Empowering Health in Aging Populations: A Multimodal Vulnerability Tool for Frail Patients

Joanna G. Kondylis KONDYLIS@MIT.EDU
Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts

Houman Javedan HJAVEDAN@BWH.HARVARD.EDU
Division of Aging, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts
Harvard Medical School, Boston, Massachusetts

Dimitris Bertsimas * DBERTSIM@MIT.EDU
Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts

Bharti Khurana * BKHURANA@BWH.HARVARD.EDU
Trauma Imaging Research and Innovation Center, Brigham and Women's Hospital, Boston, Massachusetts
Harvard Medical School, Boston, Massachusetts

Abstract

Frailty is a powerful predictor of adverse outcomes in older adults, yet its routine assessment remains limited in acute care settings due to the labor-intensive nature of the clinical Frailty Index (FI) scoring, requiring geriatric specialists and meticulous clinical assessment. We developed and externally validated the first automated multimodal vulnerability tool that provides a real-time risk assessment, integrating structured EHR data, clinical narratives, and CT imaging. Using data from two major Boston hospitals in the Mass General Brigham system, we trained models to predict six outcomes: 3- and 6-month all-cause mortality, 3- and 6-month hospital readmission, 6-month fall risk, and 1-year recurrent fall risk. Our multimodal approach achieved AUCs of 0.74-0.86, with improvements of up to 4.3% over single-modality models and 8-49% over FI's predictive power. Beyond outcome prediction, we also sought to mirror clinical practice, where discrete frailty levels guide care planning. To this end, we developed a four-tier stratification system using k-means clustering and Optimal Policy Trees. This produces interpretable decision rules that assign patients to Non-, Pre-, Moderately-, and Severely- Vulnerable categories, actionable classifications that directly inform interventions, from fall prevention to advance care planning, while adding significantly to the prognostic ability of frailty assessments.

Keywords: Multimodal Machine Learning, Prognostic Classifiers, Optimal Policy Trees, Interpretability, Aging, Frailty

Data and Code Availability Our data includes patients over 65 with frailty evaluations at Massachusetts General Hospital and Brigham and Women's Hospital. The data and code are not publicly available due to the risk of exposing protected health information and institutional restrictions. Details on data preparation and modeling implementation are in Methods. Researchers interested in more details on the methodologies can contact the first author.

Institutional Review Board (IRB) Our Protocol IRB2024P000115 focuses on building a multimodal approach to vulnerability prediction in older adult patients.

1. Introduction

Importance: Frailty assessment, based on comprehensive geriatric assessments in older adults, can be used to identify deficits, better characterize aging and disease processes, and predict adverse outcomes. Outcomes of interest include fall prevention, mortality risk reduction, and healthcare resource allocation. Existing approaches to frailty assessment have scalability and reliability challenges that limit utility. The Frailty Index (Rockwood and Mitnitski, 2011; Cooper et al., 2021) requires manual chart abstraction and

* These authors contributed equally

an extensive questionnaire administered by a geriatrician, making it resource intensive and impractical for routine use (Boreskie et al., 2022; Malmstrom et al., 2014). Manual assessment methods are prone to inter-rater reliability issues depending on physician experience (Hörlin et al., 2022), while existing automated approaches focus on single data modalities (Gilbert et al., 2018; Kochar et al., 2025), missing the comprehensive picture necessary for accurate assessment. Most critically, current tools predict limited outcomes (typically mortality alone) rather than the multiple adverse events relevant to geriatric care planning.

Approach: We have developed the first comprehensive, automated, multimodal assessment tool for vulnerability in frailty that integrates structured Electronic Health Records (EHR) data, unstructured clinical narratives (visit notes, discharge summaries, etc.), and CT imaging (abdominal/pelvic and axial brain scans). Leveraging the Holistic AI in Medicine (HAIM) framework (Soenksen et al., 2022), our approach predicts six clinically relevant outcomes after inpatient visits: 3- and 6-month all-cause mortality, 3- and 6-month hospital readmission, 6-month fall risk, and 1-year recurrent fall risk when applicable. We evaluate model performance using the Area Under the Receiver Operating Characteristic Curve (AUC), sensitivity, specificity, Precision Recall-AUC (PR-AUC), and F1 score. We compare this holistic model to single-modality baselines (e.g., tabular alone), demonstrating the power of multimodality to fully capture patient profiles.

We transformed the predicted probabilities from the outcomes of the HAIM model, along with key demographic and comorbidity features, into a per-visit vector. Using k-means clustering, we partition these visits into four vulnerability categories: Non-, Pre-, Moderately-, and Severely Vulnerable. These clusters reflect unique patient phenotypes with differing risk profiles across multiple clinical outcomes. We then train an Optimal Policy Tree (OPT) (Amram et al., 2022; Interpretable AI, 2025), to map patient features to one of these four classes. OPTs are rule-based models: at each branch, the tree checks whether a feature is above or below a threshold, making decision paths transparent and easy to interpret. Thus, we produce clear decision rules that link clinical context to category. The complete workflow is pictured in Figure 1.

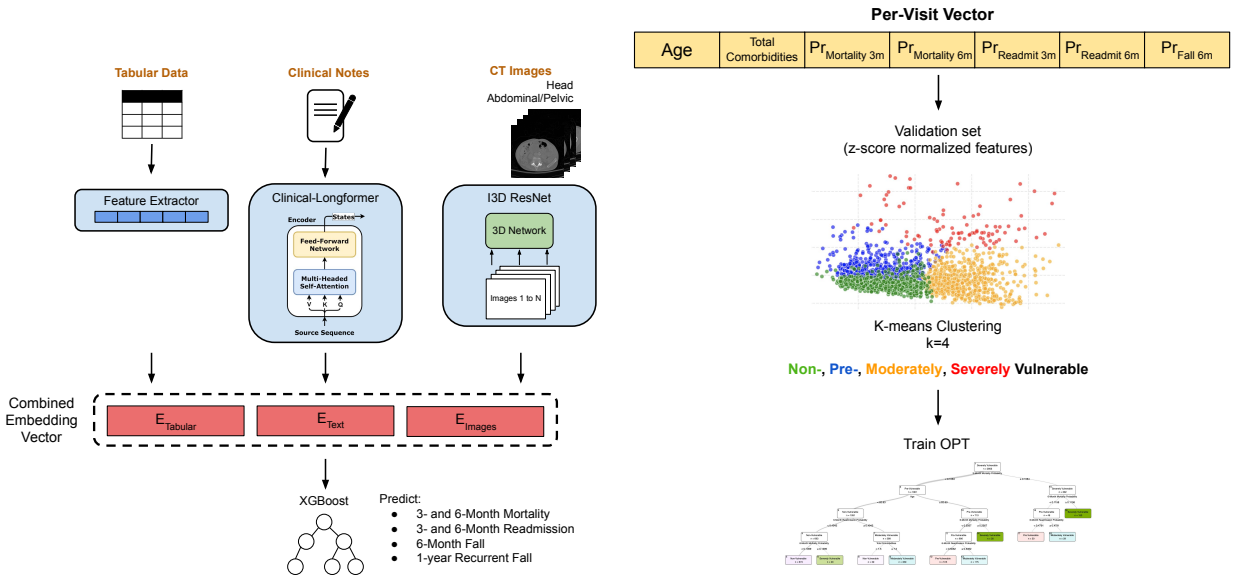
Contributions:

- Unified multimodal framework: To our knowledge, we are the first to develop an automated vulnerability tool for frail patients that simultaneously processes structured EHR data, clinical text, and medical imaging for comprehensive geriatric risk prediction.
- Multiple outcome prediction: Unlike existing tools that focus on single outcomes, we predict six adverse events that directly inform intervention planning and resource allocation decisions.
- Interpretable and automated vulnerability stratification: Via k-means clustering, we stratify patients into four buckets of vulnerability: Non-, Pre-, Moderately-, and Severely. An OPT is then trained to assign patients to these levels based on feature thresholds, producing transparent decision rules. Represented as integer labels (0–3), we show that this new stratification demonstrates significantly stronger predictive power for adverse clinical outcomes compared to the Frailty Index alone.

2. Related Work

Clinical Frailty Index: Clinical frailty is a syndrome characterized by increased susceptibility to external stressors and is driven by a cumulative deterioration of the body. Unlike chronological age alone, which has more limited predictive power for health outcomes, frailty provides a quantitative framework for modeling health risk trajectories (Schuurmans et al., 2004; Joseph et al., 2014; Makary et al., 2010).

The Frailty Index (FI) captures a patient's risk by aggregating health indicators including comorbidities, medications, functional limitations, and disability markers into a normalized score. The score is computed as the number of deficits present divided by the total evaluated. The deficit-set typically comprises 30-70 predefined variables. Scores range from 0 to 1, with an empirical upper bound around 0.7 (Searle et al., 2008). Values approaching this threshold indicate severe decline. The FI has strong predictive validity for mortality and chronic disease burden (Kojima et al., 2017; Hall et al., 2017), but notable limitations. It does not consistently predict falls or hospitalization (Si et al., 2021; Schoufour et al., 2015). Furthermore, its score captures a static snapshot of vulnerability, rather than a continuously updating target of health risk. These constraints limit its scalability and timely use in acute care and health



(*a*) First part of the pipeline, training models for outcome prediction (adapted from dvgodoy (2021))

(*b*) Second part of the pipeline, learning an OPT to assess vulnerability.

Figure 1: The complete methodology.

management. The FI is typically interpreted in four levels: very mildly frail (< 0.2), mildly frail (0.2 to < 0.3), moderately frail (0.3 to < 0.4), and severely frail (≥ 0.4) (Blodgett et al., 2021). Stratifying patients into discrete frailty levels is important because these categories directly inform prognosis and care planning. Mirroring this approach, we defined four vulnerability strata in our system to maintain the clinical interpretability.

We specifically evaluated the Comprehensive Geriatric Assessment Frailty Index (CGA-FI) (Cooper et al., 2021), administered starting in 2018, which uses the CGA questionnaire to identify patient deficits in areas such as nutrition, balance, comorbidities, medications, and cognition, that are then incorporated into the frailty score.

