Iowa Initiative for Artificial Intelligence

Final Project Report

Project title:	Predicting Deterioration Trajectories in Patients with Lung Cancer Using Machine Learning and Natural Language Processing
Principal Investigator:	Alaa AlBashayreh, PhD, MSHI, RN (College of Nursing, Assistant Professor, University of Iowa)
Prepared by (IIAI):	Mudireddy Avinash / Joonwoo Park/ Hafiza Akter Munira
Other investigators:	Co-PI: Nick Street, Muhammad Furqan, Stephanie H White
Date:	March 08, 2026
Were specific aims fulfilled:	Y
Readiness for extramural proposal?	Y

Brief summary of accomplished results:

- To accomplish the prediction of deterioration score for lung cancer patients, our team has done two major goals below.
- Aim 1: The significant functional features with the value for whole patients are extracted by a large language model (LLM) and synchronized them as a harmonized score for using it as a scaled target value on Aim 2. Among the 2,392 patients with 109,959 textual data, LLM found 53,973 harmonized score for 2,275 patients.
- Aim 2: We developed and examined several machine learning classifiers to predict lung cancer patients' harmonized deterioration scores using structured clinical variables. Among the models examined, XGBoost performed the best, with a ROC AUC of 0.76 and an accuracy of 59% over a 180-day period. The performance dropped slightly for the 90- and 30-day windows but stayed consistent. Excluding class 5 (death) helped improve the results. From the feature importance analysis, we found that diagnosis count, lung site, allergy info, and vitals were among the top predictors.

Research report:

Aims (specified by the project proposal):

Advanced lung cancer patients face high risk of deterioration, yet early prediction is challenging due to inconsistent use and variability in tools like ECOG, KPS, and PPS. Current models often miss critical structured and unstructured EHR data, including functional decline noted in clinical narratives. This study uses machine learning (ML) and natural language processing (NLP) to integrate diverse data sources, aiming to improve prediction deterioration and support timely palliative care interventions.

Our specific aims are as follows:

Aim 1: To extract the functional features from clinical notes for every medical record number (MRN). To evaluate whether extracting functional decline information from clinical notes using NLP enhances deterioration prediction, we would eventually make a harmonized performance score based on the extracted functional features. We will employ NLP methods to extract functional decline scores from clinical notes, focusing on ECOG (Range 0 to 4), KPS (Range 0 to 100), and the Palliative Performance Scale (PPS, range 0 to 100), which are common indicators of a patient's functional status. These scores will be synchronized into a **harmonized performance score (Range 0 to 5)** to assess whether they enhance the model's predictive accuracy. We hypothesize that incorporating functional decline information from clinical notes as extracted using NLP will improve the ML model performance in predicting deterioration.

Aim 2: To develop and validate a machine learning classifier for predicting harmonized deterioration scores in lung cancer patients using aggregated clinical features across multiple time windows. In this aim, we converted longitudinal clinical data into a non-longitudinal format by aggregating key features over three-time windows—180 days, 90 days, and 30 days—prior to each prediction point (T_target), selected only if at least 6 days had passed since the previous score. We extracted and summarized data from multiple clinical sources, including **labs, vitals, medications, chemotherapy records, diagnoses, and cancer-related attributes.** Feature engineering involved calculating statistical measures like mean, min, max, standard deviation, and count, along with binary indicators and time-since-last-event metrics. These aggregated features, combined with constant baseline information such as **demographics** and **diagnosis** details, were used to train and validate several machine learning models including XGBoost, Gaussian Naive Bayes, Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Logistic Regression. The goal was to determine whether structured clinical data, when summarized across meaningful time windows, could effectively predict a patient's functional deterioration score.

Data:

Aim 1: EHR was provided by the University of Iowa Hospital and Clinics (UIHC) with adult lung cancer patients (N=3,800). It includes about 26 million textual data with 3,800 textual labels, which are demographic information, cancer diagnosis, symptom information, treatment, clinical notes, etc. The following steps were taken for preprocessing.

