health Stroke Score (NIHSS), and onset time (18). These scores, such as the Acute Stroke Registry, and Analysis of Lausanne (ASTRAL) (19), among others, are based on assigning weights in the form of integers to the most relevant clinical variables pre-selected by clinicians. The weights are then aggregated to provide insight into patients' risk of functional impairment. These scores are usually based on a small set of clinical variables and usually do not incorporate scores derived from imaging that may have predictive value (20). With the advancement in AI, both traditional ML and DL algorithms have gained popularity in predictive modelling of functional outcomes. ML classifiers used in predictive modelling range from logistic regression (LR) to more complex methods such as support vector machines (SVM) and random forests (RF). LR differs from the other ML classifiers in that it can only discover linear mappings between the input data and the desired prediction. Therefore, we differentiate the other traditional ML methods that can perform linear and non-linear mappings from LR and refer to them as non-linear ML algorithms. Table 2 provides information on the non-linear ML and DL algorithms that have been used in developing predictive models for functional outcomes.

Figure 1 summarizes the different properties of the rule-based, LR, non-linear ML, and DL methods. These properties are the type of input features they can process and the complexity of the patterns they can discover. We additionally

TABLE 2 **Non-linear Machine Learning (ML) and Deep Learning (DL) algorithms that are used for functional outcome prediction in acute ischemic stroke patients.**

ML		DL	
Name	**Description**	**Name**	**Description**
SVM	SVMs perform classification by maximizing the distance between the instances of the two target classes. Although this is a linear mapping, they can also efficiently perform a non-linear classification by first applying non-linear transformations to the input.	ANN	ANNs are composed of neurons that are connected to one another in a layered structure. These neurons pass information from the input towards the output performing a non-linear mapping at each pass. They learn to optimize these mappings through backwards propagation of the prediction loss.
Decision Tree	A decision tree encodes if-then-else rules in a tree structure. Each tree node represents a feature, and each branch from a node a different value of that feature. A path from the root node to a leaf node defines a sequence of feature values that an instance must satisfy for the label associated with the leaf node to be predicted for that instance.		
RF	RFs are ensembles of decision trees where the independent prediction from each tree is averaged to obtain a final prediction. More trees give a more robust model.	CNN	CNNs are a special type of ANNs that are biologically inspired by how the human visual cortex works. CNNs are specialized for processing data that has a known grid-like topology, for example, images.
Gradient boosting	Gradient boosting is another tree-based ensemble method that uses weak decision trees one after another (i.e., sequentially) to build a stronger classifier each time.		

ANN, Artificial neural network; CNN, Convolutional neural network; DL, Deep learning; ML, Machine learning; SVM, Support vector machine; RF, Random forest.

emphasize two properties that change gradually between methods: interpretability and computational cost. Generally, the intuitive interpretation of a model's decision process decreases as complexity, in terms of its input features and mapping, increase. Additionally, the computational cost of training an algorithm increases with complexity due to the increased amount of model parameters that need training when more complex architectures and high-dimensional features are used. Finally, the arrows indicate that the advent of DL, and thus the ability to learn complex features and mappings simultaneously, has opened the path to the possibility of multi-modal learning, i.e., learning from structured and unstructured data together.

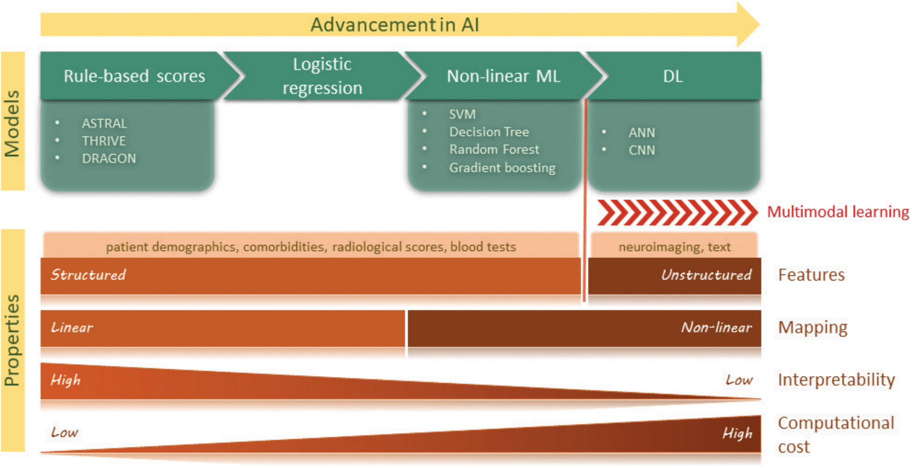

Figure 1. Types of Models. The four types of models used in functional outcome prediction studies and their properties.

