# Glossary of technical terms

## Anticoagulants

Anticoagulants are medicines that help prevent blood clots. They may be used to help prevent stroke that is caused by atrial fibrillation. Atrial fibrillation is a condition where the heart beats irregularly and can cause blood to pool in the heart. This can increase the risk of blood clots forming, which can lead to stroke. Anticoagulants are given to people with atrial fibrillation to help prevent these blood clots from forming. They work by interfering with the blood’s ability to clot, which can reduce the risk of stroke. Anticoagulants need to be carefully monitored to make sure they are working effectively and not causing any unwanted side effects.

Thrombolysis is not usually given to patients on anticoagulants, though a doctor may to choose to use it in some patients, for example, if they have tested the blood’s ability to clot, which may be measured and reported as an INR (International Normalized Ratio).

## Artificial patients

Artificial patients are patients where we make up their characteristics. We might, for example, make up a patient as follows:

* Infarction (stroke caused by clot) = Yes
* Stroke severity (NIHSS) = 15
* Precise onset time known = Yes
* Disability (mRS) before stroke = 0
* Age = 72
* Onset-to-arrival = 60 minutes
* Arrival-to-scan = 15 minutes
* Use of anticoagulants for AF = No
* Onset during sleep = No

This allows us to explore how changing any feature or features of the patient will change the likelihood that such a patient would receive thrombolysis in each hopsital.

## Explainable machine learning

Explainable Machine Learning (XAI) is a field of research that aims to make machine learning models more transparent and interpretable to humans. In plain English, it’s about making sure that the decisions made by machine learning models are understandable and trustworthy to people.

Traditional machine learning models often operate as black boxes, where it can be difficult to understand how the model arrived at its predictions or decisions. XAI aims to address this issue by providing a way to explain how a model makes decisions and what factors are contributing to those decisions.

## FAST

“FAST” is an acronym used to recognize the signs and symptoms of a stroke, and to prompt people to seek medical attention immediately. The acronym stands for:

1. Face drooping: If one side of the face droops or feels numb, it could be a sign of a stroke.
2. Arm weakness: If one arm is weak or numb, it could be a sign of a stroke.
3. Speech difficulty: If speech is slurred or difficult to understand, it could be a sign of a stroke.
4. Time to call emergency services: If any of the above symptoms are present, it’s important to call emergency services (such as 911 in the United States) immediately.

A “FAST” positive in stroke refers to the presence of one or more of these symptoms, which indicates a possibility of a stroke.

## Gradient boosting decision trees

Gradient boosting trees is a machine learning algorithm that combines multiple decision trees to make accurate predictions. In plain English, it’s a technique that allows a computer to learn from examples and make predictions about new data.

The algorithm works by creating a series of decision trees, each one building on the previous tree’s predictions. At each step, the algorithm identifies the features that are most important for making accurate predictions and creates a new tree that corrects for the errors of the previous trees.

To create the decision trees, the algorithm starts with a single tree that makes predictions based on a subset of the available features. It then calculates the error of the predictions and uses that information to build a second tree that corrects for the errors of the first tree. This process continues until a set number of trees have been created, or until the error rate is minimized.

The final model is a combination of all the decision trees, with each tree contributing its own unique perspective on the data. The algorithm uses a technique called gradient descent to optimize the model, adjusting the parameters of the decision trees to minimize the error rate.

Gradient boosting trees is a powerful machine learning algorithm that can be used for a wide range of applications, including image recognition, natural language processing, and predictive analytics. It’s particularly useful for problems where there are many variables and complex relationships between them, and where accurate predictions are critical.

We use a type of gradient boosting decision tree model called XGBoost (eXtreme Gradient Boost).

## Haemorrhage

A haemorrhage is a bleed from a blood vessel within the body.

2 out of 10 strokes are caused by a bleed in the brain. These patients must not be given thrombolysis (which is used to dissolve clots), otherwise the bleed is likely to be made much worse.

## Infarction and ischaemic stroke

An infarction is a blood clot that leads to loss of blood supply to an area of the body. In the case of stroke an infarction is a clot in a blood vessel in the brain that leads to loss of blood supply to the area of the brain fed by that blood vessel. In stroke, if the patient arrives in hospital in time they may be suitable for thrombolysis to dissolve the clot.

A stroke caused by an infarction is also known as an ischaemic stroke, which is another way to describe a stroke caused by a blockage in a blood vessel.

8 out of 10 strokes are cause by a clot.

## Machine learning

Machine learning is a type of artificial intelligence that involves teaching machines to learn from data, rather than being explicitly programmed to perform specific tasks. In plain English, machine learning is a way to teach computers how to make predictions or decisions based on patterns in data, similar to how humans learn from experience.

The process of machine learning involves training a model using a large dataset, which contains examples of inputs and corresponding outputs. The model learns from these examples and makes predictions or decisions based on new input data. As the model is exposed to more data, it can refine its predictions or decisions, improving its accuracy and reliability.

