Human-Guided Agentic AI for Multimodal Clinical Prediction: Lessons from the AgentDS Healthcare Benchmark

Lalitha Pranathi Pulavarthy, Raajitha Muthyala, Aravind V Kuruvikkattil,
Zhenan Yin, Rashmita Kudamala, and Saptarshi Purkayastha
Dept. of Biomedical Engineering and Informatics
Indiana University Indianapolis
Indianapolis, USA
{lapula, rmuthya, akuruvik, yin10, rakuda, saptpurk}@iu.edu

***Abstract*—Agentic AI systems are increasingly capable of autonomous data science workflows, yet clinical prediction tasks demand domain expertise that purely automated approaches struggle to provide. We investigate how human guidance of agentic AI can improve multimodal clinical prediction, presenting our approach to all three AgentDS Healthcare benchmark challenges: 30-day hospital readmission prediction (Macro-F1 = 0.8986), emergency department cost forecasting (MAE = $465.13), and discharge readiness assessment (Macro-F1 = 0.7939). Across these tasks, human analysts directed the agentic workflow at key decision points, multimodal feature engineering from clinical notes, scanned PDF billing receipts, and time-series vital signs; task-appropriate model selection; and clinically informed validation strategies. Our approach ranked 5th overall in the healthcare domain, with a 3rd-place finish on the discharge readiness task. Ablation studies reveal that human-guided decisions compounded to a cumulative gain of +0.065 F1 over automated baselines, with multimodal feature extraction contributing the largest single improvement (+0.041 F1). We distill three generalizable lessons: (1) domain-informed feature engineering at each pipeline stage yields compounding gains that outperform extensive automated search; (2) multimodal data integration requires task-specific human judgment that no single extraction strategy generalizes across clinical text, PDFs, and time-series; and (3) deliberate ensemble diversity with clinically motivated model configurations outperforms random hyperparameter search. These findings offer practical guidance for teams deploying agentic AI in healthcare settings where interpretability, reproducibility, and clinical validity are essential.**

***Index Terms*—Agentic AI, Human-AI collaboration, Multimodal feature engineering, Clinical prediction, Healthcare benchmark**

I. Introduction

Hospital readmissions cost the U.S. healthcare system over $26 billion annually [1], while emergency department overcrowding and suboptimal discharge timing continue to degrade patient outcomes and hospital efficiency [2]. Machine learning has shown considerable promise for predicting these clinical events. Multimodal approaches that combine structured electronic health record (EHR) data with unstructured clinical notes and time-series vital signs have demonstrated superior predictive performance over unimodal baselines, achieving AUROCs above 0.90 for 30-day readmission prediction in elderly populations [2] and enabling earlier identification of discharge-ready patients by 5-13 hours compared with physician judgment [3]. Graph neural network architectures that jointly model EHR tabular data, clinical imaging, and patient similarity have further improved readmission prediction by capturing spatiotemporal dependencies across admissions [4]. Yet despite these advances, the practical deployment of such models remains constrained by the need to balance automated optimization with clinical interpretability, reproducibility, and the integration of domain expertise, a challenge that existing benchmarks have only partially addressed.

The recent emergence of Agentic AI systems, autonomous agents capable of planning, reasoning, and executing multi-step workflows, offers a new paradigm for healthcare data science [5], [6]. Unlike conventional machine learning pipelines that require manual orchestration, agentic systems powered by large language models (LLMs) can autonomously perform data loading, feature engineering, model training, and evaluation [7]. In the clinical domain, multi-agent LLM frameworks such as MDAgents have demonstrated that adaptive collaboration structures can improve medical decision-making accuracy by up to 11.8% on diagnostic benchmarks [8], while EHRAgent has shown that LLM agents equipped with code generation capabilities can outperform traditional baselines by up to 29.6% on complex EHR reasoning tasks [9]. More broadly, surveys of LLM-based data science agents identify healthcare as a high-stakes application domain where fairness, explainability, and reproducibility requirements impose additional constraints on agent design [7], [10].

However, fully autonomous agentic workflows face well-documented limitations in clinical prediction contexts. Clinical prediction tasks demand domain expertise that purely automated approaches struggle to provide: selecting appropriate text extraction strategies for heterogeneous clinical documents, engineering physiologically meaningful features from vital sign time series, and making informed trade-offs between

model complexity and sample size [11]. Prior work on discharge readiness prediction has shown that handcrafted features informed by clinical knowledge, such as statistical representations of early warning scores and domain-specific temporal features, are critical for accurate predictions [12]. Studies on ED cost forecasting emphasize that billing code structures and cross-source feature engineering require human judgment to interpret correctly [13]. Furthermore, recent analyses of agentic AI in healthcare caution that autonomous systems require robust governance frameworks, human oversight, and interdisciplinary collaboration to ensure patient safety [5], [14]. These observations suggest that the most effective paradigm may not be fully autonomous agentic AI, but rather a human-AI collaboration loop in which domain experts guide the agent's decisions at critical junctures [15].

The AgentDS Healthcare Data Challenge provides a rigorous framework for evaluating such human-AI collaboration through three standardized clinical prediction tasks: (1) 30-day readmission prediction (binary classification, N=5,000 train), (2) 3-year ED cost forecasting (regression, N=2,000 train), and (3) day-11 discharge readiness assessment (binary classification, N=1,000 train). The benchmark evaluates human-AI collaboration using multimodal data-structured admission records, unstructured clinical notes, scanned PDF receipts, and time-series vital signs with standardized metrics (Macro-F1 for classification, MAE for regression), ensuring reproducible comparison. This design addresses a critical gap in existing benchmarks, which typically evaluate either autonomous agent capabilities in isolation [16] or purely manual modeling approaches, without systematically measuring the value added by human guidance at each pipeline stage [7].

Our approach combines domain-guided feature engineering with transparent model selection, achieving 5th place overall in the healthcare domain on the public leaderboard, including a 3rd-place finish on the discharge readiness task. Rather than complex neural architectures or extensive hyperparameter tuning, which prior work has shown to be less effective than feature engineering on modest clinical sample sizes [2], [16]. We focused on: (1) clinical knowledge-informed multimodal feature extraction, (2) task-appropriate model selection balancing performance and interpretability, and (3) rigorous validation strategies. Critically, we document all human decision points to maintain reproducibility, following the principle that in healthcare ML, transparency and auditability are as important as raw predictive performance [14], [17].

