STROKE RISK PREDICTION

BY: DANIEL JOLIN, ANDREW NICOLA, NATHAN WATERS

PROBLEM

- ~6.5 million people die of a stroke every year, accounting for 11% of all deaths worldwide - making them the second leading cause of mortality.
- The challenge lies in their sudden, unpredictable nature - when a stroke occurs, every minute counts, with 1.9 million neurons lost each minute treatment is delayed.
- Our mission: Develop an early warning system to identify high-risk individuals before a stroke occurs, enabling preventive interventions that could save millions of lives.

PROPOSED APPROACH

- The goal is to develop a machine learning model capable of predicting the risk of stroke using patient data.
- This involves processing raw healthcare data and selecting appropriate features to train the model on.
- The model will aim to provide accurate predictions to support earlier medical interventions.

ALGORITHM

- Logistic Regression: A simple and interpretable baseline model for binary classification
- Support Vector Machines (SVM): To explore linear and non-linear decision boundaries.
- Neural Networks: Leveraged ResNet-inspired architecture for capturing complex patterns in data.

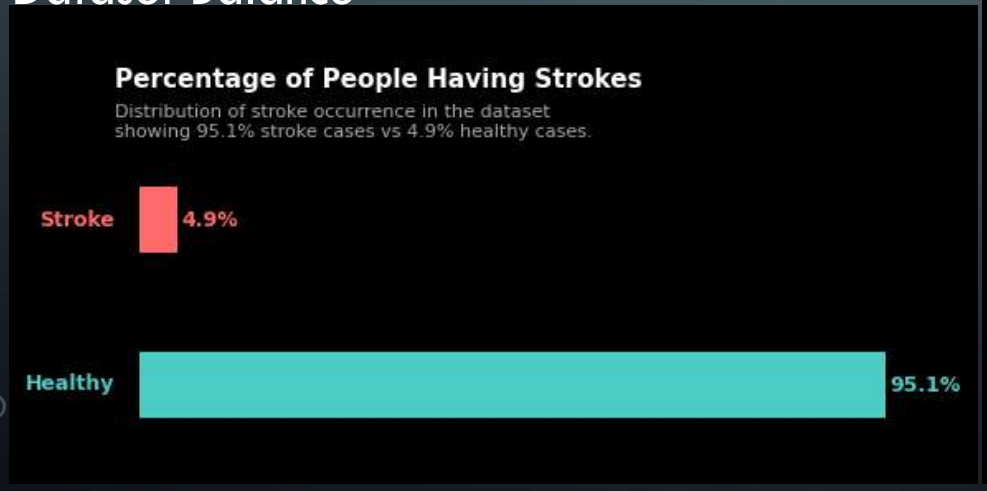
DATA SET

We used the “Stroke Prediction Dataset” found on kaggle.

