NeuraHealthNLP: An Automated Screening Pipeline to Detect Undiagnosed Cognitive Impairment in Electronic Health Records with Deep Learning and Natural Language Processing

Tanish Tyagi
ttyagi@mgh.harvard.edu
Phillips Exeter Academy
Exeter, New Hampshire, United States
Table of Contents

Table of Contents ............................................................................................................................ 1
Abstract .......................................................................................................................................... 2
1 Introduction .................................................................................................................................  3
   1.1 Background ............................................................................................................................ 3
   1.2 Research Problem ................................................................................................................... 4
   1.3 Research Goals ...................................................................................................................... 5
   1.4 Related Works ........................................................................................................................ 6
2. Unstructured Data Preparation Pipeline ........................................................................................ 6
   2.1 Dataset Description ................................................................................................................ 7
   2.2 Dataset Filtering and Keyword Selection .................................................................................. 7
   2.3 Sequence Construction and Extraction from Dataset ............................................................... 9
   2.4 Sequence Annotation ............................................................................................................ 10
3. Sequence Classification Methods .................................................................................................. 12
   3.1 TF-IDF Machine Learning Model .......................................................................................... 12
      3.1.1 TF-IDF Computations ................................................................................................... 13
   3.2 Attention-based Deep Learning ClinicalBERT Model ............................................................... 14
      3.2.1 ClinicalBERT Model Architecture ................................................................................... 15
      3.2.2 ClinicalBERT Model Computations ................................................................................. 16
4. Sequence Classification Results ................................................................................................... 17
   4.1 Comparison between TF-IDF and ClinicalBERT ................................................................. 17
5. Patient Level Predictor ............................................................................................................... 21
   5.1 Comparison with Clinical Methods ......................................................................................... 22
6. System Integration with Web Application for Real-Time diagnosis of CI by Clinicians ............... 23
7 Conclusion and Future Work ....................................................................................................... 26
8 Appendix A: Web App used for Sequence Annotation ............................................................... 28
9 References .................................................................................................................................. 29
Abstract

Dementia related cognitive impairment (CI) affects over 55 million people worldwide and is growing rapidly at the rate of one new case every 3 seconds. It is a terrifying disease that robs a person of their identity, and eventually leads to death (100% fatality rate). With a recurring failure of clinical trials, early diagnosis is crucial, but 75% of dementia cases go undiagnosed globally with up to 90% in low-and-middle-income countries. Current diagnostic methods are notoriously complex, involving manual review of medical notes, numerous cognitive tests, expensive brain scans or spinal fluid tests. Information relevant to CI is often found in the electronic health records (EHRs) and can provide vital clues for early diagnosis, but a manual review by experts is tedious and error prone. This project develops a novel state-of-the-art automated screening pipeline for scalable and high-speed discovery of undetected CI in EHRs. To understand the linguistic context from complex language structures in EHR, a database of 8,656 sequences was constructed to train attention-based deep learning natural language processing model to classify sequences. A patient level prediction model based on logistic regression was developed using the sequence level classifier. The deep learning system was tested using 343 patient EHRs and achieved 93% accuracy and AUC = 0.98 to identify patients who had no earlier diagnosis, dementia-related diagnosis code, or dementia-related medications in their EHR. These patients would have otherwise gone undetected or detected too late. The EHR screening pipeline was deployed in NeuraHealthNLP, a web application for automated and real-time CI screening by simply uploading EHRs in a browser. NeuraHealthNLP is cheaper, faster, more accessible, and outperforms current clinical methods including text-based analytics and machine learning approaches. It makes early diagnosis viable in regions with scarce health care services but accessible internet or cellular services.
1 Introduction

1.1 Background

Dementia related cognitive impairment (CI) is a neurodegenerative disorder in which cells of the central nervous system progressively stop working and there is no cure. Common dementia related diseases include alzheimer’s, parkinson’s, lewy body, huntington’s, vascular, frontotemporal, and more; affecting over 55 million people worldwide, growing rapidly at the rate of one new case every 3 seconds, with an estimated annual worldwide cost reaching USD 1.3 trillion and forecasted to reach 2.8 trillion by 2030.\textsuperscript{1,2,3} Dementia has been recognized as a public health problem by the World Health Organization for over a decade and the rapidly aging global population is only compounding this problem.\textsuperscript{4,5} In addition, over half of primary care physicians believe that they are not prepared for this growing problem.\textsuperscript{6}

Despite high prevalence and key implications for patients and families, dementia is underdiagnosed by clinicians and underreported by patients and families.\textsuperscript{7} With 75\% of dementia cases undiagnosed globally and up to 90\% in low- and middle-income countries, most do not have access to treatment, care and organized support that getting a formal diagnosis could provide.\textsuperscript{2} Even when a diagnosis is made, the patient has often reached moderate dementia and irreversible damage has already been done to the brain.\textsuperscript{8} It robs a person of their identity, and eventually leads to death (100\% fatality rate).\textsuperscript{9} If these cases could be diagnosed early, it could enable the patient an access to vital information and resources, including available drug and non-drug therapies for reversible symptoms, help them make important financial and legal decisions while the dementia is mild, and access benefits to manage the devastating effect on their life.\textsuperscript{10}
1.2 Research Problem