Multimodal Artificial Intelligence for Healthcare: Recent advances in multimodal AI show strong promise for clinical prediction by integrating diverse data sources. Soenksen et al. (2022) introduced the Holistic AI in Medicine (HAIM) framework, a unified pipeline for processing and integrating tabular EHR data, time-series measurements, clinical text, and medical images. Evaluating 14,324 models, they showed multimodal approaches consistently outperformed single-modality systems by 6-33% across various healthcare tasks, including mortality prediction and hospital length-of-stay estimation. Subsequent work has introduced Explainable HAIM (xHAIM) (Petridis et al., 2025), which uses generative AI to maintain model interpretability. Our work enhances the HAIM framework to address the challenge of vulnerability assessment in frail populations: we simultaneously model six adverse events and transform these predictions into interpretable classes through unsupervised learning, bridging the gap between predictions and clinical decision-making

Benchmark Models: While our study focuses on frail older adults, comparable prediction models in general elderly populations provide important benchmarks. Mortality prediction studies report AUC values of 0.79–0.89 for outcomes within 6 months (Manz et al., 2020; Olender et al., 2023; Haimovich et al., 2024). Fall prediction models have achieved AUCs between 0.70-0.78 for 6-month horizons (Patterson et al., 2019; Capodici et al., 2025; Oshiro et al., 2019). For readmission prediction, Davis et al. (2022) reported an AUC of 0.83 in general populations but only 0.74 in elderly subgroups, while Mohanty et al. (2022) achieved 0.79. Despite the inherent challenges of predicting outcomes in frail populations, who typically have a higher baseline risk, our models match or exceed performance in existing literature.

3. Methods

3.1. Data

We extracted electronic health records from Massachusetts General Hospital, Brigham and Women's Hospital, and partner hospitals within the Massachusetts General Brigham system. The dataset includes patients with at least one FI evaluation from 2018 onwards, yielding 8,519 eligible patients. We pulled all clinical data for these patients from January 2015 through March 2024.

Our modeling focuses on predicting adverse outcomes during inpatient admissions, as these visits generate the most comprehensive clinical data, although we incorporate features derived from all prior clinical encounters. We require at least two years of patient history to build comprehensive feature sets, so our training and prediction cohorts include inpatient visits between January 2017 and March 2024. We exclude visits during which a patient died, as these represent different clinical trajectories that would confound prediction. Finally, we filtered visits to include only patients aged 65 and older.

These constraints yielded a dataset of 6,763 unique patients across more than 19,000 inpatient visits from 2017 onward, and over 1.13 million total clinical encounters from 2015 onward, including outpatient, emergency, and other non-inpatient visits. While outcome predictions were limited to inpatient encounters, the models incorporated longitudinal data from all visits since 2015 to construct complete patient profiles. All predictions were performed at the encounter level rather than aggregated by patient. This approach captures the dynamic nature of vulnerability, as a patient may have different risk profiles across multiple admissions depending on their evolving health status.

3.2. Target Variables

All-cause mortality: All-cause mortality was assessed at 90- and 180- days post-discharge. We applied a three-step approach to capture mortality events, detailed in Appendix A.1.

All-Cause Hospital Readmission: Our outcome is all-cause readmission within 90- and 180-days after discharge.

Falls: We identified patient falls using ICD9 and ICD10 codes. For ICD9 codes, any patient diagnosed with E880-E887 during a visit is considered to have fallen. For ICD10, we include W00-W19 and R29.6 (codes representing accidental falls and related events) (Kakara et al., 2023; Centers for Disease Control and Prevention, National Center for Health Statistics, 2024). Our first outcome is a 6-month fall prediction. Our second outcome is 1-year recurrent fall prediction, identifying patients who have fallen once and are at risk of falling again within the year. For this outcome, we predict on any visit with a fall diagnosis, lifting the restriction of predicting only on inpatient visits to maintain a sufficient cohort size.

For all target variables we do not predict on visits without adequate follow-up time (up to study completion).

3.3. Feature Accumulation

We partition the data at the patient level into training (80%), validation (10%), and test (10%) cohorts, ensuring that all visits from a patient reside in only one split, eliminating any risk of data leakage. We use the same patients in each split across all outcomes.

For the tabular-only, text-only, and HAIM combined models, we train on the entire training cohort and select the best checkpoint using validation AUC.

For the imaging models, we first create an internal 80-20 patient-level split within the training cohort; hyper-parameter tuning and model selection are performed on this 20% subset, with AUC as the criterion. The external validation and test sets remain untouched throughout development, preserving their integrity for unbiased final performance assessment.

For each modality, we use time series aggregation techniques to create features capturing both acute and chronic risk factors. To do so, we create features collecting data from seven time periods: current visit, 30, 60, 90, 180 days, 1 year, and 2 years prior to admission. For example, we incorporate the number of urinary-tract infections over each period as its own feature. This multi-temporal approach recognizes that risk factors operate on different timescales, where acute conditions and medication changes in the immediate time period may have different implications than chronic comorbidities accumulated over years.

Tabular Features: Our tabular data includes demographic, encounter, diagnostic, medication, procedure, physical, and laboratory records. We extract demographic features (age and gender) and encounter patterns (frequency and duration) to establish baseline patient characteristics. Physical measurements

include BMI, temperature, pain scores, and alcohol consumption, while labs comprise complete blood count, metabolic markers (A1C, TSH), and lipid profiles.

For diagnostic data, we convert all ICD-9 codes to ICD-10 format and apply the Clinical Classifications Software Refined (CCSR) (Agency for Healthcare Research and Quality, 2023a) to transform over 18,000 diagnostic codes into 530 clinically meaningful categories organized by body system. This aggregation strategy reduces sparsity while preserving clinical interpretability. Procedural data utilizes CPT codes mapped through CCSR into 248 categories Agency for Healthcare Research and Quality (2023b).

Our medication features focus on high-risk drug classes identified through keyword matching across six categories relevant to geriatric and hospitalized populations: anticoagulants/antithrombotics, insulin and high-risk diabetes medications, opioids, diuretics, high-risk cardiac medications, and medications from the Beers Criteria (CNS-active and others) (American Geriatrics Society, 2022; Budnitz et al., 2007; Sehgal et al., 2013). We additionally capture polypharmacy burden using thresholds of ≥ 5 (standard), ≥ 10 (severe), and ≥ 15 (extreme) concurrent medications (Masnoon et al., 2017).

Text Features: For the text data, we utilize all clinical notes per patient: cardiology, discharge, progress, pulmonary, visit, radiology, endoscopy, and H&P. We embed each note using Clinical Longformer (Li et al., 2022), a transformer architecture for lengthy clinical documents, into a 768-dimensional vector.

Abdominal/Pelvic CT Scans: For the abdominal/pelvic CT scans, we excluded patient scans with derived images, fewer than 20 scans in a series, and non-axial orientation following the methodology of (Bridge et al., 2018). We use all scans from 2015 onwards for training and inference. We adapted the Merlin foundation model (Blankemeier et al., 2024), a 3D vision-language model trained on paired CT scans, EHR diagnosis codes, and radiology reports that processes 3D voxel data. Following Merlin's preprocessing pipeline, we reformatted CT scans to the right-anterior-superior (RAS) orientation, resampled in-plane axial images to 1.5mm resolution and out-of-plane slice thickness to 3mm using bilinear interpolation, mapped Hounsfield units from $-1000 : 1000$ to the $0 : 1$ range, and applied padding and center cropping to $224 \times 224 \times 160$ voxels. We utilized only the image encoder component of Merlin to extract 2048-dimensional embeddings from the CT scans, excluding text-based features. We fine-tuned the pretrained Merlin model per outcome by adding a classification head of four fully connected layers with ReLU activations and dropout regularization, yielding a 128-dimensional embedding for downstream classification. Appendix A.2 compares Merlin's predictive performance with body composition metrics (muscle, subcutaneous fat, and visceral fat mass).

Head CT Scans: For the head CT scans, we follow a similar pre-processing methodology as for the abdominal pelvic CT scans. However, for these volumes we applied a (40-80) Hounsfield Unit window to emphasize soft brain tissue. We created an I3D ResNet152 model (Carreira and Zisserman, 2017), the same base architecture as the Merlin imaging component, which inflates a pretrained 2D ResNet152 architecture to 3D by replicating its convolutional kernels along the temporal dimension, enabling the transfer of ImageNet-learned features to volumetric CT data. We added two final layers fine-tuned to produce a 128-dimensional embedding.

Feature Selection: After accumulating our predictors across all modalities and timeframes, we obtained over $12,000$ unique features. To reduce computational complexity, mitigate overfitting, and improve model interpretability, we implemented a systematic feature selection pipeline.

Prior to feature selection, missing values were addressed using median imputation for numerical features. The imputation strategy was fitted exclusively on the training set to prevent data leakage, with learned parameters applied to validation and test sets. Importantly, no patients were excluded due to missing modalities; tabular data served as the only minimally required modality (available for all patients), while missing text or imaging features were imputed to maintain consistent input dimensionality.

Elastic net regularization (Zou and Hastie, 2005) was employed for automated feature selection, with regularization strength of 0.1 and L1 ratio of 0.1 (determined via 3-fold cross-validation) to address correlated features while promoting sparsity. The model used balanced class weights to account for class imbalance. After training, features with coefficient weights > 0.01 were retained, reducing the feature space from over $12,000$ to 80-400 of the most informative predictors.

4. Experimental Results

4.1. Training and Implementation Details

Multiple machine learning algorithms for data classification were evaluated for our tabular only, text only, and multimodal models, including logistic regression, XGBoost, support vector machines (SVM), and random forest classifiers. XGBoost (Chen and Guestrin, 2016) consistently demonstrated superior performance across all modalities and was therefore selected as the primary classifier for both single modality and multimodal implementations.

We conducted comprehensive hyperparameter tuning using grid search. The hyperparameter search space varied subsample and column sample by tree from 0.6 to 0.8, L1 and L2 regularization from 1 to 10, maximum tree depth from 4 to 7, minimum child weight from 1 to 3, learning rate from 0.05 to 0.3, and the number of estimators from 200 to 300.