Predictive models based on structured data

The potential utility of ML algorithms in functional outcome prediction was demonstrated for the first time in 2014 (21). An SVM model predicted the 90-days mRS scores of patients treated with EVT, based on patient demographics, comorbidities, the baseline NIHSS, and the occluded vessels' location with 70% accuracy (21). In subsequent studies, two main approaches have improved the predictive accuracy of ML and DL models for the functional outcome prediction of ischemic stroke patients. The first approach combines clinical information on admission with information from future time points (e.g., post-treatment or discharge assessments) as inputs to the models, resulting in higher predictive performance than admission information alone in several studies (18, 22–24). For example, RF classifiers trained by progressively adding information such as NIHSS and lesions detected during CT or MRI at 2 hours, 24 hours, and seven days after stroke onset, and discharge, were shown to achieve AUC values well above 0.90 as more variables from future time points were included (18). Similarly, a study performed using the MR CLEAN registry found that by training an RF classifier on clinical variables obtained on admission (e.g., age, baseline NIHSS, comorbidities, laboratory tests, baseline ASPECTS) and post EVT assessments, including the NIHSS scores and modified thrombolysis in cerebral infarction scale (mTICI), the predictive accuracy of 90-days mRS increased to an AUC of 0.91 compared to an AUC of 0.79 using only admission variables (22).

The second approach involves using more advanced AI models, such as artificial neural networks (ANN) and gradient boosting to process the information available at admission (20, 25–27). For example, when comparing models based on patient demographics and baseline clinical information, an ANN model performed better than the rule-based ASTRAL score for predicting 90-days mRS

(AUC 0.84 vs AUC 0.89) (25). However, our work showed no difference in the predictive performance of LR models to gradient boosting and ANN models trained on patient demographics, comorbidities, baseline NIHSS and the presence of IV-rtPA treatment (27). Similarly, another study that compared LR to RF, SVM, gradient boosting, and ANN classifiers reported similar performance between all classifiers, with ANN achieving the highest AUC (0.81) (26), showing that while more advanced AI models might outperform rule-based scores, they perform comparably to traditional ML methods.

We suggest that neither of the two approaches discussed above will revolutionize functional outcome prediction. Combining information from future time points with information available at admission gives high predictive performance, especially when using ANN models with huge amounts of data (28). However, these models do not help answer which treatment option is more suitable for a patient in the acute setting since the only available information in a real-life scenario will be the information at admission. DL models such as ANNs have shown significant improvements over rule-based scores, but as indicated above, their improvements over ML algorithms are minor. Additionally, the predictive performance of both ML and DL models reported across different studies is similar, suggesting that the predictive power of AI models trained only on clinical variables may have plateaued, highlighting the need to move beyond clinical variables.

Predictive models based on neuroimaging

As a standard of care, all suspected ischemic stroke patients undergo imaging before treatment (8, 9). Non-contrast CT (NCCT) is usually the first-line imaging modality due to its widespread availability and speed and is used to rule out hemorrhage or other causes of neurological symptoms. However, MRI is becoming more commonplace in the acute stroke setting in western countries (14) as it is more sensitive to ischemia than NCCT and so can be used for direct diagnosis (29). MRI also benefits from being multiparametric, with different sequences revealing different pathophysiological changes in the ischemic brain (13).

One approach for using ML to predict functional outcomes using imaging is to use radiological variables of the images as input variables. These radiological variables initially involved human input, for example, a trained radiologist calculating ASPECTS scores (27). However, this manual approach is not necessarily feasible for a clinical setting because it is time-consuming and requires a trained neuroradiologist. This limitation has been overcome by AI-enabled inventions such as the e-ASPECTS tool (Brainomix, Oxford, UK; www.brainomix.com) for the automated use of ASPECTS which has been used in functional outcome prediction studies. For example, e-ASPECTS was found to be predictive of unfavorable outcomes (mRS 4–6) after EVT 3 months post-stroke (30).

Although using radiological variables is a promising method for leveraging neuroimaging to improve functional outcome prediction, neuroimaging can offer much more for clinical decision support in acute ischemic stroke. CNNs have been shown to excel in image processing tasks (12). The novelty of DL methods, such as CNNs, is that they do not require a priori assumptions of which image features are important. DL methods can be applied directly to raw images straight from the CT or MRI scanner, allowing them to learn useful features for making predictions that the scientist or clinician may be unaware of (5, 12).

At the time of writing in August 2021, there had been little work done on developing CNN models to process acute stroke neuroimaging for functional outcome prediction. Studies so far suggest little benefit of using imaging for functional outcome prediction of ischemic stroke patients. For example, our CNN trained on whole-brain volumes of baseline Time-of-flight Magnetic Resonance Angiography (TOF-MRA) had a low predictive performance of 90-days mRS (AUC:0.64) (31). Similarly, a CNN applied to pre-EVT CTA images from the MR CLEAN registry (32) also predicted 90-days mRS with average performance (AUC:0.71) (33). In a separate analysis of the same images, LR and RF models trained on 20 radiological variables identified by clinicians showed slightly worse performance than the CNN model (LR AUC: 0.68 and RF AUC: 0.66), demonstrating the added value of DL based models applied to raw imaging data compared to ML models based on pre-defined radiological variables. Although promising, ultimately, the predictive performance reported in both studies are not high enough for the deployment of neuroimaging-based CNN models for functional outcome prediction in a clinical setting.

