Using Deep Learning to Identify Cognitive Impairment in Electronic Health Records

Tanish Tyagi, Colin Magdamo, Ayush Noori, Mayuresh Deodhar, Zhuoqiao Hong, Dmitry Prokopenko, Rudy E. Tanzi, Deborah Blacker, Bradley T. Hyman, Shibani S. Mukerji, M. Brandon Westover, Sudeshna Das

Athinoula A. Martinos Center For Biomedical Imaging
MASSACHUSETTS GENERAL HOSPITAL
HARVARD MEDICAL SCHOOL
Massachusetts Institute of Technology

Background

Dementia is under-diagnosed by healthcare professionals—only one in four people who suffer from dementia are diagnosed. Even when a diagnosis is made, it may not be entered as a structured diagnosis code in a patient's charts. Information relevant to cognitive impairment is often found within electronic health records but manual review of clinician notes by experts is both time consuming and often prone to errors. Automated evaluation of these notes presents an opportunity to label patients with cognitive impairment (CI) in real-world data.

Methods

We selected a cohort of patients from the Mass General Brigham (MGB) healthcare system who were older than 60 years (as of July 13, 2021), had at least one unique encounter with a match of a keyword pertinent to CI, and had *APOE* genotype data available from the BioBank (N=16,428).

The sequences with the below keyword matches were used to extract sequences.

Keyword	Match Count
Memory	109218
Cognition	87655
Dementia	51034
Cerebral	45886
Cerebrovascular	36370
Cerebellar	26863
Cognitive Impairment	20267
Alzheimer	20581
MOCA	9767
Neurocognitive	7711
MCI	3889
Amnesia	3695
AD	2673
Lewy	2561
MMSE	2134
LBD	224
Corticobasal	147
Pick's	41

Table 1. Keywords used for sequence extraction

Study Sample Characteristics

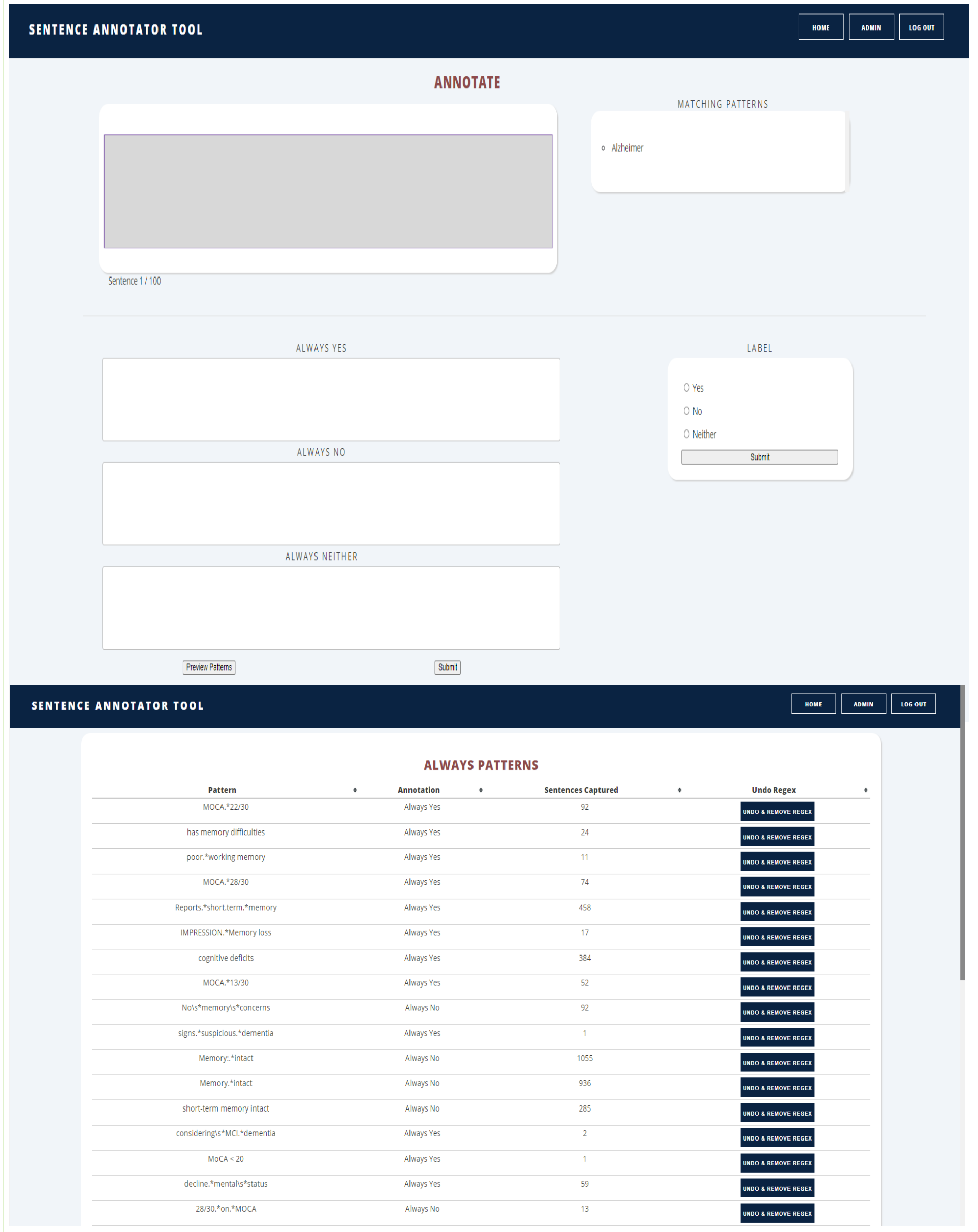
We extracted unstructured clinician notes, identified matches to dementia-related keywords, and constructed 800-character sequences from these matches.