Models were trained on the training set and hyperparameters were selected using the highest validation AUC score. This approach ensures hyperparameter selection is based on unseen validation data rather than on cross-validation training performance, which is particularly important for our multimodal models where imaging features were finetuned on portions of the training set during feature extraction. By optimizing on validation performance, the risk of overfitting to the training distribution is reduced. The final model evaluation was performed on the held out test set to provide unbiased performance estimates.

4.2. Individual Model Results

We first evaluated our models' performance on individual modalities: tabular features, text features, and images separately. Results are summarized in Figure 2. The tabular data consistently outperformed text and imaging, except for 6-month mortality for which abdominal CT is more powerful.

Text-based models showed moderate predictive capability, varying by outcome. The abdominal CT model achieved comparable AUCs to tabular data for mortality but no predictive value for readmissions and modest performance for falls (AUC 0.63), suggesting imaging captures physiological decline but not broader clinical factors. Head CT similarly predicted mortality best (AUC 0.71) with mild predictive value for falls. Complete modality results (AUC, sensitivity, specificity, PR-AUC, F1 score) appear in Appendix A.2.

4.3. Combined HAIM Model

Our multimodal models achieved notable performance improvements compared to the strongest single-modality models (Figure 2). For mortality tasks, AUC gains exceeded 0.03 (+4.3% improvement for both) demonstrating the complementary value of combining tabular, text, and imaging data. The exception was 3-month readmission, where adding modalities to tabular data provided no improvement, despite clinical notes having good predictive value, suggesting redundancy in the information captured. Sensitivity and specificity also improved (Table A.2 in Appendix A.2) relative to the best single-modality models, except for slight decreases in sensitivity in 3-month mortality and specificity in 6-month falls. For 3-month mortality the increases correspond to 2.24% and 4.94% respectively.

For mortality outcomes (3- and 6-month), age, neurocognitive disorders, and cancers dominated predictions. Dysphagia, malnutrition, and BMI also emerged as strong predictors, highlighting nutrition's critical role in prognosis. Readmission models (3- and 6-month) identified length of stay and inpatient history as highly predictive, alongside high-risk medications (diuretics, opioids, Beers criteria drugs), suggesting medication adverse effects may drive readmissions. Finally, the fall models have fall history, gait disorders, depression, and BMI among the significant predictors.

To interpret model decisions, we employed SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), with visualizations of HAIM versus tabular-only models in Appendix C.1. These plots rank features vertically by importance, with each visit shown as a colored dot positioned horizontally by its predictive impact (SHAP value).

4.4. Frailty Index Model

To compare our multi-modal approach to the existing Frailty Index's predictive power, we used the Frailty Index alone to predict outcomes (Figure 2). Because FI evaluations are not necessarily conducted during an inpatient visit, this model was not restricted to them; we predicted on every visit with a FI evaluation.

Our HAIM model significantly outperforms the existing Frailty Index for our six studied outcomes with improvements ranging from 8.15% to 49.1%. The Frailty Index shows moderate capability for predicting mortality with an AUC from 0.70 to 0.73 and

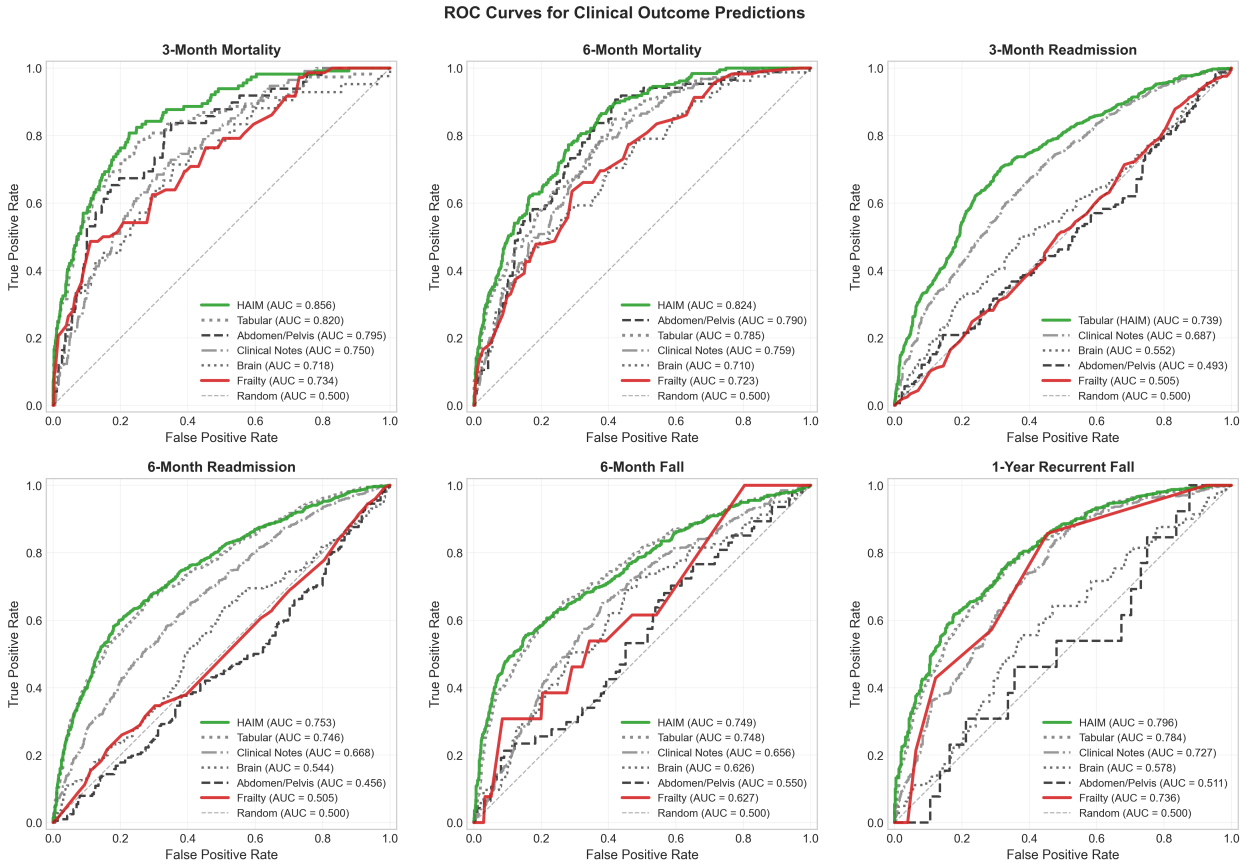


Figure 2: Test Set ROC curves of single-modality, HAIM, and Frailty Index only models.

weaker performance for falls, AUC 0.58–0.73. However, it is uninformative for assessing 3- and 6- month readmission risk, with AUCs close to random.

4.5. Methodological Considerations for Fair Comparison

The comparison in Figure 2 reports AUCs for the HAIM model and FI model on independent data points, where the former used all inpatient visits for both training and testing, and the latter was trained and tested only on visits with a recorded FI evaluation. The temporal anchor point for each prediction window was the discharge date of the inpatient visit (or, for the FI model, the date of the CGA-FI evaluation), from which 3- and 6-month outcomes were prospectively measured.

Using the most recent historical FI evaluation available for each inpatient visit would introduce confounding, as patients with more frequent evaluations would have more temporally proximate FI scores, and those recorded months before the inpatient encounter may not accurately reflect the patient's health at the time of prediction. To avoid this temporal mismatch and give the FI model its strongest footing, we evaluated it only on visits where FI scores were actually recorded. We chose inpatient visits, of which there are more than $19,000$ compared to approximately $9,000$ FI evaluations, for constructing our own vulnerability model as they represent a more common clinical setting and provide the richest clinical information.

To enable a direct comparison between models, we trained an additional HAIM model specifically for this analysis. This comparison model, identical in architecture and features to the main inpatient outcome model, was trained and tested exclusively on all visits with recorded FI evaluations (not restricted to inpatient settings). This design enables a direct comparison between the HAIM and FI models on the same

subset of visits, providing the fairest possible evaluation of their relative predictive performance. The results are summarized in Appendix B. As shown, this FI-subset HAIM model exceeds the original inpatient model in some outcomes and strongly outperforms the FI alone.

4.6. Vulnerability Phenotype Identification and Interpretable Decision Rules

Our final objective was to develop a vulnerability stratification system with clinical utility comparable to the four risk categories of the Frailty Index. Each patient visit was represented by a 7-dimensional feature vector: predicted probabilities from five HAIM models (3- and 6-month mortality, 3- and 6-month readmission, and 6-month fall risk), along with age and comorbidity count. The recurrent fall model was excluded because, unlike the other outcomes limited to inpatient visits, it was trained and validated on all encounter types with a fall, and inpatient-specific instances were sparse.

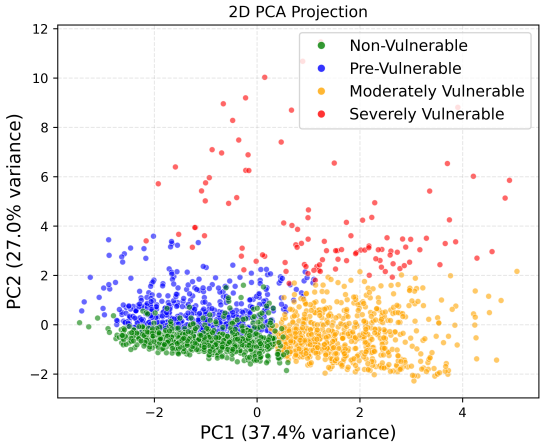


Figure 3: Characteristics of k-means clustering with k=4.

Our original training set, as defined in Methods, was reserved for training the HAIM prediction models. Since these models were already fit to this data, their predicted probabilities could, in some instances, be significantly more predictive than those in the validation and test sets. Reusing it for clustering would have meant that the clusters were based on overly optimistic risk estimates rather than the performance on unseen patients. Thus, we used the validation set ($> 2,000$ data samples) instead for the k-means clustering and for training the OPT. The held-out test set ($> 2,000$ visits) was not used for any model training or OPT evaluation. The final OPT model was applied to the test data to assign vulnerability classes.

On the validation set, we employed unsupervised k-means clustering with k=4. Prior to clustering, features were standardized using z-score normalization to ensure equal weighting across variables. Clusters are visualized using principal component analysis in Figure 3.