Considering the advances in AI that allow for multi-modal learning, a natural way to improve predictive performance is to develop DL-based models that process neuroimaging data together with clinical variables available on admission. We, therefore, developed multi-modal neural networks to jointly process whole volumes of TOF-MRA imaging together with patient demographics and clinical variables at admission for predicting 90-days mRS (Figure 2) (31). Our multimodal networks had slightly better predictive performance (AUC:0.76) compared

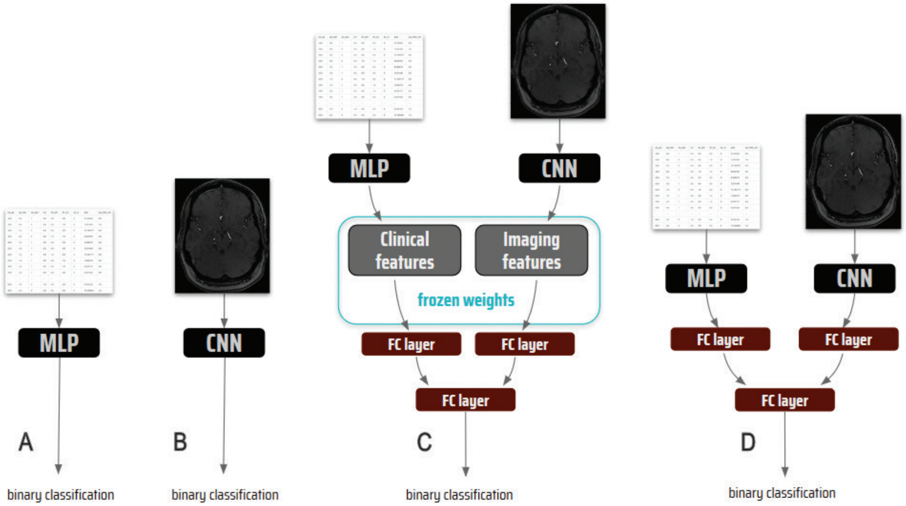

Figure 2. Neural network models. The four neural networks developed for modeling. **A,** Clinical variables; **B,** Neuroimaging; **C** and **D,** both clinical variables and neuroimaging.

to an ANN model trained on only clinical variables (AUC:0.75) and a substantially better performance than our previously mentioned TOF-MRA based CNN model (AUC:0.64). A similar study that used whole-brain NCCT images acquired on admission together with clinical variables to predict 90-days mRS demonstrated superior performance of the multi-modal neural network (AUC:0.75) compared to models separately trained on the two data modalities (AUCs:0.54 and 0.61) (34).

Overall, a multi-modal approach to predicting functional stroke outcomes can achieve better performance than imaging variables alone. However, there is a need to explore alternative fusion approaches if these models are to excel when trained using only clinical variables. Additionally, a single imaging modality such as TOF-MRA or NCCT may not be enough to capture all the information predictive of stroke outcome. Incorporating a range of images that reveal different aspects of stroke pathophysiology into the models may provide better predictive ability. For example, diffusion-based MRI for cytotoxic oedema (DWI or ADC), perfusion-based imaging for collateral flow (CT or MRI), NCCT or T2-based MRI (e.g., FLAIR, T2 relaxation) for vasogenic oedema, and angiography (CT or MRI) for vessel information. This possibility will become more feasible for research and clinical practice when AI-based MRI acquisition techniques such as magnetic resonance fingerprinting, which permits simultaneous acquisition of multiple MRI parameters at a similar speed to CT, becomes more readily available in clinical scanners (35). With the wealth of physiological information within acute stroke images and the capabilities of DL for learning new features and making predictions, we would intuitively expect that by combining ML and multiparametric MRI, the accuracy of functional outcome predictions for stroke could only improve. However, this hypothesis remains to be tested.

Moving beyond performance

In their current form, even the most advanced models do not perform well enough to be implemented in the clinical setting (4). There are several limitations to functional outcome prediction in stroke that contribute to these models' inability to go beyond a prediction accuracy ceiling. One is the lack of an established open-source data registry that combines patient information from multiple centers worldwide and different sources such as clinical assessments, brain imaging, and surveys. Another is the simplicity of using dichotomized mRS scores, whereas experts require more information from the model than a simple binary prediction for supporting their decisions (2).