Characteristic	(N = 16428)
Age (years) mean (SD)	73.01 (7.96)
Gender Male, *n* (%)	8740 (53.2)
Race, *n* (%)	
White	14896 (90.7)
Other/Not Recorded	608 (3.7)
Black	570 (3.5)
Hispanic	170 (1.0)
Asian	168 (1.0)
Indigenous	16 (0.01)
Ethnicity, *n* (%)	
Hispanic	16053 (97.8)
Non-Hispanic	375 (2.2)
APOE Genotype, *n* (%)	
APOE $\varepsilon 2$	2028 (12.3)
APOE $\varepsilon 3$	10177 (62.0)
APOE $\varepsilon 3$	4223 (25.7)
Average Speciality Visits (SD)	1.67 (4.6)
Average PCP Encounters (SD)	5.25 (5.63)

Table 2: Dataset Demographics

Annotations

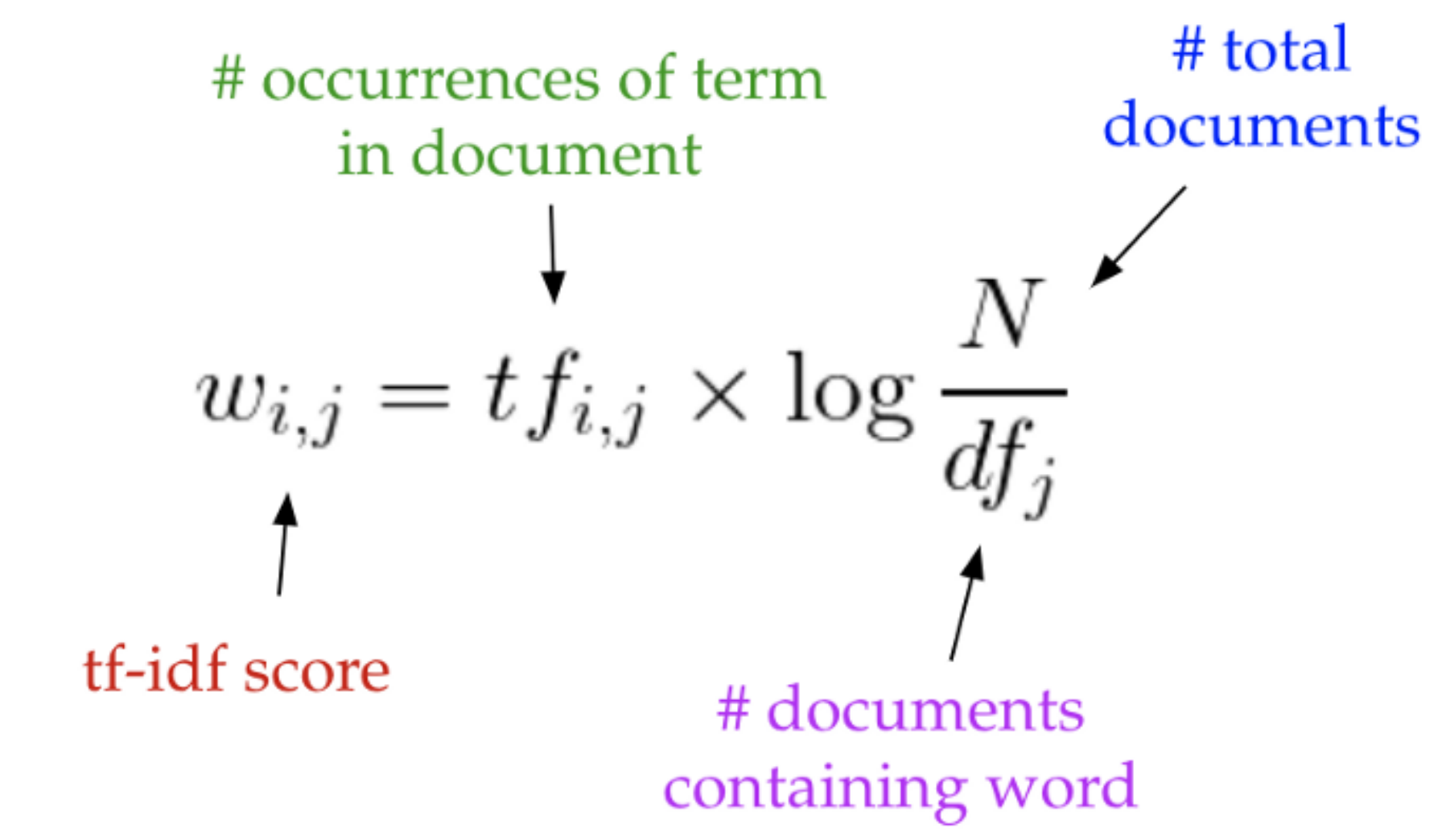
The collected sequences were then annotated by neurologists using a web-based annotation tool as 1) Yes, i.e., patient has CI; 2) No i.e., Patient does not have CI; and 3) Neither i.e., sequence has no information on patient's cognition.



Pattern	Annotation	Sentences Captured	Undo Regex