To classify new patients into these four categories, we trained an OPT that learns interpretable decision rules mapping the original 7-dimensional feature vectors (without normalizing) to one of the four cluster assignments. The OPT generates a hierarchical tree of threshold-based rules (e.g., if mortality risk $> X\%$, classify as cluster 4), providing transparent clinical logic while maintaining high accuracy. The OPT with decision rules is shown in Figure 12, with validation accuracy in Figure 11 (Appendix C.2).

Figure 4 shows cluster characteristics, with mean and standard deviation of the seven features across the four clusters. The clusters demonstrated meaningful patterns: "Non-Vulnerable" patients, on average younger, exhibited the lowest mortality risk with minimal comorbidity burden; "Severely Vulnerable" patients showed markedly elevated mortality risk; "Pre-Vulnerable" patients were older but with low mortality, fall, and readmission risk; and "Moderately Vulnerable" patients had high fall and admission risk.

The trained OPT was deployed on the test set to classify each patient into one of four vulnerability categories. To validate the prognostic value of these assignments, we performed downstream prediction tasks: cluster assignments were one-hot encoded and used as features in logistic regression models predicting each adverse outcome. Models were trained on the validation set without hyperparameter tuning to avoid test set leakage, then evaluated on the test set. The strong discriminative performance across all five outcomes (3- and 6-month mortality, 3- and 6-month readmission and 6-month falls) shown in Table 1 demonstrates that these new phenotypes encode meaningful risk information transferable to independent patient cohorts. The four categories outperformed the CGA-FI (which has two decimal precision) across all outcomes except 6-month falls. For the readmission tasks, improvements exceed 35%.

Cluster Characteristics: Mean ± Standard Deviation

Cluster	N (%)	Age	Comorbidities	Mortality 3m	Mortality 6m	Readmit 3m	Readmit 6m	Fall 6m
Non	673 (33.6%)	75.63 ± 4.49	6.19 ± 3.39	0.02 ± 0.02	0.02 ± 0.03	0.24 ± 0.10	0.32 ± 0.12	0.16 ± 0.07
Pre	536 (26.8%)	88.77 ± 4.38	5.66 ± 3.33	0.04 ± 0.04	0.06 ± 0.06	0.22 ± 0.08	0.31 ± 0.12	0.19 ± 0.09
Moderately	685 (34.2%)	80.09 ± 7.45	13.20 ± 3.33	0.04 ± 0.04	0.07 ± 0.07	0.46 ± 0.12	0.60 ± 0.12	0.29 ± 0.14
Severely	109 (5.4%)	86.13 ± 6.31	11.68 ± 3.82	0.26 ± 0.14	0.37 ± 0.15	0.44 ± 0.13	0.58 ± 0.13	0.24 ± 0.11

Figure 4: Characteristics of k-means clustering with k=4.

Table 1: Logistic regression on cluster one-hot encodings, trained on the validation (val) set and evaluated on the test set; FI Test AUC is shown for comparison.

Outcome	Val AUC	Test AUC	FI Test AUC
Mortality 3-Months	0.803	0.753	0.734
Mortality 6-Months	0.722	0.727	0.723
Readmit 3-Months	0.639	0.685	0.505
Readmit 6-Months	0.650	0.695	0.505
Fall 6-Months	0.636	0.604	0.627

5. Discussion

Our study presents a comprehensive multimodal framework for automated vulnerability assessment that addresses limitations of current frailty evaluation methods. By integrating structured EHR data, clinical narratives, and CT imaging through the HAIM framework, we achieved robust prediction across six adverse outcomes and classified patients into risk categories while maintaining clinical interpretability through optimal policy trees. The resulting four-tier vulnerability stratification system outperformed the traditional Frailty Index, particularly in readmission prediction where it achieved over 35% AUC gains compared to the FI's near-random performance. The most primary contribution of our approach is its dual capability: automating the labor-intensive frailty assessment process while providing more granular risk information than existing tools. This added vulnerability assessment can help geriatricians and non-geriatricians alike. For geriatricians, this tool enhances prognostication of mortality, falls, and readmissions within frailty subgroups, enabling more precise interventions based on the interplay between aging and disease patterns. For non-geriatricians, who typically focus on disease management, it serves as a screening tool to flag patients whose biological age diverges from chronological age, identifying those who may benefit from formal geriatric evaluation. Some limitations point to future directions. Our analysis was limited to inpatient visits, though adding outpatient encounters could allow earlier risk detection. While OPTs preserve interpretability, transparency is reduced when text is processed via Clinical-Longformer, motivating future use of xHAIM. It is also important to note that we only looked at a limited number of short term outcomes that do not cover the breadth and depth of the utility of frailty assessments, so these outcomes do not justify this tool as a replacement for frailty evaluations but rather as an additional tool available to clinicians.

References

Agency for Healthcare Research and Quality. Clinical classifications software refined (ccsr) for icd-10-cm diagnoses, version 2023.1. Technical report, Healthcare Cost and Utilization Project (HCUP), 2023a. URL https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp.

Agency for Healthcare Research and Quality. Clinical classifications software refined (ccsr) for icd-10-pcs procedures, version 2023.1. Technical report, Healthcare Cost and Utilization Project (HCUP), 2023b. URL https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp.

American Geriatrics Society. 2022 updated ags beers criteria for potentially inappropriate medication use in older adults, 2022. URL https://www.americangeriatrics.org/sites/default/files/inline-files/2022%20Updated%20AGS%20Beers%20Criteria%20For%20Comment%20Period%2011.10.22.pdf.

Maxime Amram, Jack Dunn, and Ying Daisy Zhuo. Optimal policy trees. *Machine Learning*, 111(7): 2741–2768, March 2022. ISSN 1573-0565. doi: 10.1007/s10994-022-06128-5. URL http://dx.doi.org/10.1007/s10994-022-06128-5.

Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, Christian Bluethgen, Malte Engmann Kjeldskov Jensen, Sophie Ostmeier, Maya Varma, Jeya Maria Jose Valanarasu, Zhongnan Fang, Zepeng Huo, Zaid Nabulsi, Diego Ardila, Wei-Hung Weng, Edson Amaro Junior, Neera Ahuja, Jason Fries, Nigam H. Shah, Andrew Johnston, Robert D. Boutin, Andrew Wentland, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, and Akshay S. Chaudhari. Merlin: A vision language foundation model for 3d computed tomography, 2024. URL https://arxiv.org/abs/2406.06512.

Joanna M Blodgett, Kenneth Rockwood, and Olga Theou. Changes in the severity and lethality of age-related health deficit accumulation in the usa between 1999 and 2018: a population-based cohort study. *The Lancet Healthy Longevity*, 2(2): e96–e104, February 2021. ISSN 2666-7568. doi: 10.1016/s2666-7568(20)30059-3. URL http://dx.doi.org/10.1016/S2666-7568(20)30059-3.

Kevin F. Boreskie, Jacqueline L. Hay, Patrick E. Boreskie, Rakesh C. Arora, and Todd A. Duhamel. Frailty-aware care: giving value to frailty assessment across different healthcare settings. *BMC Geriatrics*, 22(1), January 2022. ISSN 1471-2318. doi: 10.1186/s12877-021-02722-9. URL http://dx.doi.org/10.1186/s12877-021-02722-9.

Christopher P. Bridge, Michael Rosenthal, Bradley Wright, Gopal Kotecha, Florian Fintelmann, Fabian Troschel, Nityanand Miskin, Khanant Desai, William Wrobel, Ana Babic, Natalia Khalaf, Lauren Brais, Marisa Welch, Caitlin Zellers, Neil Tenenholtz, Mark Michalski, Brian Wolpin, and Katherine Andriole. *Fully-Automated Analysis of Body Composition from CT in Cancer Patients Using Convolutional Neural Networks*, page 204–213. Springer International Publishing, 2018. ISBN 9783030012014. doi: 10.1007/978-3-030-01201-4_22. URL http://dx.doi.org/10.1007/978-3-030-01201-4_22.

Daniel S. Budnitz, Nadine Shehab, Scott R. Kegler, and Chesley L. Richards. Medication use leading to emergency department visits for adverse drug events in older adults. *Annals of Internal Medicine*, 147(11):755–765, December 2007. ISSN 1539-3704. doi: 10.7326/0003-4819-147-11-200712040-00006. URL http://dx.doi.org/10.7326/0003-4819-147-11-200712040-00006.

Angelo Capodici, Claudio Fanconi, Catherine Curtin, Alessandro Shapiro, Francesca Noci, Alberto Giannoni, and Tina Hernandez-Boussard. A scoping review of machine learning models to predict risk of falls in elders, without using sensor data. *Diagnostic and Prognostic Research*, 9 (1), May 2025. ISSN 2397-7523. doi: 10.1186/s41512-025-00190-y. URL http://dx.doi.org/10.1186/s41512-025-00190-y.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. doi: 10.1109/cvpr.2017.502. URL http://dx.doi.org/10.1109/CVPR.2017.502.

Centers for Disease Control and Prevention, National Center for Health Statistics. International classification of diseases, tenth revision, clinical modification (icd-10-cm), 2024. URL https://www.cdc.gov/nchs/icd/icd-10-cm/index.html.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM, August 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

Lisa Cooper, Julia Loewenthal, Laura N. Frain, Samir Tulebaev, Kristin Cardin, Tammy T. Hshieh, Clark Dumontier, Shoshana Streiter, Carly Joseph, Austin Hilt, Olga Theou, Kenneth Rockwood, Ariela R. Orkaby, and Houman Javedan. From research to bedside: Incorporation of a cga-based frailty index among multiple comanagement services. *Journal of the American Geriatrics Society*, 70(1):90–98, September 2021. ISSN 1532-5415. doi: 10.1111/jgs.17446. URL http://dx.doi.org/10.1111/jgs.17446.

