Stroke Prediction

Mahmoud Ghanem, Michael Kalish, Snehal Desai, Kate Reva

Motivation and Objective

Stroke is the **second leading cause of death globally**, responsible for **~11%** of total deaths [1]. Every **40 seconds**, someone in the US has a stroke. Every **3.5 minutes**, someone dies of stroke.

Stroke has major negative impacts to society and economy. Understanding key factors leading to stroke can potentially help reduce risk factors and improve early diagnosis.

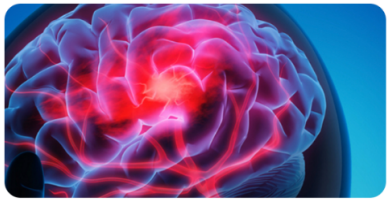
*Our **objective** is to predict a chance of stroke given health data to improve patient care.*

1. World Health Organization

Data

Stroke Prediction Dataset

11 clinical features for predicting stroke events



Source: **kaggle**

The data contains 5110 observations with 12 attributes (including patient id and stroke: Yes/No)

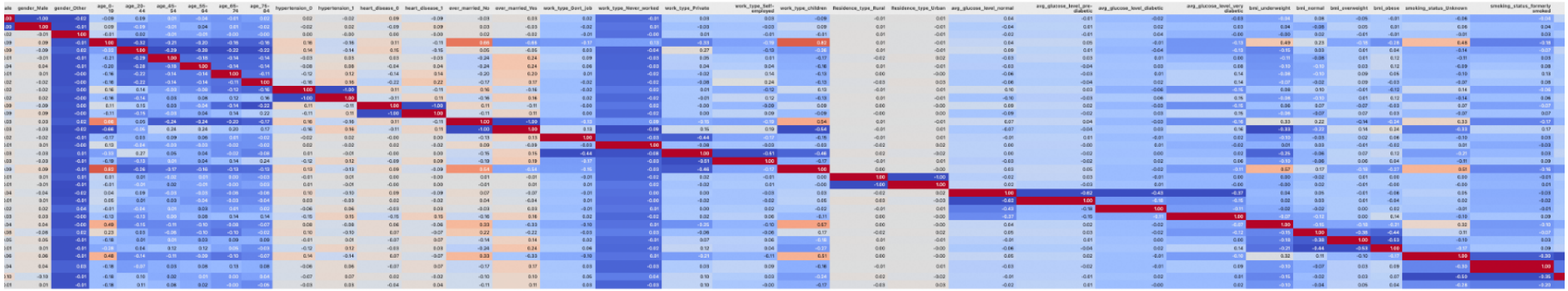
Data Pre-Processing

Update Representation

- **Fill n/a.** Save ~200 BMI nulls with KNN
- **One hot encode.** Convert categorical data columns into sparse representations.

Balance

- **Stratify.** Ensure sufficient minority class representation across train, validation & test sets.
- **Balance**. Balance train data with SMOTE (Synthetic Minority Over-sampling Technique)



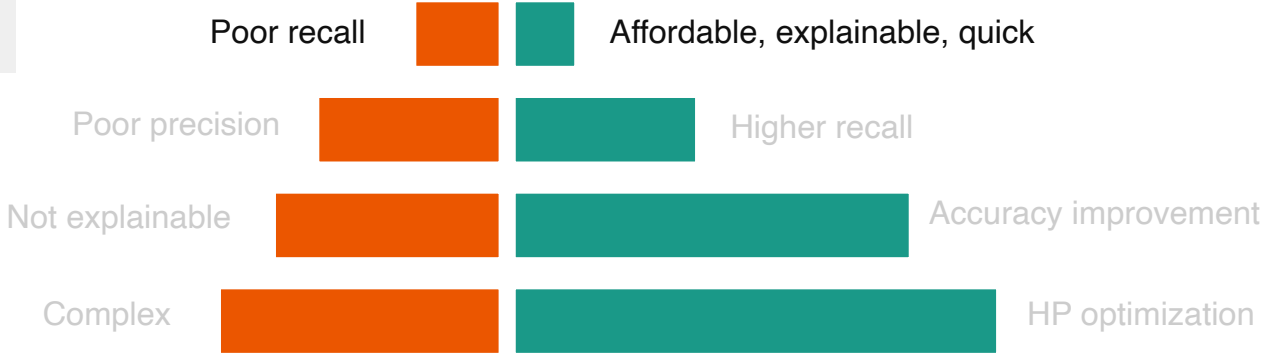
Modeling

Training and tuning of supervised ML options:

- Binary Logistic Regression
- Random Forest
- Deep Learning - Keras Sequential
- Deep Learning with Hyper Parameters (Optuna)

Modeling

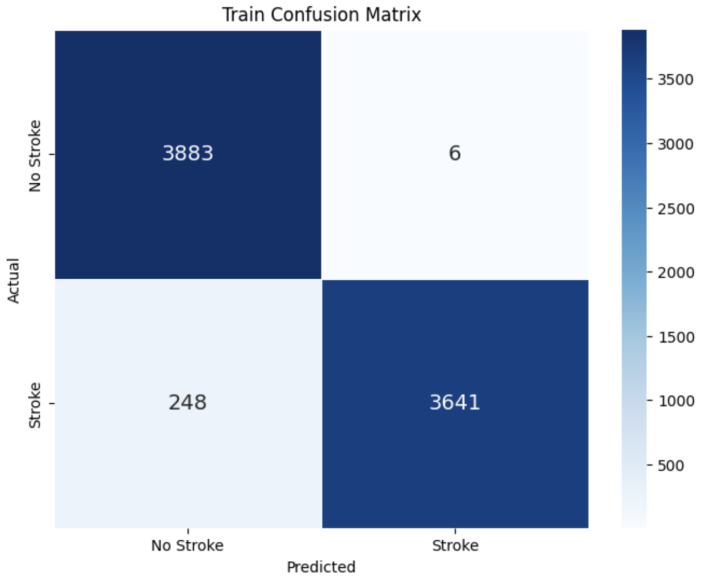
	Challenges	Opportunities
Baseline (Logistic Regression)	Poor recall	Affordable, explainable, quick
Random Forest	Poor precision	Higher recall
Keras Sequential	Not explainable	Accuracy improvement
Keras Sequential w/ Optuna	Complex	HP optimization



Binary Logistic Regression

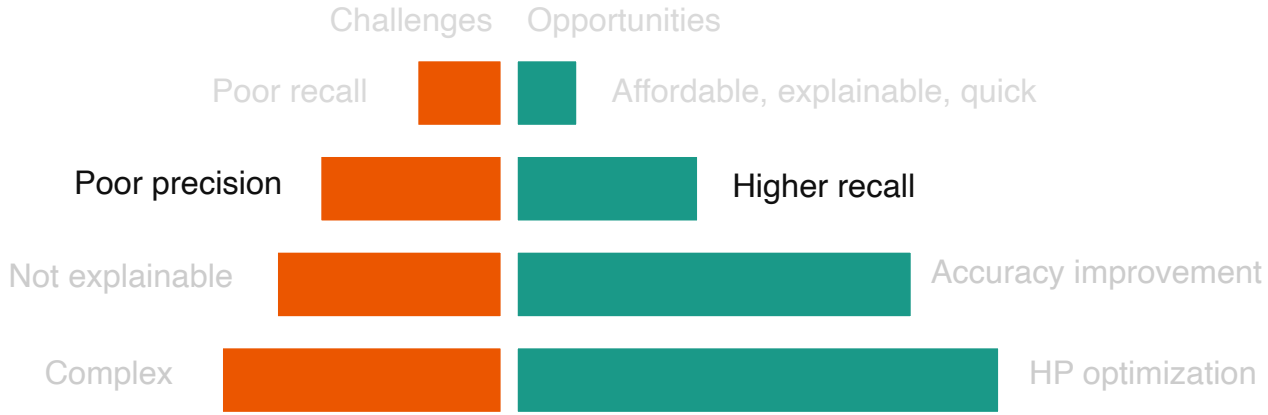
The model is biased towards missing actual stroke prediction
[**Low** False Positives at the cost of **High** False Negatives]