MOCA.*22/30	Always Yes	92	UNDO & REMOVE REGEX
has memory difficulties	Always Yes	24	UNDO & REMOVE REGEX
poor.*working memory	Always Yes	11	UNDO & REMOVE REGEX
MOCA.*28/30	Always Yes	74	UNDO & REMOVE REGEX
Reports.*short.term.*memory	Always Yes	458	UNDO & REMOVE REGEX
IMPRESSION.*Memory loss	Always Yes	17	UNDO & REMOVE REGEX
cognitive deficits	Always Yes	384	UNDO & REMOVE REGEX
MOCA.*13/30	Always Yes	52	UNDO & REMOVE REGEX
No\s*memory\s*concerns	Always No	92	UNDO & REMOVE REGEX
signs.*suspicious.*dementia	Always Yes	1	UNDO & REMOVE REGEX
Memory:.*intact	Always No	1055	UNDO & REMOVE REGEX
Memory.*intact	Always No	936	UNDO & REMOVE REGEX
short-term memory intact	Always No	285	UNDO & REMOVE REGEX
considering\s*MCI.*dementia	Always Yes	2	UNDO & REMOVE REGEX
MoCA < 20	Always Yes	1	UNDO & REMOVE REGEX
decline.*mental\s*status	Always Yes	59	UNDO & REMOVE REGEX
28/30.*on.*MOCA	Always No	13	UNDO & REMOVE REGEX