Sacha Davis, Jin Zhang, Ilbin Lee, Mostafa Rezaei, Russell Greiner, Finlay A. McAlister, and Raj

Padwal. Effective hospital readmission prediction models using machine-learned features. *BMC Health Services Research*, 22(1), November 2022. ISSN 1472-6963. doi: 10.1186/s12913-022-08748-y. URL http://dx.doi.org/10.1186/s12913-022-08748-y.

dvgodoy. Illustrations for the transformer, and attention mechanism, June 2021. URL https://github.com/dvgodoy/dl-visuals/. Licensed under CC BY 4.0.

Thomas Gilbert, Jenny Neuburger, Joshua Kraindler, Eilis Keeble, Paul Smith, Cono Ariti, Sandeepa Arora, Andrew Street, Stuart Parker, Helen C Roberts, Martin Bardsley, and Simon Conroy. Development and validation of a hospital frailty risk score focusing on older people in acute care settings using electronic hospital records: an observational study. *The Lancet*, 391(10132):1775–1782, May 2018. ISSN 0140-6736. doi: 10.1016/s0140-6736(18)30668-8. URL http://dx.doi.org/10.1016/S0140-6736(18)30668-8.

Adrian D. Haimovich, Ryan C. Burke, Larry A. Nathanson, David Rubins, R. Andrew Taylor, Erin K. Kross, Kei Ouchi, Nathan I. Shapiro, and Mara A. Schonberg. Geriatric end-of-life screening tool prediction of 6-month mortality in older patients. *JAMA Network Open*, 7(5):e2414213, May 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2024.14213. URL http://dx.doi.org/10.1001/jamanetworkopen.2024.14213.

Daniel E. Hall, Shipra Arya, Kendra K. Schmid, Mark A. Carlson, Pierre Lavedan, Travis L. Bailey, Georgia Purviance, Tammy Bockman, Thomas G. Lynch, and Jason M. Johanning. Association of a frailty screening initiative with postoperative survival at 30, 180, and 365 days. *JAMA Surgery*, 152(3):233, March 2017. ISSN 2168-6254. doi: 10.1001/jamasurg.2016.4219. URL http://dx.doi.org/10.1001/jamasurg.2016.4219.

Erika Hörlin, Samia Munir Ehrlington, Joakim Henricson, Rani Toll John, and Daniel Wilhelms. Inter-rater reliability of the clinical frailty scale by staff members in a swedish emergency department setting. *Academic Emergency Medicine*, 29(12):1431–1437, October 2022. ISSN 1553-2712. doi: 10.1111/acem.14603. URL http://dx.doi.org/10.1111/acem.14603.

LLC Interpretable AI. Interpretable ai documentation, 2025. URL https://www.interpretable.ai.

Bellal Joseph, Viraj Pandit, Bardiya Zangbar, Narong Kulvatunyou, Ammar Hashmi, Donald J. Green, Terence O'Keeffe, Andrew Tang, Gary Vercruysse, Mindy J. Fain, Randall S. Friese, and Peter Rhee. Superiority of frailty over age in predicting outcomes among geriatric trauma patients: A prospective analysis. *JAMA Surgery*, 149(8):766, August 2014. ISSN 2168-6254. doi: 10.1001/jamasurg.2014.296. URL http://dx.doi.org/10.1001/jamasurg.2014.296.

Ramakrishna Kakara, Gwen Bergen, Elizabeth Burns, and Mark Stevens. Nonfatal and fatal falls among adults aged $\geq$ 65 years — united states, 2020–2021. *MMWR. Morbidity and Mortality Weekly Report*, 72(35):938–943, September 2023. ISSN 1545-861X. doi: 10.15585/mmwr.mm7235a1. URL http://dx.doi.org/10.15585/mmwr.mm7235a1.

Bharati Kochar, David Cheng, Hanna-Riikka Lehto, Nelia Jain, Elizabeth Araka, Christine S. Ritchie, Rachelle Bernacki, and Ariela R. Orkaby. Application of an electronic frailty index to identify high-risk older adults using electronic health record data. *Journal of the American Geriatrics Society*, 73(5):1491–1497, February 2025. ISSN 1532-5415. doi: 10.1111/jgs.19389. URL http://dx.doi.org/10.1111/jgs.19389.

Gotaro Kojima, Steve Iliffe, and Kate Walters. Frailty index as a predictor of mortality: a systematic review and meta-analysis. *Age and Ageing*, 47(2):193–200, October 2017. ISSN 1468-2834. doi: 10.1093/ageing/afx162. URL http://dx.doi.org/10.1093/ageing/afx162.

Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences, 2022. URL https://arxiv.org/abs/2201.11838.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Martin A. Makary, Dorry L. Segev, Peter J. Pronovost, Dora Syin, Karen Bandeen-Roche, Purvi Patel, Ryan Takenaga, Lara Devgan, Christine G. Holzmueller, Jing Tian, and Linda P. Fried. Frailty as a predictor of surgical outcomes in older patients. *Journal of the American College of Surgeons*, 210(6):901–908, June 2010. ISSN 1072-7515. doi: 10.1016/j.jamcollsurg.2010.01.028. URL http://dx.doi.org/10.1016/j.jamcollsurg.2010.01.028.

Theodore K. Malmstrom, Douglas K. Miller, and John E. Morley. A comparison of four frailty models. *Journal of the American Geriatrics Society*, 62(4):721–726, March 2014. ISSN 1532-5415. doi: 10.1111/jgs.12735. URL http://dx.doi.org/10.1111/jgs.12735.

Christopher R. Manz, Jinbo Chen, Manqing Liu, Corey Chivers, Susan Harkness Regli, Jennifer Braun, Michael Draugelis, C. William Hanson, Lawrence N. Shulman, Lynn M. Schuchter, Nina O'Connor, Justin E. Bekelman, Mitesh S. Patel, and Ravi B. Parikh. Validation of a machine learning algorithm to predict 180-day mortality for outpatients with cancer. *JAMA Oncology*, 6 (11):1723, November 2020. ISSN 2374-2437. doi: 10.1001/jamaoncol.2020.4331. URL http://dx.doi.org/10.1001/jamaoncol.2020.4331.

Nashwa Masnoon, Sepehr Shakib, Lisa Kalisch-Ellett, and Gillian E. Caughey. What is polypharmacy? a systematic review of definitions. *BMC Geriatrics*, 17(1), October 2017. ISSN 1471-2318. doi: 10.1186/s12877-017-0621-2. URL http://dx.doi.org/10.1186/s12877-017-0621-2.

Somya D. Mohanty, Deborah Lekan, Thomas P. McCoy, Marjorie Jenkins, and Prashanti Manda. Machine learning for predicting readmission risk among the frail: Explainable ai for healthcare. *Patterns*, 3(1):100395, January 2022. ISSN 2666-3899. doi: 10.1016/j.patter.2021.100395. URL http://dx.doi.org/10.1016/j.patter.2021.100395.

Robert T. Olender, Sandipan Roy, and Prasad S. Nishtala. Application of machine learning approaches in predicting clinical outcomes in older adults – a systematic review and meta-analysis. *BMC Geriatrics*, 23(1), September 2023. ISSN 1471-2318. doi: 10.1186/s12877-023-04246-w. URL http://dx.doi.org/10.1186/s12877-023-04246-w.

Caryn E. S. Oshiro, Timothy B. Frankland, A. Gabriela Rosales, Nancy A. Perrin, Christina L. Bell, Serena H. Y. Lo, and Connie M. Trinacty. Fall ascertainment and development of a risk prediction model using electronic medical records. *Journal of the American Geriatrics Society*, 67(7): 1417–1422, March 2019. ISSN 1532-5415. doi: 10.1111/jgs.15872. URL http://dx.doi.org/10.1111/jgs.15872.

Brian W. Patterson, Collin J. Engstrom, Varun Sah, Maureen A. Smith, Eneida A. Mendonça, Michael S. Pulia, Michael D. Repplinger, Azita G. Hamedani, David Page, and Manish N. Shah. Training and interpreting machine learning algorithms to evaluate fall risk after emergency department visits. *Medical Care*, 57(7):560–566, July 2019. ISSN 0025-7079. doi: 10.1097/mlr.0000000000001140. URL http://dx.doi.org/10.1097/MLR.0000000000001140.

Periklis Petridis, Georgios Margaritis, Vasiliki Stoumpou, and Dimitris Bertsimas. Holistic artificial intelligence in medicine; improved performance and explainability, 2025. URL https://arxiv.org/abs/2507.00205.

Kenneth Rockwood and Arnold Mitnitski. Frailty defined by deficit accumulation and geriatric medicine defined by frailty. *Clinics in Geriatric Medicine*, 27(1):17–26, February 2011. ISSN 0749-0690. doi: 10.1016/j.cger.2010.08.008. URL http://dx.doi.org/10.1016/j.cger.2010.08.008.

Josje D. Schoufour, Michael A. Echteld, Luc P. Bastiaanse, and Heleen M. Evenhuis. The use of a frailty index to predict adverse health outcomes (falls, fractures, hospitalization, medication use, comorbid conditions) in people with intellectual disabilities. *Research in Developmental Disabilities*, 38:39–47, March 2015. ISSN 0891-4222. doi: 10.1016/j.ridd.2014.12.001. URL http://dx.doi.org/10.1016/j.ridd.2014.12.001.

H. Schuurmans, N. Steverink, S. Lindenberg, N. Frieswijk, and J. P. J. Slaets. Old or frail: What tells us more? *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 59(9):M962–M965, September 2004. ISSN 1758-535X. doi: 10.1093/gerona/59.9.m962. URL http://dx.doi.org/10.1093/gerona/59.9.m962.

Samuel D Searle, Arnold Mitnitski, Evelyne A Gahbauer, Thomas M Gill, and Kenneth Rockwood.

A standard procedure for creating a frailty index. *BMC Geriatrics*, 8(1), September 2008. ISSN 1471-2318. doi: 10.1186/1471-2318-8-24. URL http://dx.doi.org/10.1186/1471-2318-8-24.

Vishal Sehgal, Sukhminder JitSingh Bajwa, Rinku Sehgal, Anurag Bajaj, Upinder Khaira, and Victoria Kresse. Polypharmacy and potentially inappropriate medication use as the precipitating factor in readmissions to the hospital. *Journal of Family Medicine and Primary Care*, 2 (2):194, 2013. ISSN 2249-4863. doi: 10.4103/2249-4863.117423. URL http://dx.doi.org/10.4103/2249-4863.117423.