Contributions. (1) We demonstrate practical multimodal feature engineering strategies across text, PDFs, and time-series data with ablation studies quantifying impact, showing that human-guided multimodal extraction contributes +0.041 F1, the single largest improvement category. (2) We identify which modeling decisions generalize across tasks versus require task-specific adaptation, providing evidence that no single extraction strategy works across all clinical data types. (3) We show that systematic domain-informed decisions outperform extensive automated search, achieving competitive results (within 0.006 F1 of top-performing systems on readmission prediction) while maintaining interpretability, a finding consistent with recent evidence that tree-based models with domain features remain competitive with deep learning on moderate-scale clinical datasets [16], [18].

Our results demonstrate that principled human-AI collaboration yields strong, reproducible performance across diverse healthcare prediction tasks, with human decisions contributing a cumulative +0.065 F1 improvement over agentic baselines.

II. Methodology

Our approach follows an iterative human-AI collaboration workflow in which an Agentic AI system handled routine data science operations, data loading, preprocessing, initial model training, and hyperparameter search via Bayesian optimization, while human analysts intervened at critical decision points to redirect the pipeline. Each challenge progressed through multiple iterations: the agentic system produced a baseline, human analysts diagnosed limitations through cross-validation error analysis and clinical reasoning, and then guided the next iteration. We describe each challenge below, highlighting the specific human decision points that drove performance gains and distinguishing them from automated steps. All models across the three challenges were implemented using scikit-learn and XGBoost (Python 3.11). PDF parsing used PyPDF2 with fallback error handling, vital sign trends were computed via scipy.stats.linregress, and NLP features used regex-based partial-stem keyword matching. Code, hyperparameters, and random seeds are publicly available for full reproducibility at https://github.com/iupui-soic/agentds-challenge-2025.

A. Challenge 1: 30-Day Readmission Prediction

Task Overview. The readmission prediction task required binary classification of whether an inpatient admission would be followed by a readmission within 30 days. The training set comprised 5,000 admissions (2,479 non-readmit, 2,521 readmit; approximately balanced classes). We integrated data from three sources via left joins on patient and admission identifiers: (1) structured admission records (length of stay, acuity level, Charlson comorbidity band, 6-month ED visits, discharge weekday, primary diagnosis); (2) patient demographics (age, sex, insurance type); and (3) 5-year ED cost history (prior ED visits, prior ED costs). Each admission was also linked to an unstructured discharge note summary. The evaluation metric was Macro-F1.

Human-Guided Feature Engineering. Our feature engineering strategy progressed through two iterations informed by domain analysis. The agentic system's initial approach trained a single Optuna-tuned XGBoost (n=957, lr=0.28, depth=8) with one-hot encoding and 200 unigram TF-IDF features, achieving CV Macro-F1=0.846. Human analysis of per-class errors and clinical literature identified two limitations: (1) unigram TF-IDF missed multi-word clinical phrases critical for readmission risk, and (2) a single model could not simultaneously capture the diverse decision boundaries present

in structured versus text features. This guided the second iteration with expanded features and an ensemble architecture.

Structured features (21 engineered + one-hot dummies): We created domain-informed features organized by clinical hypothesis: (1) *temporal discharge patterns*, weekend discharge flag and specific weekday indicators (Monday, Friday), motivated by studies showing discharge timing affects readmission risk; (2) *non-linear age and comorbidity effects*: age^2, Charlson^2, and threshold flags for elderly ($\geq$65) and very elderly ($\geq$80) patients; (3) *ED utilization patterns*: cost-per-visit ratio with Laplace smoothing, recent-to-historical visit ratio (6-month / 5-year), frequent ED user flags ($\geq$2, $\geq$3 visits), and new-patient indicator; (4) *length-of-stay features*: log-transformed LOS and long-stay flag ($\geq$5 days); (5) *interaction terms*: age$\times$Charlson, age$\times$ED visits, Charlson$\times$ED visits, and LOS$\times$ED visits; and (6) *composite risk scores*: two weighted combinations mimicking clinical stratification (e.g., age/100 + 0.5$\times$Charlson + 0.3$\times$ED visits).

Text features (850 TF-IDF): We applied TF-IDF vectorization with trigrams (ngram_range=(1,3), max_features=850, min_df=3, max_df=0.95, sublinear_tf=True). Expanding from the baseline's 200 unigram features was a key human decision: domain analysis revealed that discharge notes contain multi-word clinical phrases ("ambulating without assistance", "wound care instructions") requiring higher-order n-grams and greater dimensionality to capture.

Model Architecture: Stacking Ensemble. We employed a stacking ensemble with five base learners and 8-fold stratified internal cross-validation (stack_method='predict_proba', passthrough=False):

Base Models:

- **XGBoost-Optimized** ($n = 771$, $lr = 0.030$, $depth = 7$, $subsample = 0.68$, $colsample\_bytree = 0.81$, $\gamma = 3.42$): Hyperparameters selected via Optuna Bayesian optimization; $random\_state = 42$.
- **XGBoost-Deep** ($n = 1500$, $lr = 0.015$, $depth = 6$, $subsample = 0.75$, $colsample\_bytree = 0.85$, $\gamma = 2.0$, $min\_child\_weight = 2$): Conservative learning rate with more iterations; $random\_state = 999$.
- **XGBoost-Shallow** (n=1200, lr=0.02, depth=5, subsample=0.8, colsample_bytree=0.9, γ=1.0): Shallower trees for broader decision boundaries; random_state=555.
- **Logistic Regression-L2** ($C = 0.5$, $solver =$ saga, $class\_weight =$ balanced): Preceded by *StandardScaler*; $random\_state = 42$.
- **Logistic Regression-L1** ($C = 0.3$, $solver =$ saga, $class\_weight =$ balanced): Sparse feature selection; $random\_state = 777$.

