Virginia HIMSS Data and Analytics Symposium
May 5, 2025

Analyze ICU Patients' Notes with Multiple NLP Models to Predict Diagnosis

ROLAND DEPRATTI
CENTRAL CONNECTICUT STATE UNIVERSITY

Problem

- EHR systems' cut-and-paste functions result in 17% of patients' notes having a similarity measure of > 20% and 5% having a similarity score of > 50%, which makes using patients' notes as part of diagnosis difficult (Shala, 2023).

- Researchers estimate that 15% of diagnoses are misclassified, accounting for 17% of adverse effects (Berner et al., 2008)

- Interactions between staff and patients, as captured in notes, can provide a wealth of observational data.

- A predicted diagnosis code generated from patient notes could be used during a physician's diagnosis efforts and as a trigger data point (Stockwell, 2022).

Natural Language Processing

Two main NLP approaches were examined:

1) Traditional
- Convert notes into a matrix using Term Frequency-Inverse Document Frequency (TF-IDF)
- Fit a model, Random Forest, using the matrix as input to predict diagnosis.

2) Large Language Models
- Tokenize the notes to token vectors
- Fit a model, BERT, using the vectors to predict diagnosis

Kalra et al. (2019) used text from pathology reports to classify cancer diagnoses. The highest classification accuracy achieved was 92%.

Huang et al. (2019) used patients' notes from the MIMIC-III database to predict readmission within 30 days of discharge (ROC = 67.2%).

Purpose and Research Questions

Purpose: Using the first 3 days of patients' notes, identify which NLP model best classifies the primary diagnosis code.

- Would a Random Forest model using a TFIDF matrix or a BERT model produce an accuracy that would make it medically beneficial?
- Will an ensemble method that combines the two algorithms to make a prediction provide more accurate predictions than base models?
- Does the 'closeness' of the diagnoses chosen impact the model's predictive performance?
- How best to tune the large number of BERT hyperparameters: cross-validation vs Response Surface Methodology?
- Is the predictive power of the developed models equally good regardless of the patient's ethnicity?
- What performance levels make the models medically useful?

Data and Fitted Models

Distinct Diagnoses		Similar Diagnoses	
ICD10	Description	ICD10	Description
A41.9	Sepsis unspecified organism	I21.4	Subendocardial infarction
I21.4	Subendocardial infarction	I25.10	Coronary Atherosclerosis
J96.00	Acute Respiratory Failure	I35.2	Nonrheumatic Aortic Stenosis
N17.9	Acute Kidney Failure, Unspecified	I50.9	Unspecified Congestive Heart Failure

Total Fit Models

Combining all options to test, 14 models were fitted and tested.

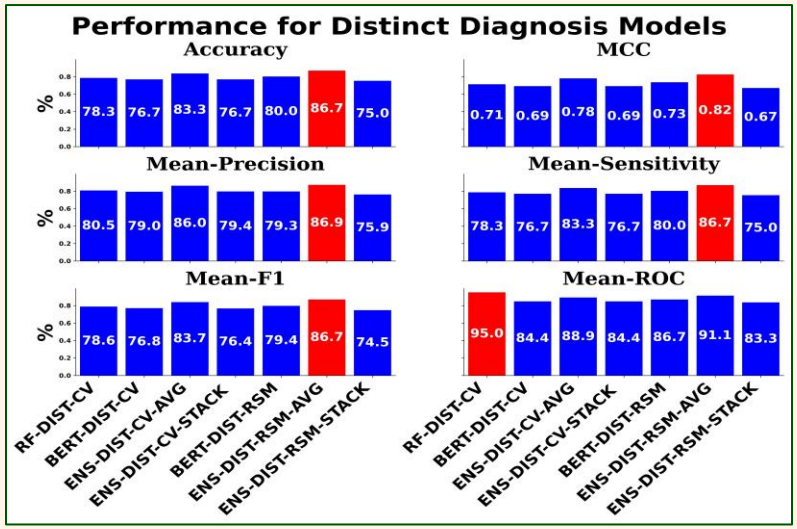
Data: MIT MIMIC-III database

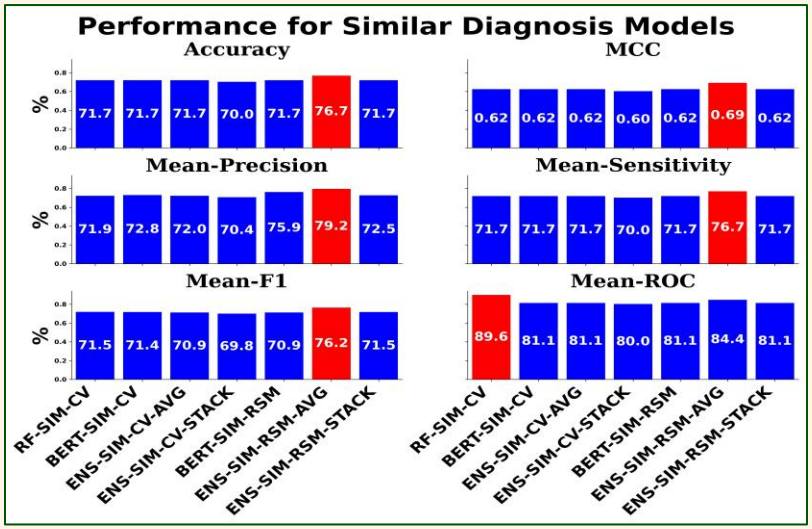
General population:
155 admissions were randomly selected for each diagnosis.

