Investigating variation in clinical decision-making with explainable AI

A project with the National Stroke Audit of England & Wales

Michael Allen [1, 2] Kerry Pearn [1, 2]

[1]University of Exeter Medical School

[2]NIHR Applied Research Collaboration South West Peninsula (PenARC)

July 2022

Acknowledgements

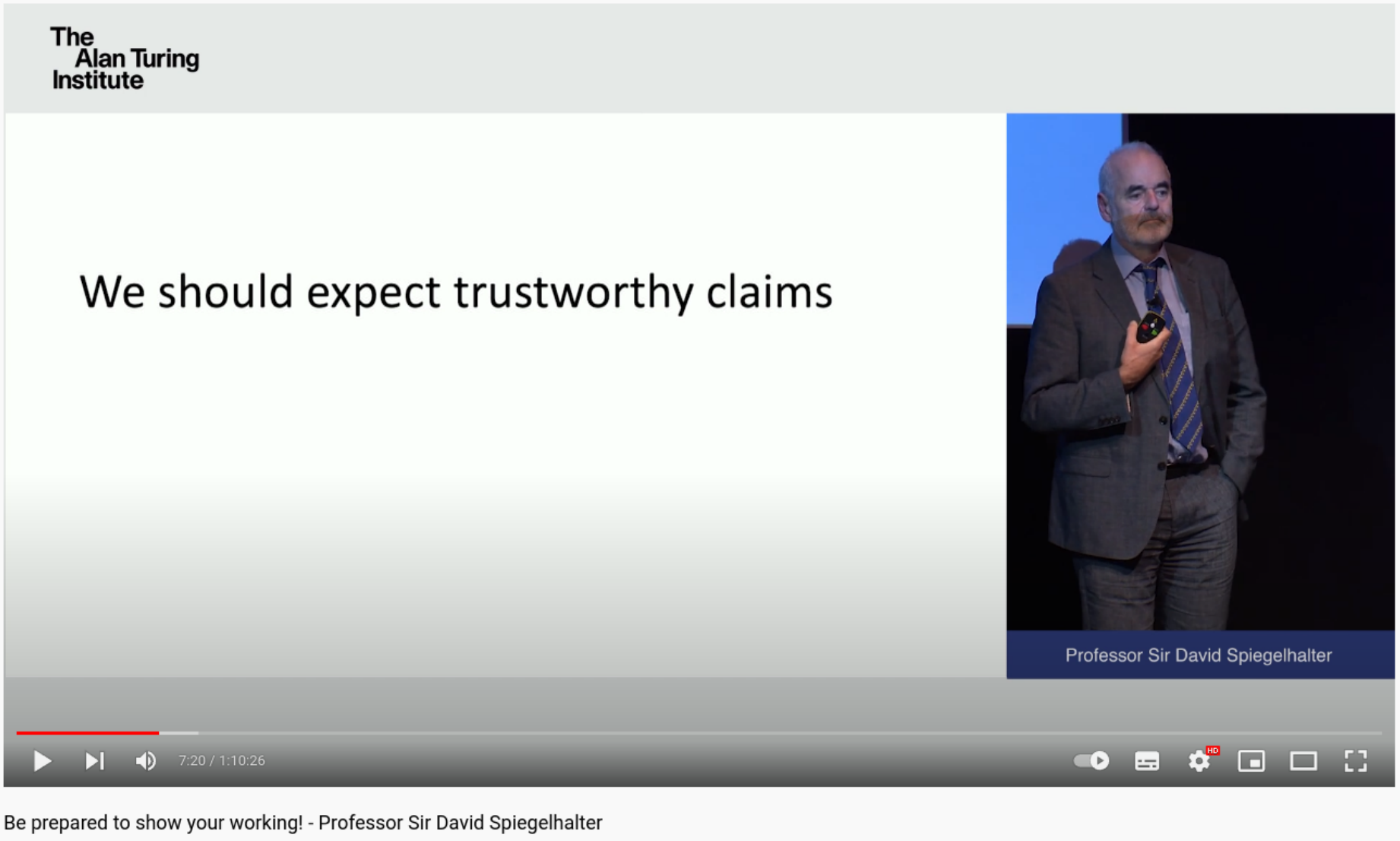
- Charlotte James (earlier machine learning work)
- Martin James (stroke physician)
- Richard Everson (machine learning advice and all-round guru)
- Kristin Liabo (qualitative work with physicians, and PPI)
- Julia Frost (qualitative work with physicians)
- Leon Farmer and Penny Thompson (PPI reps)
- Ken Stein (clinical and project management advice)
- Anna Laws (recently joined us and is working on improved outcome models and synthetic data)

This work was funded by NIHR Health Service and Delivery Research.

‘Show your workings’

‘Be prepared to show your workings’ (David Spiegelhalter)

https://youtu.be/E12_F4xeOHw



‘Be prepared to show your workings’

Models should be:

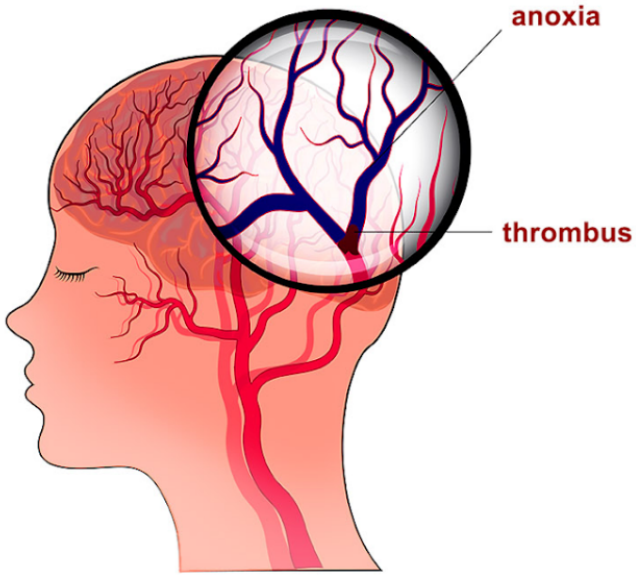
- *Trustworthy*: Can the model be trusted? Can the claims about the model be trusted?
- *Accessible*: Can people find information on it?
- *Intelligible*: Can people understand the general mechanics of the model?
- *Assessable*: Can other people check the workings?
- *Fair*: Is the model free of any unjust bias?
- *Explainable*: See panel on right.

Model explainability:

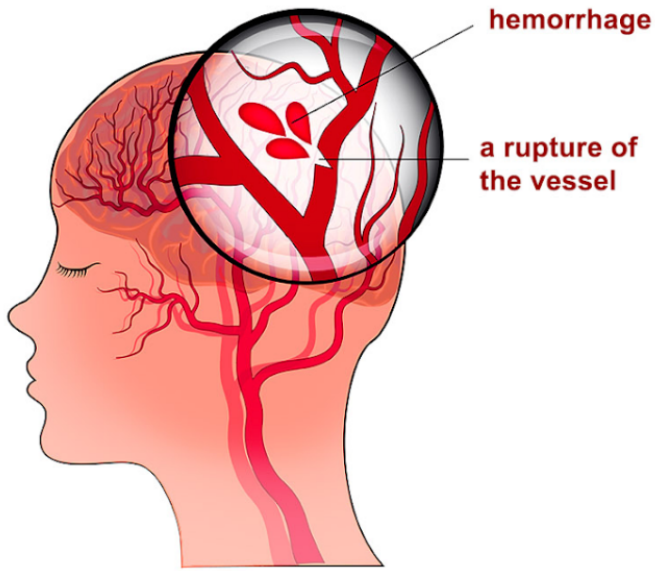
- What is the pedigree of the model type?
- What training data was used? How representative is it?
- How was accuracy and fairness evaluated? How certain are the predictions? Are the uncertainties well calibrated?
- *Global explainability*: What are the most influential items of information generally? How do they influence the model?
- *Local explainability*: What are the most influential items for any particular prediction? How do they influence the prediction?

The problem

Causes and treatment of stroke



Ischemic Stroke

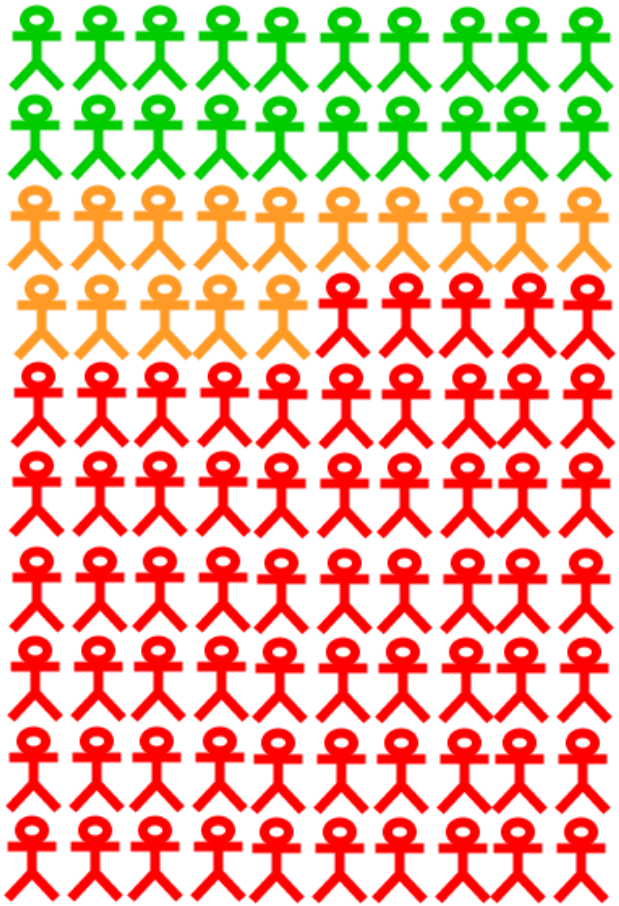


Hemorrhagic Stroke

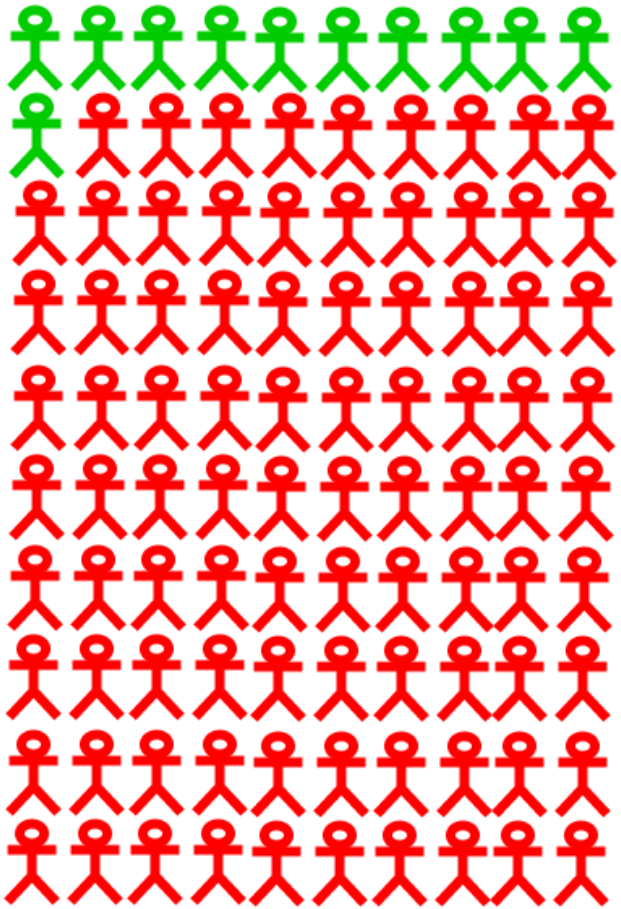
80% to 85% of strokes are *ischaemic* strokes (caused by a blood clot) and are candidates for clot-busting medicine (*thrombolysis*) if it can be given with 4 hours of stroke onset and there are no other contradictions to treatment.