Lack of transparency is identified as one of the main barriers to implementation. No matter how accurate an AI model is in its predictions, clinicians should be confident that the predictions are also trustworthy. Improving model transparency is an essential step towards trustworthy AI, along with reporting data quality and conducting external validation studies (36). Explainable AI (xAI) provides a rationale that allows clinicians to understand why an AI system has produced a particular recommendation, allowing increased model transparency. Therefore, xAI has become a popular field of research to increase the adoption of AI systems in clinical practice.

EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

xAI is an emerging field that aims to help humans understand decisions made by AI systems. The terminology around xAI is still not well-established, leading to the interchangeable use of “interpretability” and “explainability”. Here, we use interpretability to refer to a passive characteristic of a model that makes it easily understandable to a human; and explainability to refer to an active characteristic of a model that encompasses actions that aim to give details about its internal function explicitly (37). We suggest that there are two ways to achieve xAI (36):

i. *Explainable modelling* which means using models that are interpretable by design, i.e., “white-box” models. White-box models include LR classifiers and decision trees in which the internal functioning is directly accessible to the user and can be understood effortlessly (37).
ii. *Post-hoc explanations* which means converting models that are not interpretable by design, i.e., “black-box” models, into explainable ones using post-hoc explainability methods. These methods aim to enhance interpretability using text explanations, visual explanations, and feature importance explanations, among others (37). Feature importance explanations rank the explanatory power of input features on the model predictions and constitute the majority of post-hoc explainability models. For clinicians, it is helpful to learn which features are responsible for the predicted outcome to compare these with their own prior knowledge (38). The quantitative assessment of feature importance is usually made more human-readable via a visual representation of how different factors contributed to the final decision (e.g., boxplots).

xAI applications in functional stroke outcome prediction

One of the first studies investigating the relative importance of ML model variables in functional outcome prediction sorted the magnitude of model weights in descending order for each clinical variable used to train an LR model (i.e., a white-box model). They showed that the most influential predictor was baseline NIHSS, with age amongst the top four, indicating that model decisions were driven by factors compatible with prior knowledge. Following this, several studies that adopted post-hoc explainability methods with black-box models were able to reveal a similar promising outcome. For example, we applied two different feature importance explanation methods to our gradient boosting and ANN models (27). Additionally, we used the model weights of our LR models to provide a rating of features based on a white-box model for comparison. Our results (Figure 3) showed that all models rated age and baseline NIHSS consistently as the top two important features in predicting 90-days mRS. A study that compared post-hoc feature importance explanations of their RF, SVM and gradient boosting models to feature rankings of their decision trees and LR reported the same results. All models rated age and baseline NIHSS to be the top two important predictors (20).

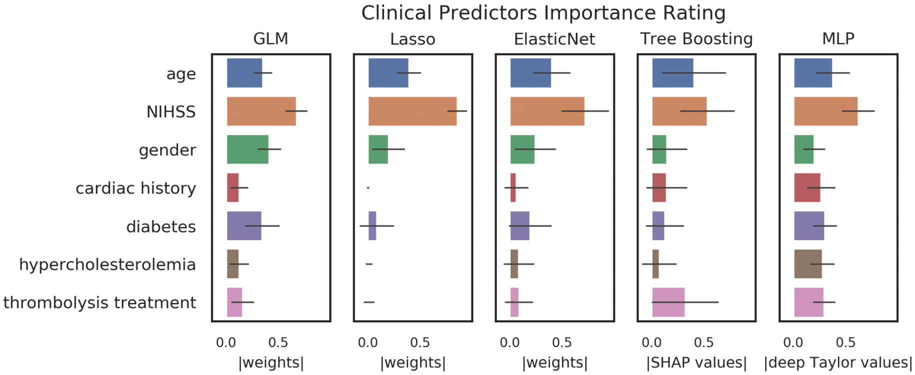

Figure 3. Importance rating. Graphical representation of the feature importance explanations derived from five AI models that were trained to predict 90-days mRS based on patient demographics, comorbidities, and presence of IV-rtPA treatment. All values were normalized to the range of [0,1] for comparability.

Furthermore, the MR CLEAN study that used baseline and post-treatment variables to train RF models provided a feature importance table which not only showed that age and baseline NIHSS was rated among the top predictors but also suggested that most top predictors did not overlap with their selection of important variables based on prior knowledge (22). This study shows the potential of post-hoc explanations in providing the researcher with previously undiscovered information. Overall, these studies indicate that black-box models can provide high predictive accuracy and reliable explanations that are compatible with prior knowledge while also helping discover new variables that may be important for functional outcome prediction.

There are relative advantages and disadvantages to using different post-hoc explainability methods for ML and DL models. The selection will depend on the type of models and the type of data used or the context in which it is used (36). In the context of healthcare, the two methods used in our study are valuable because they generate local explanations, i.e., feature relevance scores for each patient. For precision medicine, we need explanations at the patient level, where important predictors of outcome are laid out specifically for each patient.