Data frame Head

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

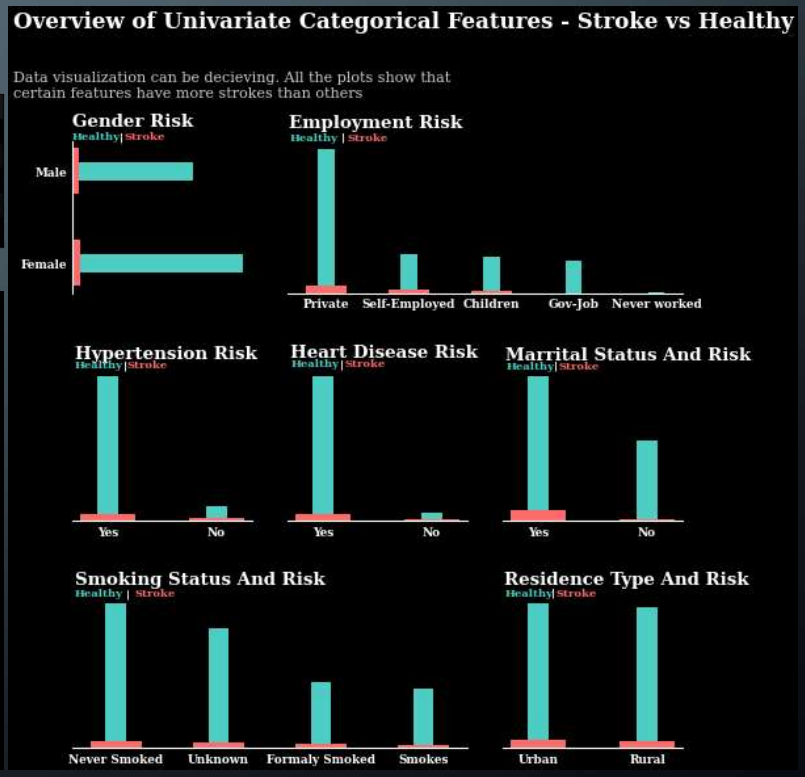
Dataset Balance



Null Values

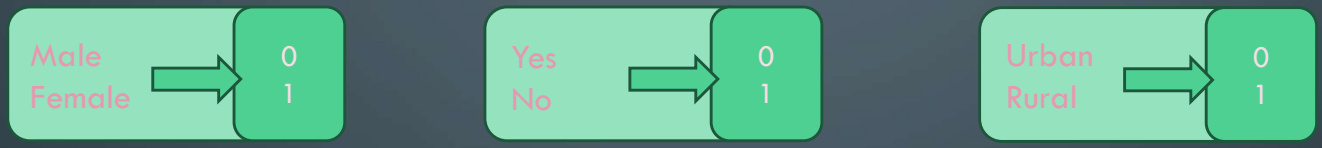
id	0
gender	0
age	0
hypertension	0
heart_disease	0
ever_married	0
work_type	0
Residence_type	0
avg_glucose_level	0
bmi	201
smoking_status	0
stroke	0

Feature Distribution

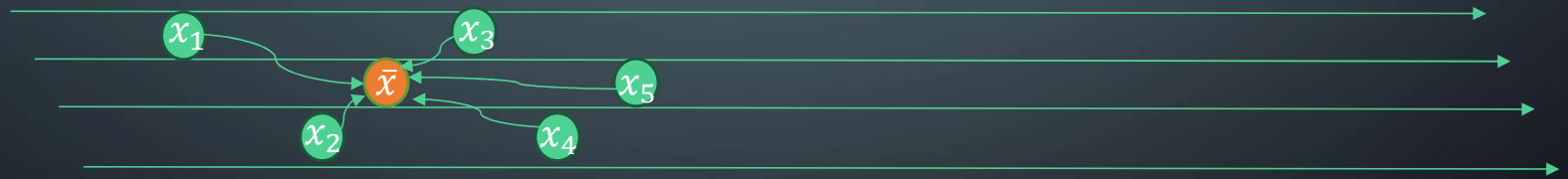


DATA PREPROCESSING

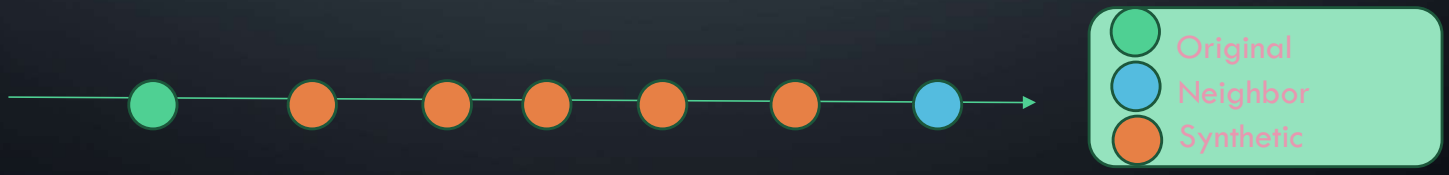
- Categorizing – All categorical labels must be converted to numerical values