Validation Dataset	
Accuracy	0.95
Precision	0.73
Recall	0.53
F-1 score	0.55



Modeling

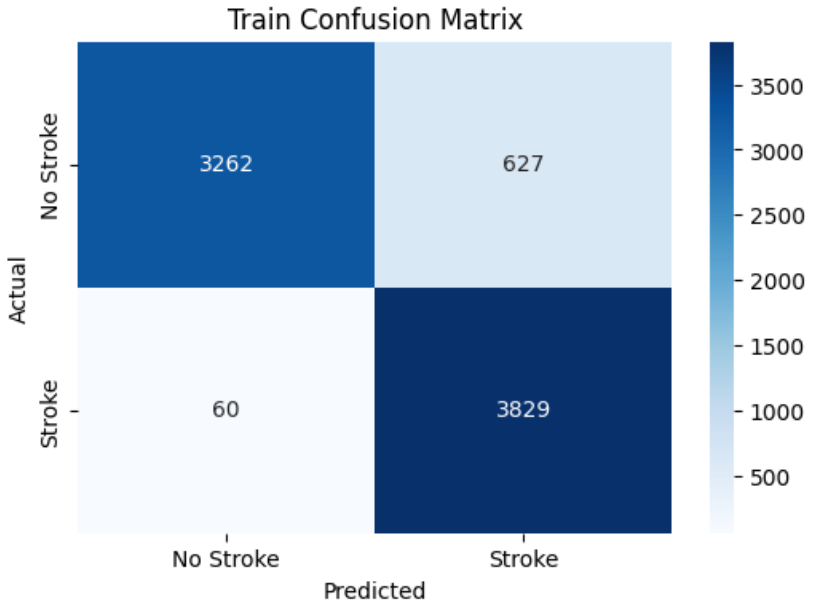
- Baseline (Logistic Regression)
- Random Forest
- Keras Sequential
- Keras Sequential w/ Optuna



Random Forest

The model improved in predicting actual stroke prediction at the cost of predicting of no stroke
[**Lower** False Negatives at the cost of **Higher** False Positives]

Validation Dataset	
Accuracy	0.80
Precision	0.56
Recall	0.71
F-1 score	0.56



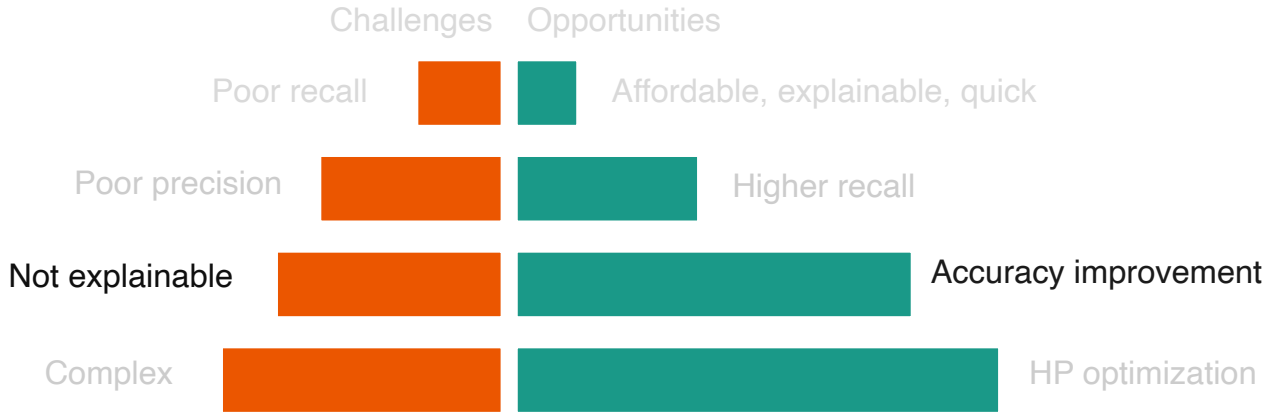
Modeling

Baseline (Logistic Regression)