All XGBoost models used `tree_method='hist'` with $scale\_pos\_weight$ computed from training class frequencies (≈ 0.98 given the nearly balanced classes). Different random seeds across variants ensured initialization diversity, while deliberately varied depth and learning rate profiles created complementary decision boundaries.

Meta-learner: Logistic regression ($C = 1.5$, class_weight = balanced, solver = lbfgs) trained on 10-dimensional out-of-fold probability vectors (two class probabilities per base model).

Rationale for Human Decisions. The human-guided second iteration lifted CV F1 from 0.846 to 0.896 (+0.050). Key choices included: (1) *multi-table integration*: the agentic system initially used only admission records; human analysts directed it to join patient demographics and ED cost history, providing cross-domain features (e.g., cost-per-visit ratio) unavailable in any single table; (2) *TF-IDF expansion*: upgrading from 200 unigrams to 850 trigrams captured multi-word medical phrases ("ambulating without assistance," "wound care instructions"); (3) *deliberate ensemble diversity*: the agentic system's single-model approach was replaced with three XGBoost variants with different complexity profiles (depth 5/6/7, learning rates 0.02/0.015/0.03) plus two regularized linear models (L1/L2), ensuring complementary feature usage; (4) *8-fold stacking CV*: reduced meta-learner training variance while maintaining sufficient per-fold training size (N$\approx$625).

Validation Strategy. We used a nested cross-validation scheme: an outer 5-fold stratified CV for score estimation, with each outer fold training the full stacking pipeline (including an 8-fold inner CV for generating stacked features). This produced 200 base model fits total (5 outer $\times$ 8 inner $\times$ 5 base models). Outer CV Macro-F1 was 0.8956 (accuracy 0.90), with per-class F1 of 0.89 (non-readmit) and 0.90 (readmit). Final test predictions achieved Macro-F1 = 0.8986 on the public leaderboard, ranking 5th overall.

B. Challenge 2: ED Cost Forecasting

Task Overview. The ED cost forecasting task required predicting total emergency department costs over the next 3 years (regression). The training set comprised 2,000 patients with structured features (primary chronic condition, prior 5-year ED visits, prior 5-year ED costs) and one PDF billing receipt per patient containing detailed line-item charges with CPT codes. The evaluation metric was Mean Absolute Error (MAE) in dollars.

Human-Guided Feature Engineering. The agentic system's initial approach used structured features with basic PDF extraction (29 features) and a single XGBoost regressor, yielding a 5-fold CV MAE of $467. Human analysts identified two shortcomings: the PDF parser captured only raw totals without clinical detail, and a single model was unstable across folds ($\pm$\$22 MAE). This guided an enhanced iteration with richer cross-source features and a weighted ensemble, yielding approximately 40 features across three groups.

Structured features ($\sim$12): Beyond raw prior cost and visit counts, we engineered: (1) *cost-per-visit ratio* with Laplace smoothing; (2) *log, square root, and squared transformations* of prior costs and visits to handle right-skewed distributions; (3) *cost$\times$visit interaction*; and (4) one-hot encoding for primary chronic condition (HF, DiabetesComp, Pneumonia).

PDF-extracted features ($\sim$16): We developed a PDF parsing pipeline using PyPDF2 with regex-based CPT code and cost extraction. Features included: (1) *total billed amount*

and *line item count*; (2) *visit complexity*, counts of high-complexity (CPT 99284/99285), medium-complexity (99283), and total ED visit codes (9928x prefix); (3) *service utilization*, laboratory test count (CPT 80000-89999) and imaging count (CPT 70000-79999); (4) *cost statistics*, average, maximum, minimum, and standard deviation of per-line costs; (5) *insurance type* extracted from receipt text (private, Medicare, Medicaid binary indicators); and (6) 3-digit ZIP code.

Cross-source and derived features (∼14): A key human decision was engineering features that cross-referenced PDF and structured data: (1) *cost consistency ratio* (PDF total / structured prior cost); (2) *cost and visit discrepancies* between PDF and structured data; (3) *complexity ratios*: high-complexity and medium-complexity visits as fractions of total visits; (4) *per-visit service rates*: lab tests and imaging per ED visit; (5) *cost variability*: line-item cost standard deviation normalized by mean; (6) *service intensity score* weighted composite of complexity, lab, and imaging counts per visit; and (7) log-transformed PDF cost features.

Model Architecture: Weighted Ensemble. We employed a five-model weighted ensemble with manually assigned weights reflecting model reliability:

- **XGBoost** (35%): $n = 700$, $lr = 0.03$, $depth = 8$, $subsample = 0.8$, $colsample\_bytree = 0.8$; *StandardScaler*
- **Gradient Boosting** (30%): $n = 500$, $lr = 0.03$, $depth = 7$, $subsample = 0.8$; *StandardScaler*
- **Random Forest** (15%): $n = 500$, $depth = 20$, $min\_samples\_split = 6$; unscaled features
- **Extra Trees** (15%): $n = 500$, $depth = 20$, $min\_samples\_split = 6$; unscaled features
- **Ridge Regression** (5%): $\alpha = 5.0$; *StandardScaler*

Human decision rationale: (1) *Weighted averaging over stacking*: the agentic system initially proposed a stacking meta-learner, but human analysts overrode this: with only N=2,000 samples, a learned meta-learner risked overfitting; manually assigned weights based on hold-out MAE provided more stable combination. (2) *Higher weight for boosting models*: hold-out validation showed XGBoost (MAE $435) and Gradient Boosting (MAE $450) consistently outperformed RF ($466), ET ($468), and Ridge ($507), justifying 65% combined weight for boosting. (3) *Selective scaling*: tree-based models (RF, ET) were trained on raw features since they are scale-invariant, while boosting and linear models used StandardScaler. (4) *Cross-source features*: human-designed features linking PDF and structured data (cost consistency ratio, visit discrepancies) captured billing patterns invisible to either source alone.

Validation Strategy. We used both 80/20 hold-out validation for rapid iteration and 5-fold CV (KFold, shuffle=True) for final estimation. CV MAE was $458.51 ± $19.93, with per-fold MAEs ranging from $435 to $478. The final ensemble achieved MAE = $465.13 on the public leaderboard, ranking 6th.

