Physician-Augmented AI: Enhancing Machine Learning Detection of NSCLC and Associated Clinical Indicators

Nayyar A, Roy A, Somani M, Verma V, Markan R, Sachdev A, Sethi A, Goyal R, Brooks L, Seligman M, Webster D

Objective

This study aims to determine the effectiveness of incorporating physician inputs into AI/ML models for detecting NSCLC and associated clinical indicators from unstructured data, and to compare the results with models trained exclusively on pre-annotated data.

Methodology

- In this study, we utilized Optum® de-identified clinical notes from 2007 to 2023 to confirm NSCLC diagnoses using ICD-9/10 diagnosis codes and textual references of NSCLC and its variations.
- To ensure the models were trained on various clinical complexities, patients were divided into two cohorts: Cohort 1, consisting of patients diagnosed exclusively with NSCLC, and Cohort 2, comprising patients with NSCLC and other types of cancers.
- Notes were segmented into training, validation, and test sets to support model training and performance assessment.
- The identified concepts included NSCLC, symptoms, risk factors, and family history.
- We developed two AI/ML models: one trained on pre-annotated data (Model A) and another enhanced with physician reviews (Model B).
- Both models employed a Named Entity Recognition (NER) architecture consisting of Char CNNs, BiLSTM, and a CRF layer to identify NSCLC, risk factors, family history, and symptoms.
- The performance of these models was evaluated based on precision, recall, and F1-scores to compare the effectiveness of physician-augmented data in improving model accuracy.

Results

- The total number of patients and notes identified during the time frame were 78,926 and 1,046,347, respectively.
- The notes were split into training (660), test (100), and validation sets (240) (**Figure 1**).
- Models trained on clinically reviewed data outperformed those trained on pre-annotated data across all the metrics.

Char CNN: Character-level Convolutional Neural Networks
BiLSTM: Bidirectional Long Short-Term Memory
CRF: Conditional Random Field

- NSCLC identification saw a 20% increase in the F1 score, with improvements in both precision and recall.
- NSCLC symptoms showed the most significant improvement, with a 65% increase in the F1 score, primarily driven by a 108% rise in precision.
- NSCLC risk factor recognition showed a 47% improvement in the F1 score, with a notable increase in precision, while NSCLC family history achieved high accuracy with a 16% improvement in the F1 score.

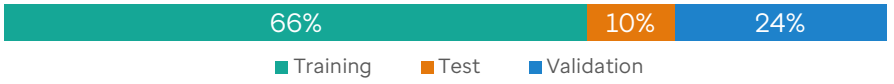

Figure 1. Note split for model training

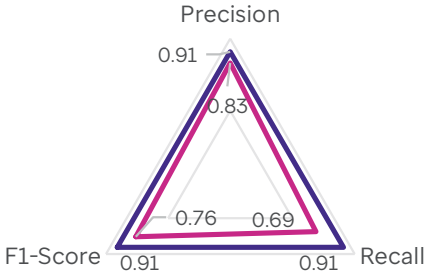

Figure 2. NSCLC Detection

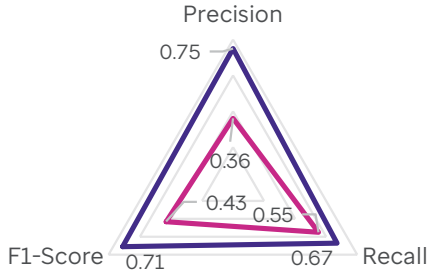

Figure 3. NSCLC Symptoms

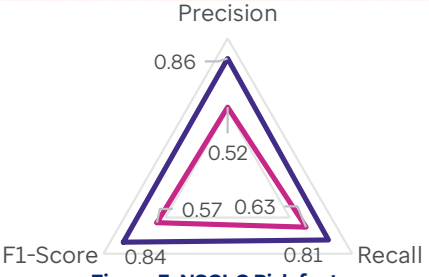

Figure 3. NSCLC Risk factors

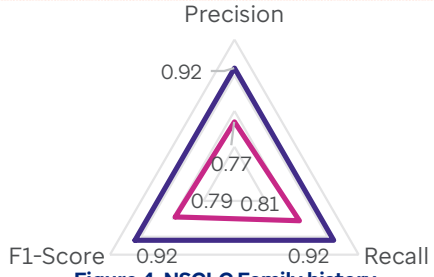

Figure 4. NSCLC Family history

Conclusions

Clinically reviewed notes significantly enhance the accuracy and reliability of NLP models in extracting unstructured clinical data.
These findings support integrating clinical guided review in healthcare NLP pipelines to improve the real-world applicability.

Optum