A feature importance explanation method that is specifically designed for imaging-based CNN models and can provide local explanations is Gradient-weighted Class Activation Mapping (Grad-CAM) (39). Grad-CAM calculates the rate of change in the prediction of a target class regarding a change in the input features (i.e., pixel/voxels). By allowing the preservation of spatial information, Grad-CAM identifies regions of interest in the input image that are important for the prediction rather than individual pixel/voxels. These regions of interest can be then visualized as heat maps allowing for qualitative analysis of the important areas through visual inspection. We applied Grad-CAM on a CNN model that was trained on MRI Apparent Diffusion Coefficient (ADC) maps of 40 hyperacute ischemic stroke patients to predict 90-days mRS (AUC:0.91) (40). Figure 4 shows the heatmaps generated from the quantitative Grad-CAM analysis for eight

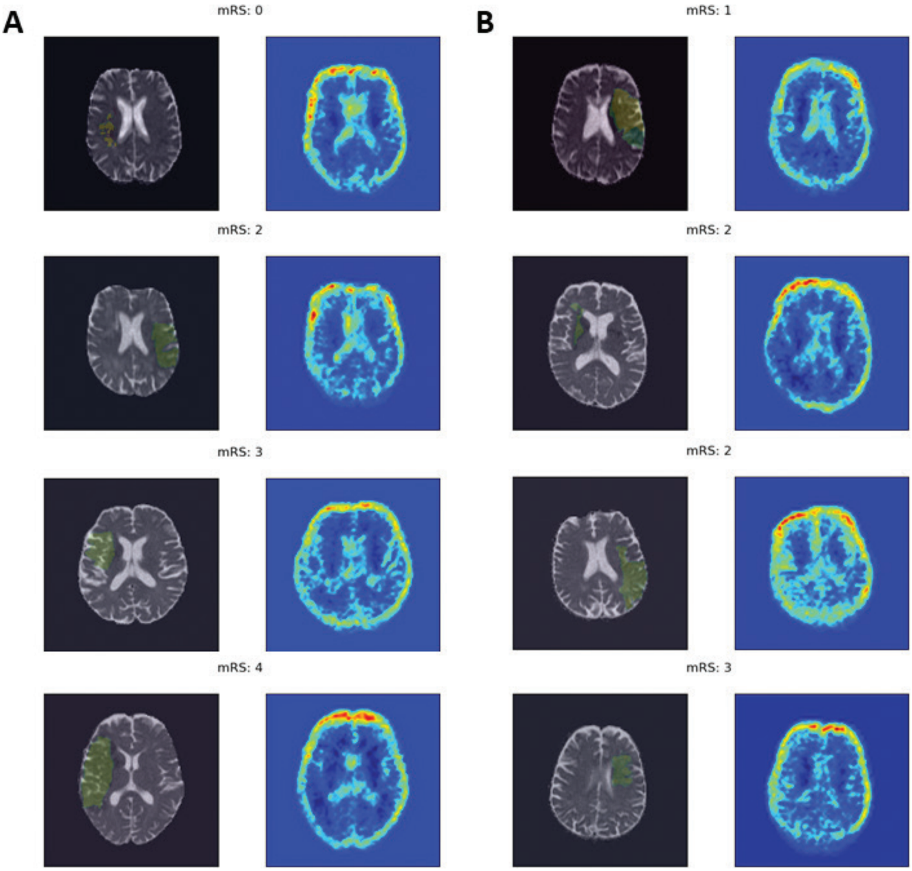

Figure 4. Activation maps. Illustration of class activation maps for correctly predicted patients from the **A,** training, and **B,** validation data. For each patient, a slice from the ADC (apparent diffusion coefficient) maps is shown with the lesion mask overlayed. The generated heatmap for that slice is shown beside the original image, where red and yellow areas indicate regions of interest.

example patients, all classified correctly by the model. The heatmaps showed that the model did not focus on the visible ischemic regions in the ADC maps but consistently focused on the boundaries of the brain. This may suggest that the model's predictions were likely based on MR artefacts rather than pathophysiological information represented in the ischemic regions of ADC maps. On the other hand, by focusing on the boundaries, the model may have discovered atrophy related to the patient's age. These results highlight that (i) high performing models are not necessarily reliable, and (ii) when the explanations do not identify imaging features that are known to be predictive of functional outcome, it is hard to determine why.

We suggest that a multidisciplinary process is needed to understand better the properties of existing models and what is clinically valid. Our study (27) showed that, complex algorithms such as neural networks correctly identified the most important clinical variables that drive the prediction of functional stroke outcome in alignment with existing literature (41). Similar work is needed for DL models applied to neuroimaging for functional stroke outcome, i.e., consensus between features found important using post-hoc explainability and the imaging properties already known to be important. This can be achieved by acknowledging the importance of information exchange between clinicians and developers and making these interactions a part of model development.