The gap between clinical targets and reality

What should be happening?



What is happening?

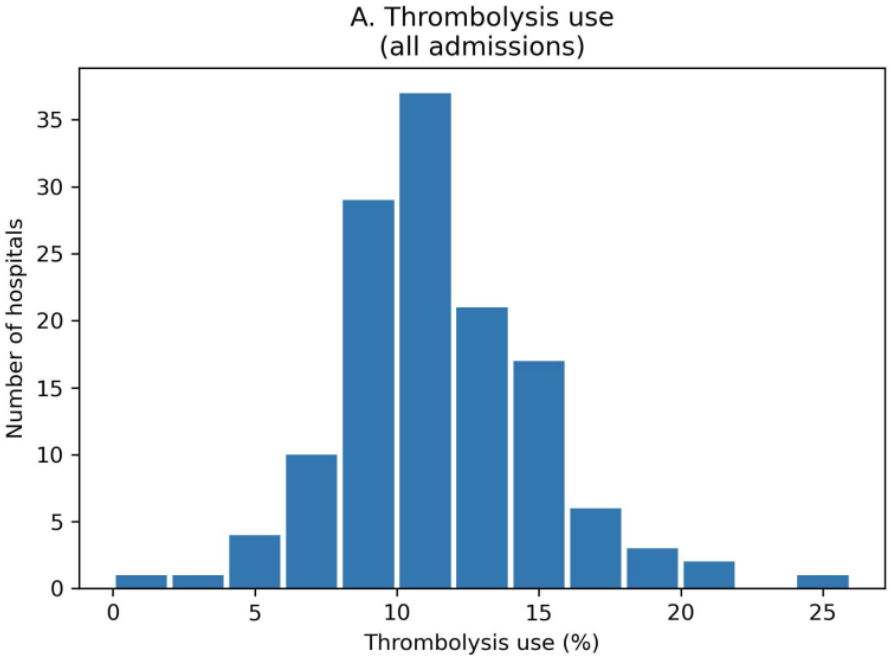


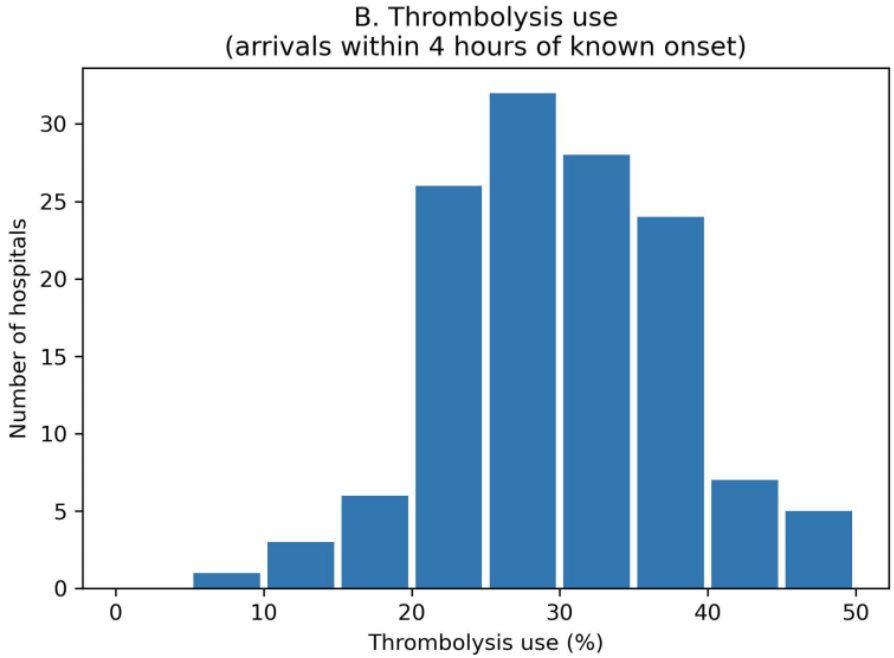
Suitable for thrombolysis?

- Yes, recieves
- Possibly
- No

Source: National Stroke Audit

Hospital thrombolysis rates vary significantly

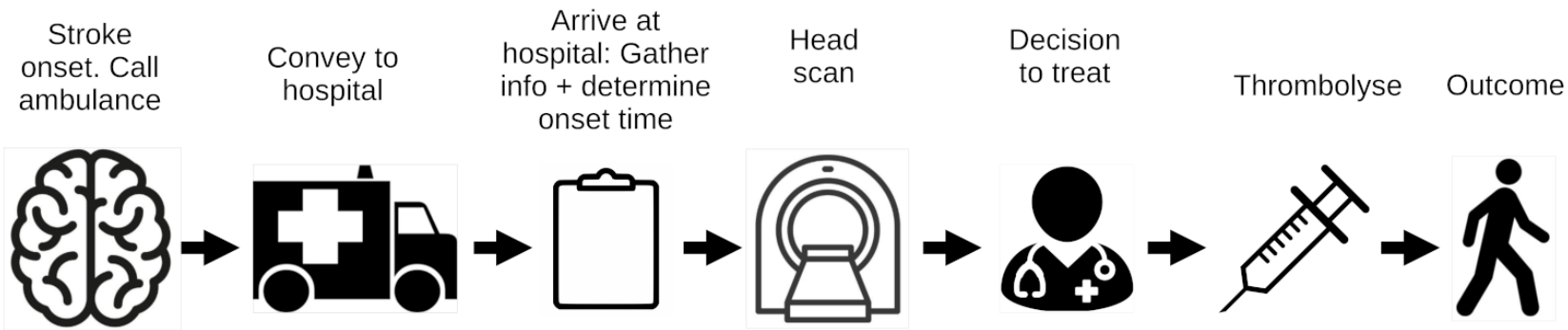




What causes this variation?

- Variation in patients?
- Variation in pathway processes?
- Variation in clinical decision-making?

Breaking down the emergency stroke pathway into key steps

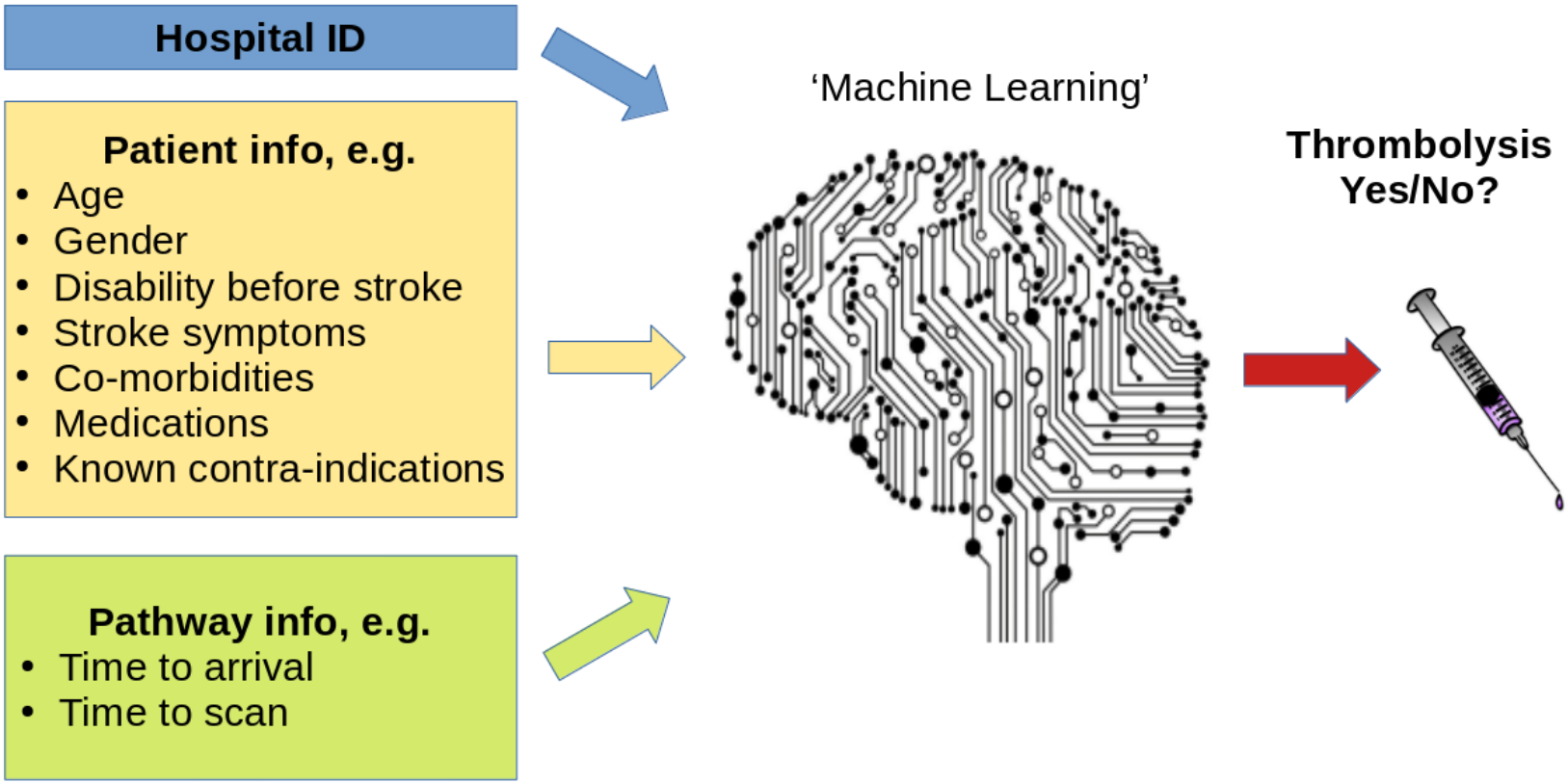


We can model key changes to pathway:

- What if the pathway were faster?
- What if hospital determined the stroke onset time in more patients?
- What if clinical decision-making was like that of *benchmark* hospitals? (Predict what treatment a patient would receive at other hospitals).