Purpose	Admissions
Train	440 (110 each diagnosis)
Validate-1	60 (15 each diagnosis)
Validate-2	60 (15 each diagnosis)
Test	60 (15 each diagnosis)
Minority Test	100 admissions

Model	Diagnoses Group	Hyper parameter	Algorithm	Ensemble	Ensemble Type
RF-Dist-CV	D	CV	RandomF	N	
RF-Sim-CV	D	CV	RandomF	N	
BERT-Dist-CV	D	CV	BERT	N	
BERT-Dist-RSM	D	RSM	BERT	N	
BERT-Sim-CV	S	CV	BERT	N	
BERT-Sim-RSM	S	RSM	BERT	N	
Ens-Dist-CV-Avg	D	CV	Both	Y	Avg
Ens-Dist-CV-Stack	D	CV	Both	Y	Stacked
Ens-Dist-RSM-Avg	D	RSM	Both	Y	Avg
Ens-Dist-RSM-Stack	D	RSM	Both	Y	Stacked
Ens-Sim-CV-Avg	S	CV	Both	Y	Avg
Ens-Sim-CV-Stack	S	CV	Both	Y	Stacked
Ens-Sim-RSM-Avg	S	RSM	Both	Y	Avg
Ens-Sim-RSM-Stack	S	RSM	Both	Y	Stacked

Results





1. Using all other performance measurements besides mean ROC, the ensemble method using average probability performed best.
   Distinct -> ROC = 83% accuracy = 86.7%,
   Similar -> ROC = 81% accuracy = 76.7%.

2. The RF-TFIDF model had the better mean ROC.
   Distinct -> ROC = 95%, Accuracy = 78%
   Similar -> ROC = 90%, Accuracy = 71.7%

3. The ROC of all ensemble models exceeded the ROC, reported Huang et al., for their readmission prediction models.

Minority Testing

There are lots of horror stories of models trained on biased data that result in poor predictive results for subsets of a population.

The MIMIC-III database is comprised of 70% Caucasian admissions vs 58% in the general population, and only 19.9% minority population vs 42% in the general population.

An additional set of 100 minority admissions were sampled, not previously randomly selected, with a distinct diagnosis. These 100 admissions were then used to predict diagnoses using the distinct diagnosis models, and the prediction performance was assessed.

The minority ensemble performance was slightly smaller than the general population, but the RF-TFIDF minority model exceeded all other models for the general population.

Model	Accuracy	MCC	Mean Prec	Mean Sens	Mean F1	Mean Roc
RF-DIST-CV	89.0%	0.853	91.0%	87.6%	89.0%	97.5%
BERT-DIST-CV	75.0%	0.667	75.1%	77.6%	75.3%	84.6%
ENS-DIST-CV-AVG	82.0%	0.760	82.8%	81.9%	81.7%	87.9%
ENS-DIST-CV-STACK	79.0%	0.720	80.5%	81.8%	80.9%	88.2%
BERT-DIST-RSM	75.0%	0.667	75.6%	75.6%	74.8%	83.6%
ENS-DIST-RSM-AVG	81.0%	0.747	80.2%	80.4%	79.9%	87.0%
ENS-DIST-RSM-STACK	82.0%	0.760	86.6%	82.9%	84.0%	88.2%

Conclusions

The number of misclassified diagnoses is estimated as 15%. These misclassifications account for 17% of adverse effects.

Fourteen models were developed using the first three days of nurses' and doctors' patient notes as input. The models varied in algorithm, hyperparameter tuning, stand-alone or ensemble, and diagnosis codes.

An ensemble model trained in distinct diagnosis codes using an average probability technique that used Response Surface Methodology to develop hyperparameters for BERT predicted distinct diagnoses with an accuracy of 86.7% and a Matthews Correlation Coefficient (MCC) of 0.8222.

The same ensemble model trained on similar diagnosis codes predicted similar diagnoses with an accuracy of 76.7% and a Matthews Correlation Coefficient (MCC) of 0.6889.

The same ensemble model trained on a minority population had an 82% accuracy with an MCC > .7600. A stand-alone RF-TFIDF model performed better with an accuracy of 89% and MCC > .8533.