THE NEED FOR XAI FOR CDSS USED IN ACUTE STROKE CARE

The healthcare domain has unique ethical and legal challenges as decisions may considerably impact a patient's physical and mental health and financial well-being. xAI can pave the way towards trusted decisions by enforcing the deployment of explainable models. We elaborate on four questions of concern for the deployment of AI models in the clinic, how these concerns are translated to the acute stroke setting and how they may be answered through xAI (2, 6, 36).

i. *Why and when does the system fail?*
   A global understanding of the model's decisions can highlight possible confounding factors or inappropriate features that may have driven the decision. For example, our study using ADC maps (40) to predict functional outcome demonstrated that a high performing model, contrary to expectations, may have focused on imaging artefacts, or factors related to age, instead of the ischemic region. The use of xAI tools can help prevent a phenomenon like this while the model is still in the development phase, i.e., before it is adopted in another hospital setting, allowing a medical expert to detect and correct misguided decisions.

ii. *How can the model advance our understanding of the underlying neurobiological mechanisms?*
   As discussed, the primary benefit of DL algorithms is that they can intrinsically learn complex patterns from the data, eliminating the need to hand-select features based on domain knowledge (12). Thus, DL algorithms have no intrinsic constraints related to pathological plausibility or validity, potentially leading to AI systems that learn from confounding factors instead of plausible biomarkers. On the other hand, this freedom from constraints may help promote discovery, thus helping researchers build new hypotheses and theories inspired by AI models (42).

iii. *To which individuals or subgroups does the model apply?*
   Algorithmic bias is present in some of the AI models that have been applied to healthcare systems against under-represented populations (43). Explainability can help reveal certain systematic biases in an AI system and allow developers to detect and correct for these biases which may pave the way towards impartiality in decision making.

iv. *Who is responsible?*
There are important ethical and legal discussions around clinicians' accountability and patients' autonomy concerning decisions made by a computer system. In 2021, there is no consensus on whether disclosing the use of a black-box AI system should be mandatory for informed consent. Similarly, there is no precise legal regulation requiring explainability during the development of AI systems that inform medical treatment (6). However, understanding the reasoning behind an AI system's decision will enable shared decision making between the patient and the clinician, allowing for an equal share of responsibility and increased patient autonomy. In the acute stroke setting, shared decision making may be difficult due to clinicians having to make time-pressuring treatment decisions and the possibility of patients not having the required cognitive abilities, which further emphasizes the need for transparent systems.

Ultimately, xAI is essential when developing CDSSs for acute stroke care. Whether xAI should be acquired using white-box models that may have limited predictive power but offer complete transparency or using high-performing, black-box models with suitable post-hoc explainability techniques depends on the context of the CDSS design. Regarding functional outcome prediction, some studies have favored the use of white-box models when possible (18, 20), especially when using only clinical variables where performances between different ML/DL models do not vary much. However, our findings suggest that the predictive power of ML models has reached a natural limit due to the type of data they can process. The field needs to move toward a data-driven multi-modal learning approach, which can only be achieved by using DL algorithms that capture rich patterns from unstructured data. Therefore, we are in favor of using post-hoc explainability methods together with black-box DL models and encourage potential users of CDSSs to understand that post-hoc explanations are only approximations of the inner mechanisms of DL and cannot provide full transparency (36).

CONCLUSION

In this chapter, we have identified studies that applied AI models to predict functional stroke outcome in terms of 90-days mRS using a variety of structured and unstructured data. We showed that transparency in ML/DL models is possible to a certain level and feature importance explanations is a popular way to achieve this. We suggest that xAI is essential for developing models that can overcome the limitations of the predictive performance ceiling while providing reliable decisions.

Acknowledgment: This research was supported by the PRECISE4Q project, funded through the European Union's Horizon 2020 research and innovation program under grant agreement No. 777107, and the ADAPT Research Centre, funded by Science Foundation Ireland (Grant 13/RC/2106_P2) and is co-funded by the European Regional Development fund.

Conflict of Interest: The authors declare no potential conflicts of interest with respect to research, authorship and/or publication of this manuscript.

Copyright and Permission Statement: The authors confirm that the materials included in this chapter do not violate copyright laws. Where relevant, appropriate permissions have been obtained from the original copyright holder(s), and all original sources have been appropriately acknowledged or referenced.