Current clinical methods to diagnose dementia utilize techniques such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) scans.¹¹ These scans are very expensive and often not available to many people. CT scans cost $3,275, MRI scans cost $3,500, and PET scans cost upwards of $5,750.¹²,¹³,¹⁴ Often, these tests need to be repeated, costing patients even more money and time, which is extremely of the essence in situations where it is likely the patient has dementia. This not only delays early detection and is cost prohibitive, but also, yields only 77% accuracy.¹⁵

Other methods such as cognitive and neuropsychological tests (NPT) evaluate a patient’s thinking ability through testing memory, reasoning, problem-solving, language skills, visual and spatial skills, and other abilities related to mental functioning. However, NPT is time-consuming, requires special training, and usually occurs during a separate appointment with a neuropsychologist, in addition to an earlier visit to a regular neurology doctor.¹⁶ For early stages of dementia, NPT tests achieve only 80% accuracy.¹⁷

A review of patient’s electronic health records (EHRs) shows that clinicians may chart symptoms of cognitive issues in unstructured notes, but they may not make a formal diagnosis by entering it as a structured International Classification of Diseases (ICD) diagnosis code in a patient’s EHR, refer to a specialist, or prescribe a medication.¹⁸,¹⁹,²⁰,²¹,²²,²³

With 89% adoption rate of EHRs, its examination can provide vital clues for early detection of CI, which is essential to ensure patients get the right care and treatment to improve clinical outcomes.²⁴,²⁵ But a review of EHRs by clinicians is manually intensive,
time-consuming and error prone. These factors combined with a lack of time or expertise, patient resistance, or limited treatment options lead to dementia being severely underdiagnosed. Artificial intelligence (AI) based tool that can effectively and efficiently analyze medical records for warning signs of CI and recommend patients to follow up with a specialist is necessary in the fight against dementia.

1.3 Research Goals

This project develops a novel, state-of-the-art automated screening pipeline for scalable and high-speed discovery of undetected CI in EHRs using deep learning and natural language processing (NLP) techniques to win the fight against undetected dementia. Specifically, my main contributions are as follows:

- I developed a machine learning and an attention-based deep learning NLP model to understand the linguistic context from complex language structures to automatically detect signs of CI in sequences (snippets) of EHR. The novel deep learning NLP model outperforms other state-of-the-art methods for automated detection because of its high dimensionality.

- I developed a patient level prediction model based on logistic regression using the sequence level classifier to predict if a patient has CI. This model outperforms other clinical methods based on dementia-related ICD codes / medications because of the high performing sequence level predictions.

- I integrate the deep learning NLP based sequence classifier and patient level predictor into a web application which can perform rapid and automated discovery of undetected CI and can be deployed at scale in medical facilities.
1.4 Related Works

Prior works have used NLP techniques to detect various diseases from EHR. (Rajkomar et al., 2018) used recurrent neural networks (long short-term memory (LSTM)) among others to predict inpatient mortality using EHR data from the University of California, San Francisco (UCSF) from 2012 to 2016, and the University of Chicago Medicine (UCM) from 2009 to 2016. (Glicksberg et al., 2018) performed phenotyping for diseases such as Attention Deficit Hyperactivity Disorder (ADHD) by clustering on word2vec embeddings from EHR of the Mount Sinai Hospital (MSH) in New York City. These studies have shown that the application of NLP techniques to EHR have improved disease detection, and that NLP techniques can be applied to dementia detection to achieve similar results. Currently, dementia detection works have utilized simple text-based analytics, which struggle in understanding the nuances of diagnosing CI. My work uses novel state-of-the-art deep learning NLP techniques, which has achieved impressive results when applied to general text due to the use of word embeddings and attention-based models (Vaswani et al., 2017; Mikolov et al., 2013; Pennington et al., 2014; Wawer et al., 2018; Devlin et al., 2018), but have had limited applications in healthcare, and have not been hitherto applied to dementia detection.

2. Unstructured Data Preparation Pipeline

A database of ≈40,000 patients consisting of 10 million EHRs was filtered to match a dementia related keyword, resulting in extraction of 279,224 EHRs from 16,428 unique patients. Further processing was done to construct a custom database of 8,656 sequences from 2,487 patients to train an attention-based deep learning NLP model using word embedding algorithms to understand the linguistic context from
complex language structures in EHR. Sequences were annotated by neurology physicians from Massachusetts General Hospital and always-patterns were developed to automate the annotation process for generating training data for the deep learning NLP model.

2.1 Dataset Description

The dataset originally consisted of ≈ 40,000 patients from the Partners BioBank Database. The Partners BioBank is a Mass General Brigham (MGB) HealthCare (formerly Partner’s Healthcare, comprising two major academic hospitals, community hospitals, and community health centers in the Boston area) initiative that houses genotype data for patients in the MGB Healthcare system. The genotype of interest for this project was the Apolipoprotein E. (APOE) genotype, which is the biggest genetic risk factor for dementia.\textsuperscript{33} The APOE genotype has 3 alleles: ε2, ε3, ε4. The ε2 allele is the rarest form of APOE and reduces the risk of developing dementia by up to 40%. ε3 is the most common allele and does not influence risk of dementia. The ε4 allele increases the risk for dementia and lowers the age of onset.\textsuperscript{34} The APOE genotype data was used to ensure that the study consisted of diverse patients.