C. Challenge 3: Discharge Readiness Assessment

Task Overview. The discharge readiness task required predicting whether a patient would be ready for discharge by day 11 of their hospital stay (binary classification). The training set comprised 1,000 hospital stays (656 discharge-ready, 344 not ready; 65.6%/34.4% class split) with structured features (unit type, admission reason) and multimodal time-series data: 10 days of vital signs (HR, SBP, DBP, temperature, respiratory rate) plus daily progress notes. We additionally merged patient demographics, admission records, and ED cost history via shared identifiers. The evaluation metric was Macro-F1.

Human-Guided Feature Engineering. This challenge required the most extensive human-AI iteration, progressing through five rounds. The agentic system's baseline used only two categorical features (unit type, admission reason) with a Random Forest, achieving CV F1 = 0.49. Human-guided iterations added feature groups incrementally: (i) vital sign statistics (+0.18 F1 to 0.67), (ii) NLP keyword features (no gain, prompting strategy revision), (iii) derived physiological composites (+0.01), (iv) cross-dataset integration of admission records and ED cost history (+0.05 to 0.73). The final model used ∼120 features. Missing numeric values were imputed with training-set medians.

Structured features (∼15): One-hot encoding for unit_type, admission_reason, sex, insurance, primary diagnosis, and primary chronic condition, plus numeric age, prior 5-year ED visits, and prior 5-year ED costs. Merging across stays, patients, admissions, and ED cost tables, a human decision to exploit cross-dataset signal, added features unavailable in any single source.

Vital sign features (∼57): For each of five vital signs across 10 days, we computed: (1) *endpoint values*: first and last day readings; (2) *distributional statistics*: mean, median, standard deviation, min, max, and range; (3) *higher-order statistics*: coefficient of variation, skewness, and kurtosis for HR and temperature; (4) *trend indicators*, linear regression slope via scipy.stats.linregress; (5) *change metrics*: absolute and percent change (last minus first); (6) *recent-window features*: last-3-day means for key vitals; (7) *volatility*: rolling 3-day standard deviation for HR; and (8) *clinical threshold crossings*: days with HR>100, HR<60, temperature>37.5°C, and SBP>140.

Derived physiological features (∼22): We computed clinically meaningful composites: (1) *hemodynamic features*: mean and last pulse pressure (SBP−DBP), pulse pressure change, mean arterial pressure (DBP + PP/3); (2) *stability composites*: vitals stability score (negative sum of HR, SBP, temperature, and RR standard deviations) and volatility score (sum of HR and temperature CVs); (3) *binary improving flags*: whether HR, temperature, SBP, and overall vital trends were improving; (4) *normal-range flags*: whether last-day HR (60-100), temperature (36.5-37.5°C), and SBP (90-140) fell within normal limits, plus a count of vitals in normal range; and (5) *age interactions*: age×HR, age×stability, age×temperature, capturing how age modulates vital sign significance.

NLP features (∼37): We extracted clinical indicators from the 10 daily progress notes using domain-specific keyword matching (with partial stemming, e.g., "ambulat" for ambulatory/ambulating). Feature groups included: (1) *positive/negative sentiment*, keyword counts across 15 positive terms (stable, improved, ambulatory, afebrile, tolerating, etc.) and 15 negative terms (confused, fever, unstable, worsening, decline, etc.); (2) *mobility and independence*, counts for 7 mobility and 5 independence keywords, with last-day, mean, and trend (linear slope) aggregations; (3) *discharge readiness*, mentions of discharge-related terms (discharge, home, ready, cleared, planning); (4) *complication indicators*, mentions of complication, setback, concern, deterioration; (5) *therapy progress*, physical/occupational therapy mentions; (6) *pain trajectory*, pain keyword trend; (7) *cognitive status*, mean score from alert/oriented/responsive keywords; (8) *temporal aggregations*, last-3-day sentiment, note length trend, and composite scores combining sentiment with mobility; and (9) *meta-features*, total note count and average note length.

Human decision: The agentic system initially applied generic TF-IDF to the daily notes, as it had for discharge summaries in Challenge 1. Human analysts observed that this yielded no improvement (iteration ii above) because daily notes were short (10-20 words) and lacked the lexical diversity that TF-IDF exploits. They redirected the system to use domain-specific clinical keyword matching with partial stemming ("ambulat," "discharge") targeting discharge criteria terminology, which improved recall and, when combined with temporal aggregation, contributed to the final gains.

Model Architecture: Five-Model Stacking Ensemble. We employed a stacking classifier with five base models and a logistic regression meta-learner, trained via 5-fold stratified internal CV:

- **XGBoost-A** (n=400, lr=0.02, depth=7, subsample=0.85, colsample_bytree=0.85): Primary gradient boosting model; random_state=42
- **XGBoost-B** (n=300, lr=0.03, depth=9, subsample=0.8, colsample_bytree=0.9): Deeper variant with slightly inflated scale_pos_weight ($\times$1.1) for minority class; random_state=43
- **Random Forest** (n=400, depth=25, min_samples_split=3, class_weight=balanced): Robust bagging baseline
- **Extra Trees** (n=300, depth=20, class_weight=balanced): Randomized splitting for diversity
- **Gradient Boosting** (n=300, lr=0.03, depth=7, subsample=0.85): Complementary sequential learner

Meta-Learner: Logistic regression (class_weight=balanced, max_iter=1000) trained on out-of-fold predictions via predict_proba.

Rationale for Human Decisions. The five iterations moved CV F1 from 0.49 to 0.73 (+0.24), with human decisions at each checkpoint: (1) *Statistical aggregation over raw time series* (iteration i): the agentic system proposed feeding raw 10$\times$5 vital matrices to an LSTM; human analysts redirected to statistical aggregation, recognizing that N=1,000 would cause sequence model overfitting. This single decision yielded the largest gain (+0.18 F1). (2) *NLP strategy pivot* (iteration ii$\rightarrow$iii): when TF-IDF produced no improvement, human analysts diagnosed the failure (short notes, specialized vocabulary) and substituted clinical keyword matching, ultimately contributing via temporal aggregation in later iterations. (3) *Cross-dataset integration* (iteration iv): merging admission and ED cost data added predictive context (diagnosis, prior utilization) absent from stays data alone, this was the key insight of the final iteration (+0.05 F1). (4) *Physiological composites:* Mean arterial pressure, pulse pressure, and vitals stability scores encode clinical knowledge about hemodynamic status that raw vitals do not capture individually. (5) *Five-model stacking:* With class imbalance (34.4% minority), diverse models with balanced class weights and probability-based stacking improved minority class recall from 0.58 to 0.67.