We figured out the labels among EHR we should keep extracting the functional features of ECOG, KPS, and PPS. For preserving every visit of each MRN, the 3 textual labels have remained, which are clinical notes, dates, and types of notes. Afterward removing the patient not having at least one clinical note, that means there is no evidence to extract the functional feature, the **final patient number was 2,392** with **109,959 textual data**.

Aim 2: For Aim 2, we selected clinical data (non-text) pertaining to patients who had available harmonized deterioration scores and metrics. We constructed a dataset with **53,973** prediction instances from **2,275** different patients. Each instance was associated with a harmonized score calculated at a certain Note Date (T_target) and included only if there was a minimum of 6 days since the last score for that patient.

To train ML models, we transformed longitudinal data into **non-longitudinal** features across **30**-, **90**-, and **180**-day windows before each T_target. Feature sets included labs (CBC, liver function), vitals (HR, RR), medications (active use of opioids, steroids), chemotherapy (exposure to platinum-based drugs), diagnoses (Calderon categories), cancer details (stage, histology), and demographics. Statistical descriptions (mean, min, max, std, count) and event features (days since last event) were calculated. A 30% completeness criterion kept ~11,000 informative features from the 180-day window. The outcome label (harmonized score 0-5), the ground truth for model evaluation based on Aim 1 NLP outputs, was used for model validation.

AI/ML Approach:

Aim 1: The LLaMA-3.1-8B-Instruct model has been used to extract the functional features from clinical notes. Based on our aim, a text generation model is the optimal way to generate the score either the clinical notes have the feature implicitly or explicitly. The **59-performance lexicon in a prompt** written by PI has been used to infer the functional features from clinical notes if the feature, the value, or both are not mentioned explicitly in clinical notes.

Aim 2: Using structured clinical data, we formulated a machine learning pipeline to predict harmonized deterioration scores (0-5). The preprocessing phase included removing extraneous identifiers, categorical feature one-hot encoding, and filling gaps with MRN-level group means

followed by KNN imputation. No rows were removed after outlier detection via the IQR method since over 99% of records were flagged, to avert losing meaningful clinical variation. The overall workflow of the approach is shown in Figure 1.

We trained multiple classifiers which include XGBoost, a custom PyTorch MLP, Gaussian Naive Bayes (GNB), Linear SVM, and Logistic Regression (LR). XGBoost was chosen as a key model because it performs well with high-dimensional structured data and can handle missing values effectively.

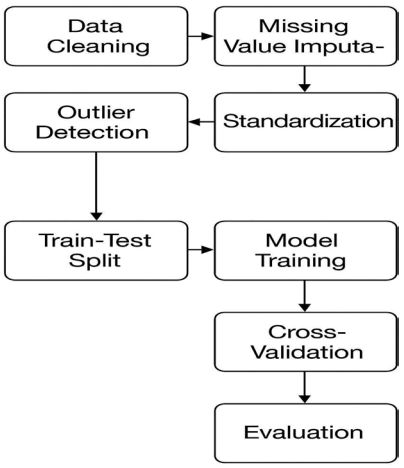

Figure 1: Workflow of the Machine Learning

Approach Experimental methods, validation approach:

Aim 1: Several NLP models such as Mistralai, Meta-LLaMA, and Qwen are experimented to find the model has the best performance. To validate whether the NLP model we selected is trustful or not, the small sample experiment has been conducted. 320 golden datasets, which have explicit notes of 214 and inferred of 106, were used as a validation. After fixing the model that shows the highest accuracy, the hyperparameter study has been done as well. The hyperparameters considered this study are temperature, top_p, top_k, maximum token, and number of beams. When the model is configured with **a temperature of 0.3, top_p of 0.9, top_k of 50, and a maximum token of 150,** the performance has the best performance. Zero-shot and few-shot learning comparisons have also been made to compute the accuracy and time efficiency of each. Finally, zero-shot learning was used in this study because it has the advantage of time efficiency with a similar performance of few-shot learning.