2.2 Dataset Filtering and Keyword Selection

The first step was to filter for patients who were older than 60 years of age (as of July 13, 2021) and who had an allele of the APOE genotype available in the BioBank, which resulted in an initial selection of ≈ 20K patients. We then developed a list of 18 dementia-related keywords (available in Table 1).
| Number | Keyword           | Match Count | Number | Keyword       | Match Count |
|--------|------------------|-------------|--------|---------------|-------------|
| 1      | Memory           | 109218      | 10     | Neurocognitive| 7711        |
| 2      | Cognition        | 87655       | 11     | MCI           | 3889        |
| 3      | Dementia         | 51034       | 12     | Amnesia       | 3695        |
| 4      | Cerebral         | 45886       | 13     | AD            | 2673        |
| 5      | Cerebrovascular  | 36370       | 14     | Lewy          | 2561        |
| 6      | Cerebellar       | 26863       | 15     | MMSE          | 2134        |
| 7      | Cognitive Impairment | 20267     | 16     | LBD           | 224         |
| 8      | Alzheimer        | 20581       | 17     | Corticobasal  | 147         |
| 9      | MOCA             | 9767        | 18     | Picks         | 41          |

Table 1: Keyword List indicative of Cognitive Impairment

These keywords were based on careful literature review of established methods for identifying patients with dementia using EHR. Expert neurologists at Massachusetts General Hospital (MGH) ensured that these keywords comprehensively capture evidence of CI, and that it would be exceedingly rare to describe CI (or the lack thereof) in EHR notes without using one of these keywords. Note that the presence of any of these keywords does not always indicate that the patient has CI.

We used this list of keywords to further prune our dataset to only include patients who had at least one clinician note with a dementia-related keyword, which resulted in a final dataset consisting of 16,428 unique patients. Table 2 shows the demographics of the cohort of patients.

| Characteristic       | (N = 16,428) |
|----------------------|--------------|
| 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)|

**APOE Genotype, n(%)**

| APOE Genotype |       |       |
|---------------|-------|-------|
| APOE ε2       | 2028  | (12.3)|
| APOE ε3       | 10177 | (62.0)|
| APOE ε4       | 4223  | (25.7)|

**Average Specialty Visits (SD)**

| Parameters         |       |       |
|--------------------|-------|-------|
| Average Specialty  | 1.67  | (4.6) |
| Encounters (SD)    | 5.25  | (5.63)|

Table 2: Dataset Demographics

**2.3 Sequence Construction and Extraction from Dataset**

For each patient in our dataset, we extracted unstructured clinician notes and identified matches with the dementia-related keywords (Table 1). Sequences were extracted from the note text spanning each of these matches. The below preprocessing steps were followed to produce sequences that could be easily interpreted by humans and the models:

1. Removed all empty lines and multiple white spaces.
2. Computed context windows of 100 characters before start of keyword match and 100 characters after.
3. For notes that had multiple keyword matches, the context windows were merged.
4. Constructed sequences by extracting note text from computed context windows.
5. Tokenized extracted sequences into BERT tokens (where 1 token = 1 word) and extended context windows for all sequences that were less than 512 tokens.

6. Cleaned up spaces and other special characters through regular expression substitutions to make the sequence more readable for human annotators.

Our final cohort of 16,428 patients had 279,224 sequences with dementia-related keywords in total. Table 3 shows the demographics of the sequences.

| Characteristic                        | (N = 279,224) |
|---------------------------------------|---------------|
| Average Sequence Length (SD)          | 910 (485)     |
| Average Keyword Count (SD)            | 1.97 (1.62)   |
| % Sequences with 1 Keyword Match      | 54.5          |
| % Sequences with 2 Keyword Matches    | 24.2          |
| % Sequences with 3 Keyword Matches    | 9.30          |
| % Sequences with 4+ Keyword Matches   | 12.0          |

Table 3: Sequence Demographics

2.4 Sequence Annotation

The selected sequences were annotated (labelled) by experts for indication of CI. In this context, CI was defined as evidence of either Mild CI (MCI), where one cognitive domain is involved, or dementia, where more than one cognitive domain is involved, and activities of daily living are affected. Concern from the family of the patient or the patient was not considered as CI. Each sequence was labeled with one of three classes:

1. Positive, i.e., patient has CI
2. Negative, i.e., patient does not have CI
3. Neither, i.e., sequence does not contain information pertinent to a patient’s cognitive status
Sequences were annotated using two approaches: one, manually by neurology physicians from Massachusetts General Hospital and second, by utilizing an always-pattern scheme to automate the annotation process. An always pattern is defined as a phrase or regex expression that in any context indicates the phrase will be labeled with a particular class (i.e. positive, negative, or neither). Once an always pattern is defined, all other sequences that match the pattern are automatically labeled with that always pattern's class. Figure 2 contains examples of sequences and always patterns for all three classes.