Learning clinical decision-making at hospitals

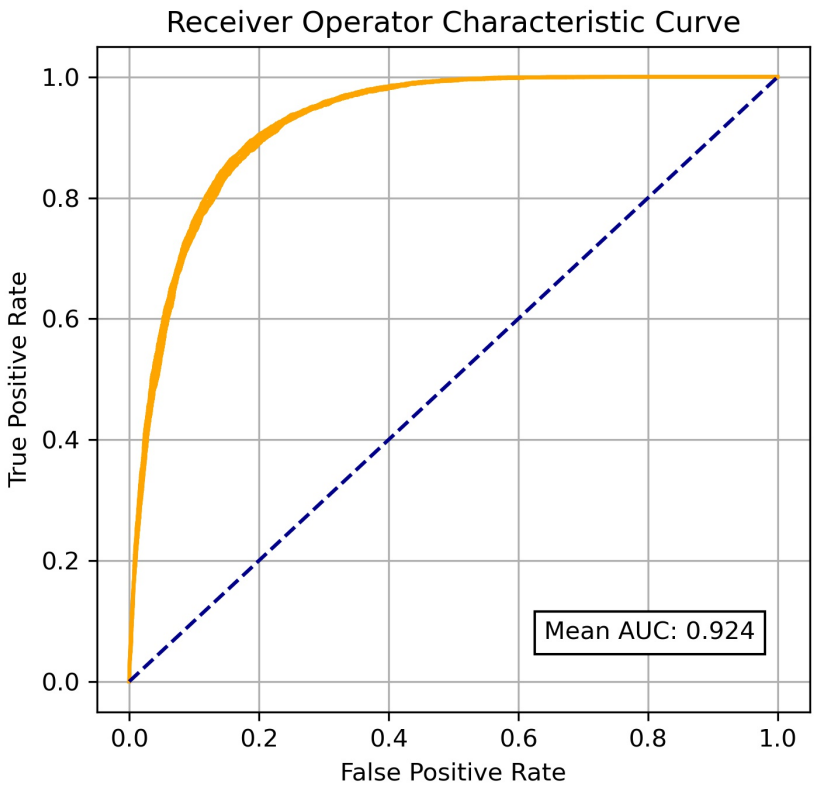
Machine learning overview

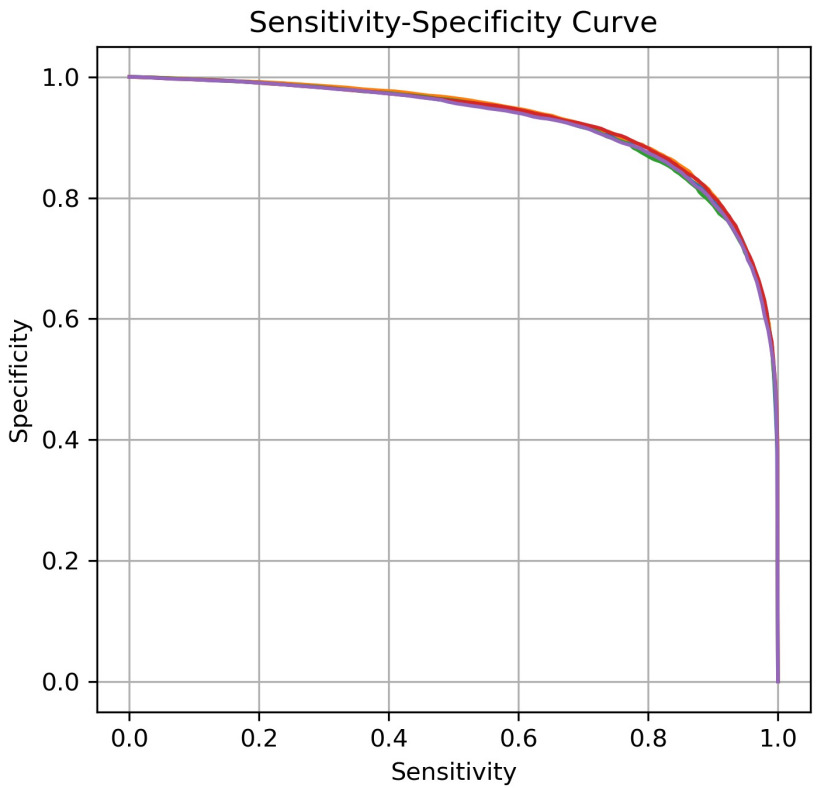


Model types used: logistic regression; random forest, XGBoost, neural networks (fully connected, and compartmentalised), ensemble.

We accessed 240,000 emergency stroke admissions in England and Wales over three years.

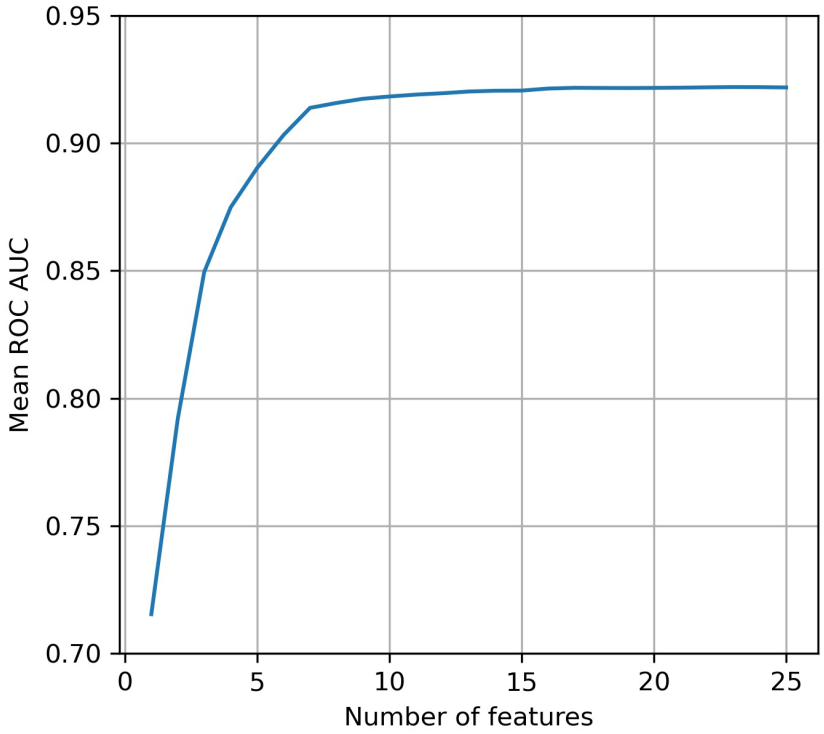
XGBoost model for predicting use of thrombolysis - accuracy