Validation Strategy. We used 5-fold stratified CV with Macro-F1 throughout all iterations. The final model achieved CV F1 = 0.732 $\pm$ 0.018 (per-class: 0.65 not-ready, 0.81 ready). Feature importance analysis confirmed that avg_note_length (0.194), HR-related features, and sentiment trends dominated, validating the multimodal engineering approach. Final test predictions achieved Macro-F1 = 0.7939 on the public leaderboard, ranking 3rd.

III. Results

We present performance results across all three AgentDS Healthcare challenges, ablation studies quantifying the impact of human decisions, and cross-task analysis revealing generalizable patterns in human-AI collaboration for healthcare prediction.

A. Overall Performance

Table I summarizes our final leaderboard performance across all three challenges. To contextualize these results, we contrast them with the agentic system's unguided baselines: Challenge 1 improved from CV F1 = 0.846 (single Optuna-tuned XGBoost) to 0.896 after human-guided feature expansion and ensemble design; Challenge 2 from CV MAE \$467 (single model, basic PDF parsing) to \$459 with enriched cross-source features and weighted ensembling; and Challenge 3 from CV F1 = 0.49 (two categorical features) to 0.73 through four rounds of human-directed iteration. These gains, +0.050 F1, -\$8 MAE, and +0.24 F1 respectively, quantify the value added by human guidance at each stage.

Key Observations: (1) *Cross-task consistency:* We ranked 5th overall in the healthcare domain, with per-challenge rankings of 5th, 6th, and 3rd, respectively, demonstrating that the same human-guided workflow, iterative error analysis, domain-informed feature engineering, deliberate ensemble diversity, transfers across classification, regression, and multimodal data types without task-specific redesign. (2) *Competitive with minimal automation:* Our readmission prediction (0.8986) was within 0.0058 F1 of the top score (0.9044), and discharge

TABLE I
PERFORMANCE ON AGENTDS HEALTHCARE PUBLIC LEADERBOARD

Challenge	Metric	Our Score	Rank	1st Place
Challenge 1: Readmission	Macro-F1	0.8986	5th	0.9044
Challenge 2: ED Cost	MAE (USD)	$465.13	6th	$448.75
Challenge 3: Discharge	Macro-F1	0.7939	3rd	0.8006
Domain Score	**Aggregate**	**0.8430**	**5th**	**0.9426**

readiness (0.7939) ranked 3rd, despite using only tree-based models and manual feature engineering, no neural architectures or automated feature search. (3) *Validation alignment:* Test scores closely matched or exceeded cross-validation estimates (Challenge 1: CV 0.896 vs test 0.899; Challenge 2: CV $459 vs test $465; Challenge 3: CV 0.732 vs test 0.794), confirming that human-guided validation strategies (stratified CV, appropriate metrics) prevented overfitting.

B. Ablation Studies: Impact of Human Decisions

To isolate the value of each human intervention, we conducted counterfactual ablation studies: for each decision category, we reverted to what the agentic system would have done without human guidance and measured the resulting performance change. Table II presents 5-fold cross-validation results on training sets.

TABLE II
ABLATION STUDY: QUANTIFYING IMPACT OF HUMAN DECISIONS

Configuration	Ch1 (F1)	Ch2 (MAE)	Ch3 (F1)
Full Approach	**0.895**	**$458**	**0.785**
Multimodal Feature Ablations:			
Text/PDF/TS features	0.858	$478	0.720
Only multimodal	0.812	$495	0.680
Domain Engineering:			
Interaction terms	0.887	$463	0.780
Risk scores	0.889	$461	0.782
Automated selection	0.881	$469	0.770
Model Architecture:			
Ensemble (single)	0.889	$463	0.773
Default hyperparams	0.883	$471	0.768
Validation Strategy:			
5-fold (vs 8-fold)	0.893	$460	0.783
No stratification	0.887	$458	0.772

Note: Ch1/Ch3 = Macro-F1, Ch2 = MAE (USD). Negative deltas indicate degradation.

Multimodal Data Integration: Removing the human-directed multimodal features, trigram TF-IDF (Challenge 1), enriched PDF parsing (Challenge 2), and vital sign statistics (Challenge 3), resulted in the largest performance drop (-0.037 F1 / +$20 MAE average). This corresponds directly to the single highest-gain human decisions in each challenge: expanding TF-IDF from 200 unigrams to 850 trigrams, adding CPT-code-level PDF extraction, and redirecting from raw time series to statistical aggregation. Using *only* multimodal features without structured data also performed poorly (-0.083 F1 / +$37 MAE), confirming that human-guided integration of *both* modalities is critical.

Domain Feature Engineering: Human-designed interaction terms (age×Charlson, chronic condition×cost) and composite risk scores contributed modestly but consistently (-0.006 to -0.008 F1 / +$3-5 MAE). Replacing human-guided feature selection with automated selection (SelectKBest) degraded performance by -0.014 F1 / +$11 MAE on average, illustrating that clinical intuition for feature relevance outperformed statistical filtering on these sample sizes.

Ensemble Architecture: Reverting from the human-designed ensembles to the agentic system's single-model baselines degraded performance by -0.006 F1 / +$5 MAE, with the largest impact on Challenge 3 (N=1,000) where ensemble diversity was most valuable for minority class recall. Human-tuned hyperparameters outperformed agentic defaults by -0.012 F1 / +$13 MAE, though notably this was the smallest human contribution category, consistent with our recommendation to prioritize feature engineering over hyperparameter tuning.

C. Cross-Task Analysis: Generalizable Patterns

Table III shows top-5 features by importance across all three challenges, revealing consistent patterns.