| Positive Sequences | Positive Always Patterns |
|--------------------|-------------------------|
| 1. Patient MOCA is 22/30. | 1. (?)[0-9][0-9]*s*[0-9][0-9]s*[0-9][0-9]s*[0-9][0-9]| |
| 2. Patient with past medical history of dementia. | 2. (?)[0-9][0-9]s*[0-9][0-9]s*[0-9][0-9]s*[0-9][0-9](dementia) | |

| Negative Sequences | Negative Always Patterns |
|--------------------|-------------------------|
| 1. Patient memory is intact. | 1. (?)[0-9][0-9]s*[0-9][0-9]s*[0-9][0-9]s*[0-9][0-9]| |
| 2. No memory concerns. | 2. (?)[0-9][0-9]s*[0-9][0-9]s*[0-9][0-9]s*[0-9][0-9](concerns) | |

| Neither Sequences | Neither Always Patterns |
|--------------------|-------------------------|
| 1. History: Father has Alzheimer’s Disease | 1. (?)[0-9][0-9]s*[0-9][0-9]s*[0-9][0-9]s*[0-9][0-9](disease) |
| 2. Patient attends anticoagulation therapy daily. | 2. (?)[0-9][0-9]s*[0-9][0-9]s*[0-9][0-9]s*[0-9][0-9](anticoagulation) |

Figure 2: Example Sequence and Always Patterns

Both, manual and automated annotations were carried out through a web-based annotation tool. The tool was constructed using Python-based open-source Django web development framework with a SQLite database. Data Models were established for the selected sequences, clinician notes, user account creation and authentication, and sequence assignment to individual or multiple annotator accounts. User interface (UI) templates (i.e., pages) were created to present the data in an integrated fashion for annotation, as shown in Appendix A.

The final dataset of 8,656 annotated sequences from 2,487 unique patients was split between train (90%) and holdout test (10%) sets. Validation datasets were split from the
train set using techniques described in the Section 3 (Sequence Classification Methods). The train, validation, and test sets were stratified across label and proportion of sequences annotated manually and through always patterns. No patients were featured in multiple sets.

3. Sequence Classification Methods

Two models were used to perform the task of sequence classification: a baseline term frequency-inverse document frequency (TF-IDF) machine learning model and a fine-tuned attention-based state-of-the-art deep learning model.

3.1 TF-IDF Machine Learning Model

TF-IDF vectorization was performed on the annotated sequences and feature selection was based on a term’s Pearson correlation coefficient (PCC) with the CI label.\textsuperscript{38,39} L1 Regularized Logistic Regression was applied with the annotated CI labels.\textsuperscript{40} 10-Fold Cross Validation was used to identify optimal hyperparameters.\textsuperscript{41} Figure 3 depicts the procedure for the machine learning model.

![Figure 3: Machine Learning Model Overview Diagram](image)

First, annotated sequences are converted into TF-IDF vectors. This is done through TF-IDF vectorization, which converts the text into a vector by counting the occurrence of words in a document. TF-IDF vectors take this a step further as they contain insights about the less relevant and more relevant words in a document, which is of great significance.
3.1.1 TF-IDF Computations

For a particular word, the TF-IDF value is the product of the term frequency (TF) and inverse document frequency (IDF). TF is of the frequency of a word \( w \) in a document \( d \). \( \text{TF}(w, d) = \frac{\text{frequency of } w \text{ in } d}{\text{total # of words in } d} \)

IDF measures the importance of each word, and provides a weightage based on the frequency of a particular word \( w \) in the corpus (collection of documents) \( c \). \( \text{IDF}(w, c) = \ln \left( \frac{\text{total # of documents in } c}{\text{# of documents containing } w} \right) \). Therefore, \( \text{TF} - \text{IDF}(w, d, c) = \text{TF}(w, d) \times \text{IDF}(w, c) \). TF-IDF creates a vector for each document in the corpus that has dimensions 1 * length of vocabulary (total words in corpus).

A limitation of TF-IDF is that it can be computationally expensive for large vocabularies. To combat this, we eliminated word features that were deemed to have little correlation to the cognitive impairment label using the PCC. PCC is the measure of correlation between two sets of data and ranges from 0 - 1. It is defined as the ratio between the covariance of two variables and the product of their standard deviations, and is defined below: For paired data \{ (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) \}, \( \text{PCC}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \).

Once the PCC was computed for each of the word features, a L1 Regularized Logistic Regression model was regressed on the TF-IDF vectors. Logistic Regression is a regression technique is an adaption of linear regression to create a classification model. It is defined as: \( h_\theta(x) = \frac{1}{1 + e^{-\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + ... + \theta_n x_n}} \) with a cost function defined as

\[
J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} y^i \log \left( h_\theta \left( x^i + (1 - y^i) \right) \right) \log \left( 1 - h_\theta \left( x^i \right) \right)
\]
plus a \( \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2 \) regularization parameter to reduce overfitting. 10-fold cross validation was used to tune for a PCC threshold that removed the optimal amount of word features to maximize the performance metrics of a L1 Regularized logistic regression model that was regressed on the TF-IDF vectors where no element had a PCC score less than the arbitrary threshold.

In the 10-fold cross validation loop, the training data (7,487 annotated sequences) was split into 10 subsets. A holdout procedure then commenced for 10 iterations, where for each iteration 1 of the subsets was chosen as a validation set while the other 9 formed the training set. The validation set was used to tune the hyperparameters for the L1 Regularized logistic regression model, specifically the \( \lambda \) value, which controls the impact of the regularization parameter on the cost function, and PCC threshold. The model trained on the 9 subsets and evaluated itself on the validation set, adjusting the aforementioned hyperparameters.