Huaxin Si, Yaru Jin, Xiaoxia Qiao, Xiaoyu Tian, Xinyi Liu, and Cuili Wang. Predictive performance of 7 frailty instruments for short-term disability, falls and hospitalization among chinese community-dwelling older adults: A prospective cohort study. *International Journal of Nursing Studies*, 117:103875, May 2021. ISSN 0020-7489. doi: 10.1016/j.ijnurstu.2021.103875. URL http://dx.doi.org/10.1016/j.ijnurstu.2021.103875.

Luis R. Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M. Wiberg, Michael L. Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digital Medicine*, 5(1), September 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00689-4. URL http://dx.doi.org/10.1038/s41746-022-00689-4.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 03 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x. URL https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Appendix A. Supplementary Material

A.1. Definition of Mortality Outcomes:

For the 3− and 6−month mortality outcomes, we implement a three-step method to generate the y-variable. First, we identified patients with documented death dates in the demographics data that occurred within the respective post-discharge windows. Second, we captured deaths through subsequent inpatient visits that ended with "Expired" discharge dispositions within the specified timeframes. Third, for patients with incomplete mortality documentation but a recorded age at death in the demographics records, we implemented an age-gap inference rule. This rule subtracts the age at date of death (always integer) from the exact age at time of visit (date of visit - date of birth), using discrepancies of ≥ 0.75 years to infer death and ≤ -0.25 years to infer survival for 90-day mortality, and ≥ 0.5 years and ≤ -0.5 years respectively for 180-day mortality. For example, if the patient's recorded integer age of death is 90 years and the last inpatient date of visit is recorded at 90.55 years old, then we can conclude that the patient dies within 6-months of the visit. Visits for which 3− or 6− month mortality status could not be definitively determined through any of these methods were excluded from the analysis to ensure data integrity.

A.2. Complete Results

Table 2: Complete results for the HAIM Model.

Outcome	**Number of visits**	**Event Rate (%)**	**True Train (80% split of full dataset) AUC**	**True Validation (10% split of full dataset) AUC**	**True Test (10% split of full dataset) AUC**
Mortality 3-Months	19,874	5.21%	0.9653	0.8646	0.8554
Mortality 6-Months	19,095	8.51%	0.9849	0.8340	0.8240
Readmit 3-Months	19,549	32.87%	0.7730	0.6819	0.7385
Readmit 6-Months	18,912	44.62%	0.7758	0.7019	0.7527
Fall 6-Months	18,004	22.14%	0.8217	0.76506	0.7493
Falls 1-year Recurrent	12,211	60.04%	0.8629	0.7948	0.7959

Table 3: Complete results for the Tabular-Only Model.

Outcome	**Number of visits**	**True Train (80% split of full dataset) AUC**	**True Validation (10% split of full dataset) AUC**	**True Test (10% split of full dataset) AUC**
Mortality 3-Months	19,874	0.8975	0.8422	0.8198
Mortality 6-Months	19,095	0.9219	0.8222	0.7849
Readmit 3-Months	19,549	0.7730	0.6819	0.7385
Readmit 6-Months	18,912	0.7720	0.7084	0.7458
Fall 6-Months	18,004	0.7863	0.7644	0.7481
Falls 1-year Recurrent	12,211	0.8425	0.8005	0.7838

Table 4: Complete results for the Abdominal/Pelvic CT Model. The number of images is the number of images used in total across the training, validation, and testing. The images are split according to patient identifier, as specified in the data section.

Outcome	**Number of images**	**Train (80% of true train set) AUC**	**Validation (20% subset of true train set) AUC**	**True Validation (10% split of full dataset) AUC**	**True Test (10% split of full dataset) AUC**
Mortality 3-Months	7,477	0.7937	0.8321	0.8254	0.7953
Mortality 6-Months	7,292	0.7679	0.7537	0.7936	0.7897
Readmit 3-Months	3,990	0.5574	0.5706	0.5583	0.4932
Readmit 6-Months	3,862	0.5577	0.5267	0.5136	0.4558
Fall 6-Months	6,680	0.5753	0.5362	0.5493	0.5495
Falls 1-year Recurrent	1,040	0.5320	0.5253	0.4566	0.5111

Table 5: Comparing Merlin model to a Body Composition (muscle, subcutaneous and visceral fat area) regression model, reporting AUC.

Outcome	**Merlin**		**Body Composition**	
	Val	Test	Val	Test
3-Month Mortality	0.8254	0.7953	0.7758	0.7642
6-Month Mortality	0.7936	0.7897	0.7352	0.7286
3-Month Readmission	0.5583	0.4932	0.4948	0.5254
6-Month Readmission	0.5136	0.4558	0.4976	0.5020
6-Month Fall	0.5493	0.5495	0.6118	0.4910
1-year Recurrent Fall	0.4566	0.5111	0.5729	0.5629

Table 6: Complete results for the Head CT Model.

Outcome	**Number of images**	**Train (80% of true train set) AUC**	**Validation (20% subset of true train set) AUC**	**True Validation (10% split of full dataset) AUC**	**True Test (10% split of full dataset) AUC**
Mortality 3-Months	8,332	0.8324	0.6432	0.6484	0.7176
Mortality 6-Months	7,744	0.8035	0.6221	0.6663	0.7101
Readmit 3-Months	5,284	0.6915	0.5318	0.5334	0.5522
Readmit 6-Months	5,143	0.6350	0.5209	0.5235	0.5444
Fall 6-Months	7,525	0.6354	0.5458	0.5157	0.6259
Falls 1-year Recurrent	3,289	0.6589	0.5784	0.5864	0.5777

Table 7: Complete results for text modalities. The best validation AUC is in bold with its corresponding test AUC.

3-Month Mortality

Note Type	Train AUC	Val AUC	Test AUC
Cardiology	0.7901	0.6359	0.6020
Discharge	0.8908	0.6982	0.6790
Endoscopy	0.7588	0.5404	0.3757
H&P	0.7588	0.7067	0.6698
Progress	0.8874	**0.7605**	**0.7497**
Pulmonary	0.9980	0.7037	0.6960
Radiology	0.9776	0.6835	0.6619
Visit	0.8221	0.7425	0.7009

6-Month Mortality

Note Type	Train AUC	Val AUC	Test AUC
Cardiology	0.9633	0.7184	0.7012
Discharge	0.7470	0.6964	0.7039
Endoscopy	0.7785	0.7328	0.68967
H&P	0.9633	0.7184	0.7012
Progress	0.9180	**0.7438**	**0.7592**
Pulmonary	0.7873	0.7308	0.6961
Radiology	0.8837	0.7116	0.6991
Visit	0.7785	0.7328	0.6896

3-Month Readmission

Note Type	Train AUC	Val AUC	Test AUC
Cardiology	0.7878	0.6158	0.6478
Discharge	0.6843	0.5928	0.6336
Endoscopy	0.7686	0.5763	0.5746
H&P	0.7382	0.6053	0.6786
Progress	0.7631	**0.6180**	**0.6871**
Pulmonary	0.8172	0.5999	0.5816
Radiology	0.6986	0.6102	0.6610
Visit	0.8329	0.6081	0.6646

6-Month Readmission

Note Type	Train AUC	Val AUC	Test AUC
Cardiology	0.7984	0.6345	0.6711
Discharge	0.7050	0.6312	0.6468
Endoscopy	0.8175	0.5581	0.5431
H&P	0.7536	0.6508	0.6958
Progress	0.7448	0.6372	0.6912
Pulmonary	0.6954	0.5651	0.5907
Radiology	0.9999	**0.6589**	**0.6678**
Visit	0.7408	0.6268	0.6662

6-Month Falls

Note Type	Train AUC	Val AUC	Test AUC
Cardiology	0.7554	0.6051	0.6054
Discharge	0.6774	0.6131	0.6503
Endoscopy	0.8144	0.6240	0.4820
H&P	0.6880	0.6105	0.6380
Progress	0.9434	0.6188	0.6535
Pulmonary	0.9317	0.6257	0.5169
Radiology	0.7131	**0.6635**	**0.6556**
Visit	0.6668	0.6275	0.6342

1-Year Recurrent Fall

Note Type	Train AUC	Val AUC	Test AUC
Cardiology	0.7233	0.6759	0.6495
Discharge	0.7424	0.7131	0.7119
Endoscopy	0.8181	0.6462	0.5511
H&P	0.6972	0.6794	0.6750
Progress	0.8114	**0.7226**	**0.7273**
Pulmonary	0.8325	0.5741	0.5545
Radiology	0.7588	0.7152	0.7186
Visit	0.7950	0.7208	0.7350

Table 8: Sensitivity, specificity, F1 score, and PR-AUC on validation and test sets for each outcome. For each outcome, we determine the optimal threshold on the validation set by identifying the point on the ROC curve where the true positive rate (TPR) most closely matches the true negative rate (TNR), $\arg\min_t |\mathrm{TPR}(t) - \mathrm{TNR}(t)|$. We report results for the best single-modality model in addition to HAIM and frailty models.