REFERENCES

1. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271–97. https://doi.org/10.1016/S2589-7500(19)30123-2
2. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. Appl Sci. 2021;11(11):5088. https://doi.org/10.3390/app11115088
3. Mouridsen K, Thurner P, Zaharchuk G. Artificial Intelligence Applications in Stroke. Stroke. 2020;51(8):2573–9. https://doi.org/10.1161/STROKEAHA.119.027479
4. Goyal M, Ospel JM, Kappelhof M, Ganesh A. Challenges of Outcome Prediction for Acute Stroke Treatment Decisions. Stroke. 2021;52(5):1921–8. https://doi.org/10.1161/STROKEAHA.120.033785
5. McGarry BL, Kauppinen RA. Timing the Ischemic Stroke by Multiparametric Quantitative Magnetic Resonance Imaging. In: Dehkharghani S, editor. Stroke [Internet]. Brisbane (AU): Exon Publications; 2021;79–96. https://doi.org/10.36255/exonpublications.stroke.timingischemicstroke.2021
6. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1):310. https://doi.org/10.1186/s12911-020-01332-6
7. World Health Organization [Internet]. [cited 2021 Sep 14]. Available from: https://www.who.int/classifications/icf/icfbeginnersguide.pdf
8. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, et al. Guidelines for the Early Management of Patients With Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. Stroke. 2019;50(12):e344–418. https://doi.org/10.1161/STR.0000000000000211
9. Berge E, Whiteley W, Audebert H, De Marchis G, Fonseca AC, Padiglioni C, et al. European Stroke Organisation (ESO) guidelines on intravenous thrombolysis for acute ischaemic stroke. Eur Stroke J. 2021;6(1):I-LXII. https://doi.org/10.1177/2396987321989865
10. Saver JL. Time Is Brain-Quantified. Stroke. 2006;37(1):263–6. https://doi.org/10.1161/01.STR.0000196957.55928.ab
11. Barber PA, Demchuk AM, Zhang J, Buchan AM. Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy. ASPECTS Study Group. Alberta Stroke Programme Early CT Score. Lancet Lond Engl. 2000;355(9216):1670–4. https://doi.org/10.1016/S0140-6736(00)02237-6
12. Kelleher JD. Deep learning. Cambridge, Massachusetts: The MIT Press; 2019. 280 p. (The MIT press essential knowledge series).
13. Kauppinen RA. Multiparametric magnetic resonance imaging of acute experimental brain ischaemia. Prog Nucl Magn Reson Spectrosc. 2014;80:12–25. https://doi.org/10.1016/j.pnmrs.2014.05.002
14. McGarry BL. A preclinical and clinical investigation into quantitative magnetic resonance imaging as a tool for estimating onset time in hyperacute ischaemic stroke. University of Bristol, UK. 2020; PhD thesis.

15. Kelleher JD, Mac Namee B, D'Arcy A. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. Cambridge, Massachusetts: The MIT Press; 2015. 595 p.
16. Broderick JP, Adeoye O, Elm J. Evolution of the Modified Rankin Scale and Its Use in Future Stroke Trials. Stroke. 2017;48(7):2007–12. https://doi.org/10.1161/STROKEAHA.117.017866
17. Quinn TJ, Dawson J, Walters MR, Lees KR. Exploring the reliability of the modified rankin scale. Stroke. 2009;40(3):762–6. https://doi.org/10.1161/STROKEAHA.108.522516
18. Monteiro M, Fonseca AC, Freitas AT, Pinho e Melo T, Francisco AP, Ferro JM, et al. Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients. IEEE/ACM Trans Comput Biol Bioinform. 2018;15(6):1953–9. https://doi.org/10.1109/TCBB.2018.2811471
19. Ntaios G, Faouzi M, Ferrari J, Lang W, Vemmos K, Michel P. An integer-based score to predict functional outcome in acute ischemic stroke: The ASTRAL score. Neurology. 2012;78(24):1916–22. https://doi.org/10.1212/WNL.0b013e318259e221
20. Alaka SA, Menon BK, Brobbey A, Williamson T, Goyal M, Demchuk AM, et al. Functional Outcome Prediction in Ischemic Stroke: A Comparison of Machine Learning Algorithms and Regression Models. Front Neurol. 2020;11:889. https://doi.org/10.3389/fneur.2020.00889
21. Asadi H, Dowling R, Yan B, Mitchell P. Machine Learning for Outcome Prediction of Acute Ischemic Stroke Post Intra-Arterial Therapy. Gómez S, editor. PLoS ONE. 2014;9(2):e88225. https://doi.org/10.1371/journal.pone.0088225
22. van Os HJA, Ramos LA, Hilbert A, van Leeuwen M, van Walderveen MAA, Kruyt ND, et al. Predicting Outcome of Endovascular Treatment for Acute Ischemic Stroke: Potential Value of Machine Learning Algorithms. Front Neurol. 2018;9:784. https://doi.org/10.3389/fneur.2018.00784
23. Xie Y, Jiang B, Gong E, Li Y, Zhu G, Michel P, et al. JOURNAL CLUB: Use of Gradient Boosting Machine Learning to Predict Patient Outcome in Acute Ischemic Stroke on the Basis of Imaging, Demographic, and Clinical Information. Am J Roentgenol. 2019;212(1):44–51. https://doi.org/10.2214/AJR.18.20260
24. Brugnara G, Neuberger U, Mahmutoglu MA, Foltyn M, Herweh C, Nagel S, et al. Multimodal Predictive Modeling of Endovascular Treatment Outcome for Acute Ischemic Stroke Using Machine-Learning. Stroke. 2020;51(12):3541–51. https://doi.org/10.1161/STROKEAHA.120.030287
25. Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke. Stroke. 2019;50(5):1263–1265. https://doi.org/10.1161/STROKEAHA.118.024293
26. Ramos LA, Kappelhof M, van Os HJA, Chalos V, Van Kranendonk K, Kruyt ND, et al. Predicting Poor Outcome Before Endovascular Treatment in Patients With Acute Ischemic Stroke. Front Neurol. 2020;11:580957. https://doi.org/10.3389/fneur.2020.580957
27. Zihni E, Madai VI, Livne M, Galinovic I, Khalil AA, Fiebach JB, et al. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. Stoean R, editor. Plos One. 2020;15(4):e0231166. https://doi.org/10.1371/journal.pone.0231166
28. Lin C-H, Hsu K-C, Johnson KR, Fann YC, Tsai C-H, Sun Y, et al. Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. Comput Methods Programs Biomed. 2020;190:105381. https://doi.org/10.1016/j.cmpb.2020.105381
29. Chalela JA, Kidwell CS, Nentwich LM, Luby M, Butman JA, Demchuk AM, et al. Magnetic resonance imaging and computed tomography in emergency assessment of patients with suspected acute stroke: a prospective comparison. Lancet. 2007;369(9558):293–8. https://doi.org/10.1016/S0140-6736(07)60151-2
30. Pfaff J, Herweh C, Schieber S, Schönenberger S, Bösel J, Ringleb PA, et al. e-ASPECTS Correlates with and Is Predictive of Outcome after Mechanical Thrombectomy. Am J Neuroradiol. 2017;38(8):1594–9. https://doi.org/10.3174/ajnr.A5236
31. Zihni E, Madai V, Khalil A, Galinovic I, Fiebach J, Kelleher J, et al. Multimodal Fusion Strategies for Outcome Prediction in Stroke: In: Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies [Internet]. Valletta, Malta: SCITEPRESS - Science and Technology Publications; 2020 [cited 2021 Sep 8]. p. 421–8. Available from: https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0008957304210428 https://doi.org/10.5220/0008957304210428