### 3.2 Attention-based Deep Learning ClinicalBERT Model

The architecture of a pre-trained language model called ClinicalBERT was used.\(^4^2\) The model was programmed using the implementation available in the Huggingface Transformers and Simpletransformers packages.\(^4^3, 4^4\) After text preprocessing, input texts were tokenized with the default tokenizer and converted to embeddings. The model was initialized with pre-trained parameters and later fine-tuned on our labeled training set. We used the Adam Optimizer and Optuna was used to perform a 20-trial study and tune the learning rate, Adam epsilon, and the number of train epochs on the held-out validation set.\(^4^5, 4^6\) An early stopping rule was used to prevent overfitting by ensuring that training stopped if the loss did not change substantially over 3 epochs. Figure 4 depicts the procedure for the deep learning model.
3.2.1 ClinicalBERT Model Architecture

Figure 5 shows the proposed model architecture for the ClinicalBERT model. ClinicalBERT has a transformer architecture, which enables models to process text in a bidirectional manner, from start to finish and from finish to start. This design overcomes the limits of previous state-of-the-art models such as Long short-term memory (LSTM) models, which could only process text from start to finish. The scaled dot-product attention and multi-head attention layers capture the relationships between each word in a sequence with every other word, which allows ClinicalBERT to achieve higher performance levels than the TF-IDF approach.

Source: 1706.03762.pdf (arxiv.org)
3.2.2 ClinicalBERT Model Computations

The input into the scaled dot-production attention layer consists of queries (Q) and keys (K) of dimension $d_k$, and values (V) of dimension $d_v$. The dot products of the query with all keys are computed, divided by $\sqrt{d_k}$, and followed by the application of the softmax ($\sigma$) function to obtain the weights on the values. \[ \sigma(\hat{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}, \]

\[ \text{attention}(Q, K, Z) = \sigma(\frac{QK^T}{\sqrt{d_k}})V. \]

The Multi-head Attention layer is a module for attention mechanisms which runs through an attention mechanism several times in parallel. The independent attention outputs are then concatenated and linearly transformed into the expected dimension. Multiple attention heads allow for attending to parts of the sequence differently. \[ \text{MultiHead}(Q, K, V) = [\text{head}_0, \text{head}_1, \text{head}_2, \ldots, \text{head}_h]W_0 \text{ where } \text{head}_i = \text{attention}(QW_i^Q, KW_i^K, VW_i^V), W = \text{learnable parameter matrices}. \] For more details regarding the transformer architecture, see (Vaswani et. al, 2017).

ClinicalBERT was initialized from the transformer model and trained on the MIMIC II database containing EHR records from ICU patients. This training allowed the model to develop an understanding on clinical terminology. Since the attention mechanism in the Transformer allows ClinicalBERT to model any downstream task, we fine-tuned it on our training set so that it could develop an understanding of terminology relevant to CI in particular.

The held-out validation set was used to tune the hyperparameters learning rate, Adam epsilon, and the number of train epochs. To tune these hyperparameters, the Bayesian hyperparameter optimization library Optuna was used. Optuna employs a pruning strategy that constantly checks for algorithm performance during training and terminates a trial if a combination of hyperparameters does not yield good results, and
a sampling algorithm for selecting the best hyperparameter combination, concentrating on hyperparameters which yield good results and ignoring those that do not. I created a 20 trial Optuna study designed to maximize accuracy with Tree-structured Parzen Estimator (TPE) sampling algorithm. The learning rate and adam epilson were tuned from ranges of $[1e^{-8}, 1e^{-4}]$, and number of training epochs was tuned between 1 and 3. ClinicalBERT model training took place on a Linux cloud cluster on two 16GB NVIDIA Graphic Processing Units (GPUs) over the course of 25 hours.

4. Sequence Classification Results

We evaluated each model based on sequence level class assignments. Model performance for each model on the held-out test set are shown in Table 4. To compute each metric, we used the threshold that maximized accuracy.

4.1 Comparison between TF-IDF and ClinicalBERT

| Model       | AUC  | Accuracy | Sensitivity | Specificity | Weighted F1 |
|-------------|------|----------|-------------|-------------|-------------|
| TF-IDF      | 0.94 | 0.85     | 0.83        | 0.92        | 0.84        |
| ClinicalBERT| 0.98 | 0.93     | 0.91        | 0.96        | 0.93        |

Table 4: Model Performance

The TF-IDF model achieved an AUC of 0.94, accuracy of 0.85, sensitivity of 0.83, specificity of 0.92, and weighted F1 of 0.84. Hyperparameters were selected by finding the combination of the $\lambda$ value and PCC threshold that maximized the average accuracy over the 10 CV folds. The optimal $\lambda$ value and PCC threshold were 10 and 0.01, respectively. Word features related to memory and CI had the highest coefficients in the model. The 20 words with the highest correlation coefficients using TF-IDF word vectorization are shown in Table 5. Figure 6 shows the one vs. all ROC curve for the TF-IDF model.
| Number | Word          | Correlation | Number | Word       | Correlation |
|--------|---------------|-------------|--------|------------|-------------|
| 1      | Intact        | 0.5573      | 11     | Homicidal  | 0.3610      |
| 2      | Oriented      | 0.4233      | 12     | Observation| 0.3602      |
| 3      | Concentration | 0.4157      | 13     | Knowledge  | 0.3598      |
| 4      | Orientation   | 0.4029      | 14     | Insight    | 0.3561      |
| 5      | Perceptions   | 0.3959      | 15     | Associations| 0.3538     |
| 6      | Sensorium     | 0.3954      | 16     | Abstract   | 0.3524      |
| 7      | Judgement     | 0.3851      | 17     | Suicidal   | 0.3514      |
| 8      | Fund          | 0.3733      | 18     | Attention  | 0.3433      |
| 9      | Experiences   | 0.3693      | 19     | Content    | 0.3396      |
| 10     | Ideation      | 0.3612      | 20     | Thought    | 0.3385      |