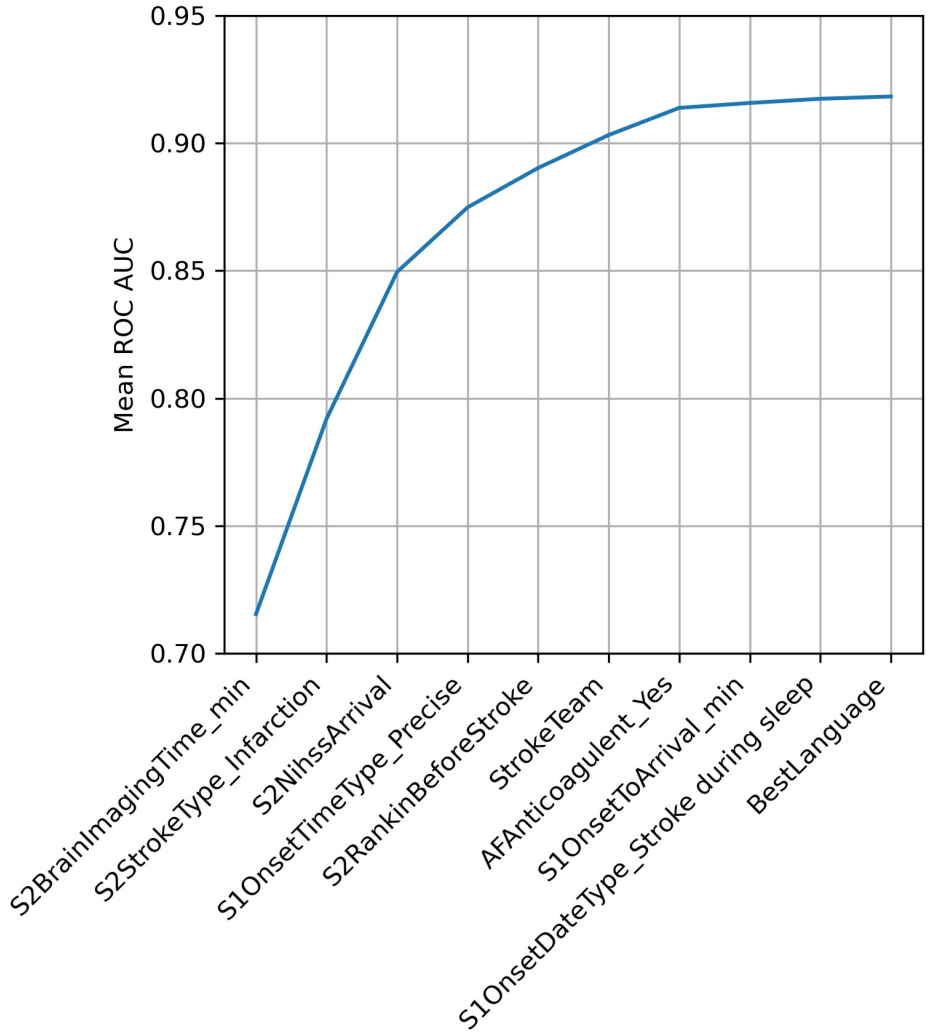




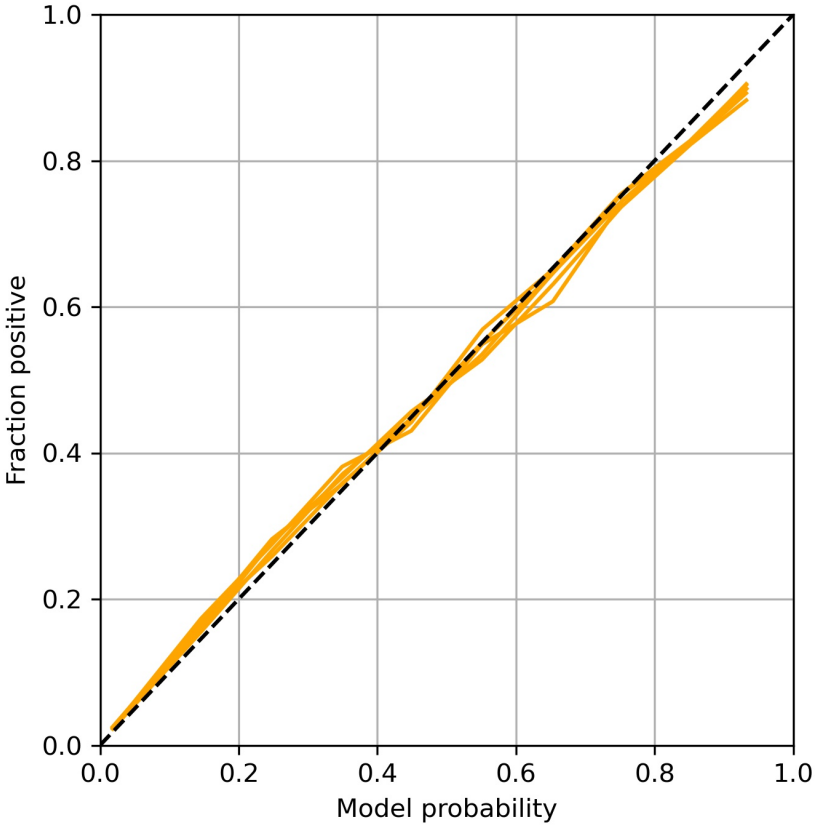
Model accuracy = 85% (and can achieve 85% sensitivity and specificity simultaneously)

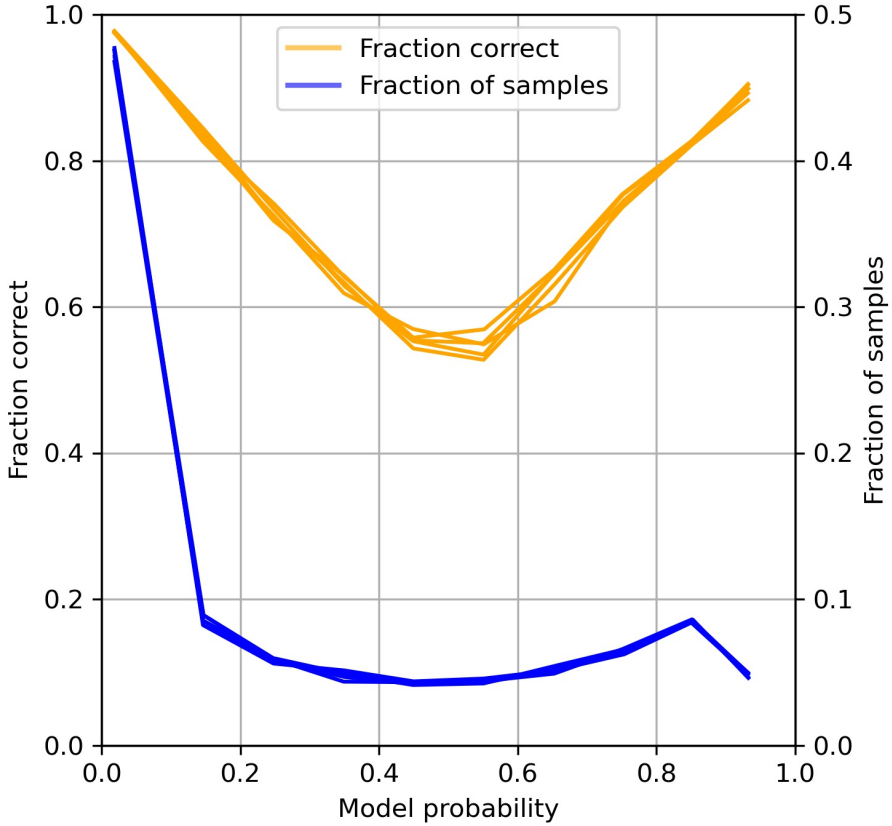
Ease explainability by simplifying model





