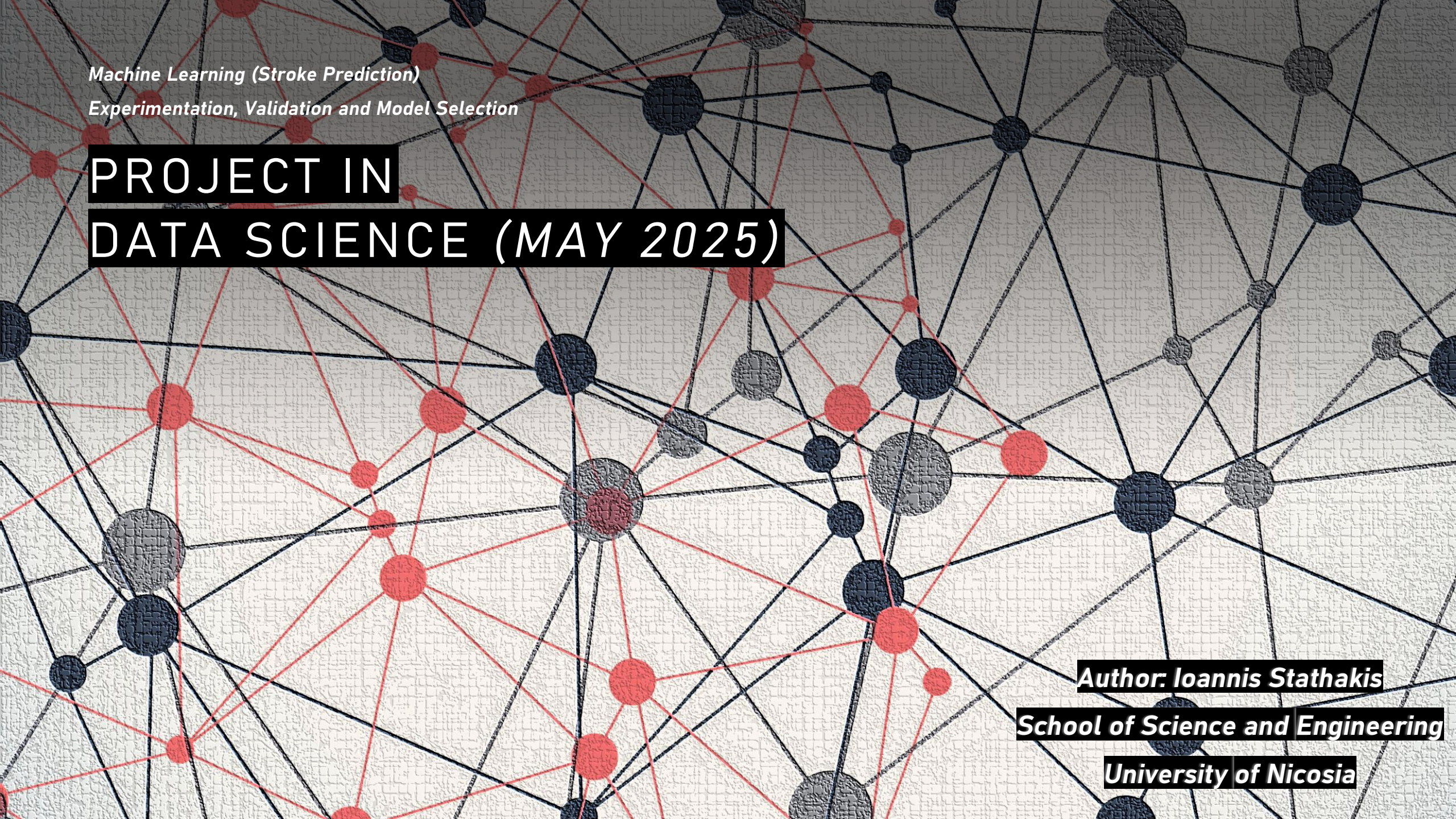


Machine Learning (Stroke Prediction)