Table 5: Top 20 TF-IDF Word Features and their PCC

![TF-IDF ROC Curve](image)

Figure 6: TF-IDF ROC Curve

While TF-IDF was able to identify the presence of a keyword or always pattern in a sequence, it was unable to leverage the context around each keyword match. The context of the keywords and the agents within the sentence often contained useful
information regarding a patient's cognitive status. For example, the sequence "Patient is caregiver for wife who has dementia" has the keyword dementia, but it does not pertain to the patient's cognitive diagnosis but instead their wife's. This led the baseline TF-IDF model to incorrectly predict sequences as evidence of cognitive impairment, resulting in a large count of false positives, as shown by the precision matrix in Figure 7.

![TF-IDF Precision Matrix](image)

Figure 7: TF-IDF Precision Matrix

ClinicalBERT, with its more complex architecture as discussed in Section 3.2, was able to leverage the context of the keyword matches within the sequences and overcome the aforementioned issues. This was evident in the results, as the fine-tuned ClinicalBERT model achieved an AUC of 0.98 and substantially improved accuracy to 0.93 as well as specificity of 0.96, sensitivity of 0.91, and weighted F1 of 0.93. The precision matrix and ROC curve for ClinicalBERT can be found in Figures 8 and 9, respectively.
Additionally, when using a small dataset of manually annotated sequences (N = 150) which did not match an always pattern, ClinicalBERT was able to accurately discriminate between all three classes (see Figure 10).
5. Patient Level Predictor

To make this project fully applicable in a clinical setting, it would need to return an overall prediction regarding whether the patient had CI or not. However, annotators had only annotated whether a particular sequence showed signs of a patient having CI in the BioBank dataset. To get ground truth labels on the patient level, we utilized another in-house dataset curated in (Hong, 2020). In (Hong, 2020), each patient’s EHR record between 01/01/2018 – 12/31/2018 was reviewed by an expert clinician (neurologist, psychiatrist, or geriatric psychiatrist) to label patients with presence or absence of any cognitive impairment. After running the Data Preparation Pipeline described in section 2.2, a gold-standard patient level dataset contained 46,650 sequences from 921 unique patients was created. These sequences / patients were not part of the sequence level dataset that ClinicalBERT was trained and evaluated on.
ClinicalBERT was then applied to this dataset to generate the sequence level predictions. Using these predictions, four structured features were generated per patient: percent sequences predicted positive, percent sequences predicted negative, percent sequences predicted neither, and total sequence count. Data was split from train (90%) and holdout test (10%) sets.

Validation datasets were split from the train set using techniques described in the Section 3. A L1 Regularized logistic regression model was regressed on these features with the patient level CI label as the outcome. To tune hyperparameters, specifically the $\lambda$ value, a 10-fold cross validation loop was used.

5.1 Comparison with Clinical Methods

When evaluated on the held-out test set, the patient level model achieved an accuracy of 0.85, AUC of 0.89, Sensitivity of 0.88, Specificity of 0.88, and weighted F1 of 0.87. The precision matrix and ROC curve for the patient level model can be found in Figures 11 and 12, respectively.

These results are a major improvement over current clinical methods, which are only able to achieve 77% accuracy.$^{15}$ As shown, the patient level model was able to identify a significant proportion of patients that went undetected by current clinical methods, highlighting the utility of such a tool in a clinical setting.
6. System Integration with Web Application for Real-Time diagnosis of CI by Clinicians

The NeuraHealthNLP pipeline was then deployed as a web application designed for real-world use by both doctors and patients. The Django web framework was once again used to construct the backend and frontend components of the application. To use the application, a user only needs to upload a file containing their EHR notes. The application first runs the sequence extraction pipeline to extract sequences of
interest, which are then feed to the fine-tuned ClinicalBERT model, which generates sequence level predictions. These predictions are converted to features (as described in Section 4.2) and feed into the patient level model for overall prediction. Following this, the user receives an exhaustive report detailing their cognitive impairment status and a recommendation from the model on future next steps. The report includes the overall patient level CI probability, the percent of sequences with a high probability of CI (>50%), a scatterplot showing the correlation between their sequences and probability of CI, and model predictions / three-class probabilities for each sequence, which is displayed with its keywords highlighted for convenience.

Figure 13 shows the process of using the NeuraHealthNLP web application. The web application is able to process a patient’s EHR notes and return detailed results within 5 seconds, which is considerably faster than any manual analysis by a team of neurological experts.
Figure 13: NeuraHealthNLP Web Application
7 Conclusion and Future Work

I applied NLP algorithms to identify patients with cognitive impairment in EHR and compared a baseline TF-IDF model with an attention based deep learning model on performance of sequence level class assignment predictions. The deep learning model’s performance was significantly better than the TF-IDF model as it was able to fully leverage the context of sequences. My work illustrates the need of more complex, expressive language models for the nuanced task of detecting dementia in electronic health records. I then created a patient level machine learning model that used the deep learning model to make predictions that patient’s cognitive status as a whole.