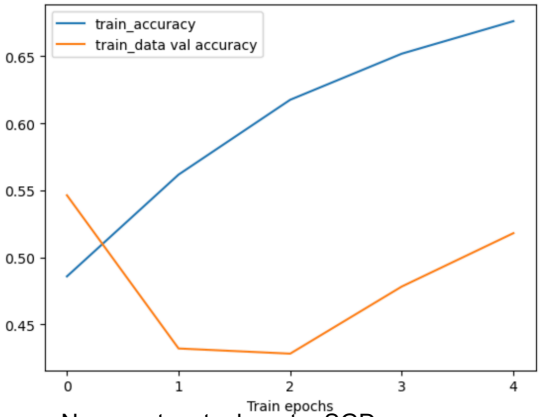
Random Forest

Keras Sequential

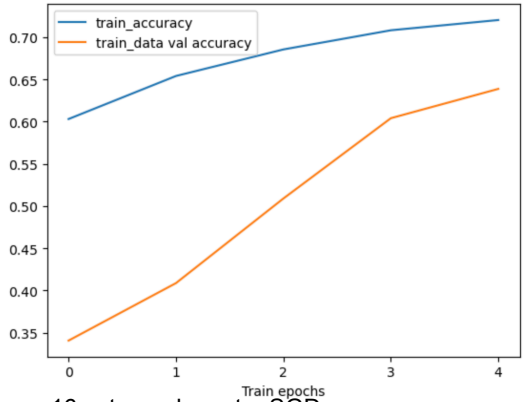
Keras Sequential w/ Optuna



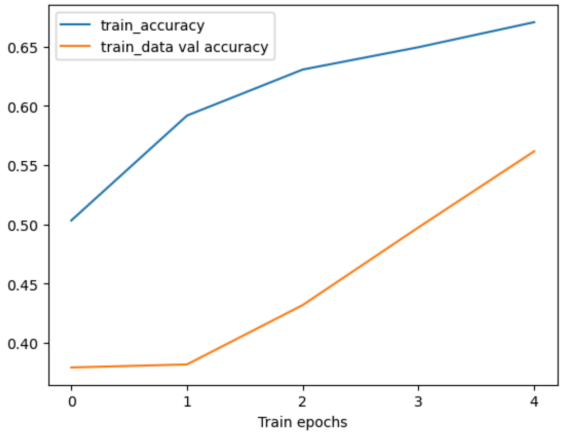
Deep Learning (improved accuracy experimentally)



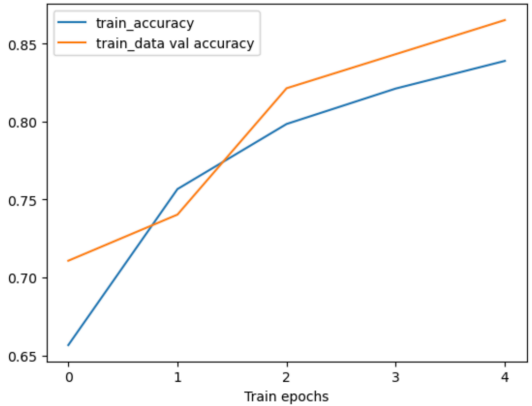
Hidden layers = None, actv = tanh, opt = SGD, learning_rate=0.01, num_epochs=5