Aim 2: For each 30, 90, and 180-day time window, we used aggregated clinical characteristics (labs, vitals, medicines, chemo, diagnoses, cancer details, demographics) to predict harmonized score (0-4; lower is better, class 5 signifying death was removed owing to class imbalance and very poor performance issue). The data (N=53,973) was divided 80/20 (train/test) using stratified sampling. Standard deviation normalization was conducted on numerical features while filtration of zero-variance columns was performed, and features showing high dimensionality versus reduced diversity were set at 0.8 and 0.9 for correlation thresholds, revealing a balance between correlated feature reduction and retained diversity. For all models, 5-fold stratified cross-validation was performed to ensure reliable performance with class-imbalanced splits. The accuracy, macro-ROC AUC, and permutation-based feature importance were all used to evaluate the system. Table 1 shows the best hyperparameter of the models that were selected based on Grid and RandomizedSearchCV.

Table 1. Best Hyperparameter Selection of the classifiers

Classifier	Best Hyperparameter
Logistic Regression	{'C': 0.1, 'penalty': 'l1'}
Gaussian Naive Bayes	{'var_smoothing': 1e-06}
XGBoost	'subsample': 0.6, 'reg_lambda': 1.0, 'reg_alpha': 0.1, 'n_estimators': 200, 'min_child_weight': 5, 'max_depth': 9, 'learning_rate': 0.05, 'gamma': 0}
SVM	{'base_estimator__C': 0.01}
MLP	'optimizer_weight_decay': 0.0, 'optimizer_lr': 0.001,'num_hidden_layers': 1, 'hidden_units': 256, 'dropout_rate': 0.2, 'activation': 'relu', 'batch_size': 128}

Results:

Aim 1: Regarding the overall accuracy of NLP model, zero-shot learning was 76.5%, compared to few-shot learning of 79.8% (**Table 2**). To evaluate two types of learnings, three comparisons were extracted respectively. First, scale name is that the model extracts or infers the scale name correctly compared to the reference. It is quite a simple task, and the accuracy of few-shot learning (74.7%) is remarkably increased than the accuracy of zero-shot learning (61.6%). On the other hand, scale value, which extracts the scale name and value together, accuracy of few-shot learning was not distinct from the zero-shot learning. The accuracy of few-shot learning was only 1.1%p higher than the accuracy of zero-shot learning. Finally, an acceptable range comparison is conducted. Here acceptable range means that each scale value gets clinical tolerance (±1 for ECOG, ±10 for KPS/PPS). The accuracy for both is slightly changed (3.3%p increment).

Table 2. Accuracy comparison between Few-shot vs. Zero-shot learning

Scale Name Comparison		
Scale Type	**Few-shot**	**Zero-shot**
ECOG	65.4%	82.7%
KPS	84.5%	74.6%
PPS	76.6%	61.0%
Overall	74.7%	61.6%
Scale Value Comparison		
Scale Type	**Few-shot**	**Zero-shot**
ECOG	64.7%	72.8%
KPS	73.9%	67.6%
PPS	71.1%	62.3%
Overall	69.2%	68.1%
Acceptable Range Comparison		
Scale Type	**Few-shot**	**Zero-shot**
ECOG	**65.7%**	**80.6%**
KPS	**88.4%**	**78.9%**
PPS	**90.8%**	**68.8%**
Overall	**79.8%**	**76.5%**

Regarding that perspective, the harmonized score from the actual clinical notes was extracted by the LLM with zero-shot learning. Among 2,392 patients with 109,959 textual data, **53,973 harmonized scores with 2,275 unique patients** were finally generated.

