Abstract

Because of the subtle onset with symptoms closely resembling that of normal aging, dementia related cognitive impairment is difficult to detect by health care professionals. With only one in four patients getting diagnosed, dementia's underdiagnosis causes a significant public health concern, as millions are left behind without the necessary care and support for their chronic condition. Information relevant to dementia related cognitive impairment is often found in the electronic health records (EHR), but a manual review by physicians is time consuming and error-prone.

In my research, I developed natural language processing (NLP) models to create an automated EHR scanning pipeline that can detect patients with dementia related cognitive impairment. The deep learning model understands the linguistic context in the EHRs and outperforms current clinical methods to identify patients who had no earlier diagnosis, dementia-related diagnosis code, or dementia-related medications in their EHR. These cases would otherwise have gone undetected or been detected too late.

To make the EHR scanning pipeline accessible and affordable, I also developed a web application that can be used on mobile or desktop devices by primary care physicians for accurate and real-time detection.

With 55 million dementia patients worldwide and growing rapidly at the rate of one new case every 3 seconds, early intervention is the key to reducing financial burden and improving clinical outcomes. My research tackles this global public health challenge and provides mechanisms for early detection of dementia and its related diseases, including Alzheimer's, Parkinson's, Lewy Body and others.

Introduction

Background Information

Because of the subtle onset with symptoms closely resembling that of normal aging, dementia related cognitive impairment is difficult to detect by health care professionals. Dementia is the most common neurodegenerative disease affecting older adults, affecting more than 55 million people worldwide and expected to affect 135 million by 2050.[1] The World Health Organization (WHO) have deemed dementia a health crisis as the global population rapidly continues to age.[2] Despite high prevalence and important implications for patients and families, dementia is underdiagnosed by clinicians and underreported by patients and families.[3] Even when a diagnosis is made, the patient has often reached moderate dementia and irreversible damage has already been done to the brain.[4] Additionally, a diagnosis may not be entered as a structured International Classification of Diseases (ICD) diagnosis code in a patient's EHR. However, a review of EHR shows that clinicians may chart symptoms of cognitive issues in unstructured notes but may not make a formal diagnosis, refer to a specialist, or prescribe a medication due to factors such as lack of time or expertise, patient resistance, and limited treatment options.[5,6,7,8,9,10]

The examination of EHR is key to guarantee early detection of dementia related cognitive impairment, which is essential to ensure patients get the right care and treatment and will lead to an improvement in clinical outcomes.[11] An early diagnosis can also help the families of the patients make important financial and legal decisions to better prepare themselves for future challenges. Yet, current clinical methods to analyze EHR are often time consuming and prone to errors. A tool that can efficiently and effectively analyze medical records for warning signs of dementia and recommend

patients for follow up with a specialist could be critical to obtaining an early diagnosis and paving the way for the fight against dementia.

Over half of primary care physicians believe that they are not prepared for this growing problem (1)

1. 2020 Alzheimer's disease facts and figures. Alzheimers Dement. 2020. Epub 2020/03/12. doi: 10.1002/alz.12068. PMID: 32157811.

Current Clinical Methods

An individual's ability to accommodate, compensate, or even deny his or her symptoms in the early stages should also be considered. The individual's family may also have noticed difficulties in communication and personality or mood changes; family concern is of particular importance.[9] Increasing frequency of patients' visits to their general practice, missed appointments, or confusion over drugs may also be warning signs. (Robinson, L., Tang, E., & Taylor, J. P. (2015). Dementia: timely diagnosis and early intervention. Bmj, 350.)

Studies have begun to build sociodemographic and medical profiles of individuals living with undiagnosed dementia [2–5, 10–12]. However, methodology has been limited in important ways. Lack of awareness of a diagnosis, based on patient/proxy report of clinician diagnosis, has been used as a surrogate for undiagnosed dementia [3, 13]. Conversely, documentation of diagnosis by providers in medical records [4, 5, 10–12] does not necessarily translate into patient or family awareness of the diagnosis. Examining either reported diagnosis or documentation provides an incomplete picture of dementia diagnosis and awareness; both elements are necessary for patients and families to understand patient cognitive and functional limitations and prognosis. No studies have examined documented dementia diagnosis together with patient or family awareness of a dementia

diagnosis. In addition, few studies [3, 13] have examined nationally representative populations. (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.)

Research Objective

The goal of this project was to develop a novel, accurate, affordable, and real-time EHR Scanning Pipeline to perform early detection of cognitive impairment that can outperform current clinical methods. The steps taken to accomplish this goal are as follows:

1. Develop a machine learning and deep learning NLP model designed to automatically detect signs of cognitive impairment in EHR. Compare both models to see which one was more proficient at the classification task.
2. Deploy a deep learning-based web application to put my pipeline in the hands of primary care physicians to perform rapid and automated detection of cognitive impairment.