32. Berkhemer OA, Fransen PSS, Beumer D, van den Berg LA, Lingsma HF, Yoo AJ, et al. A Randomized Trial of Intraarterial Treatment for Acute Ischemic Stroke. N Engl J Med. 2015;372(1):11–20. https://doi.org/10.1056/NEJMoa1411587
33. Hilbert A, Ramos LA, van Os HJA, Olabarriaga SD, Tolhuisen ML, Wermer MJH, et al. Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. Comput Biol Med. 2019;115:103516. https://doi.org/10.1016/j.compbiomed.2019.103516
34. Bacchi S, Zerner T, Oakden-Rayner L, Kleinig T, Patel S, Jannes J. Deep Learning in the Prediction of Ischaemic Stroke Thrombolysis Functional Outcomes. Acad Radiol. 2020;27(2):e19–23. https://doi.org/10.1016/j.acra.2019.03.015
35. Ma D, Gulani V, Seiberlich N, Liu K, Sunshine JL, Duerk JL, et al. Magnetic resonance fingerprinting. Nature. 2013;495(7440):187–92. https://doi.org/10.1038/nature11971
36. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform. 2021;113:103655. https://doi.org/10.1016/j.jbi.2020.103655
37. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012
38. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. ArXiv190505134 Cs Stat [Internet]. 2019 Aug 7 [cited 2021 Sep 14]; Available from: http://arxiv.org/abs/1905.05134
39. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV) [Internet]. Venice: IEEE. https://doi.org/10.1109/ICCV.2017.74 2017 [cited 2021 Sep 11]. p. 618–26. Available from: http://ieeexplore.ieee.org/document/8237336/
40. Zihni E, Kelleher JD, McGarry B. An Analysis of the Interpretability of Neural Networks trained on Magnetic Resonance Imaging for Stroke Outcome Prediction. 2021 [cited 2021 Sep 15]; Available from: https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1327&context=scschcomcon&preview_mode=1&force=yes
41. Weimar C, König IR, Kraywinkel K, Ziegler A, Diener HC. Age and National Institutes of Health Stroke Scale Score Within 6 Hours After Onset Are Accurate Predictors of Outcome After Cerebral Ischemia: Development and External Validation of Prognostic Models. Stroke. 2004;35(1):158–62. https://doi.org/10.1161/01.STR.0000106761.94985.8B
42. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Appl. 2020 Dec;32(24):18069–83. https://doi.org/10.1007/s00521-019-04051-w
43. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53. https://doi.org/10.1126/science.aax2342