XGBoost model for predicting use of thrombolysis - calibration





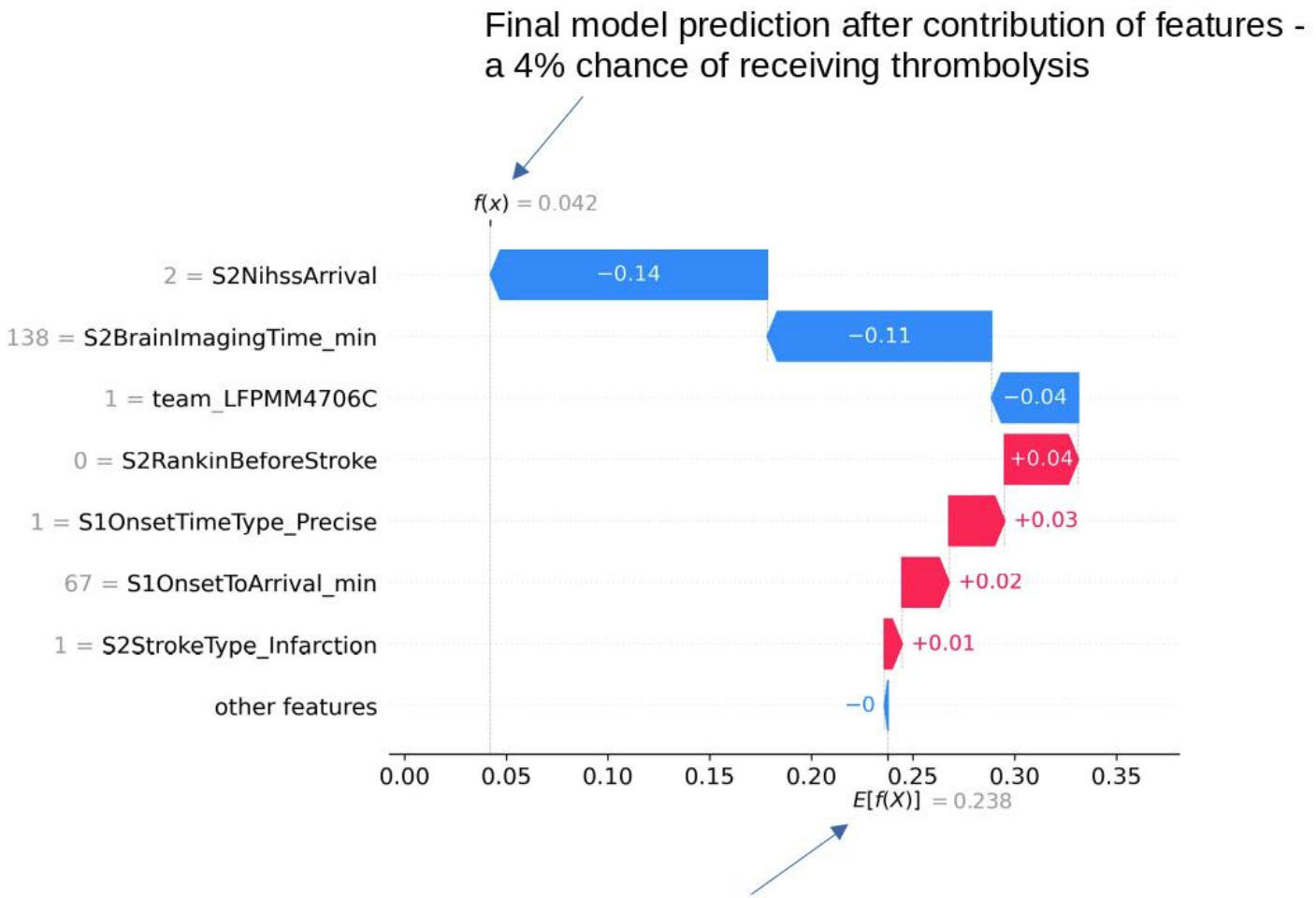
Shapley values

Shapley values

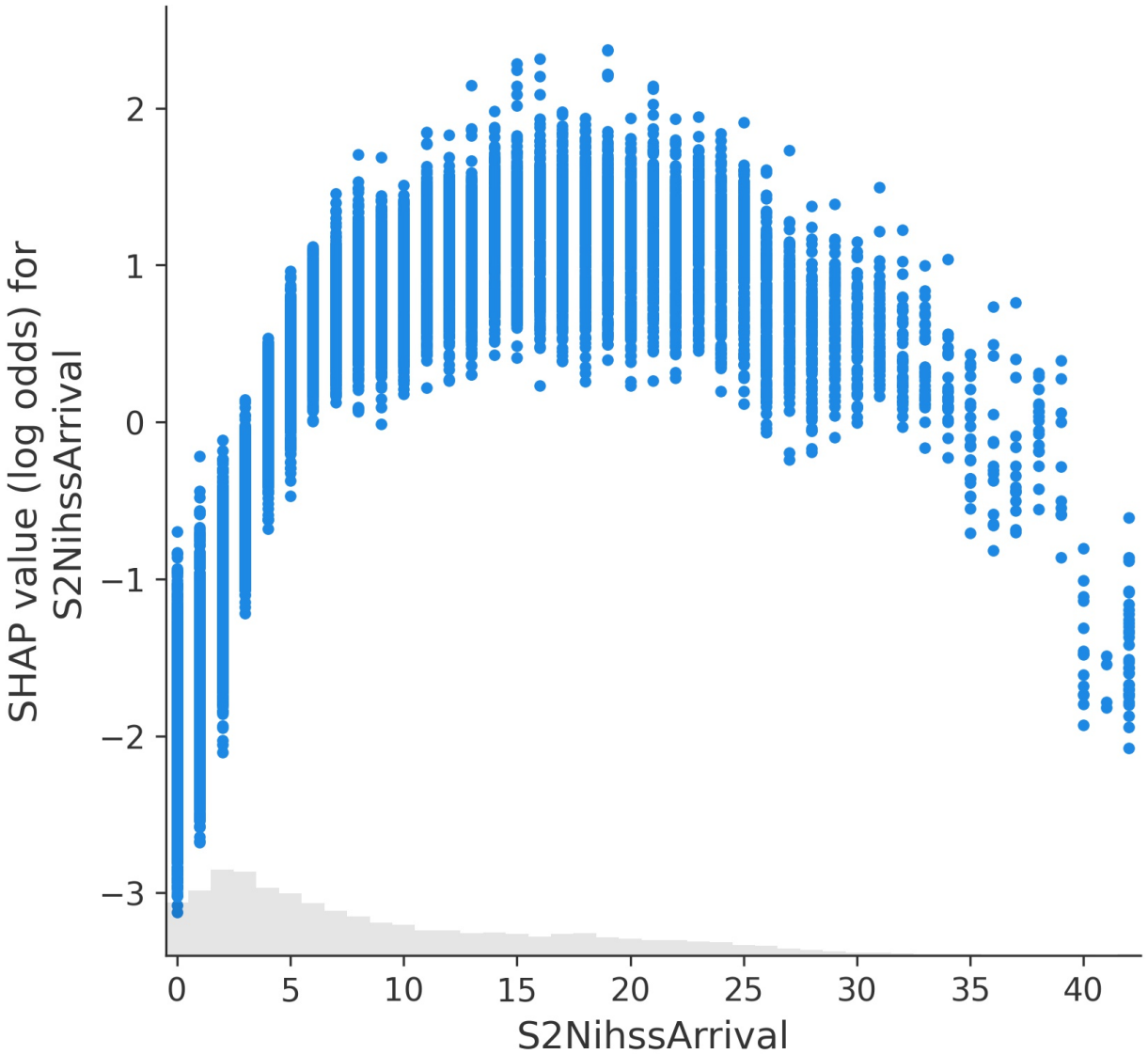
Lloyd Shapley
(Nobel Prize Winner)

'The average expected marginal contribution of one player after all possible combinations have been considered'

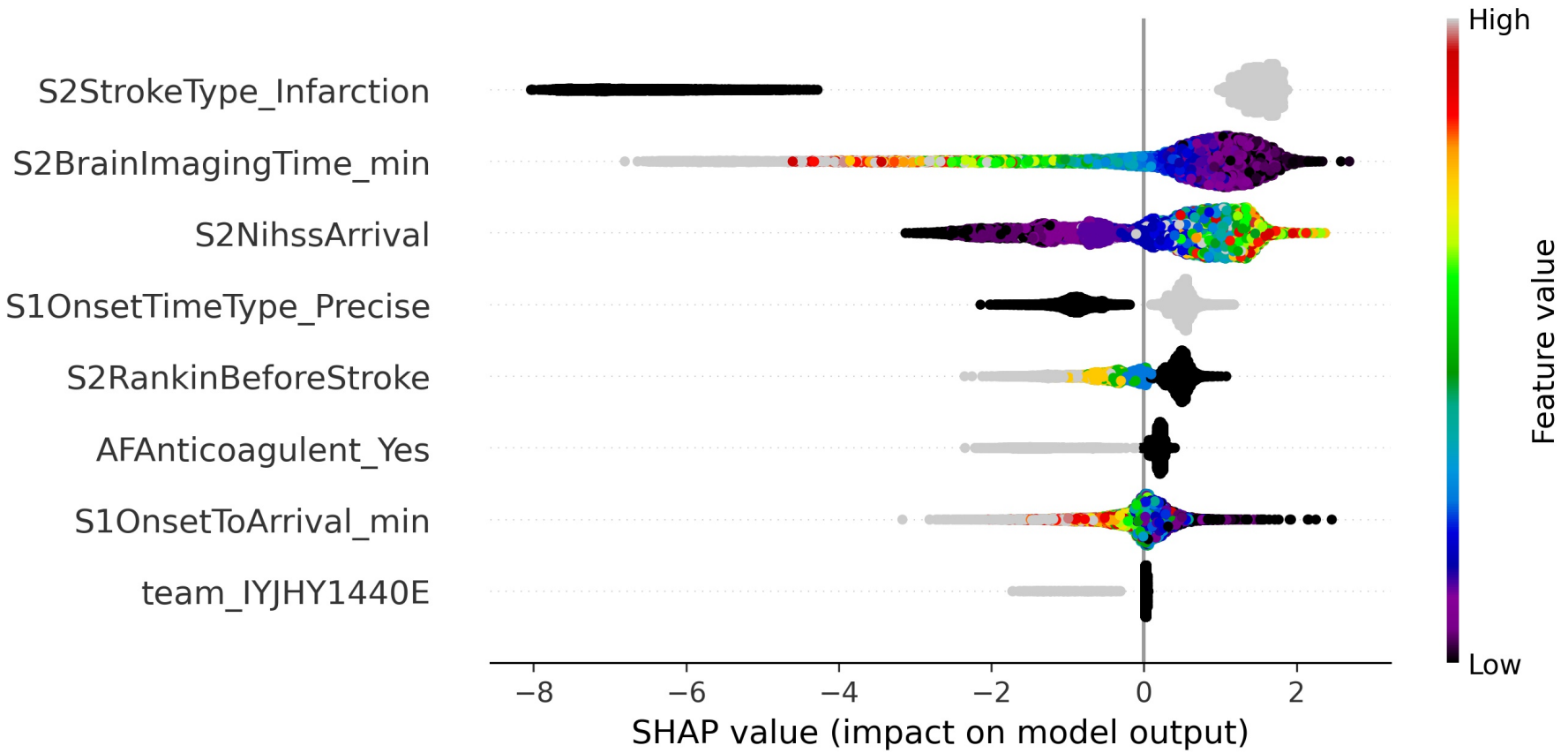
Shap plot for an individual case (low probability of receiving thrombolysis)



Shap plot for all instances of a single feature (stroke severity)

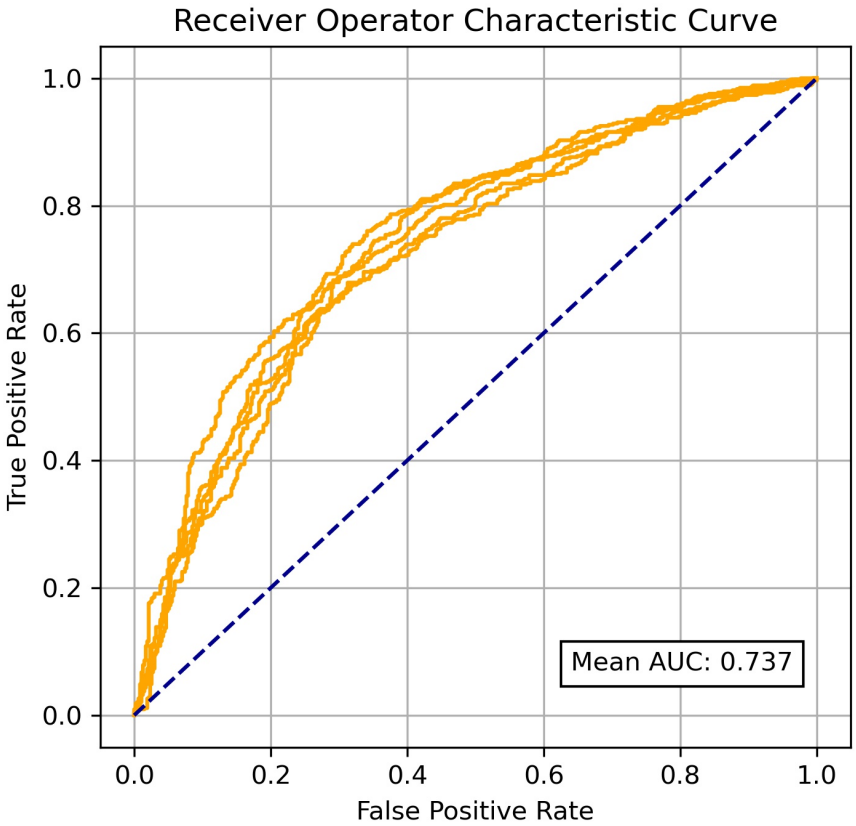


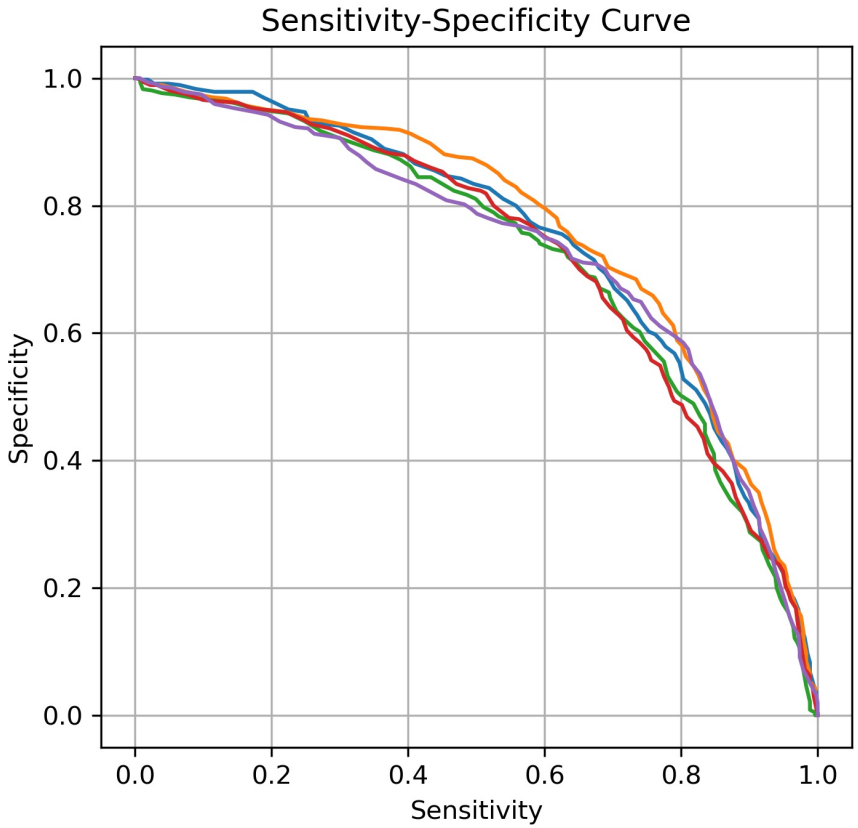
Shap plot for all instances of most significant features