Figures 1 and 2: Pictures of Annotation Tool

TF-IDF Word Vectorization



$$w_{i,j} = tf_{i,j} \times \log \frac{N}{df_j}$$

# occurrences of term in document
# total documents
tf-idf score
# documents containing word

Source: https://nanonets.com/blog/topic-modeling-with-lsa-plsa-lda-lda2vec/

We performed TF-IDF (term frequency-inverse document frequency) vectorization on the annotated sequences and selected features based on a term's Pearson correlation coefficient with the outcome of cognitive impairment.

Word	Corr	Word	Corr
Alzheimer	0.352	Relation	0.271
Deficits	0.341	Disease	0.271
Cognitive	0.325	Onset	0.266
MOCA	0.321	Age	0.263
Short	0.316	Reports	0.237
Family	0.311	Social	0.230
Mother	0.295	Learning	0.222
History	0.294	Impairment	0.212
Father	0.286	Difficulties	0.211
Memory	0.276	Term	0.207

Table 2. TF-IDF Weights of Top Words Correlated with Prediction Outcome

Regularized Logistic Regression

Regularized logistic regression was applied with the annotated cognitive impairment labels. We used different correlation coefficients as thresholds to select features and iterated over different lambda values to determine the optimal lambda value and correlation coefficient threshold.

Results

The regularized logistic regression model was trained on a dataset comprised of 8,363 annotated sequences from (N=2,417) unique patients using 10-fold Cross Validation stratified across the two CI labels.

Below are the results on the held-out test set (293 annotated sequences from N=77 unique patients) with a probability threshold of 0.89.

AUCROC	ACC	Sensitivity
0.94	0.90	0.73
Specificity	**Lambda**	**Correlation**
0.94	10	0.07

Table 3: Results from TF-IDF with Regularized Logistic Regression

We then applied our model on the other 186,730 sequences from (N=13,941) unique patients that were not part of the training loop. Below are the percentages of patients who were predicted to have CI and the percent who had at least 1 Medication or ICD code relevant to CI stratified by APOE allele.

	Count	Yes (%)	No/Ntr (%)	Med/ICD Code (%)
APOE $\varepsilon 2$	1754	0.38	0.62	0.11
APOE $\varepsilon 3$	8751	0.38	0.62	0.11
APOE $\varepsilon 4$	3436	0.40	0.60	0.17