## Modified Rankin Scale (MRS)

The Modified Rankin Scale (MRS) is a widely used clinical tool to measure the degree of disability or dependence in patients who have suffered a stroke or other neurological disorders.

The MRS scores range from 0 to 6, with 0 indicating no symptoms and 6 indicating death. The scale includes six levels of disability, with each level reflecting a different degree of dependence:

1. No symptoms.
2. No significant disability, able to carry out all usual duties and activities.
3. Slight disability, able to look after oneself but unable to carry out all previous activities.
4. Moderate disability, requires some help but able to walk unassisted.
5. Moderately severe disability, unable to walk unassisted and requires constant care.
6. Severe disability, bedridden and requires constant nursing care.
7. Death.

The MRS is commonly used in clinical practice, research studies, and clinical trials as a measure of functional outcome after stroke or other neurological disorders. It provides a simple and standardized way to assess the patient’s level of disability or dependence and can be used to track the patient’s progress over time.

The MRS is a useful tool for healthcare professionals in making treatment decisions and for patients and their families in understanding the severity of the disability and the potential for recovery. It also helps to standardize the assessment of disability across different clinical settings and research studies.

## NIHSS

NIHSS stands for the National Institutes of Health Stroke Scale. It is a standardized tool used by healthcare professionals to assess and measure the severity of neurological deficits in patients who have suffered a stroke.

The NIHSS includes 15 different assessments that evaluate different aspects of the patient’s neurological function, such as their level of consciousness, vision, movement, speech, and language abilities. Each assessment is scored on a scale of 0 to 4 or 0 to 3, depending on the assessment, with higher scores indicating more severe deficits.

The total score on the NIHSS can range from 0 to 42, with higher scores indicating more severe stroke symptoms. The NIHSS is often used by healthcare professionals to determine the appropriate treatment plan for stroke patients, as well as to track the patient’s progress over time.

The NIHSS is a widely recognized and accepted tool for evaluating stroke patients and is used in clinical practice, research studies, and clinical trials. It helps healthcare professionals to quickly and objectively assess the patient’s neurological function and provide appropriate treatment based on the severity of the stroke.

## Receiver operating characteristic (ROC) curve

A Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classification model. In plain English, it shows how well a model is able to distinguish between two classes by plotting the trade-off between sensitivity (the ability of the model to correctly identify positive cases) and specificity (the ability of the model to correctly identify negative cases).

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds. The TPR represents the proportion of positive cases that are correctly identified by the model, while the FPR represents the proportion of negative cases that are incorrectly classified as positive by the model.

A perfect model would have a TPR of 1 and an FPR of 0, resulting in a point in the top left corner of the ROC curve. A random model, on the other hand, would have a diagonal line from the bottom left to the top right of the graph, indicating that the TPR and FPR are roughly equal.

The area under the ROC curve (AUC) is often used as a summary statistic to compare the performance of different models. An AUC of 0.5 represents a random model, while an AUC of 1.0 represents a perfect model. In general, a higher AUC indicates better discrimination between the two classes.

## R-squared

R-squared, also known as the coefficient of determination, is a statistical measure that indicates how well a regression model fits the data. It is a number between 0 and 1, where 1 represents a perfect fit and 0 represents no relationship between the predictor variables and the response variable.

In plain English, R-squared measures how much of the variation in the response variable can be explained by the predictor variables in the model. For example, if R-squared is 0.80, it means that 80% of the variability in the response variable can be explained by the predictor variables in the model, and the remaining 20% is unexplained and may be due to random variation or other factors not accounted for in the model.

Another way to think of R-squared is as a measure of the goodness of fit of the model. A higher R-squared indicates that the model is better at predicting the response variable, while a lower R-squared indicates that the model is not a good fit for the data and may need to be revised or improved. However, it is important to note that a high R-squared does not necessarily mean that the model is a good predictor of future outcomes or that it is the best model for the data. Other factors, such as model complexity and overfitting, also need to be considered when evaluating a regression model.

## Scan (CT scan)

In emergency stroke treatment, a computed tomography (CT) scan is typically used as the first imaging test to quickly evaluate the brain and determine whether a person has had an ischemic or hemorrhagic stroke.

A CT scan is a non-invasive imaging test that uses X-rays to create detailed images of the brain. It is a fast and widely available test that can help identify the type and location of the stroke, which is important in determining the best treatment approach.

If the CT scan shows that the stroke is caused by a blood clot, the patient may be a candidate for treatment with a clot-busting medication called tissue plasminogen activator (tPA). In some cases, a magnetic resonance imaging (MRI) scan may also be used to provide additional information about the stroke and help guide treatment decisions.

## Sensitivity ans specificity

In machine learning, sensitivity and specificity are two important measures of the performance of a model that is classifying examples into two categories.

*Sensitivity* is a measure of how well a model can correctly identify *positive* instances, also known as true positives. It is the proportion of true positive results among all actual positive instances. In other words, sensitivity measures the ability of a model to correctly identify the instances that belong to a certain class. For example, in a medical diagnosis problem, sensitivity would measure the percentage of patients who truly have a certain disease that are correctly identified by the model. High sensitivity means that few people with a disease will be missed.

