Deep Learning Models – Final Project

By: David Wilczynski

1- Main Objective

The main objective of this analysis is to use deep learning models to predict the likelihood of stroke in individuals. Stokes are one of the leading causes of death worldwide, and being able to predict its occurrence could potentially save lives through medical screening and preventative treatments for at risk patients. Utilizing a patients demographic and lifestyle variables for known risk factors, 3 different deep learning models will be used to predict the likelihood of stroke, while evaluating and comparing the models. These models could have an exponential impact in a clinical setting by revolutionizing screening techniques and personalized preventative and treatment plans. The three models utilized in this analysis are Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). The three models will then be compared on their outputs and accuracy score to see which model is the best fit.

2- Dataset Description and Preprocessing

The dataset used in this analysis comes from the UC Irvine Machine Learning Repository and contains information on 5110 patients. The target variable is a binary variable which encodes whether a patient has a stroke or not (1=yes) for the training and test sets. The dataset also contains 10 other covariates which contain various demographic and lifestyle information of known stroke risk factors. The covariates are as follows: sex, age, hypertension, heart disease, marital status, residence type, type of employment, average glucose level, BMI, and smoking status. All of the variables were used within the model has they contain information relevant to the individual's health.

The data then underwent cleaning and preprocessing. First, the dataset was checking for missing data which then would have been removed from the dataset. However, there was no missing information, and therefore, the categorical variables like smoking status and sex were recoding using one-hot encoding in order to fit them into the models. The numerical variables on the other hand were scaled using StandardScaler in order to help with convergence during model training which resulted in the following data frame.

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke	gender_Male	gender_Other	smoking_status_formerly smoked
0	9046	67.0	0	1	228.69	36.6	1	True	False	True
2	31112	80.0	0	1	105.92	32.5	1	True	False	False
3	60182	49.0	0	0	171.23	34.4	1	False	False	False
4	1665	79.0	1	0	174.12	24.0	1	False	False	False
5	56669	81.0	0	0	186.21	29.0	1	True	False	True

3- Training the Models

As previously mentioned, three different deep learning techniques were utilized in order to train and test the dataset. The three models used in this analysis are a basic Recurrent Neural Network (RNN), a Long-Short-Term Memory (LSTM) model, and a Gated Recurrent Unit (GRU) model. The models were all run on the same dataset and used the same training/test data split.

First, the basic RNN model was fit which consists of an RNN with 50 units, followed by a dropout layer to reduce the risk of overfitting. The final output layer consists of a single neuron with a sigmoid activation function in order to predict the probability of a stroke. The RNN layer processes the input data sequentially, and as this model is the simplest, it serves as a sort of baseline for the analysis.

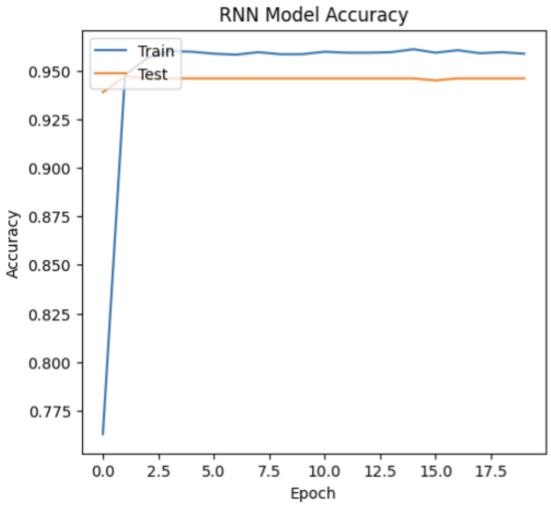
Secondly, the more advance LSTM model was run which is designed to capture long-term dependencies in the data. The LSTM layer also consists of 50 units, followed by a dropout layer to prevent overfitting. The last model run was a GRU model which is more complex than an RNN, but less complex than an LSTM as to serve as a sort of middle ground for the analysis.

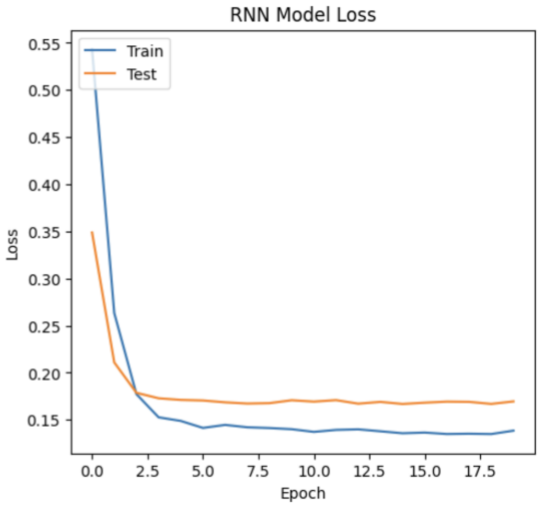
Each of the models was trained using the Adam optimizer as it is well suited for deep learning models with a binary cross-entropy loss function as this analysis is a binary classification predicting stroke or no stroke. The chosen primary evaluation metric was accuracy. Each model was trained for 20 epochs with a batch size of 32, and the performance was evaluated on the test set after each epoch. The training process was monitored for signs of overfitting or underfitting by comparing training and validation loss/accuracy curves.