Outcome	**Validation**				**Test**			
	Sens.	Spec.	F1	PR-AUC	Sens.	Spec.	F1	PR-AUC
3-Month Mortality								
HAIM	0.7767	0.7769	0.4126	0.3899	**0.7982**	**0.7746**	**0.3713**	**0.3659**
Tabular	0.7573	0.7549	0.3889	0.3283	0.7807	0.7381	0.3404	0.3493
Frailty	0.6395	0.6565	0.2573	0.1706	0.6389	0.6382	0.2893	0.2420
6-Month Mortality								
HAIM	0.7647	0.7682	0.3828	0.3522	0.6703	**0.7721**	**0.3771**	**0.3464**
Abdominal/Pelvic	0.7115	0.7178	0.300	0.1931	**0.7326**	0.7033	0.3469	0.2681
Frailty	0.6486	0.6581	0.2902	0.2026	0.6609	0.6469	0.2896	0.2641
3-Month Readmission								
Tabular	0.6398	0.6402	0.5551	0.5074	**0.7157**	**0.6555**	**0.5999**	**0.6114**
Frailty	0.5024	0.5159	0.3721	0.2359	0.5146	0.4979	0.3688	0.2247
6-Month Readmission								
HAIM	0.6533	0.6547	0.6603	0.6191	**0.7090**	**0.6563**	**0.6822**	**0.7306**
Tabular	0.6533	0.6537	0.6643	0.6390	0.7013	0.6418	0.6785	0.7252
Frailty	0.5940	0.4393	0.4658	0.3225	0.6065	0.3829	0.4821	0.3308
6-Month Fall								
HAIM	0.6875	0.6869	0.4944	0.4688	**0.6913**	0.6439	0.5162	**0.5642**
Tabular	0.6957	0.6974	0.4926	0.4785	0.6871	**0.6748**	**0.5236**	0.5499
Frailty	0.5000	0.6208	0.1538	0.1003	0.5385	0.6084	0.0000	0.0267
1-year Recurrent Fall								
HAIM	0.7038	0.7031	0.8117	0.8440	0.7268	**0.6905**	0.7959	**0.8670**
Tabular	0.7204	0.7205	0.8159	0.8463	0.7244	0.6883	**0.8134**	0.8569
Frailty	0.6364	0.5263	0.1111	0.0775	**0.8571**	0.5463	0.1132	0.0766

Appendix B. Performance on Visits with Recorded Frailty Assessments

We present the results for the frailty-matched HAIM and single-modality models. These models are trained, validated, and tested on all visits with recorded FI evaluations, ensuring that both the HAIM and FI models operate on an identical cohort for a fair comparison.

Table 9: AUC comparison of HAIM and the best single-modality models on visits with recorded FI evaluations.

Outcome	**Best Single-Modality Test AUC**	**HAIM Test AUC**
Mortality 3-Months	0.797	0.853
Mortality 6-Months	0.812	0.834
Readmit 3-Months	0.744	0.780
Readmit 6-Months	0.765	0.765
Fall 6-Months	0.722	0.729
Recurrent Fall 1-Year	0.761	0.761

We used the output probabilities from our HAIM models to construct 7-dimensional feature vectors for each visit and applied k-means clustering to identify distinct vulnerability phenotypes. The resulting clusters were translated into interpretable decision rules using an OPT. The one-hot cluster encodings were used as predictors in logistic regression models for each outcome, and Table 10 reports the corresponding AUCs alongside FI for comparison.

Table 10: Logistic regression on cluster one-hot encodings, trained on the validation (val) set and evaluated on the test set; FI Test AUC is shown for comparison.

Outcome	**Val AUC**	**Test AUC**	**FI Test AUC**
Mortality 3-Months	0.723	0.836	0.734
Mortality 6-Months	0.730	0.770	0.723
Readmit 3-Months	0.695	0.732	0.505
Readmit 6-Months	0.687	0.720	0.505
Fall 6-Months	0.696	0.641	0.627
Recurrent Fall 1-Year	0.760	0.696	0.736

Appendix C. Visualizations

C.1. SHAP Plots

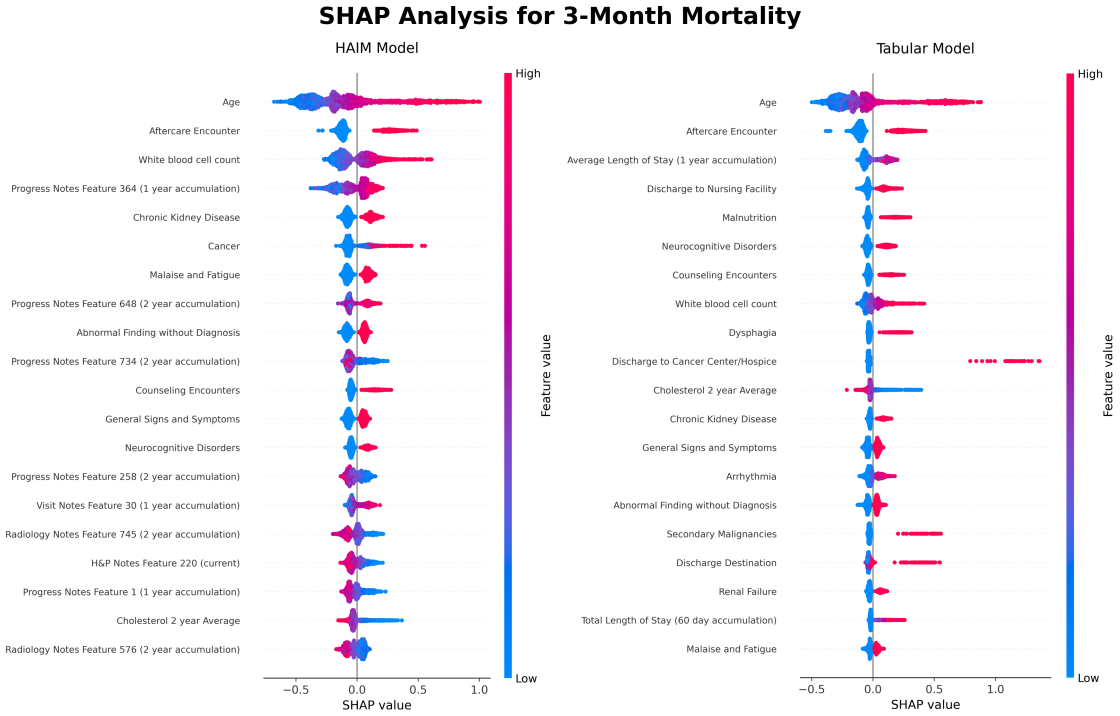


Figure 5: SHAP plots for 3-month mortality, comparing the HAIM combined model to the single modality tabular model.

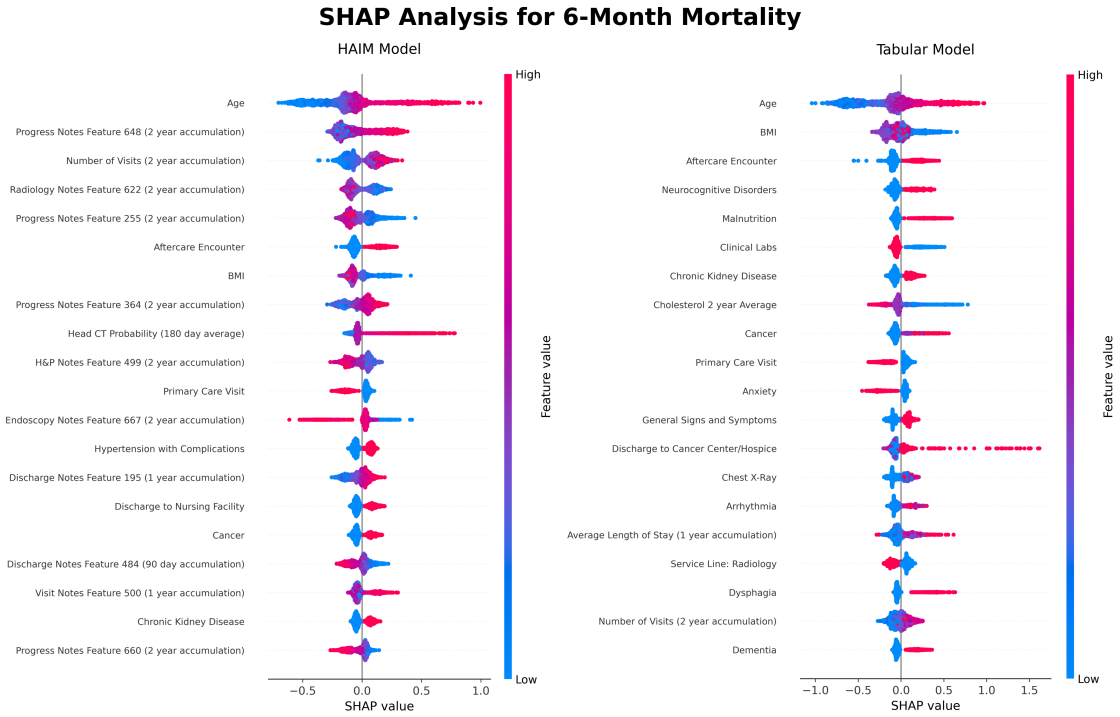


Figure 6: SHAP plots for 6-month mortality, comparing the HAIM combined model to the single modality tabular model.

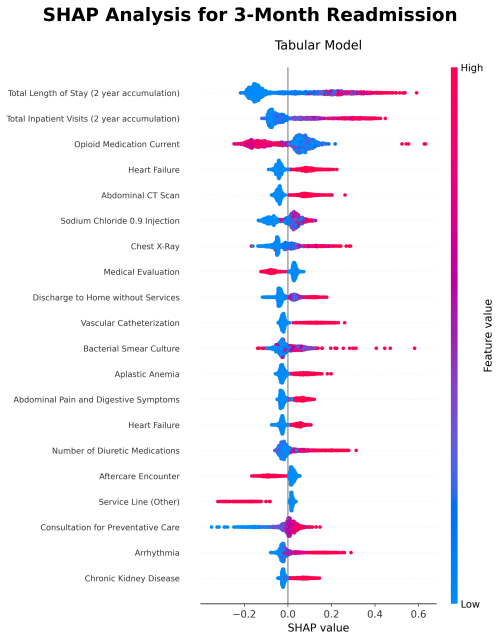


Figure 7: SHAP plot for 3-month readmission task. For this outcome, the tabular-only model had the best performance.