My work was able to outperform current clinical methods by ≈ 10%, in addition to being cheaper and faster. This work can help address the underdiagnosis of dementia and alert primary care physicians to do a formal cognitive evaluation or refer to specialists. Such a tool can be used to facilitate real-world data research to generate cohorts for dementia research studies to identify risk and protective factors of dementia as well as recruit patients into observational studies or clinical trials.

This work also highlights the future potential that deep learning NLP techniques have when used to analyze EHR. Having sequence labels that are generated by manual review are extremely hard to come by, as they require a highly dedicated team of expert neurologists and hundreds of man hours. To combat this, I devised always patterns as a method to quickly accumulate annotated data. And despite being trained on data that was not manually labeled, ClinicalBERT was able to generalize and make accurate predictions on sequences that were manually labeled.

To further improve upon this work, I plan to gather manual labels for 6000 sequences that do not match an always pattern and up sample sequences from notes that...
do not contain any keyword matches to further improve the generalizability of my model. I also am in the process of implementing an active learning loop that will be used to pick particular patients and sequences by using entropy scores to label uncertain cases and UMAP clustering of ClinicalBERT word embeddings on the sequences of the N = 13,941 patients not included in the training, validation, or test sets. This active learning loop will be used to label the ≈ 45K patients in the MGH Accountable Care Organization (ACO) system. All clinical adjudication will be performed by a team of 10 expert neurologists.

In conclusion, the establishment of an automated screening pipeline to perform early detection of CI in EHR through my project provides a tool that significantly outperforms current clinical methods and can allow for the initiation of appropriate treatment to prevent complications related to dementia and save lives. Additionally, this pipeline can be easily repurposed to perform early detection of other diseases, even those unrelated to the nervous system. Therefore, this project opens an entire new avenue of prevention for many dangerous diseases and its implementation in healthcare systems can yield extremely impactful results.
Figure 14: Web App for Sequence Annotations
9 References

1 Robinson, L., Tang, E., & Taylor, J. P. (2015). Dementia: timely diagnosis and early intervention. Bmj, 350.
2 Alzheimer’s Disease International. (2020). ADI - Dementia statistics. Alzheimer’s Disease International. https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/
3 Alzheimer’s Association. (2016). What Is Dementia? Alzheimer’s Disease and Dementia; Alzheimer’s Association. https://www.alz.org/alzheimers-dementia/what-is-dementia
4 Dementia: a public health priority. (n.d.). Www.who.int. https://www.who.int/publications/i/item/dementia-a-public-health-priority
5 Prince, M., Albanese, E., Guerchet, M., & Prina, M. (2014). World Alzheimer Report 2014: Dementia and risk reduction: An analysis of protective and modifiable risk factors.
6 2020 Alzheimer’s disease facts and figures. Alzheimers Dement. 2020. Epub 2020/03/12. doi: 10.1002/alz.12068. PMID: 32157811.
7 Amjad, H., Roth, D. L., Sheehan, O. C., Lyketsos, C. G., Wolff, J. L., & Samus, Q. M. (2018). Underdiagnosis of dementia: an observational study of patterns in diagnosis and awareness in US older adults. Journal of general internal medicine, 33(7), 1131-1138.
8 DementiaCareCentral.com. (2020, April 24). Stages of alzheimer’s & dementia: Durations & scales used to measure progression (GDS, Fast & Cdr). Dementia Care Central. Retrieved November 26, 2021, from https://www.dementiacarecentral.com/aboutdementia/facts/stages/.
9 Penn Medical Ethics and Health Policy | The harrowing new reality for Alzheimer’s patients. (n.d.). Medicalethicshealthpolicy.med.upenn.edu. Retrieved January 12, 2022, from https://medicalethicshealthpolicy.med.upenn.edu/in-the-news/the-harrowing-new-reality-for-alzheimer-s-patients
10 Benefits of Early Diagnosis. (n.d.). Wisconsin Alzheimer’s Institute. Retrieved January 12, 2022, from https://wai.wisc.edu/benefits-of-early-diagnosis/
11 Khachaturian, Z. S. (1985). Diagnosis of Alzheimer’s disease. Archives of neurology, 42(11), 1097-1105
12 How much does a CT scan cost? American Health Imaging. (2018, December 21). Retrieved December 29, 2021, from https://americanhealthimaging.com/blog/how-much-does-a-ct-scan-cost/
13 Deleon, M. (n.d.). How much does an MRI cost? Bankrate. Retrieved December 29, 2021, from https://www.bankrate.com/personal-finance/how-much-does-an-mri-cost/#:~:text=In%20general%2C%20MRIs%20range%20in,detect%20and%20diagnose%20neurological%20conditions.
14 Poslunsky, C. (2018, July 31). How much should your PET scan cost? New Choice Health Blog. Retrieved December 29, 2021, from https://www.newchoicehealth.com/pet-scan/cost#:~:text=The%20average%20PET%20scan%20cost,or%20an%20outpatient%20surgery%20center.
15 Sabbagh, M. N., Lue, L.-F., Fayard, D., & Shi, J. (2017). Increasing Precision of Clinical Diagnosis of Alzheimer’s Disease Using a Combined Algorithm Incorporating Clinical and Novel Biomarker Data. Neurology and Therapy, 6(Suppl 1), 83–95. https://doi.org/10.1007/s40120-017-0069-5
16 Cognitive and neuropsychological tests. (n.d.). Stanfordhealthcare.org. https://stanfordhealthcare.org/medical-conditions/brain-and-nerves/dementia/diagnosis/cognitive-neuropsychological-tests.html