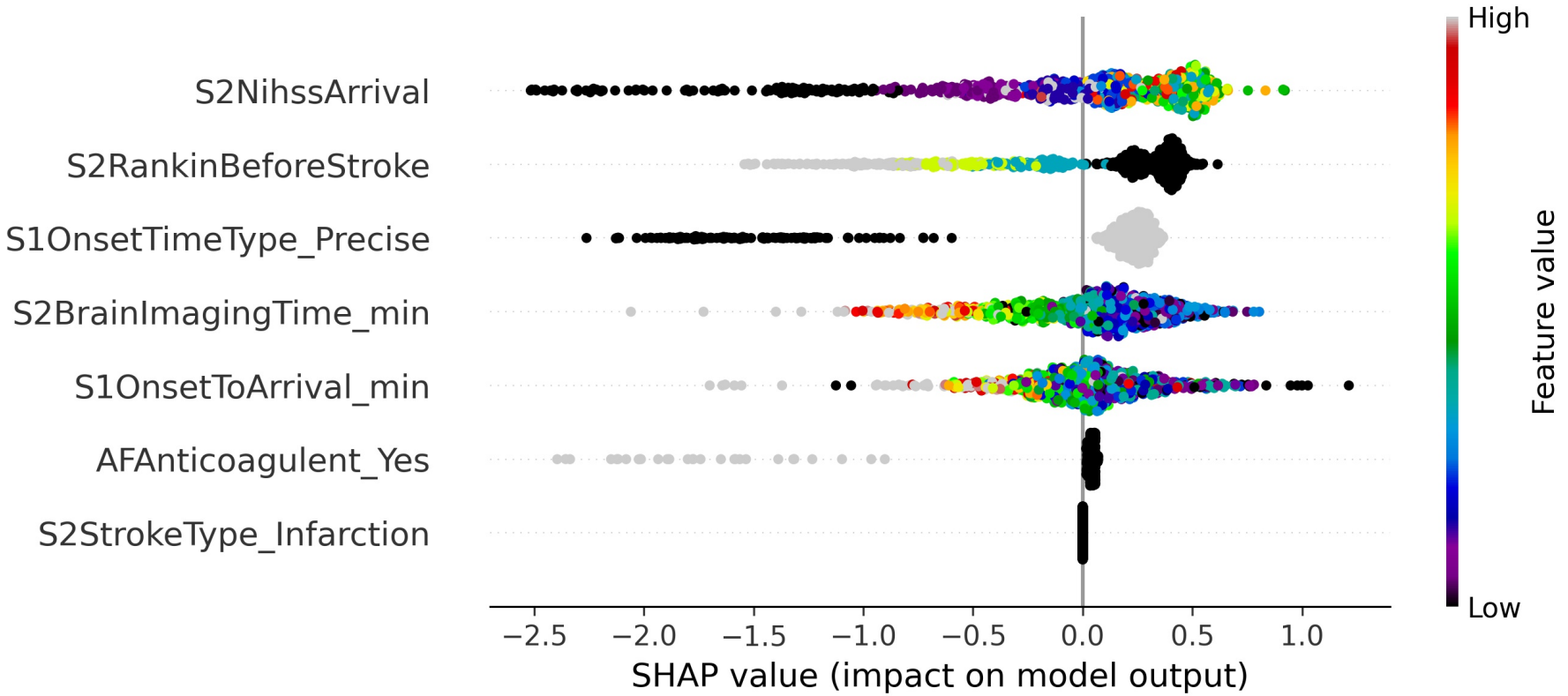
Comparing decision-making between high and low users of thrombolysis

Build a model to identify patients that are thrombolysed at high thrombolysis use hospitals but not at low thrombolysis hospitals



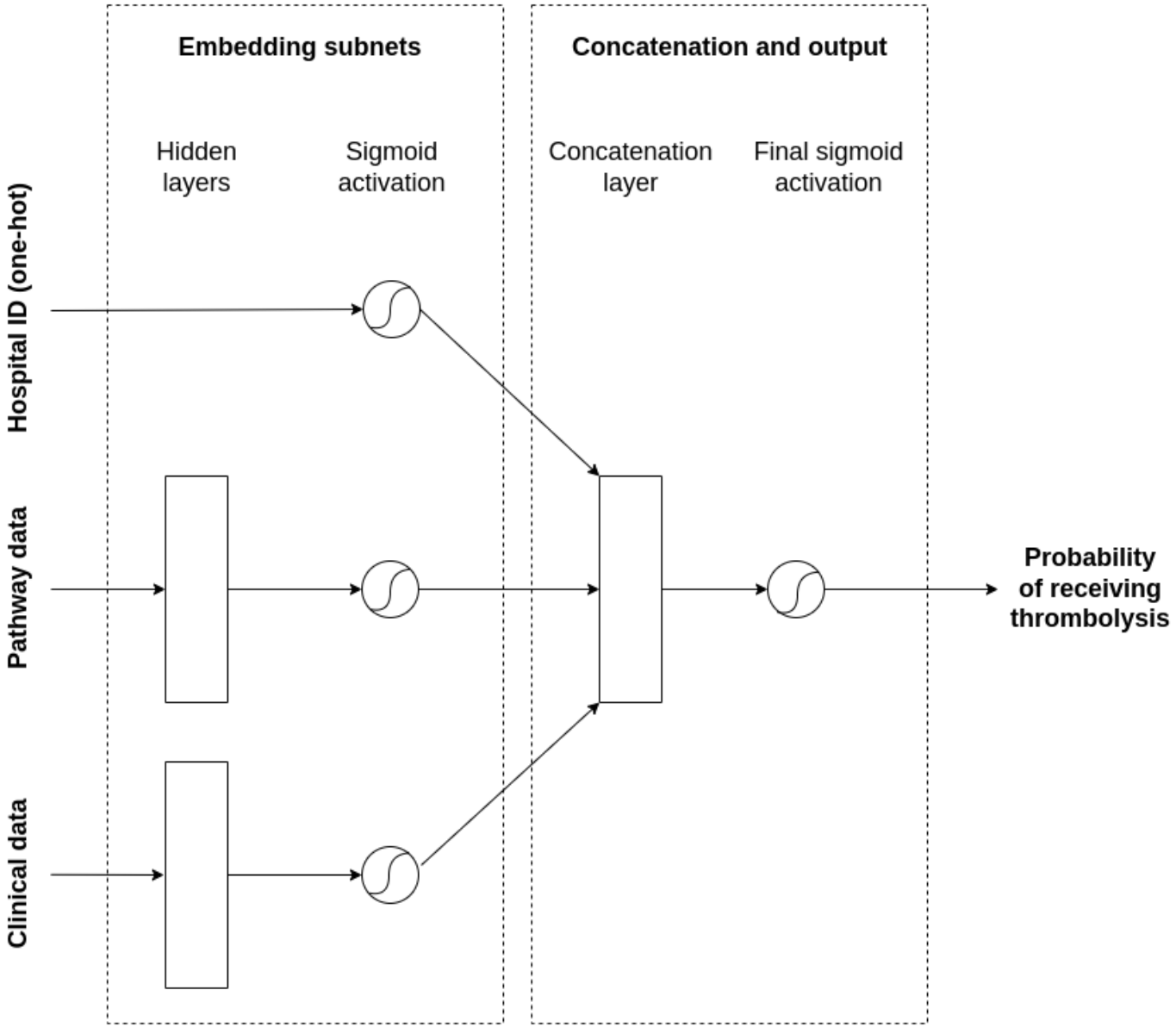


Shap plot to show which features most influence prediction of differences in thrombolysis decisions

