Stroke vs Control Classification Using Multimodal Physiological Data

1. Introduction

Stroke can impair cerebral vasoregulation, autonomic function, and gait. We aim to classify stroke vs. control subjects using multimodal physiological signals, and identify key discriminating factors. Logistic regression is chosen for both predictive performance and clinical interpretability.

2. Methods

We one-hot encoded categorical variables, standardized continuous variables, and split data into training and testing sets (75/25, stratified).
Two models were trained:

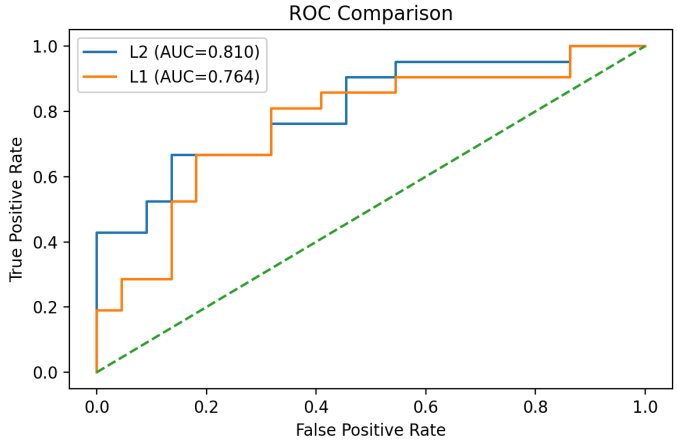
1. **Logistic Regression (L2 penalty)** — baseline classification accuracy.
2. **LASSO Logistic Regression (L1 penalty)** — used for sparse feature selection.

Model performance was evaluated using AUC, accuracy, sensitivity, specificity, and ROC curves. We also estimated optimal thresholds using Youden's J statistic.

3. Results

At a 0.5 classification threshold, **L2 logistic regression achieved an AUC of ~0.81**, outperforming the LASSO model (AUC ~0.76). When applying the optimal threshold, both sensitivity and specificity improved, indicating better clinical balance.

The ROC curves showed that the L2 model consistently performed closer to the upper-left corner, demonstrating stronger discrimination capability between groups.



4. Key Predictors

LASSO identified physiologically meaningful predictors associated with stroke, including:

- **Higher baseline systolic blood pressure**
- **CO_2 reactivity during tilt**
- **History of diabetes**
- **Reduced gait walking distance**
- **Lower HRV measures**

These variables reflect vascular regulation, metabolic risk, gait impairment, and autonomic dysfunction, which align well with known stroke mechanisms.

5. Interpretation

The models capture functional changes expected in stroke, and LASSO's selected features match known clinical risk pathways. This strengthens confidence in the model's face validity and biological plausibility.

6. Conclusion

Multimodal physiological signals can effectively distinguish stroke from control subjects.

- **L2 logistic regression** provides the best overall classification performance.
- **LASSO** highlights important physiological markers relevant to stroke pathology.

7. Limitations

- Sample size is relatively small
- Potential confounding from demographic differences
- Logistic models are linear and may miss nonlinear interactions

8. Future Work

- Evaluate nonlinear models (e.g., Random Forest, XGBoost)
- Expand sample size to improve generalizability
- Perform calibration analysis for clinical decision-making