Hidden layers = 16 actv = relu, opt = SGD, learning_rate=0.01, num_epochs=5



Hidden layers = 8, actv = relu, opt = SGD, learning_rate=0.01, num_epochs=5

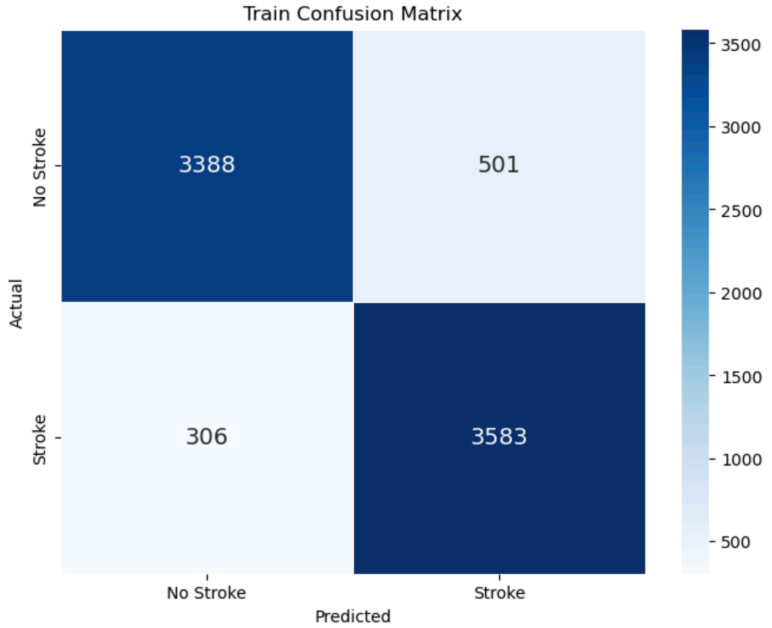


Hidden layers = 16, actv = tanh, opt = Adam, learning_rate=0.01, num_epochs=5

Keras Sequential

Deep Learning improved the key metrics, over RF, on the training and validation datasets (especially *accuracy*, 0.90 on the training data)

Validation Dataset	
Accuracy	0.86
Precision	0.59
Recall	0.76
F-1 score	0.62

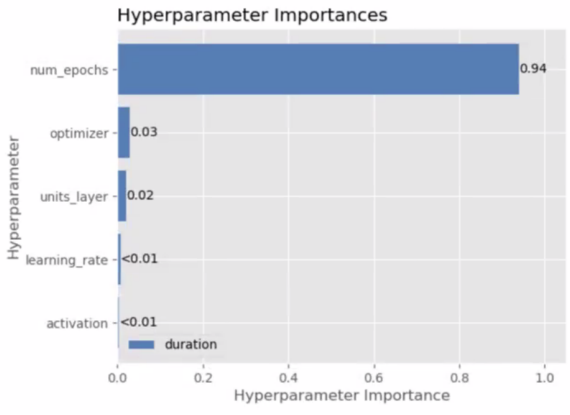


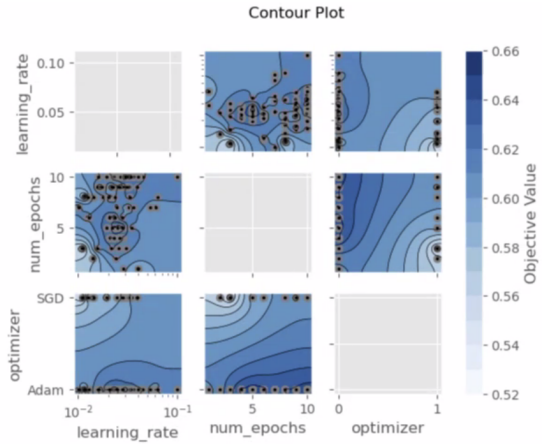
Modeling

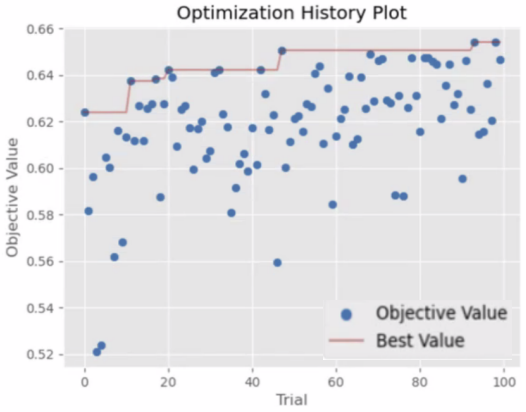

Challenges
Opportunities
Baseline (Logistic Regression)
Poor recall
Affordable, explainable, quick
Random Forest
Poor precision
Higher recall
Keras Sequential
Not explainable
Accuracy improvement
Keras Sequential (Optimized)
Complex
HP optimization