TABLE III
TOP-5 MOST IMPORTANT FEATURES BY CHALLENGE

Challenge 1	Challenge 2	Challenge 3
Charlson band	Prior 5y cost	Avg note length
Prior ED (6m)	Prior 5y visits	HR last
Age×Charlson	Line item count	Temp. last
Length of stay	Cost/visit ratio	Sentiment trend
"wound" keyword	High-complexity	Mobility trend

Common Patterns: (1) Historical utilization dominates (prior visits/costs, comorbidity); (2) Multimodal features appear in top-5 despite being $<$20% of features; (3) Domain interactions (age×Charlson) rank above raw features; (4) Trends outweigh absolutes (HR slope vs raw HR).

Interpretation: Human decisions contributed approximately 7-8% relative improvement (+0.065 cumulative F1), moving our approach from agentic-only baselines (mid-tier performance) to competitive top-5 results. While individual contributions appear modest (0.008-0.041), they compound across pipeline stages, precisely the pattern predicted by our

TABLE IV
QUANTIFIED IMPACT OF HUMAN DECISIONS

Human Decision Category	Avg Gain
Multimodal feature extraction	+0.041 F1
Domain interactions	+0.010 F1
Ensemble diversity	+0.008 F1
Clinical keywords	+0.015 F1
Hyperparameter choices	+0.014 F1
Total contribution	**+0.065 F1**

iterative methodology. The ordering of impact (multimodal extraction > clinical keywords > hyperparameters > domain interactions > ensemble diversity) suggests that early-stage decisions about *what data to extract* matter more than late-stage decisions about *how to model it*.

IV. DISCUSSION

Our participation in the AgentDS Healthcare benchmark demonstrates that systematic human-AI collaboration can achieve competitive performance (5th overall) across diverse clinical prediction tasks while maintaining interpretability and reproducibility. We discuss key lessons, limitations, and implications for healthcare ML practice, situating our findings within the broader literature on human-AI collaboration, model selection, and clinical deployment.

A. Key Lessons Learned

1. Domain Knowledge Pays Dividends. Clinical intuition for feature engineering (age×Charlson interactions, mobility keywords, vital sign trends) consistently outperformed automated feature selection by 1-3% (Table II). This advantage compounds across multiple decision points: our cumulative human contribution (+0.065 F1 / -$38 MAE, Table IV) moved us from mid-tier to top-5 performance. *Implication:* Healthcare ML practitioners should prioritize domain expertise in early pipeline stages (feature engineering) over extensive hyperparameter tuning in later stages.

2. Multimodal Integration Requires Task-Specific Strategies. While TF-IDF succeeded for discharge notes (Challenge 1), it failed for short daily notes (Challenge 3); clinical keyword extraction proved more effective. Similarly, PDF parsing required a regex tailored to the billing receipt structure (Challenge 2). *Lesson:* No single approach generalizes across all unstructured healthcare data types; human judgment is required to select appropriate extraction methods based on data characteristics (document length, structure, domain-specific terminology).

3. Ensemble Diversity Over Complexity. Our stacking ensembles (5 base models each) achieved competitive performance using only tree-based models and regularized linear models. The key was *deliberate diversity*: in Challenge 1, three XGBoost variants with different depth/learning rate profiles (depth 5/6/7) plus L1/L2 logistic regression; in Challenge 2, the human decision to use fixed weights instead of a learned meta-learner prevented overfitting on N=2,000 samples. *Insight:* Manually-configured diverse ensembles outperform random search, generating similar models.

4. Small Samples Favor Interpretability. With modest training sizes (N=1,000-5,000), our statistical feature aggregation (Challenge 3: vital trends, stability) and domain features outperformed deep learning approaches (LSTM, BERT), which require larger samples. This sample efficiency enabled faster iteration and model interpretability through feature importance analysis. *Recommendation:* For healthcare datasets with N<10,000, prioritize feature engineering and tree-based models over deep learning.

5. Validation Strategy Impacts Reliability. Stratified CV was critical for imbalanced classification tasks (Δ = -0.007 F1 when removed), while 8-fold vs 5-fold showed a negligible difference (Δ = -0.002). *Guideline:* Invest effort in stratification and appropriate metrics (Macro-F1 for imbalance, MAE for skewed costs) rather than excessive fold counts.

6. Reproducibility Through Documentation. Our approach's primary strength is not novelty but *systematic documentation* of every human decision with rationale. This transparency enables: (1) peer review and critique, (2) replication by other teams, (3) trust-building for clinical deployment. *Principle:* In healthcare ML, reproducibility and interpretability matter as much as raw performance.

B. Cross-Task Insights

What Generalized: Tree-based ensemble methods (XGBoost, Gradient Boosting, RF, Extra Trees), domain-informed interactions, and multimodal feature extraction proved universally effective. Historical utilization features (prior visits, costs, comorbidity) dominated importance rankings across all tasks.

What Required Adaptation: Text processing (TF-IDF vs keywords), time-series handling (raw values vs aggregates vs trends), and class imbalance strategies varied by task. Rigid application of a single methodology across all tasks would have degraded performance by 5-10%.

Human vs Automated Roles: Our iteration logs reveal a consistent division of labor. The agentic system excelled at *how much*: Optuna-based hyperparameter search (Challenge 1), automated CV estimation, and rapid model retraining. Humans excelled at *what* and *why*: which tables to join (cross-dataset integration in Challenge 3, +0.05 F1), which NLP strategy to use (keyword matching over TF-IDF for short notes), and when to override the system's proposals (fixed ensemble weights over learned stacking in Challenge 2). The highest-value human decisions were early-stage choices about data representation; the agentic system efficiently optimized the resulting models.

C. Limitations

1. Synthetic Data Constraints. The AgentDS benchmark uses synthetic data matching real distributions but lacking clinical noise (missing values, measurement errors, data entry inconsistencies). Our performance estimates may not generalize to messier real-world EHR data. *Mitigation:*

External validation on real EHR datasets (MIMIC-IV, eICU) is needed before clinical deployment.

2. Generalization vs Task-Specific Optimization. Our unified methodology prioritized consistency across tasks over single-task performance, potentially leaving gains on the table. We ranked between 3rd and 6th on individual challenges (5th overall by domain score), suggesting room for task-specific optimization. *Trade-off:* Generalization enables knowledge transfer across healthcare problems, but domain-specific competitions may require specialized approaches.