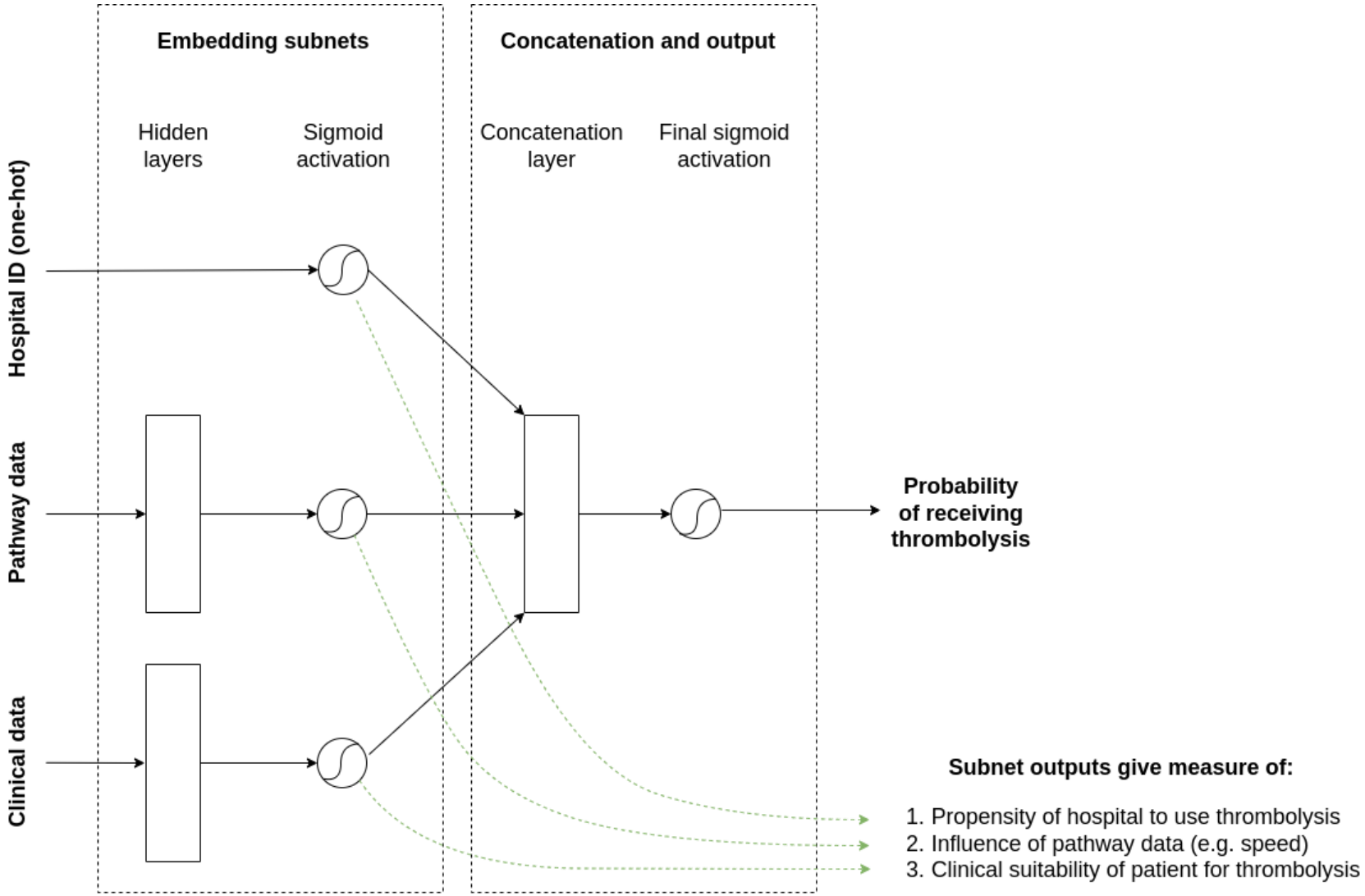


Designing explainable neural network architectures

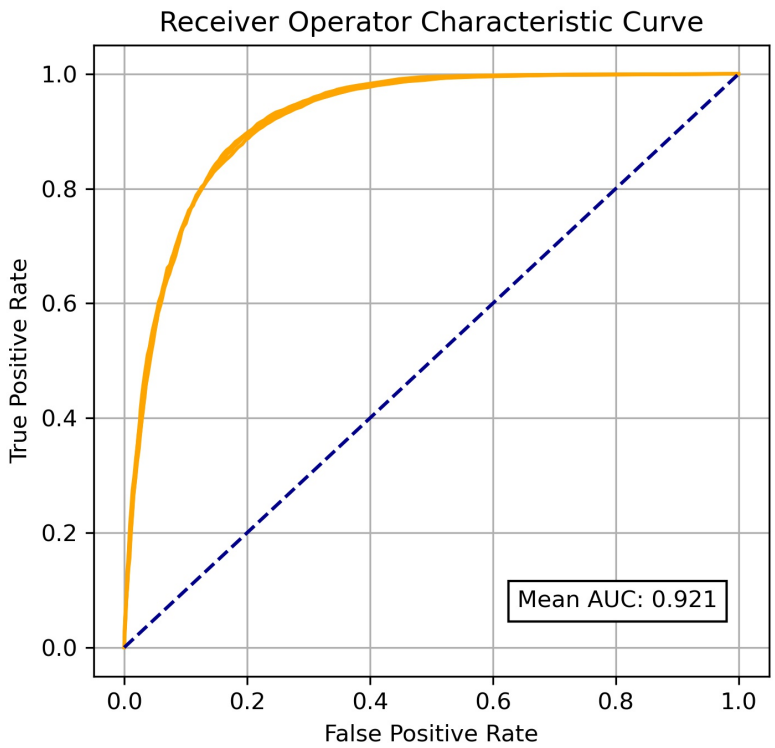
Structure of an embedding neural network

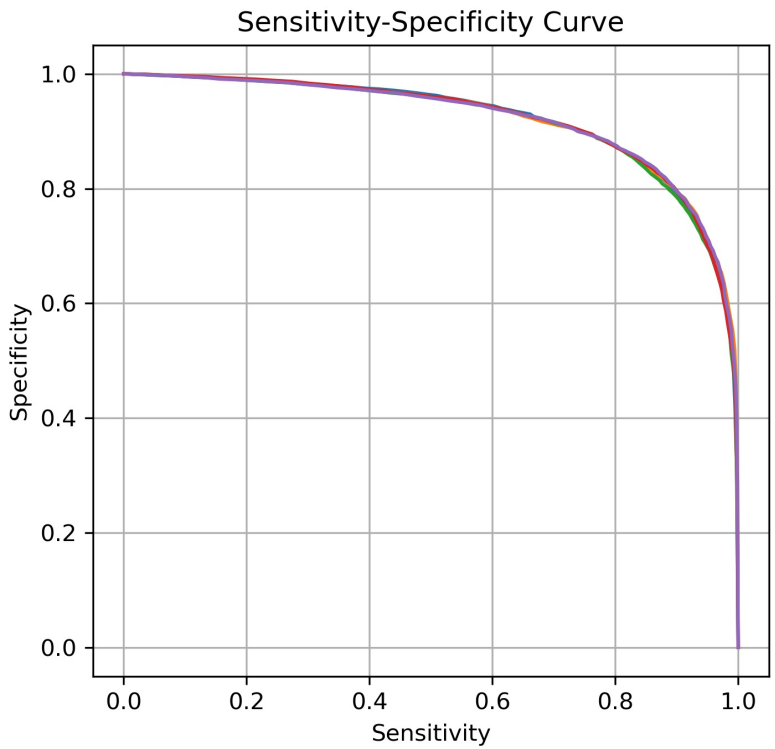


Subnet outputs give insight into predictions



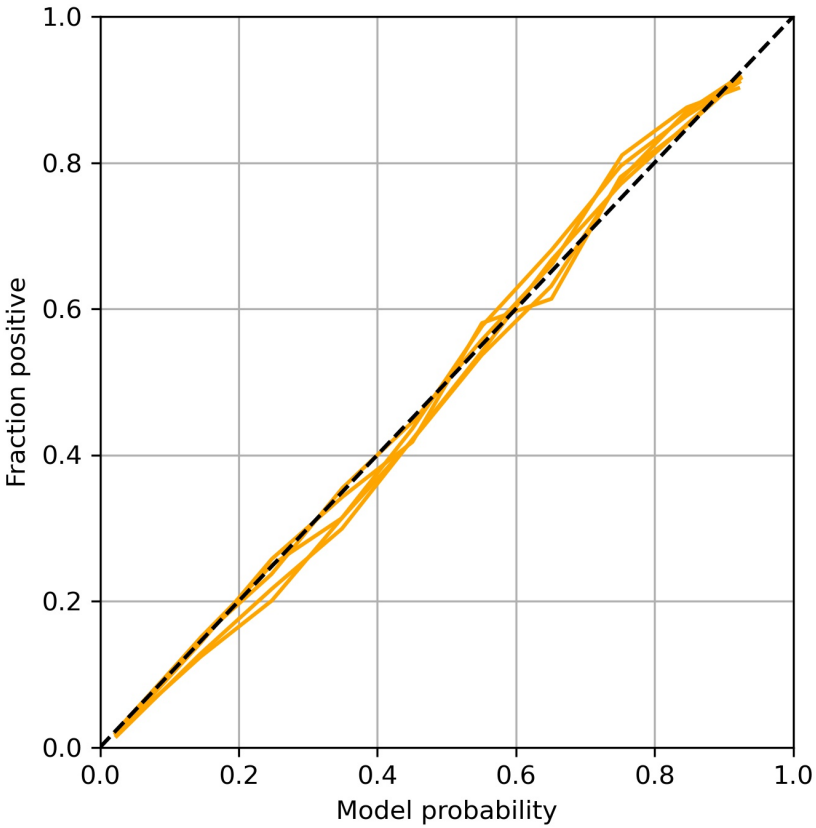
Embedding neural network model for predicting use of thrombolysis - accuracy