*Specificity* is a measure of how well a model can correctly identify *negative* instances, also known as true negatives. It is the proportion of true negative results among all actual negative instances. For example, in a medical diagnosis problem, specificity would measure the percentage of patients who do not have a certain disease that are correctly identified by the model. High specificity will mean that few patients who do not have a disease will be told that they may have that disease.

In our case, *sensitivity* reports the proportion of patients who received thrombolysis that our model predicted would receive thrombolysis, *specificity* reports the proportion of patients who did not receive thrombolysis that our model predicted would not receive thrombolysis.

## Shapley values and SHAP

Shapley values are used to explain the contribution of each feature or variable to the prediction of a model. Shap values are therefore a type of Explainable Machine Learning (XAI).

Essentially, the contribution of each feature or variable is measured by comparing the prediction with and without that feature or variable included. For example, if we have a model that predicts the price of a house based on several features such as the number of bedrooms, square footage, and location, we can use Shapley values to determine how much each feature contributes to the predicted price. By comparing the prediction with and without each feature included, we can calculate the Shapley value for each feature and determine which features have the most impact on the prediction.

In practice Shapley values are a little more complicated to calculate as the method looks at many combinations of used and non-used features for each instance (rather than only leaving a sngle feature out at a time). This makes Shapley values quite hard and time-consuming to calculate. SHAP is a package that estimates Shapley Values efficiently.

## Stratified k-fold cross-validation

When dividing your dataset into a train and test set, how much of the result is due to that particular division of the instances between these two sets? Stratified k-fold cross-validation is a method to divide your data into training and test sets a number of times, such that each instance is in one, but only one, test set. By fitting a model to each of these train-test sets allows the model performance to be measured under different splits, and the results robustness can be determined by the consistency of the results.

## Stroke

A stroke, also known as a cerebrovascular accident (CVA), is a medical condition that occurs when blood flow to a part of the brain is interrupted or reduced, leading to brain cell damage or death. The interruption or reduction in blood flow can be caused by a blockage or rupture of a blood vessel in the brain.

There are two main types of stroke: ischemic stroke and hemorrhagic stroke. Ischaemic stroke is the most common type and occurs when a blood clot blocks a blood vessel in the brain. Hemorrhagic stroke, on the other hand, occurs when a blood vessel in the brain ruptures or leaks.

Symptoms of a stroke may include sudden weakness or numbness in the face, arm, or leg, especially on one side of the body; sudden confusion or trouble speaking or understanding speech; sudden vision problems in one or both eyes; sudden dizziness or loss of balance or coordination; and severe headache with no known cause.

A stroke is a medical emergency, and rapid treatment is crucial to minimize brain damage and prevent long-term disability or death. Treatment may include medications to dissolve blood clots or control bleeding, surgery to repair ruptured blood vessels or remove blood clots, and rehabilitation to regain lost skills and functions.

## Threshold (for a two-class classification model)

Many machine learning models, such as XGBoost, return a probability of recieving thrombolysis, and we use a nominal threshold value to turn this probability into a prediction. By default, this is the midpoint, 0.5. A different threshold can be used. By choosing a lower threshold value (between 0.0-0.5) fewer instances will be classified as 0 (not recieve thrombolysis) and more will be classified as 1 (recieve thrombolysis). By doing so we will create more false alarms (false positives), but also classify more true positives. By choosing a higher threshold value (between 0.5-1.0) more instances will be classified as 0 (not recieve thrombolysis) and fewer will be classified as 1 (recieve thrombolysis). By doing so will create more undetected positives (false negatives), but also identify more true negatives. Choosing a threshold value is balancing these two cases.

## Thrombolysis

Thrombolysis in stroke is a medical treatment aimed at dissolving blood clots that block blood vessels in the brain, causing an ischemic stroke. The goal of thrombolysis is to restore blood flow to the affected area of the brain as quickly as possible to minimize brain damage and improve the chances of a good recovery.

The treatment involves administering a medication called a thrombolytic agent, which is typically given through an IV line in the arm. The most commonly used thrombolytic agent is tissue plasminogen activator (tPA), which works by breaking down the blood clot and restoring blood flow to the brain.

Thrombolysis is considered a time-critical treatment, and it is most effective when given within the first few hours after the onset of stroke symptoms. However, the treatment also carries some risks, such as bleeding in the brain, so it is not suitable for all patients. The decision to use thrombolysis must be made by a healthcare professional based on careful evaluation of the individual’s medical history, symptoms, and other factors.

## XGBoost

XGBoost stands for eXtreme Gradient Boost. It is a type of machine learning algorithm which uses many decision trees to arrive at a prediction for an instance. A decision tree is a way of making decisions based on a set of conditions or rules. Think of it like a game of 20 Questions, where you start with a broad question and keep asking more specific questions until you arrive at an answer.