What is Optuna

Navigating and visualizing the Hyperparameter space





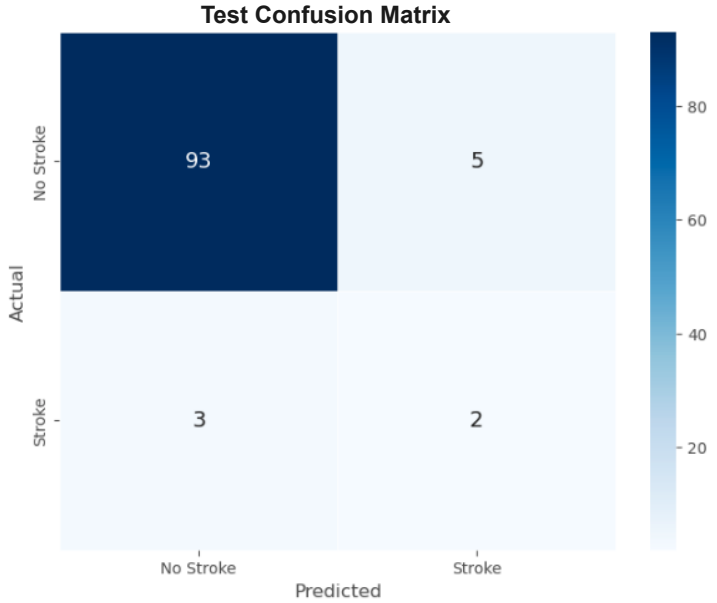


Keras Sequential (Optimized)

Improved model's accuracy, precision, recall and f-1 score, compared to other models.

Test Dataset	
Accuracy	0.92
Precision	0.63
Recall	0.67
F-1 score	0.65

Optuna Best Parameters
Hidden layers = 97, **actv** = tanh, **opt** = Adam, **learning_rate**=0.025342599583490992, **num_epochs**=5



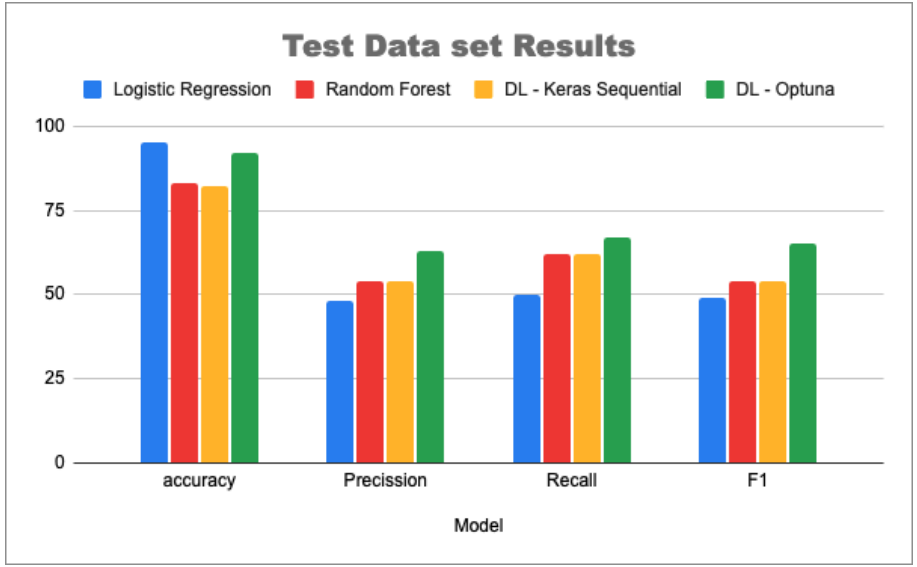
Conclusion

Models predicted stroke with at least over 80% accuracy.

Similar performance of **Random Forest** and **Keras Sequential** models across all metrics.

Logistic Regression model predicted stroke with 95% accuracy, at cost of all other metrics.