Experimentation, Validation and Model Selection

PROJECT IN DATA SCIENCE *(MAY 2025)*

Author: Ioannis Stathakis

School of Science and Engineering

University of Nicosia

PROJECT MOTIVATION/OBJECTIVES

- ***Stroke is a leading cause of death and long-term disability***
- ***Early detection can drastically reduce risk and improve outcomes***
- ***Machine Learning offers tools to support clinical decision-making***
- ***We will explore and compare multiple machine learning models, selecting the most effective based on specific valuation metrics***

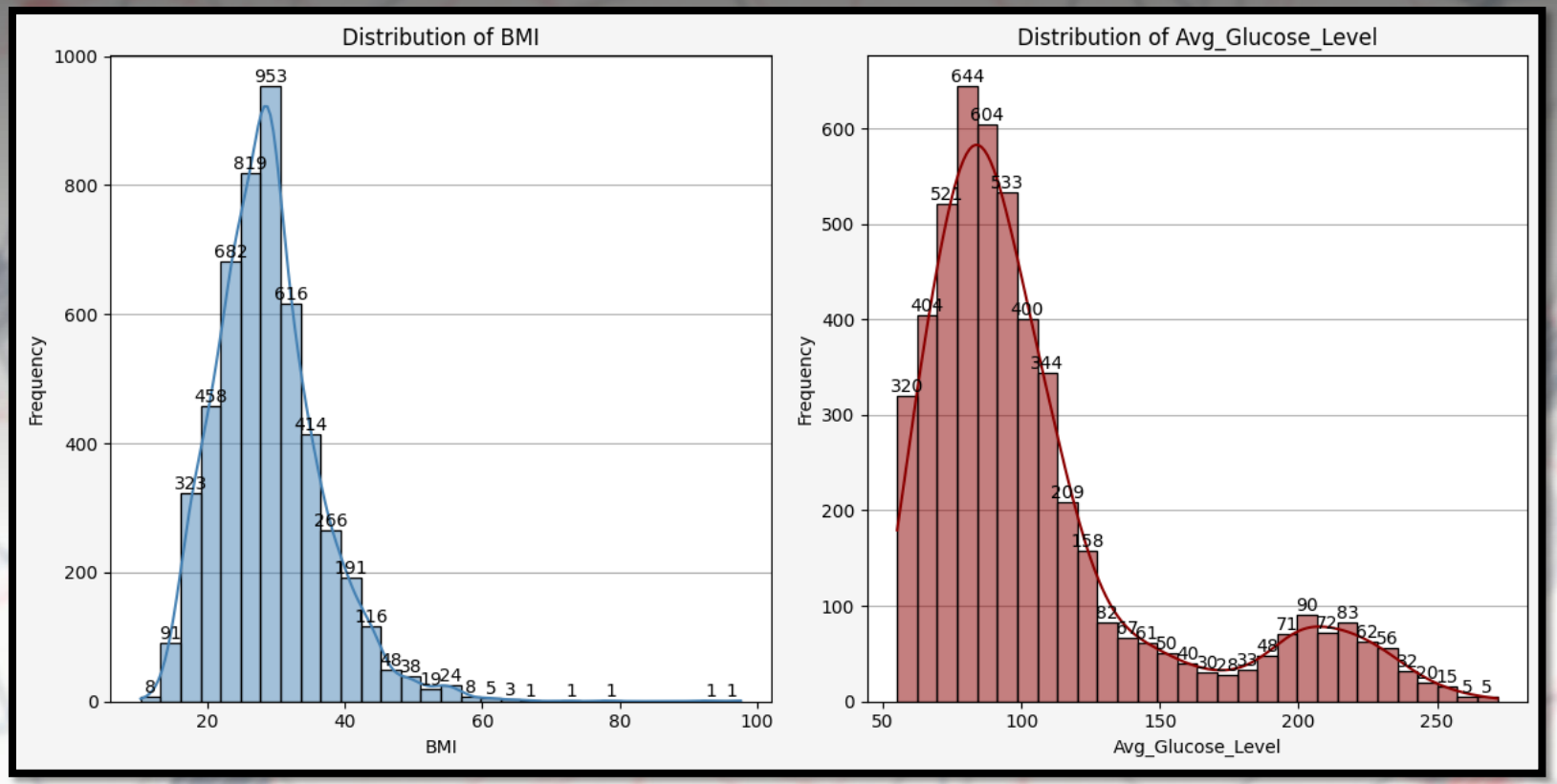
DATASET OVERVIEW

- ***Publicly available dataset – 5,110 patient records***
- ***Features include age, gender, smoking status, BMI, average glucose level, hypertension, and others***
- ***Mixed feature types (categorical – numerical)***
- ***Target variable: Stroke Occurrence (binary)***
- ***Severe target class imbalance: stroke cases ≅ 5% of total***