Aim 2: Among all classifiers tested, XGBoost consistently achieved the best performance across different time windows (180, 90, and 30 days). For the 180-day window, **XGBoost** achieved the **highest** accuracy (59%) and ROC AUC (0.76), followed by the MLP with 56.7% accuracy and 0.73 ROC AUC. ROC AUC (Receiver Operating Characteristic - Area Under Curve) measures the model's ability to distinguish between classes, with higher values indicating better discriminatory power. As the prediction window narrowed, model performance declined slightly, with XGBoost maintaining the lead at 57.9% accuracy and
0.75 ROC AUC (90-day), and 55.9% accuracy with 0.72 ROC AUC (30-day). Other models like Logistic Regression, SVM, and Decision Tree showed moderate performance, while Gaussian Naive Bayes underperformed. Notably, excluding class 5 (death) from the training set improved model calibration and ROC AUC scores across all models. All the model's performance comparisons at different time windows is illustrated in Figure 1.

Permutation-based feature importance revealed consistent and clinically relevant predictors across all three-time windows (180, 90, and 30 days) (Figure 2). For the 180-day window, key features included LungSiteCategory_OtherSite, Unmapped_Diag_Count_180d, and vital_Height_max, indicating the importance of anatomical cancer location and longitudinal diagnostic patterns.

In the 90-day window, Unmapped_Diag_Count_90d, LungSiteCategory_OtherSite, and ageAtDiagnosis remained prominent, while short-term vital signs like vital_SBP_min and vital_SpO2_count gained influence. For the 30-day window, top predictors such as Unmapped_Diag_Count_30d, LungSiteCategory_OtherSite, and stage_Stage II suggested that recent diagnostic activity and cancer stage dominate immediate deterioration risk.

Overall, diagnosis frequency, lung site, allergy history, and basic vitals/labs repeatedly emerged as influential across windows, reinforcing their role in capturing short- and long-term clinical decline.