- Missing Data – NaN's are replaced through mean imputation

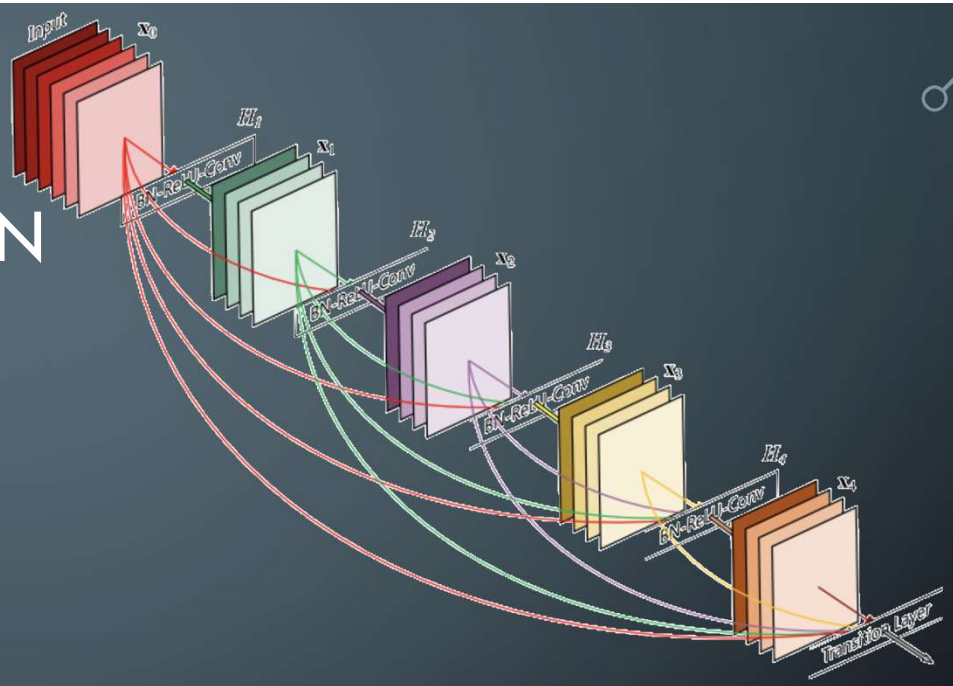


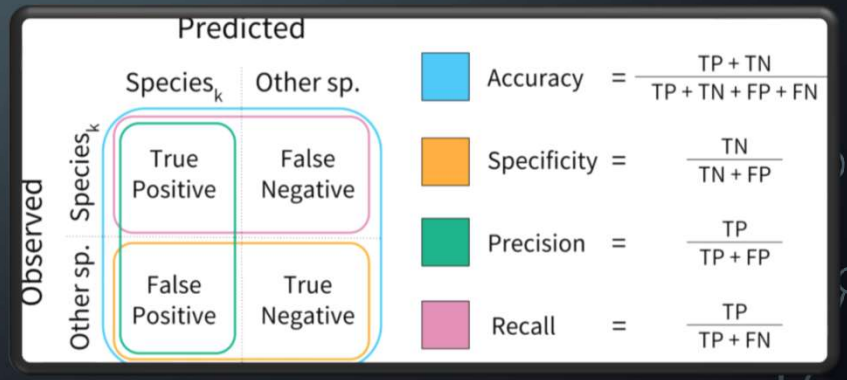
- Target Imbalance – Synthetic Minority Oversampling Technique (SMOTE)



TRAINING AND EVALUATION

- 80/20 split with batch size of 32
- Early stoppage to prevent over fitting
- Adam optimizer
- Model performance monitored through loss convergence, F1 score progression, and classification accuracy
- Multiple hyperparameters tested to optimize model





RESULTS

- Logistic Regression achieves highest accuracy (93.9%) but fails to detect strokes
- Neural Network shows balanced performance
- SVM maintains moderate recall for stroke cases
- All models struggle with stroke prediction precision
- Trade-off between accuracy and stroke detection visible