3. Limited Deep Learning Exploration. We did not extensively explore transformer models (BioClinicalBERT for text, LSTMs for vitals) due to sample size constraints (N<5,000). Larger datasets might favor end-to-end deep learning over manual feature engineering. *Scope:* Our findings apply to common healthcare ML scenarios (N=1,000-10,000 samples, tabular+text+time-series); different conclusions may hold for massive EHR databases (N>100,000).

4. Computational Resources. Our GPU-accelerated training (Challenge 1: 8 minutes) required hardware not available to all practitioners. CPU-only training extends to 30-60 minutes, still acceptable but not trivial. *Accessibility:* Cloud-based training (Google Colab, AWS) democratizes access, but cost and technical barriers remain.

5. Multimodal Feature Engineering Labor. Extracting PDF billing features, parsing clinical notes, and computing vital trends required 10-15 hours of human effort per challenge. This investment is worthwhile for research or high-stakes applications but may not scale to rapid prototyping scenarios. *Pragmatism:* Practitioners must balance effort vs performance gains.

D. Implications for Healthcare ML Practice

For Clinicians: Our feature importance rankings (Table III) align with established clinical risk factors (Charlson index, prior utilization, vital trends) [19], building trust that models capture medically-meaningful relationships. The interpretability of tree-based methods enables clinician review of model decisions, critical for clinical deployment.

For ML Engineers: Start with domain-guided feature engineering before investing in complex architectures. Our ablations show that multimodal features (+0.041 F1) and interactions (+0.010) contributed 4-10× more than ensemble complexity (+0.006-0.008). Simple ensembles with good features outperform complex ensembles with default features.

For Healthcare Systems: Benchmark performance (5th place, 0.8986 F1 readmission, $465 ED cost MAE, 0.7939 F1 discharge) suggests these models approach clinical utility thresholds. However, external validation on real EHR data, prospective trials, and fairness audits (race, age, insurance bias) are prerequisites before deployment.

For Researchers: The AgentDS benchmark successfully evaluates human-AI collaboration through: (1) multimodal data requiring domain expertise, (2) standardized evaluation preventing overfitting, (3) public leaderboards enabling comparison. We recommend similar benchmarks for other healthcare domains (oncology, radiology, mental health).

E. Future Work

Short-term Extensions:

- *Transformer-based text models:* BioClinicalBERT embeddings for discharge notes (Challenge 1) and daily notes (Challenge 3) may capture richer semantic content than TF-IDF/keywords. Requires >5,000 samples for stable fine-tuning.
- *Uncertainty quantification:* Calibrated probability predictions (isotonic regression, temperature scaling) enable clinicians to assess model confidence, critical for high-stakes decisions.
- *Fairness analysis:* Stratify performance by patient demographics (age, sex, insurance) to identify potential biases in learned models.

Long-term Directions:

- *External validation:* Apply our approach to real EHR datasets (MIMIC-IV, eICU) to assess generalization to noisy clinical data.
- *Prospective trials:* Deploy readmission or discharge models in live hospital settings (with clinician oversight) to measure impact on patient outcomes and costs.
- *Automated multimodal feature learning:* Investigate whether end-to-end deep learning (multimodal transformers) can match or exceed manual feature engineering at scale (N>50,000).
- *Continual learning:* Adapt models to distribution shift (COVID-19 pandemic, policy changes) through online learning or periodic retraining.

V. Conclusion

We have presented a human-guided agentic AI approach for multimodal clinical prediction, evaluated across three AgentDS Healthcare benchmark challenges: 30-day readmission prediction (Macro-F1 = 0.8986, 5th), emergency department cost forecasting (MAE = $465.13, 6th), and discharge readiness assessment (Macro-F1 = 0.7939, 3rd), achieving 5th place overall in the healthcare domain. Our central contribution is not the final models themselves but the *iterative workflow* through which human analysts guided an agentic AI system from naive baselines to competitive performance.

The iteration evidence is concrete. In Challenge 1, human-directed expansion of text features (200 unigrams to 850 trigrams) and the switch from a single model to a stacking ensemble lifted CV F1 from 0.846 to 0.896. In Challenge 2, human analysis of per-model hold-out MAEs justified fixed ensemble weights over the agentic system's proposed learned meta-learner, improving stability on N=2,000 samples. In Challenge 3, our strongest result (3rd place), four rounds of human intervention moved CV F1 from 0.49 to 0.73, with the largest single gain (+0.18) coming from a human decision to redirect the agentic system from raw time-series modeling to statistical vital sign aggregation. Across all challenges,

ablation studies attribute +0.065 cumulative F1 to human decisions, with early-stage choices about data representation (multimodal feature extraction: +0.041) contributing far more than late-stage model tuning (ensemble diversity: +0.008).

These results support three principles for deploying agentic AI in healthcare settings:

1) *Human guidance compounds:* Individual decisions appear modest (0.008-0.041 F1), but they accumulate across pipeline stages. Domain-informed feature engineering at each stage outperformed extensive automated search, consistent with the observation that the agentic system excelled at optimizing *within* a representation but not at choosing *which* representation to optimize.
2) *No single extraction strategy generalizes across clinical data types:* TF-IDF succeeded for discharge summaries (Challenge 1) but failed for short daily notes (Challenge 3); the human decision to pivot to clinical keyword matching was essential. PDF parsing required billing-receipt-specific regex (Challenge 2). Task-specific human judgment for unstructured data remains a bottleneck that current agentic systems do not resolve autonomously.
3) *Interpretability is a prerequisite, not a trade-off:* Our top features (Charlson band, prior ED utilization, vital sign trends, sentiment trajectory) align with established clinical risk factors, enabling clinician trust. Tree-based ensembles with documented decision rationale maintained this transparency while achieving competitive performance.

Our approach has clear limitations: synthetic benchmark data lacks real-world clinical noise as our sample sizes (N=1,000-5,000) favor feature engineering over deep learning, a conclusion that may not hold for larger EHR databases; and the human effort required (10-15 hours per challenge) constrains scalability. Future work should validate on real EHR data (MIMIC-IV, eICU), explore whether transformer-based multimodal models can reduce human effort at scale, and conduct prospective trials measuring clinical impact.