Table 4: Results from Sample on rest of patients

Patient CI	Med/ICD Code	Count
1	1	1665
1	0	6999
0	1	84
0	0	5193

Table 5: Confusion Matrix from Sample

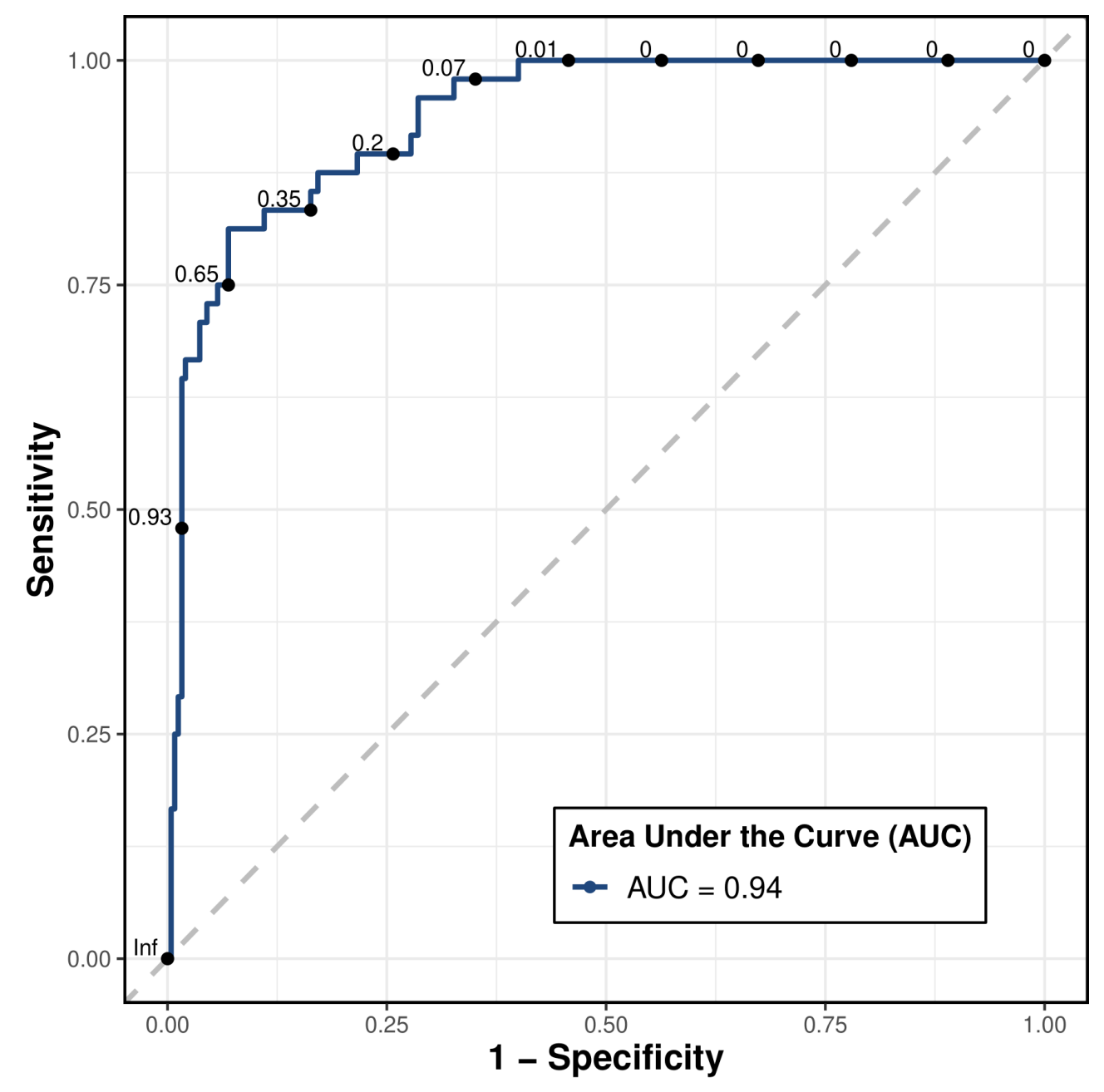


Figure 3: AUROC Curve for TF-IDF

Conclusion and Future Plans

We developed a machine learning tool to identify potential patients with cognitive impairment. Our work can help combat the issue of underdiagnosis for dementia. Future plans include reducing false positives by gathering more annotated sequences and using deep learning natural language processing (NLP) techniques (currently in development). Currently, many of our false positives are from the model identifying the presence of a keyword but struggling to understand the context around the keyword match; many of these contexts negate the presence of CI for the patient in question. Deep Learning can leverage contextual information, making it promising for this task.

References

1. Karlson, Elizabeth W et al. "Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations." *Journal of personalized medicine* vol. 6,1 2. 14 Jan. 2016, doi:10.3390/jpm6010002
2. Amjad H, Roth DL, Sheehan OC, Lyketsos CG, Wolff JL, Samus QM. Underdiagnosis of Dementia: an Observational Study of Patterns in Diagnosis and Awareness in US Older Adults. J Gen Intern Med. 2018;33(7):1131-8. Epub 2018/03/07. doi: 10.1007/s11606-018-4377-y. PubMed PMID: 29508259; PubMed Central PMCID: PMCPMC6025653.
3. Bradford A, Kunik ME, Schulz P, Williams SP, Singh H. Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. Alzheimer Dis Assoc Disord. 2009;23(4):306-14. Epub 2009/07/02. doi: 10.1097/WAD.0b013e3181a6bebc. PubMed PMID: 19568149; PubMed Central PMCID: PMCPMC2787842.
4. Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, 1996, pp. 267–288. *JSTOR*, www.jstor.org/stable/2346178.