Logistic Regression Accuracy: 93.9%

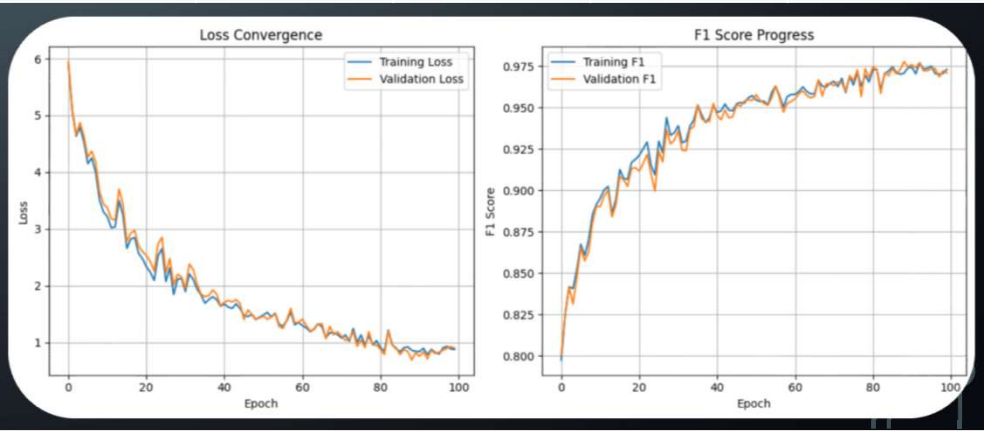
	precision	recall	F1-score	support
0	.94	1.00	.97	960
1	0	0	0	62

SVM Accuracy: 74%

	precision	recall	F1-score	support
0	.98	.74	.84	972
1	.12	.66	.20	50

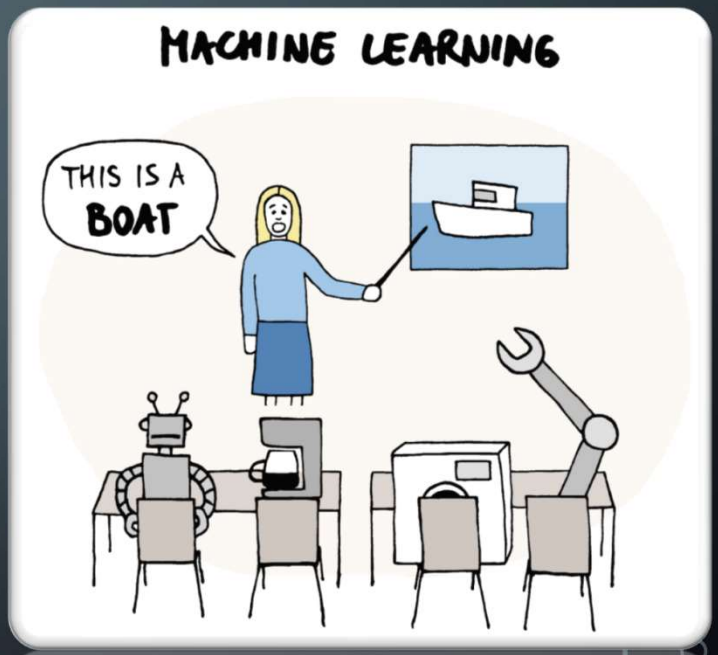
Neural Network Accuracy: 85%

	precision	recall	F1-score	support
0	.95	.88	.91	960
1	.14	.31	.19	62



LESSONS LEARNED

- Model performance shows a clear trade-off between accuracy and recall, highlighting challenges with imbalanced medical datasets
- Clinical deployment would require improved recall rates to be practically useful for stroke prediction
- With a larger dataset and deeper models, we could potentially find stronger correlations in the data



INDIVIDUAL CONTRIBUTIONS

Daniel Jolin: Logistic regression, neural network, result interpretation, Presentation.

Andrew Nicola: Finding Data set, Data visualization, SVM, Presentation.

Nathan Waters: Finding Data set, Data preprocessing, Presentation.

THANK YOU

ANY QUESTIONS?