The AgentDS benchmark demonstrates that standardized multimodal healthcare tasks can rigorously evaluate human-AI collaboration. Our results suggest that the most effective current paradigm is neither fully autonomous agentic AI nor traditional manual analysis, but an iterative loop in which human domain expertise shapes the problem representation and agentic systems efficiently optimize within it.

ACKNOWLEDGMENTS

We thank the AgentDS benchmark organizers for providing a rigorous evaluation framework for human-AI collaboration in healthcare data science.

REFERENCES

[1] E. G. Oh, S. Oh, S. Cho, and M. Moon, "Predicting readmission among high-risk discharged patients using a machine learning model with nursing data: Retrospective study," *JMIR Medical Informatics*, vol. 13, no. 1, p. e56671, 2025.

[2] R. Loutati, A. Ben-Yehuda, S. Rosenberg, and Y. Rottenberg, "Multimodal machine learning for prediction of 30-day readmission risk in elderly population," *The American Journal of Medicine*, vol. 137, no. 7, pp. 617–628, 2024.

[3] C. P. Wu, R. B. Shirley, A. Milinovich, K. Liu, E. Mireles-Cabodevila, H. Khouli, A. Duggal, and A. Bhattacharyya, "Exploring timely and safe discharge from icu: a comparative study of machine learning predictions and clinical practices," *Intensive Care Medicine Experimental*, vol. 13, no. 1, p. 10, 2025.

[4] S. Tang, A. Tariq, J. A. Dunnmon, U. Sharma, P. Elugunti, D. L. Rubin, B. N. Patel, and I. Banerjee, "Predicting 30-day all-cause hospital readmission using multimodal spatiotemporal graph neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 4, pp. 2071–2082, 2023.

[5] N. Karunanayake, "Next-generation agentic ai for transforming healthcare," *Informatics and Health*, vol. 2, no. 2, pp. 73–83, 2025.

[6] V. G. Hinostroza Fuentes, H. A. Karim, M. J. T. Tan, and N. AlDahoul, "Ai with agency: a vision for adaptive, efficient, and ethical healthcare," *Frontiers in Digital Health*, vol. 7, p. 1600216, 2025.

[7] M. Rahman, A. Bhuiyan, M. S. Islam, M. T. R. Laskar, R. Mahbub, A. Masry, S. Joty, and E. Hoque, "Llm-based data science agents: A survey of capabilities, challenges, and future directions," *arXiv preprint arXiv:2510.04023*, 2025.

[8] Y. Kim, C. Park, H. Jeong, Y. S. Chan, X. Xu, D. McDuff, H. Lee, M. Ghassemi, C. Breazeal, and H. W. Park, "Mdagents: An adaptive collaboration of llms for medical decision-making," *Advances in Neural Information Processing Systems*, vol. 37, pp. 79 410–79 452, 2024.

[9] W. Shi, R. Xu, Y. Zhuang, Y. Yu, J. Zhang, H. Wu, Y. Zhu, J. C. Ho, C. Yang, and M. D. Wang, "Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 22 315–22 339.

[10] P. Naliyatthaliyazchayil, R. Muthyala, J. W. Gichoya, and S. Purkayastha, "Evaluating the reasoning capabilities of large language models for medical coding and hospital readmission risk stratification: zero-shot prompting approach," *Journal of medical Internet research*, vol. 27, p. e74142, 2025.

[11] P. N. Srinivasu, G. L. Aruna Kumari, S. Ahmed, and A. Alhumam, "Exploring agentic ai in healthcare: A study on its working mechanism and swot analysis," *Frontiers in Medicine*, vol. 12, p. 1753443, 2026.

[12] J. A. Bishop, H. A. Javed, R. El-Bouri, T. Zhu, T. Taylor, T. Peto, P. Watkinson, D. W. Eyre, and D. A. Clifton, "Improving patient flow during infectious disease outbreaks using machine learning for real-time prediction of patient readiness for discharge," *Plos one*, vol. 16, no. 11, p. e0260476, 2021.

[13] S. Jahangiri, M. Abdollahi, E. Rashedi, and N. Azadeh-Fard, "A machine learning model to predict heart failure readmission: toward optimal feature set," *Frontiers in Artificial Intelligence*, vol. 7, p. 1363226, 2024.

[14] N. Maslej, L. Fattorini, R. Perrault, Y. Gil, V. Parli, N. Kariuki, E. Capstick, A. Reuel, E. Brynjolfsson, J. Etchemendy *et al.*, "Artificial intelligence index report 2025," *arXiv preprint arXiv:2504.07139*, 2025.

[15] S. Purkayastha, R. Isaac, A. Veldandi, R. Saxena, P. Singh, P. Vaswani, E. A. Krupinski, J. A. Danish, and J. W. Gichoya, "Measuring impact of radiologist-ai collaboration: Efficiency, accuracy, and clinical impact," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024, pp. 1–4.

[16] A. Ashfaq, A. Sant'Anna, M. Lingman, and S. Nowaczyk, "Readmission prediction using deep learning on electronic health records," *Journal of biomedical informatics*, vol. 97, p. 103256, 2019.

[17] S. Davis, J. Zhang, I. Lee, M. Rezaei, R. Greiner, F. A. McAlister, and R. Padwal, "Effective hospital readmission prediction models using machine-learned features," *BMC Health Services Research*, vol. 22, no. 1, p. 1415, 2022.

[18] Y. Deng, S. Liu, Z. Wang, Y. Wang, Y. Jiang, and B. Liu, "Explainable time-series deep learning models for the prediction of mortality, prolonged length of stay and 30-day readmission in intensive care patients," *Frontiers in Medicine*, vol. 9, p. 933037, 2022.

[19] S. N. Kasthurirathne, P. G. Biondich, S. J. Grannis, S. Purkayastha, J. R. Vest, and J. F. Jones, "Identification of patients in need of advanced care for depression using data extracted from a statewide health information exchange: a machine learning approach," *Journal of medical Internet research*, vol. 21, no. 7, p. e13809, 2019.