DATA PREPARATION

- ***Formatting issues – dropping of redundant features***
- ***Feature engineering: "Age Group" feature addition to help with stratified analysis***
- ***Missing values imputation: BMI feature (median BMI per Age Group)***
- ***Detection of BMI and Average Glucose Level outliers***
- ***Bimodality of Average Glucose Level also present***

EXPLORATORY DATA ANALYSIS (EDA)

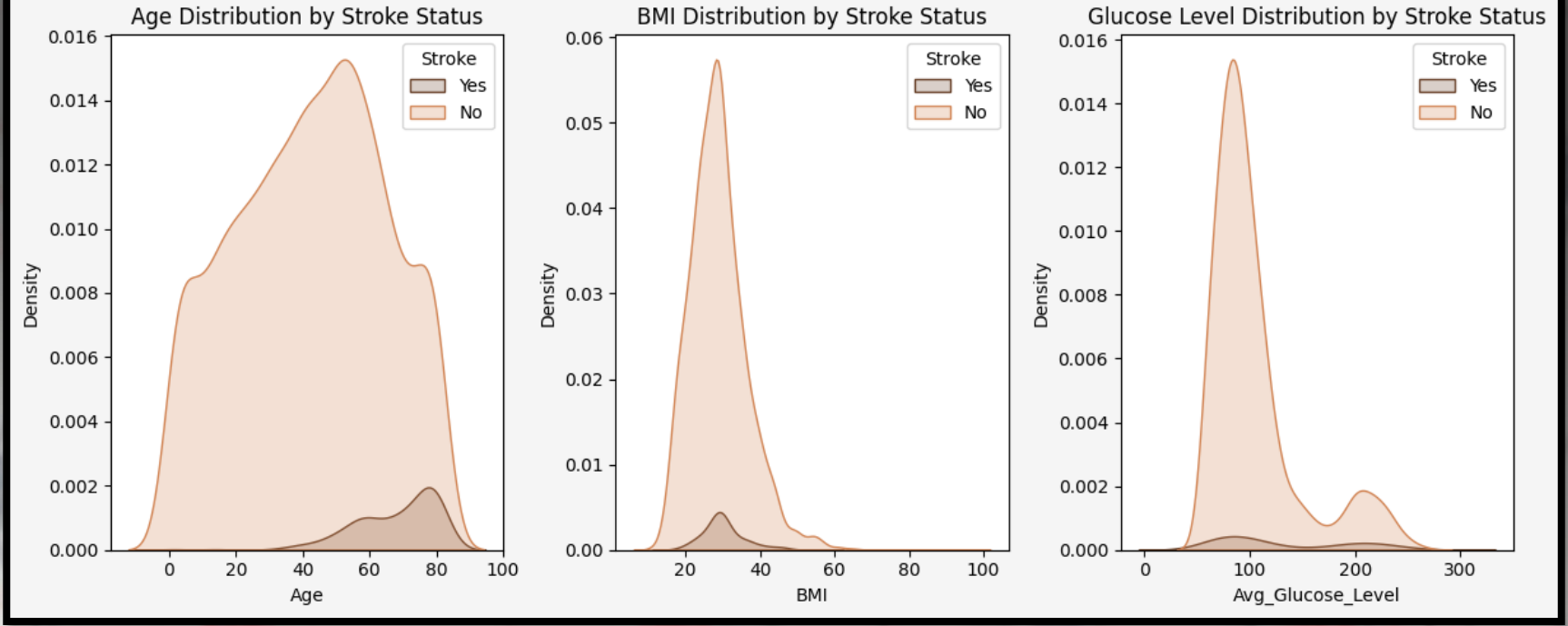


- *All non-binary numerical features were not normally distributed*
- *Outliers:*