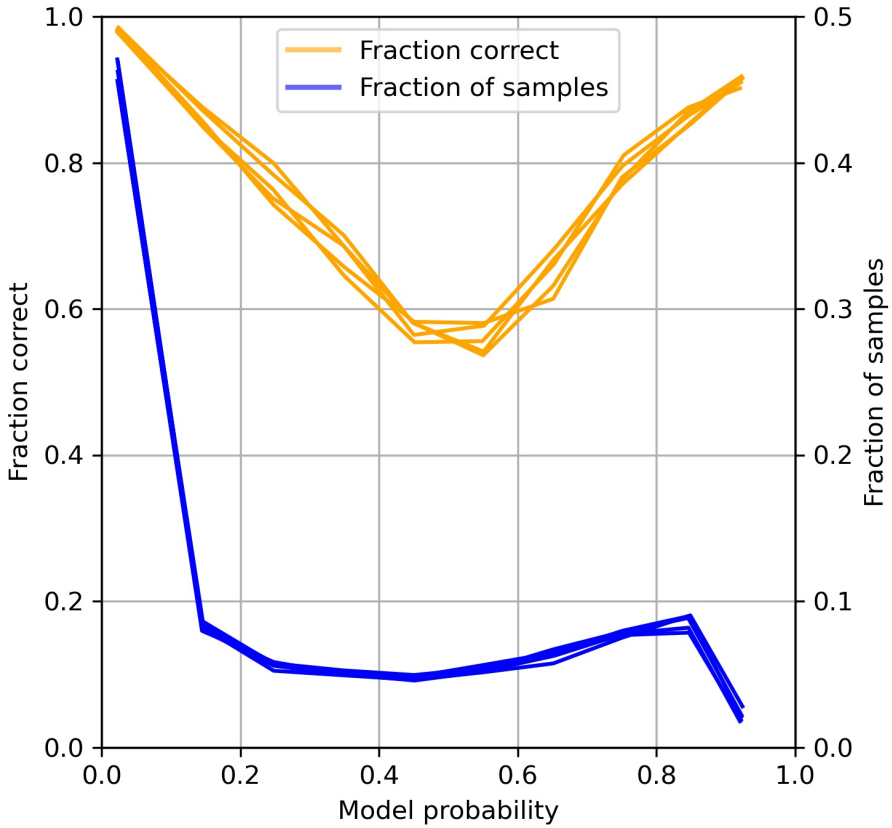




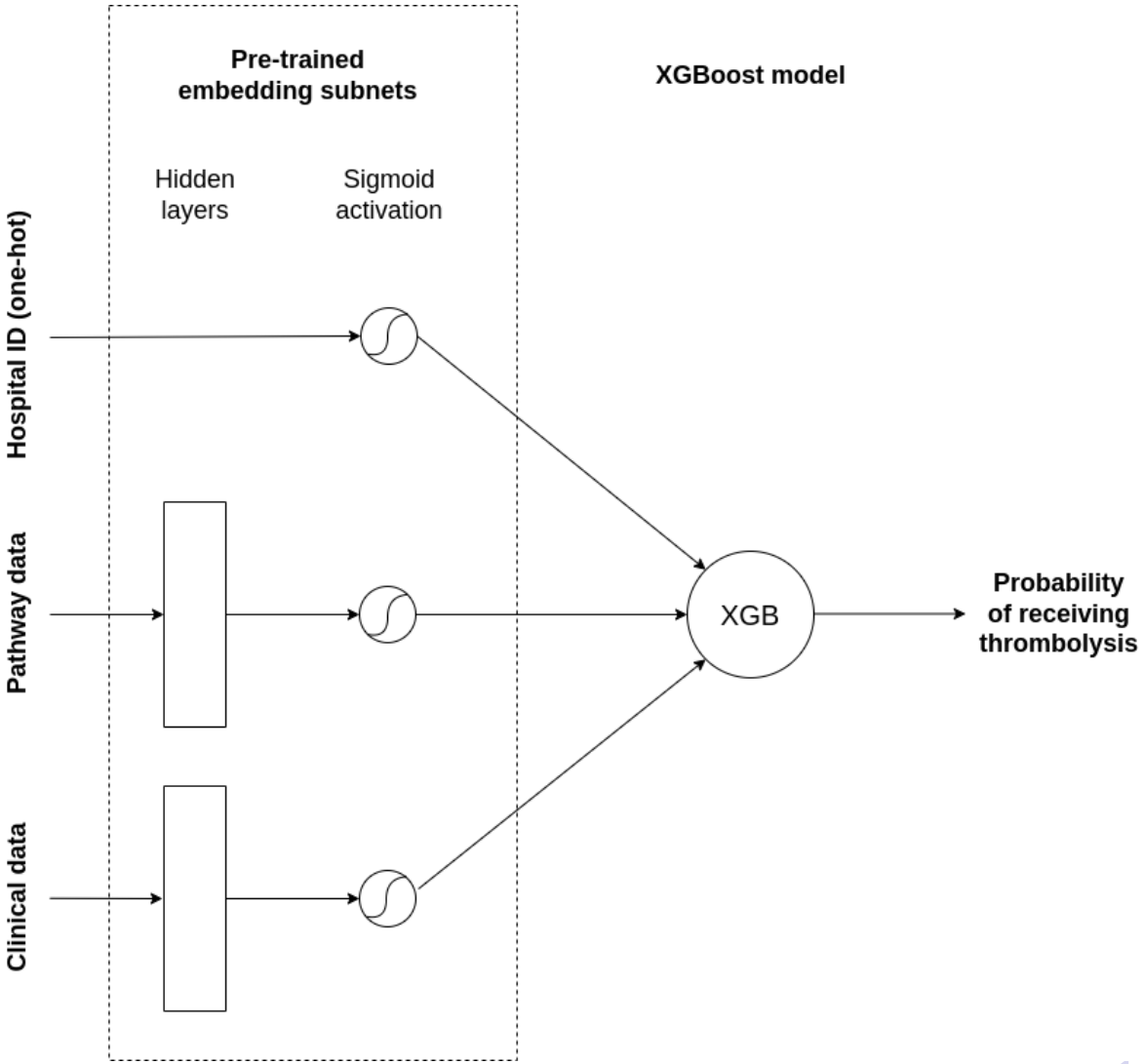
Model accuracy = 85% (and can achieve 85% sensitivity and specificity simultaneously)

Embedding neural network model for predicting use of thrombolysis - calibration



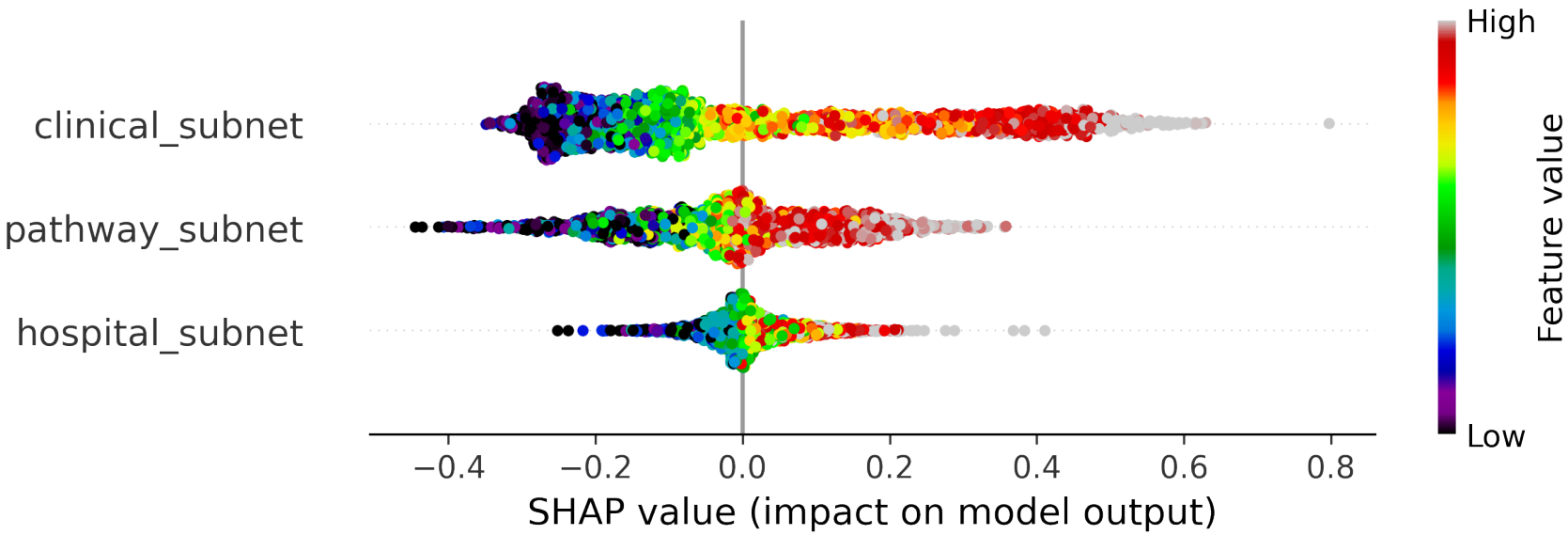


Using embedding outputs passed to a XGBoost model to predict thrombolysis

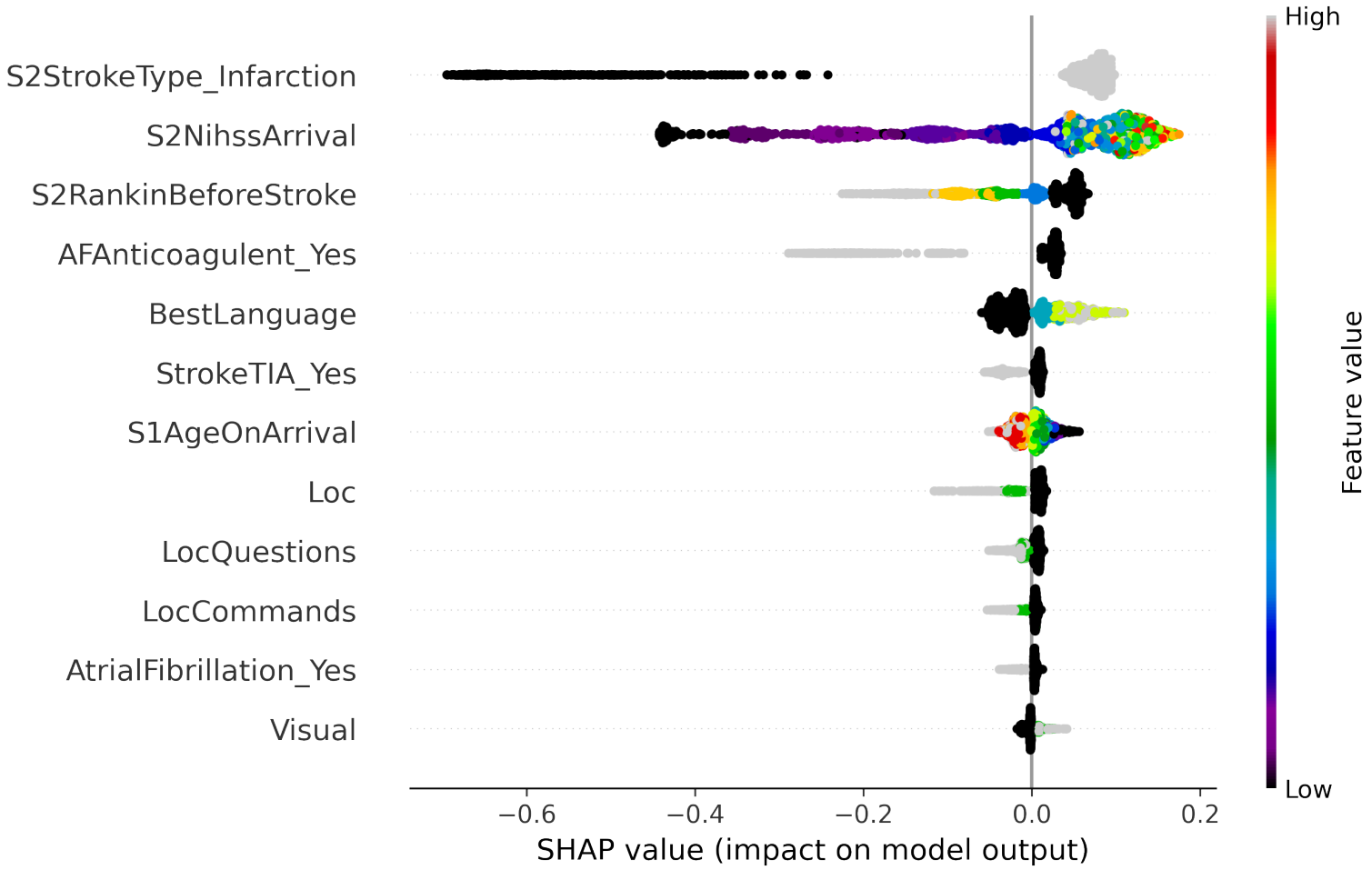


ROC AUC = 0.90

Using Shap to show the influence of the three feature types



Using an XGBoost regressor and Shap to model the contribution of features to the clinical subnet output



Lessons learned

Some lessons learned

- Explainability changes with audience.
- Simplifying models may help with explainability
- Explaining model results to a patient and carer involvement group forces clear explanations (and good understanding).
- We have found Shap is easily understood.
- Shap works best (fastest) on XGBoost.
- Neural network architectures can be used to *enhance* explainability - they do not need to be black boxes.
- Using Shap has led to better understanding of models.
- Many model types work natively in odds, but probability plots are more understandable to most.

For more, see our online Jupyter Books:

First project (complete, but no Explainabale AI):

`https://samuel-book.github.io/samuel-1/`

Some work on Shap:

`https://samuel-book.github.io/shap_short_paper_1/`