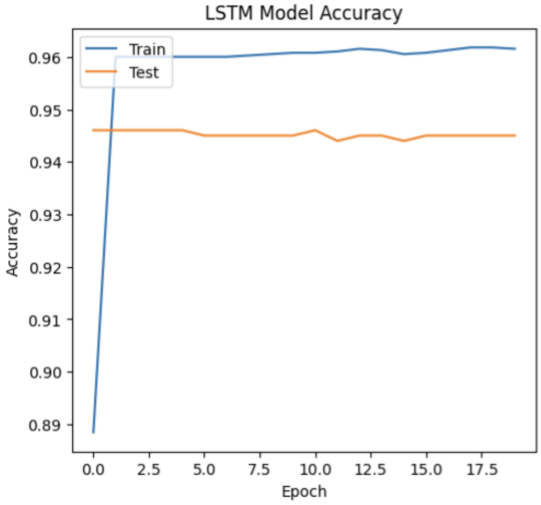
4- Model Evaluation

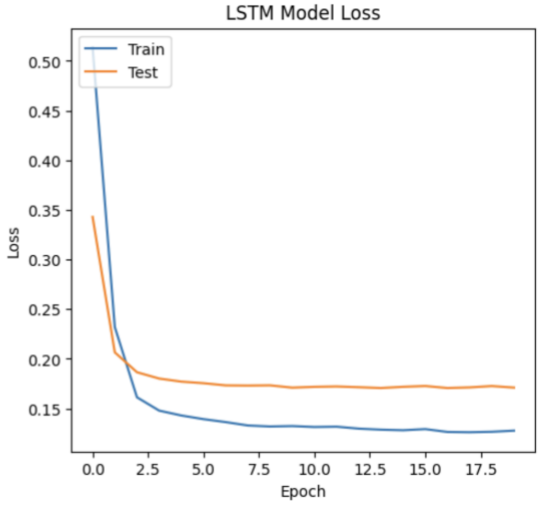
As previously stated, the models were compared based on accuracy and model loss which can be seen below for all three models.

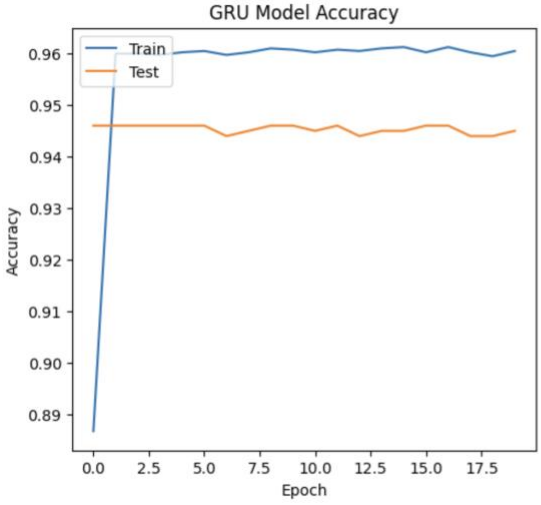
```
31/31 ━━━━━━━━━━━━━━━━━━━━ 0s 5ms/step - accuracy: 0.9467 - loss: 0.1748
31/31 ━━━━━━━━━━━━━━━━━━━━ 0s 5ms/step - accuracy: 0.9453 - loss: 0.1789
31/31 ━━━━━━━━━━━━━━━━━━━━ 0s 5ms/step - accuracy: 0.9463 - loss: 0.1766
RNN Accuracy: 0.9460
LSTM Accuracy: 0.9450
GRU Accuracy: 0.9450
```

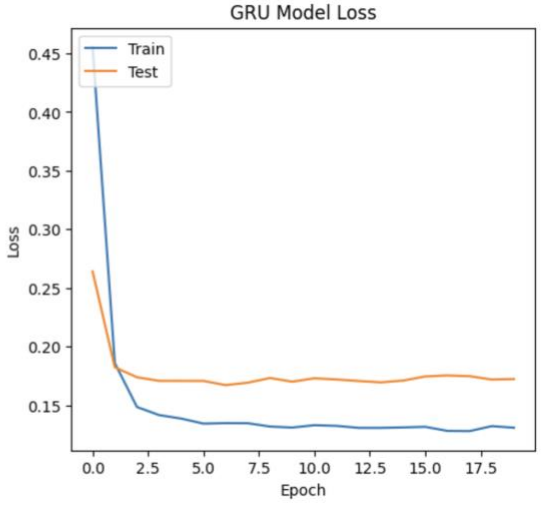












After evaluating and comparing the models, the best model for this dataset is the RNN model. The RNN model resulted in the highest accuracy and lowest loss as well as being the simplest of the three models. If this analysis were to be run again, it would be beneficial to run it on a larger set of data with more clinically relevant information like medical history to more accurately predict a patient's risk of stroke. The models could also be tested with other optimizers and hyperparameter tuning which would be necessary on a larger dataset. In conclusion, this analysis demonstrated the application of deep learning models to predict stroke risk which could have clinical significance for medical screening and personalized preventative and treatment plans. Further analysis could focus on refining the models, experimenting with other architectures, and incorporating a larger more diverse dataset.