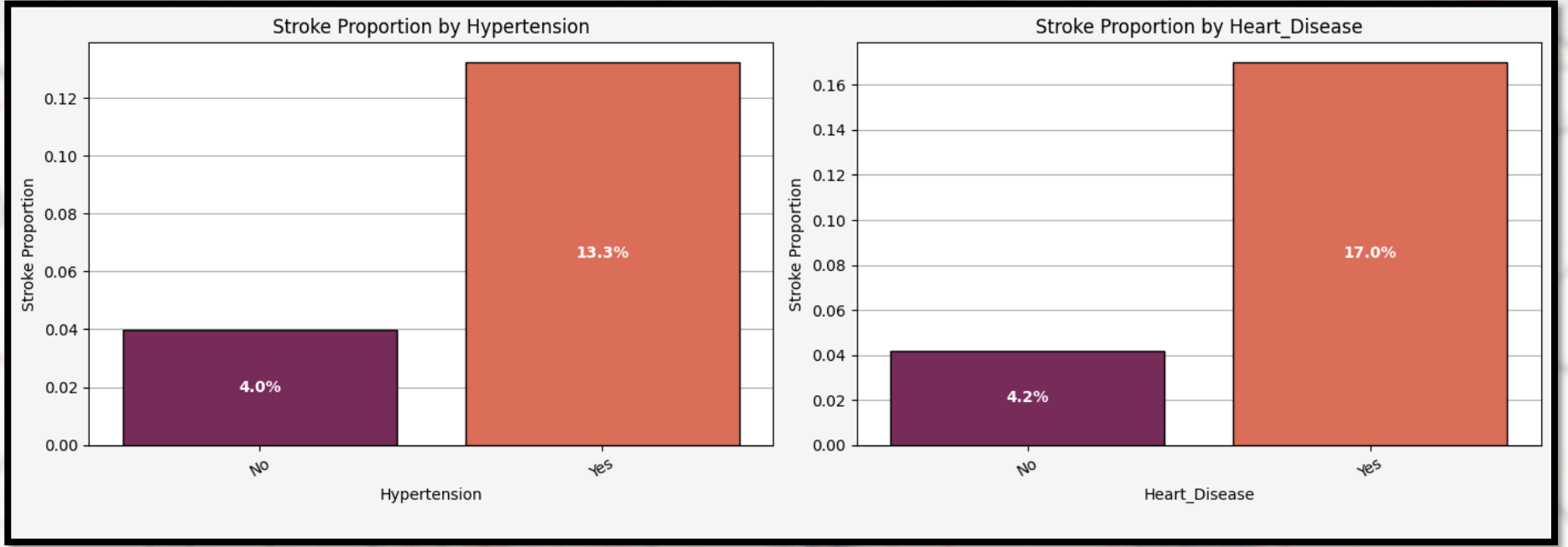
  $\cong 2.5\%$ *for BMI and*

  $\cong 12.5\%$ *for Glucose*
- *Outliers and glucose bimodality were retained with medical relevance/importance in mind*

- *Statistical Testing: Mann-Whitney U Test, Chi-Squared Test for Independence*
- *Tests confirmed significance for all differences noted in the visualizations*
- *Multicollinearity: Spearman Rank Correlation (due to non-normality)*
- *Final Feature Selection: Visuals, Statistical Tests, Mutual Information, Random Forest Importance*



- ***KDE-plots revealed continuous feature distribution differences per target subclass***



- ***Proportional differences were also evident in categorical variables, per target subclass***

MODELING STRATEGY

- ***Four different sampling strategies employed:***
  - ***Full SMOTE Oversampling (50-50 class balance)***
  - ***Reduced SMOTE Oversampling (40:60 class ratio)***
  - ***Downsampling (reducing majority class)***
  - ***Class-Weighting (inherent model class weighting scheme)***
- ***Main metrics taken into consideration:***