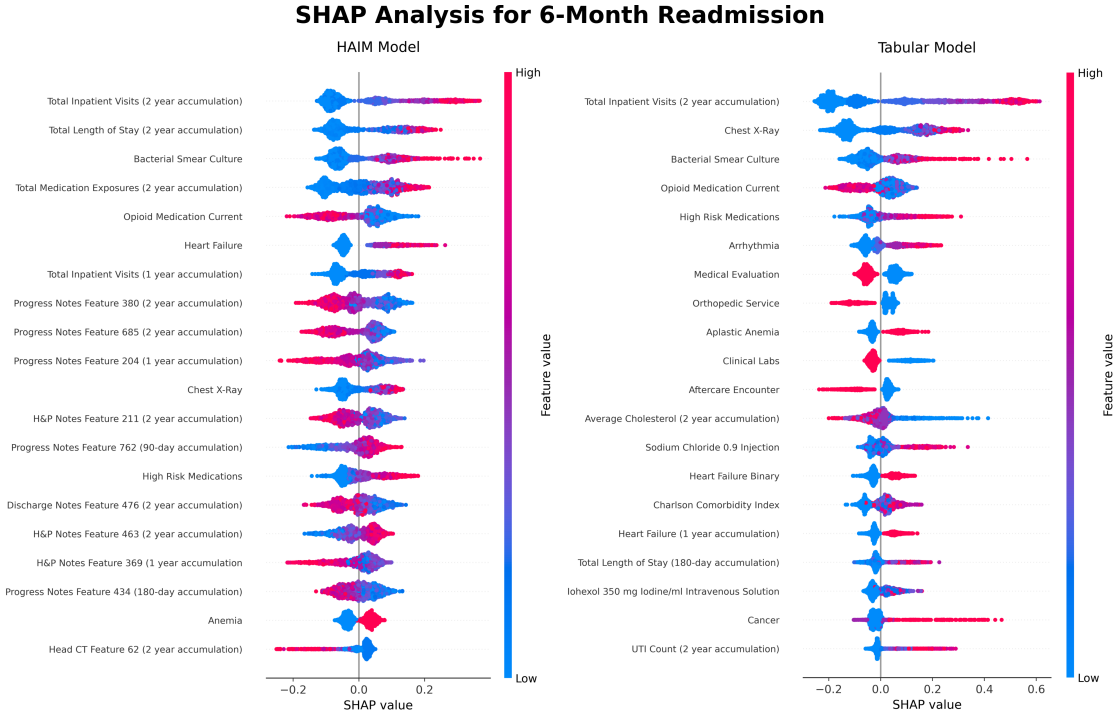


Figure 8: SHAP plots for the 6-month readmission task.

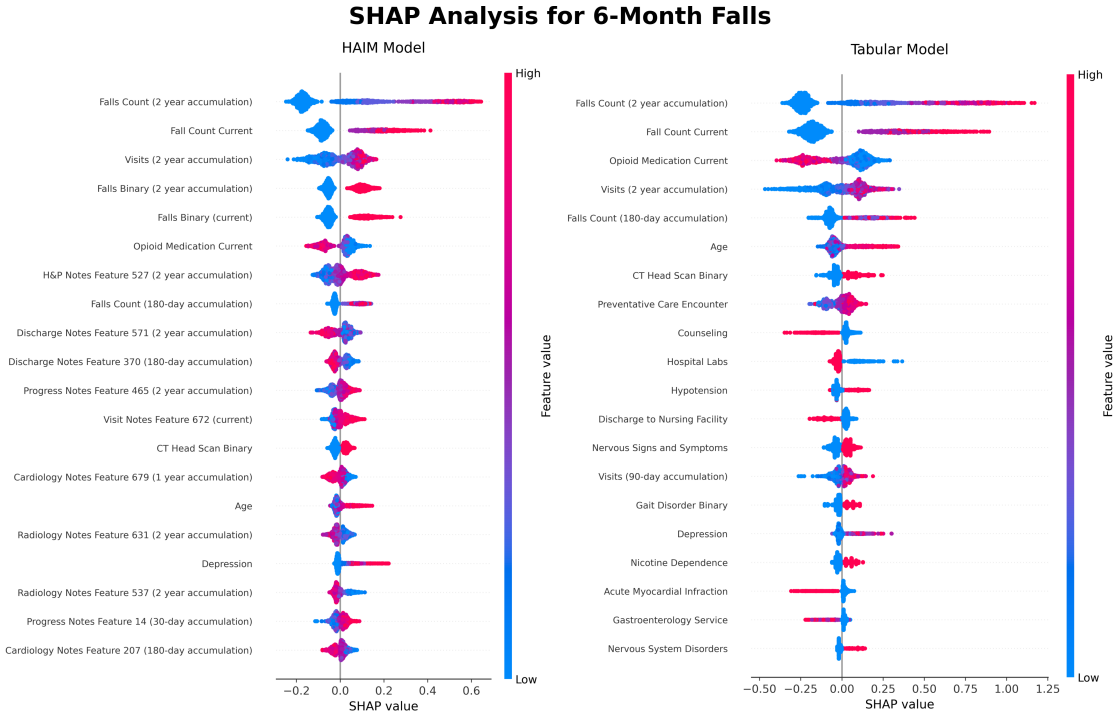


Figure 9: SHAP plots for 6-month fall prediction.

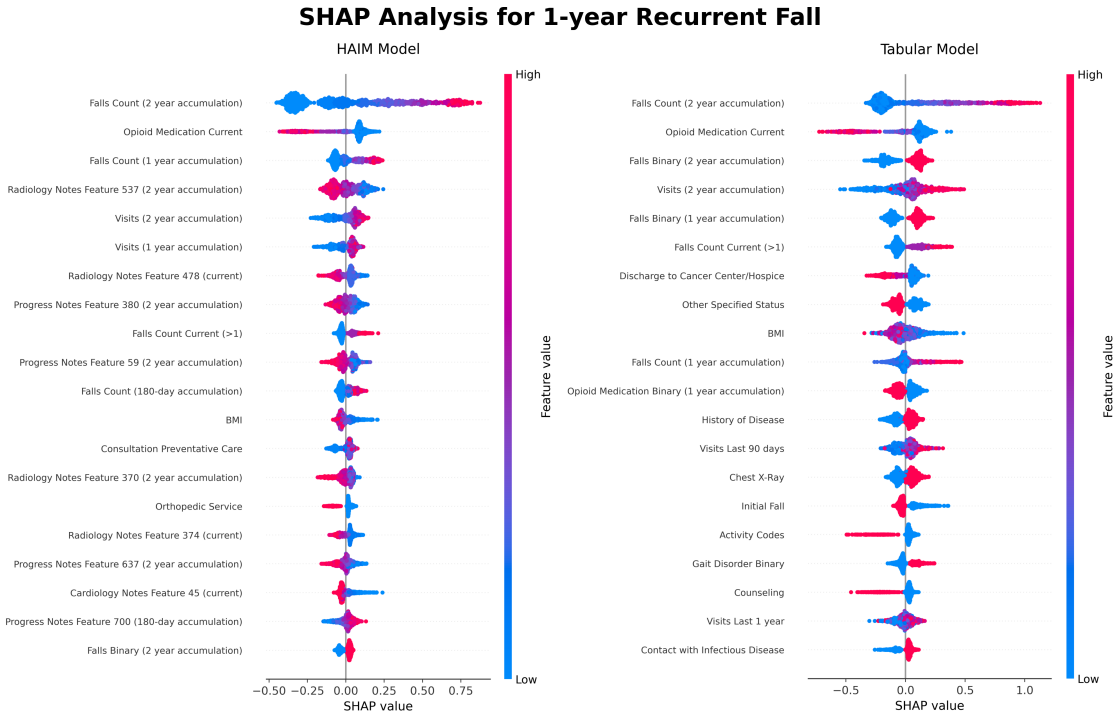


Figure 10: SHAP plots for 1-year recurrent fall prediction.

C.2. Optimal Policy Tree

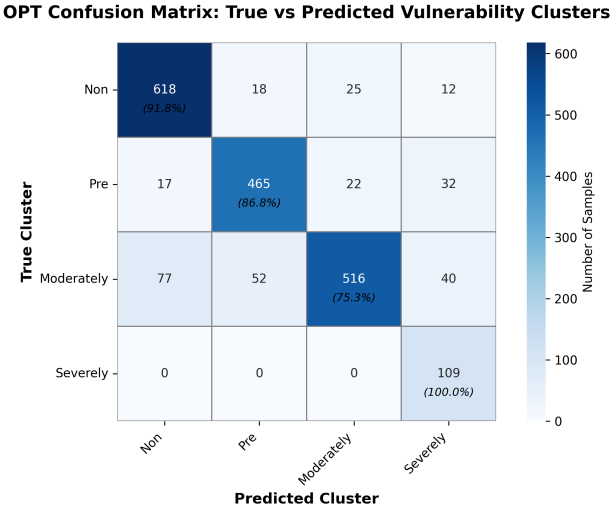


Figure 11: Confusion matrix for the OPT's predictions on the validation set.

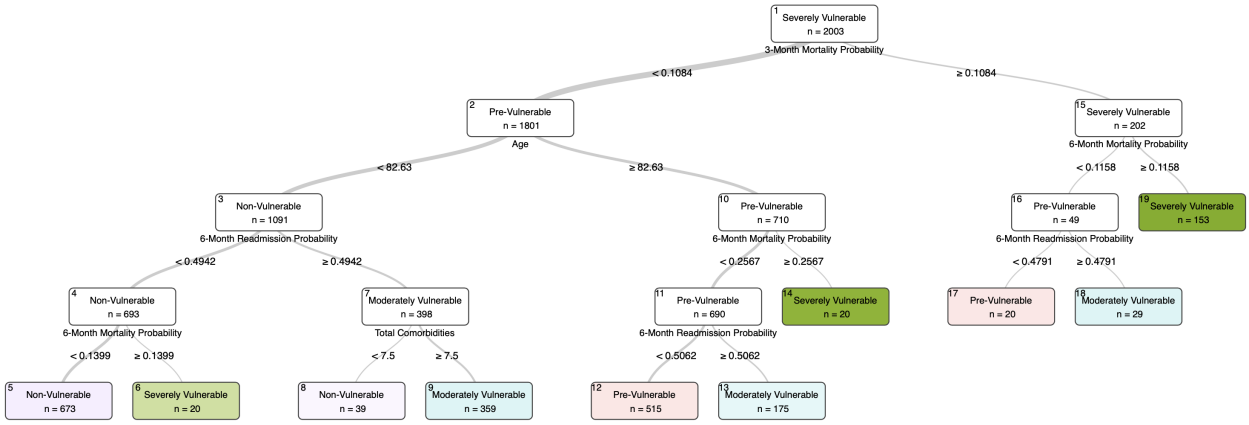


Figure 12: The trained Optimal Policy tree used to make predictions for vulnerability classification.