Keras Sequential model + demonstrated the best performance.



Conclusion

Average cost of hospitalization of patients with stroke per year, per patient in the United States is nearly **$60,000**.

Preventative measures include:

- Keep Average Glucose Level in normal range
- Be active
- Eat healthy
- Keep BMI in normal range

Thank you!

Contributions

Michael - Michael set a direction for the team and got started with data pre-processing and binary logistic regression model. He continuously provided valuable feedback and proposed creative ideas.

Mahmoud - Implemented Random Forest model, Feature Permutation Importance , Feature Importance SHAP, Code review , Model Analysis & Discussion , Holdout Dataset Testing for Random Forest Model, and updated presentation slides. I led team working sessions and sent out calendar invites for team meetings.

Snehal - Completed Deep Learning (Keras Sequential code), Deep Learning Optimization (running experiments with Optuna), Code review (entire code base), Model Analysis & Discussion (all models), Holdout Dataset Testing, Slides. Kept the team organized and led team working sessions.

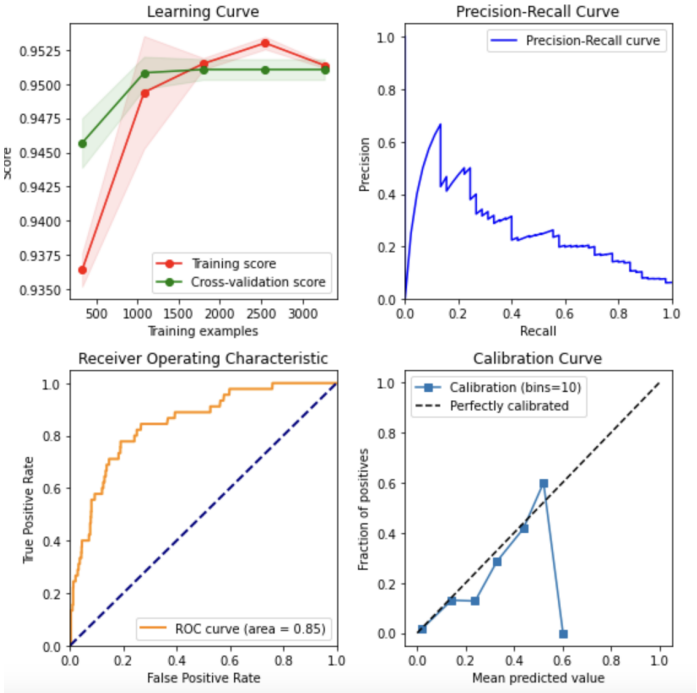
Kate - Worked on the Hypertuning for the Keras Sequential model & review Optuna visualization output, performed peer code review and optimised code by adding functions, Led team working sessions and designed & filled-in the slides.

Appendix: Conclusions [Holdout Dataset]

Model	Binary Logistic Regression	Random Forest	Deep Learning - Keras Sequential	Deep Learning with Hyper Parameters (Optuna)
Accuracy	0.95	0.83	0.82	0.92
Precision	0.48	0.54	0.54	0.63
Recall	0.50	0.62	0.62	0.67
F-1 score	0.49	0.54	0.54	0.65

Appendix

Binary Logistic Regression



Training Classification Report:

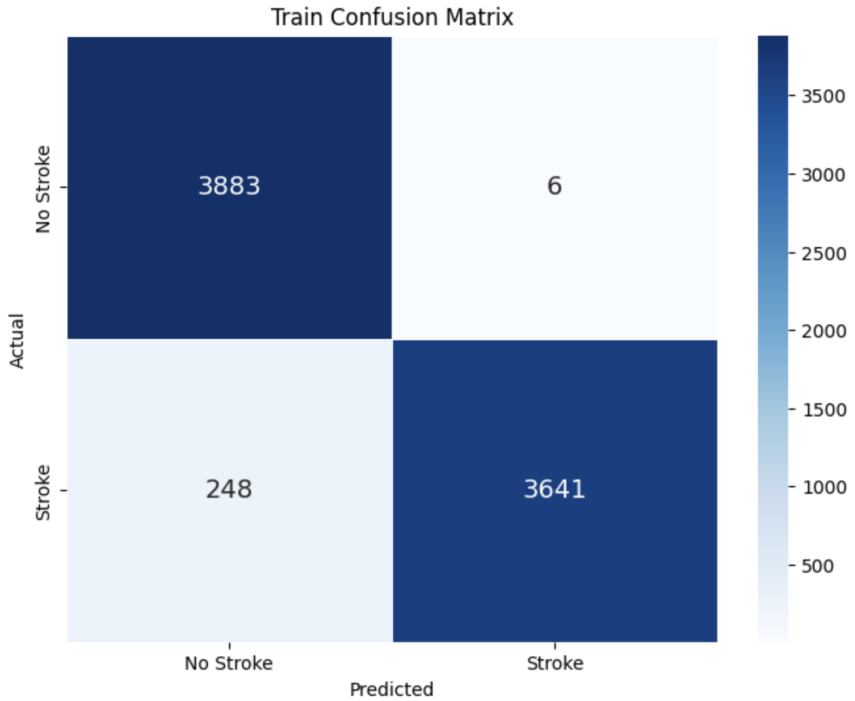
	precision	recall	f1-score	support
0	0.94	1.00	0.97	3889
1	1.00	0.94	0.97	3889
accuracy			0.97	7778
macro avg	0.97	0.97	0.97	7778
weighted avg	0.97	0.97	0.97	7778

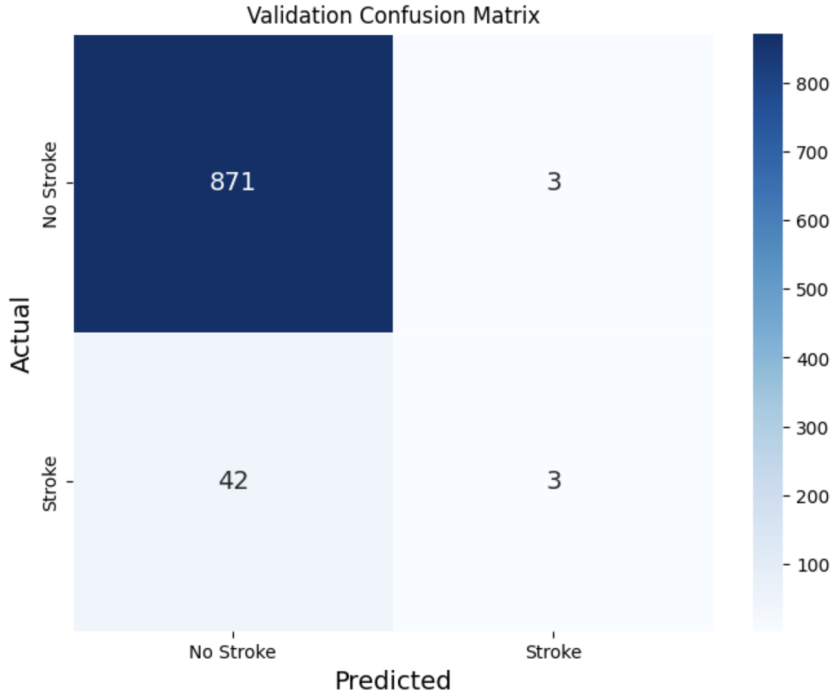
Validation Classification Report:

	precision	recall	f1-score	support
0	0.95	1.00	0.97	874
1	0.50	0.07	0.12	45
accuracy			0.95	919
macro avg	0.73	0.53	0.55	919
weighted avg	0.93	0.95	0.93	919

Appendix

Binary Logistic Regression





Appendix

Random Forest

Best params: {'bootstrap': False, 'criterion': 'gini', 'max_depth': 8, 'max_features': 'auto', 'n_estimators': 500}
Best estimator: RandomForestClassifier(bootstrap=False, max_depth=8, max_features='auto', n_estimators=500)
Best AUC: 0.91