  Recall (Sensitivity), Specificity, ROC-AUC

- ***Models used (categorized):***

1. ***Baseline Models: Logistic Regression, Naïve Bayes, K-Nearest Neighbors***
2. ***SVMs: Support Vector Machines (Linear & RBF Kernel)***
3. ***Tree-Based Models: Random Forest, XGBoost***
4. ***Neural Network Models: Two-Class Averaged Perceptron, Feedforward Neural Network (FFNN)***
5. ***Deep Learning/Transfer Learning Model: TabNet***

PERFORMANCE COMPARISONS - RESULTS

Model	Accuracy	F1 Score	Recall	Specificity	Precision	AUC
Logistic Regression	0.808	0.235	0.6	0.819	0.146	0.773
Naive Bayes	0.759	0.169	0.5	0.773	0.102	0.72
KNN	0.78	0.206	0.58	0.79	0.125	0.743
SVM (Linear)	0.798	0.226	0.6	0.808	0.139	0.767
SVM (RBF)	0.664	0.196	0.82	0.662	0.111	0.833
Perceptron	0.805	0.238	0.62	0.815	0.148	0.718
FFNN	0.806	0.215	0.54	0.82	0.134	0.777
TabNet	0.558	0.151	0.8	0.545	0.083	0.758

- ***Summary table of model performance under the Full SMOTE sampling strategy***

Model	Accuracy	F1 Score	Recall	Specificity	Precision	AUC
Logistic Regression	0.834	0.207	0.44	0.854	0.135	0.775
SVM (Linear)	0.838	0.187	0.38	0.862	0.124	0.753
SVM (RBF)	0.747	0.237	0.8	0.744	0.139	0.84
Perceptron	0.869	0.249	0.44	0.892	0.173	0.666
FFNN	0.81	0.179	0.42	0.831	0.114	0.761
TabNet	0.766	0.19	0.56	0.777	0.115	0.791

➢ ***Summary table of model performance under the Reduced SMOTE sampling strategy***

Model	Accuracy	F1 Score	Recall	Specificity	Precision	AUC
Logistic Regression	0.78	0.258	0.78	0.78	0.155	0.839
SVM (Linear)	0.757	0.254	0.84	0.753	0.149	0.838
SVM (RBF)	0.724	0.226	0.82	0.719	0.131	0.84
Perceptron	0.852	0.249	0.5	0.87	0.166	0.685
FFNN	0.859	0.265	0.52	0.876	0.178	0.803
TabNet	0.939	0.0	0.0	0.988	0.0	0.411

➢ ***Summary table of model performance under the Downsampling strategy***

Model	Accuracy	F1 Score	Recall	Specificity	Precision	AUC
Logistic Regression (Weighted)	0.7	0.22	0.86	0.692	0.126	0.845
SVM (Linear - Weighted)	0.687	0.216	0.88	0.677	0.123	0.846
SVM (RBF - Weighted)	0.67	0.196	0.82	0.662	0.111	0.833

- ***Summary table of model performance under the Class-Weighting strategy***

Sampling Strategy	Best Model	Recall	Specificity	AUC
Full SMOTE	SVM (RBF)	0.82	0.662	0.833
Reduced SMOTE	SVM (RBF)	0.8	0.744	0.84
Downsampling	SVM (RBF)	0.82	0.719	0.84
Weighted	SVM (Linear - Weighted)	0.88	0.677	0.846