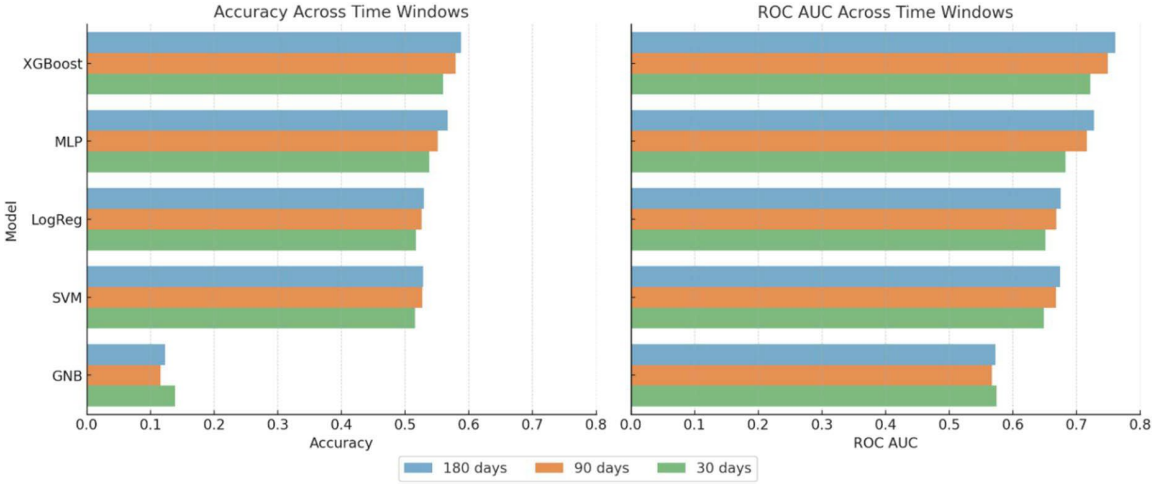

Figure 2: Model performance comparison across different time windows

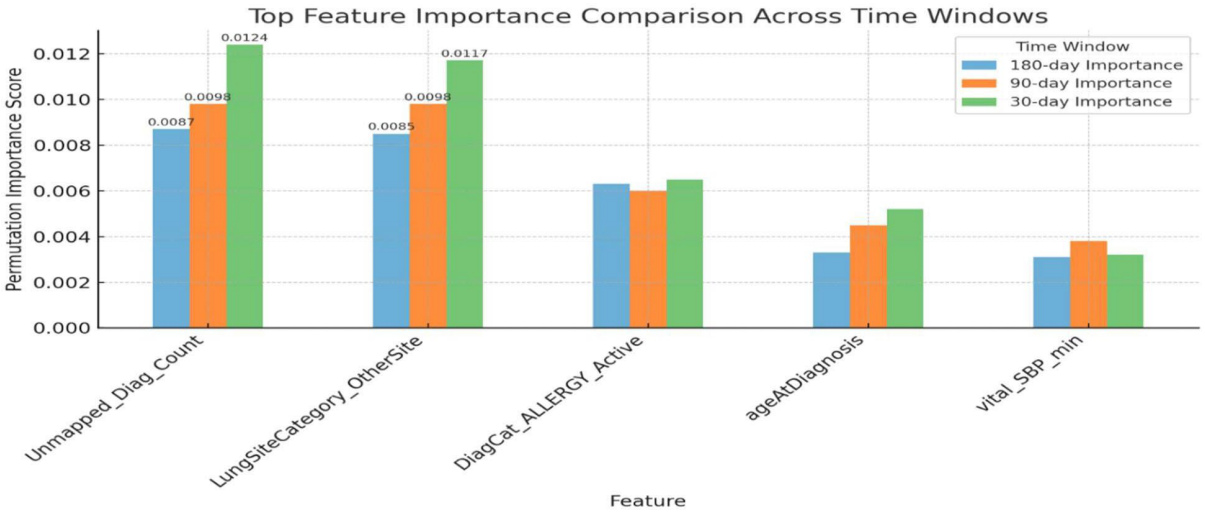

Figure 3: Top feature comparison across different time windows

Discussion:

Aim 1: We evaluated several instruction-tuned LLMs for feature and value extraction via text generation, and the LLaMA model demonstrated the best overall performance. Although few-shot prompting achieved a slightly higher accuracy (79.8%) than zero-shot (76.5%) on a 320-sample test set, we adopted the zero-shot setup due to its significantly lower inference time. When applied to 109,959 real-world clinical notes, the model extracted 53,973 non-null instances of the functional lung deterioration feature. This outcome highlights both the scalability and practical utility of the model, despite the inherent limitations of missing values and complex input length variability.

Aim 2: This aim focused on predicting functional deterioration scores using structured clinical features and machine learning. Among all tested classifiers, XGBoost consistently delivered the best performance, with a ROC AUC of 0.76 and accuracy of 0.59 using a 180-day input window. However, class imbalance, particularly underrepresentation of class 5 (death), affected early performance. Restricting prediction to classes 0–4 significantly improved the model's ability to distinguish patient conditions. Performance varied across time windows: shorter windows (30 days) better reflected acute changes, while longer windows (180 days) captured overall trends, suggesting different clinical utilities.

Despite achieving promising results, the model faced challenges during feature engineering. Inconsistent formats, sparse data, and high feature correlation required extensive cleaning and dimensionality reduction. The final input shape was reduced to 170 features for 180 days' time window improving scalability using
0.8 correlation threshold. Still, removal of longitudinal structure may have reduced sensitivity to progression patterns. While the ML pipeline worked effectively, future improvements in data standardization and richer temporal modeling could enhance both predictive efficiency and clinical relevance.

Ideas/aims for future extramural project:

Aim 1: We strongly believe that few-shot learning usually has the potential to infer the specific feature compared to zero-shot learning, and it is also dependent on the structure of prompt. Hence, we plan to reinforce the prompt for gathering more information from the baseline.
Aim 2: We plan to extend this work by using longitudinal data with RNN-based models to better capture patient trajectories over time, which were lost in the current aggregated approach. Future directions also include combining structured features with NLP-extracted functional scores to enhance predictive accuracy.