Longitudinal Modeling of ALS Severity Using Machine Learning

Daniel Kang

February 1, 2026

1 Objective

The objective of this project is to build a predictive model that estimates how Amyotrophic Lateral Sclerosis (ALS) progresses over time using patient data. The project focuses on modeling changes in disease severity using clinical measurements, demographic information, and available genetic or biomarker data. The goal is to better understand the progression patterns and identify factors that are associated with faster or slower decline.

2 Outcome

The final outcome of this capstone will be a reproducible predictive modeling pipeline implemented in Python, structured as a modular analysis workflow suitable for reuse and extension. The project will produce trained and evaluated models that predict ALS progression using longitudinal data. Additional deliverables will include exploratory data analysis, visualizations of disease trajectories, and a written report describing the modeling choices, results, and limitations.

3 Impact

This project may benefit healthcare providers and researchers who study ALS by offering a data-driven way to analyze disease progression. More accurate progression estimates help support clinical understanding and decision making, and improve patient grouping in research studies. The project may also help students and researchers better understand how machine learning methods can be applied to real-world clinical data.

4 Literature Review

Prior research has explored ALS progression using clinical trial and observational data [?]. Existing studies show that ALSFRS-R scores are a strong indicator of disease severity over time and are commonly used in predictive models [?]. Other work has demonstrated the value of combining clinical data with genetic and biomarker information to better capture patient heterogeneity [?, ?]. Large research initiatives have also emphasized the importance of coordinated clinical and biological data collection to improve ALS research and treatment development [?, ?]. Recent clinical trials further highlight how progression metrics are used to evaluate treatment effects [?].

1. Cedarbaum et al. introduced the ALS Functional Rating Scale (ALSFRS) as a standardized clinical measure for assessing disease severity and functional decline in ALS patients [?].

2. Rooney et al. used longitudinal ALSFRS data to model disease progression and survival outcomes, demonstrating the value of repeated clinical measurements in ALS progression analysis [?].

3. Su et al. conducted a large meta-analysis identifying clinical, demographic, and biomarker-related factors associated with ALS survival, highlighting the limitations of relying on single predictors [?].

4. Tam et al. showed that ALS exhibits significant biological heterogeneity by identifying distinct molecular subtypes, supporting the need to consider genetic and biological factors alongside clinical data [?].

5. National research initiatives such as ACT for ALS emphasize coordinated collection of clinical, genetic, and biomarker data to improve ALS research and therapeutic development [?].

6. The CDC ALS Annual Meeting Summary Report highlights the role of large registries and standardized data collection efforts in supporting ALS research and longitudinal analysis [?].

7. Miller et al. demonstrated how progression metrics such as changes in ALSFRS-R scores are used in clinical trials to evaluate treatment effects, reinforcing their importance in ALS outcome modeling [?].

5 Novelty

This project builds on prior ALS research by focusing on longitudinal prediction while integrating multiple types of patient data. While many studies analyze disease progression using a single dataset or emphasize either clinical or biological features alone, this capstone combines clinical measures with available genetic and biomarker information to better reflect patient variability. The project also emphasizes reproducible modeling and transparent evaluation, which extends prior research by framing ALS progression modeling within a clear and reusable data science pipeline. By prioritizing interpretability and communication, this work adds value beyond descriptive or cross-sectional analysis commonly found in the literature.

6 Data Source(s)

The primary data sources for this project include the PRO-ACT database, Answer ALS, and the Target ALS Data Engine. These datasets contain de-identified patient data such as demographic variables, ALSFRS-R scores, treatment history, and genetic or biomarker information. The data is widely used in ALS research and is considered reliable, though missing values and inconsistent formats are expected. All datasets are publicly available for research use and are fully de-identified, which reduces privacy concerns while still presenting realistic challenges associated with healthcare data.

7 Approach

The project will be completed in these steps:

1. Acquire and load ALS datasets into Python.

2. Clean and preprocess the data, including handling missing values.
3. Perform exploratory data analysis to understand trends and distributions.
4. Engineer features relevant to ALS progression.
5. Train predictive models such as random forest regression or survival models.
6. Evaluate model performance using appropriate metrics.
7. Summarize results and document findings.

8 Timeline

- **Week 1: Data Acquisition and Setup**
  - Obtain access to ALS datasets (PRO-ACT, Answer ALS, Target ALS)
  - Download and organize raw files
  - Set up project repository and folder structure
  - Load initial datasets into Python
  - Perform basic data inspection and documentation
- **Week 2: Data Cleaning and Preprocessing**
  - Handle missing values and inconsistent formats
  - Standardize variable names and units
  - Merge datasets where appropriate
  - Create clean analysis-ready tables
  - Document preprocessing pipeline
- **Week 3: Exploratory Data Analysis**
  - Analyze distributions of key variables
  - Visualize ALSFRS-R progression trajectories
  - Explore demographic and clinical patterns
  - Identify potential predictive variables
  - Summarize initial insights
- **Week 4: Feature Engineering**
  - Construct longitudinal features (time-based variables)
  - Create derived progression metrics
  - Process genetic and biomarker variables
  - Encode categorical variables
  - Prepare final modeling dataset
- **Week 5: Baseline Model Development**

  - Implement baseline models using clinical features only
  - Train regression and/or survival models
  - Evaluate initial performance
  - Identify modeling challenges
  - Document baseline results

- **Week 6: Extended Model Development**
  - Incorporate genetic and biomarker features
  - Train extended predictive models
  - Compare against baseline models
  - Perform feature importance analysis
  - Refine modeling workflow

- **Week 7: Model Evaluation**
  - Select appropriate longitudinal evaluation metrics
  - Compare predictive accuracy across models
  - Perform cross-validation
  - Analyze trajectory prediction performance
  - Document evaluation methodology

- **Week 8: Interpretation and Robustness**
  - Interpret model results
  - Assess impact of genetic/biomarker features
  - Perform sensitivity analyses
  - Evaluate model stability
  - Generate summary visualizations

- **Week 9: Results Refinement**
  - Finalize figures and tables
  - Refine written interpretations
  - Validate reproducibility of pipeline
  - Prepare presentation materials
  - Draft final report sections

- **Week 10: Final Deliverables**
  - Complete final report
  - Prepare poster and presentation slides
  - Finalize documentation and code repository
  - Conduct final review and edits

9 Possible Challenges

One challenge is dealing with missing or incomplete clinical data. This will be addressed using careful preprocessing and imputation methods. Another challenge is avoiding overfitting due to limited sample sizes for certain patient groups. Model validation and transparent reporting of limitations will be used to mitigate this issue. Another challenge may be selecting appropriate evaluation metrics for longitudinal predictions, which will be addressed by comparing multiple metrics and clearly documenting modeling assumptions. Interpreting results responsibly is also important given the sensitive nature of healthcare data.

References