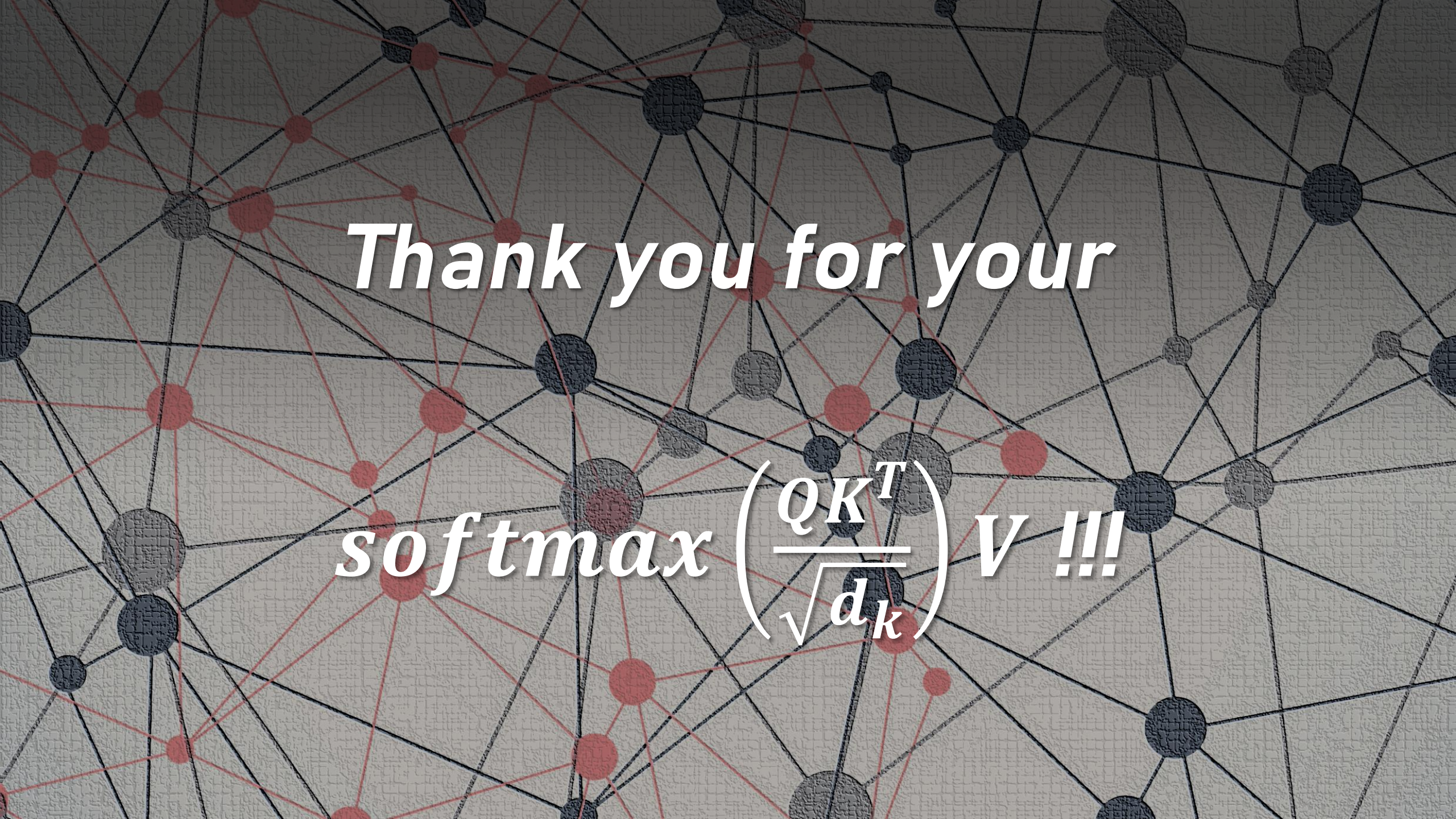
- ***Best model per sampling strategy employed (universal results)***

FINAL MODEL SELECTION

- ***Final model of choice*: SVM (Linear Kernel) under the Weighted sampling strategy**
- ***Reasoning*:**
  - **Achieved the highest recall (0.88)**
  - **Solid specificity (0.68) and strong AUC (0.85)**
  - **Did not require synthetic data / extensive hyperparameter tuning**
  - **Showed no signs of overfitting**
  - **Consistent results across various sampling strategies**
  - **Simple, fast and easy to interpret and deploy**

KEY TAKEAWAYS

- ***Data preprocessing and sampling strategies matter a lot, especially with imbalanced medical datasets***
- ***Recall (sensitivity) should take priority in a medical context when the cost of a false negative (like missing a stroke) is high***
- ***Combining EDA, Statistical Testing, and model metrics gives a complete picture***
- ***Simple models can outperform Deep Learning approaches if tuned well and paired with the right data strategy***
- ***Having a clear model selection process and modeling strategy makes the final choice easier to justify***


Thank you for your
$softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$!!!