17 Jacova, C., Kertesz, A., Blair, M., Fisk, J. D., & Feldman, H. H. (2007). Neuropsychological testing and assessment for dementia. Alzheimer’s & Dementia, 3(4), 299–317. https://doi.org/10.1016/j.j.alz.2007.07.011

18 Boustani M, Callahan CM, Unverzagt FW, Austrom MG, Perkins AJ, Fultz BA, Hui SL, Hendrie HC. Implementing a screening and diagnosis program for dementia in primary care. J Gen Intern Med. 2005;20(7):572-7. Epub 2005/07/30. doi: 10.1111/j.1525-1497.2005.0126.x. PMID: 16050849; PMCID: PMC1490164.

19 Yarnall KS, Pollak KI, Ostbye T, Krause KM, Michener JL. Primary care: is there enough time for prevention? Am J Public Health. 2003;93(4):635-41. Epub 2003/03/28. doi: 10.2105/ajph.93.4.635. PMID: 12660210; PMCID: PMC1447803.

20 Association As. 2019 Alzheimer’s disease facts and figures. Alzheimers Dement. 2019;15(3):321-87.

21 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. PMID: 19568149; PMCID: PMC2787842.

22 Boustani M, Perkins AJ, Fox C, Unverzagt F, Austrom MG, Fultz B, Hui S, Callahan CM, Hendrie HC. Who refuses the diagnostic assessment for dementia in primary care? Int J Geriatr Psychiatry. 2006;21(6):556-63. Epub 2006/06/20. doi: 10.1002/gps.1524. PMID: 16783796.

23 Fowler NR, Frame A, Perkins AJ, Gao S, Watson DP, Monahan P, Boustani MA. Traits of patients who screen positive for dementia and refuse diagnostic assessment. Alzheimers Dement (Amst). 2015;1(2):236-41. Epub 2015/08/11. doi: 10.1016/j.dadm.2015.01.002. PMID: 26258162; PMCID: PMC4527161.

24 Why early diagnosis of dementia is important. Social Care Institute for Excellence. (2020). Retrieved November 26, 2021, from https://www.scie.org.uk/dementia/symptoms/diagnosis/early-diagnosis.asp.

25 Future of Electronic Medical Records | EMR Trends For 2020. (2019, November 10). Selecthub.com. https://www.selecthub.com/medical-software/emr/electronic-medical-records-future-emr-trends/.

26 Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1(1), 1-10.

27 Glicsberg, B. S., Miotto, R., Johnson, K. W., Shameer, K., Li, L., Chen, R., & Dudley, J. T. (2018). Automated disease cohort selection using word embeddings from electronic health records. In PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium (pp. 145-156).

28 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

29 Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

30 Pennington, J., Socher, R., & Manning, C. D. (2014, October). GloVe: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

31 Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., & Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. Psychiatry Research, 304, 114135.
32 Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
33 Mahley, R. W., & Rall Jr, S. C. (2000). Apolipoprotein E: far more than a lipid transport protein. Annual review of genomics and human genetics, 1(1), 507-537.
34 Mahley, R. W., & Rall Jr, S. C. (2000). Apolipoprotein E: far more than a lipid transport protein. Annual review of genomics and human genetics, 1(1), 507-537.
35 Gilmore-Bykovskyi, A. L., Block, L. M., Walljasper, L., Hill, N., Gleason, C., & Shah, M. N. (2018). Unstructured clinical documentation reflecting cognitive and behavioral dysfunction: toward an EHR-based phenotype for cognitive impairment. Journal of the American Medical Informatics Association, 25(9), 1206-1212.
36 Reuben, D. B., Hack Barth, A. S., Wenger, N. S., Tan, Z. S., & Jennings, L. A. (2017). An automated approach to identifying patients with dementia using electronic medical records. Journal of the American Geriatrics Society, 65(3), 658-659.
37 Amra, S., O’Horo, J. C., Singh, T. D., Wilson, G. A., Kashyap, R., Petersen, R., ... & Gajic, O. (2017). Derivation and validation of the automated search algorithms to identify cognitive impairment and dementia in electronic health records. Journal of critical care, 37, 202-205.
38 Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, No. 1, pp. 29-48).
39 Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In Noise reduction in speech processing (pp. 1-4). Springer, Berlin, Heidelberg.
40 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.
41 Refaelzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. Encyclopedia of database systems, 5, 532-538.
42 Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.
43 Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface’s transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.
44 Rajapakse, T. (2019). Simple transformers.
45 Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
46 Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623-2631).
47 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
48 Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. Advances in neural information processing systems, 24.
49 Hong, Z., Magdano, C. G., Sheu, Y. H., Mohite, P., Noori, A., Ye, E. M., ... & Das, S. (2020). Natural Language Processing to Detect Cognitive Concerns in Electronic Health Records Using Deep Learning. arXiv preprint arXiv:2011.06489.
50 McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.