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.[12] (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.[13] 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 text-based analytics. Our 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.[14,15,16,17,18]

Moreover, research studies relying on diagnosis codes and dementia-specific medications to define a dementia outcome may suffer from biases and inaccuracies in these data.
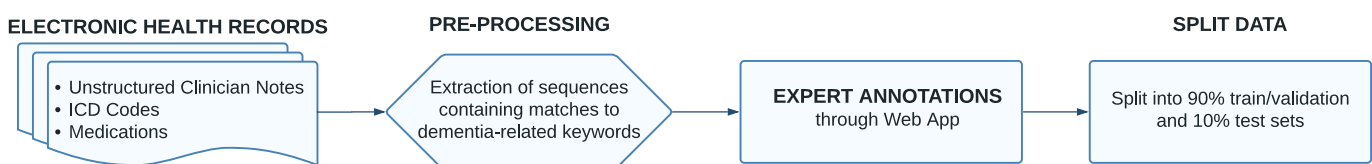
Data Preparation Pipeline

Figure 1: Data Preparation Pipeline Overview Diagram

Dataset Extraction

The dataset originally consisted of $\approx$ 40K 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.[19] 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.[20] APOE genotype data was used to ensure that the study consisted of diverse patients.

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.[21] 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.[22,23] 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 ε2	2028 (12.3)
APOE ε3	10177 (62.0)
APOE ε4	4223 (25.7)
Average Speciality Visits (SD)	1.67 (4.6)
Average PCP Encounters (SD)	5.25 (5.63)

Table 2: Dataset Demographics

Preprocessing

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. Created sequences by extracting note text from computed context windows.

5. Tokenized extracted sequences into BERT tokens (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 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

Data Labeling

We selected 5,000 diverse sequences containing at least one match to every keyword from 5,000 unique patients for labeling. These sequences were annotated by experts for indication of cognitive impairments. CI was defined as evidence of 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 cognitive impairment. 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 cognition

To expedite annotations, we utilized an "always pattern" scheme. 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 1 contains examples of sequences and always patterns for all three classes.

Positive Sequences
  1. Patient MOCA is 22/30.
  2. Patient with past medical history of dementia.
	Positive Always Patterns
	  1. (?i)\bMOCA\s*([0-9]|[12][0-5])\s*/\s*30
	  2. (?i)\bpast\s*medical\s*history\s*[^.]*(dementia)

Negative Sequences
  1. Patient memory is intact.
  2. No memory concerns.
	Negative Always Patterns
	  1. (?i)Memory.*intact
	  2. (?i)No\s*memory\s*concerns

Neither Sequences
  1. History: Father has Alzheimer's Disease
  2. Patient attends anticoagulation therapy daily.
	Neither Always Patterns
	  1. (?i)Father.*Alzheimer's\s*disease
	  2. (?i)anticoagulation

Figure 2: Example Sequence and Always Patterns

Annotations were carried out through a web-based annotation tool. The tool was constructed using the 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. 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.
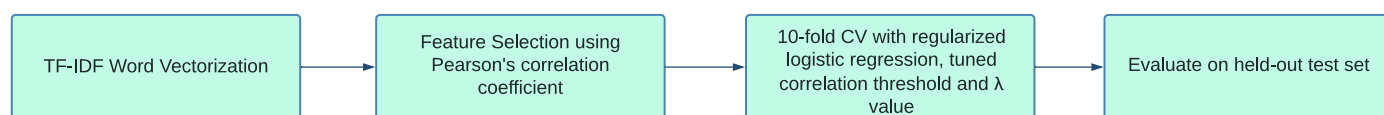
Methodology

| TF-IDF Word Vectorization | → | Feature Selection using Pearson's correlation coefficient | → | 10-fold CV with regularized logistic regression, tuned correlation threshold and λ value | → | Evaluate on held-out test set |

Figure 3: Machine Learning Model Overview Diagram

Overview

Term frequency-inverse document frequency vector (TF-IDF) was performed on the annotated sequences and feature selection was based on a term's Pearson correlation coefficient (PCC) with the cognitive impairment label.[24, 25] L1 Regularized Logistic Regression was applied with the annotated cognitive impairment labels.[26] 10-Fold Cross Validation was used to identify optimal hyperparameters.[27]

Model Procedure

Figure 3 depicts the procedure for the machine learning model. First, annotated sequences are converted into TF-IDF vectors. TF-IDF vectorization is a technique based on a Bag of Words (BoW) model, 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. 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). $TF(w, d) = \frac{frequency\ of\ w\ in\ d}{total\ \#\ of\ words\ in\ d}$

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). $IDF(w, c) = ln\ (\frac{total\ \#\ of\ documents\ in\ c}{\#\ of\ documents\ containing\ w})$. Therefore, $TFIDF(w, d, c) = TF(w, d) * IDF(w, c)$. 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* $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}, PCC(x, y) =$ $\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$.

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

$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^i\ log\left(h_\theta\left(x^i + (1 - y^i)\right)\right) log\ (1 - h_\theta(x^i))\right]$ 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 λ 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.
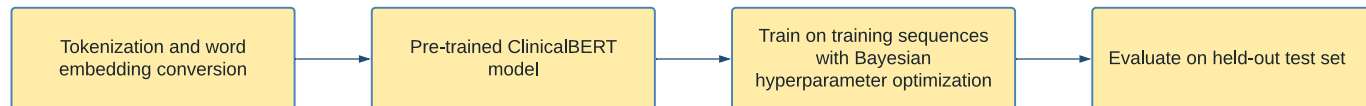
Deep Learning Model

| Tokenization and word embedding conversion | → | Pre-trained ClinicalBERT model | → | Train on training sequences with Bayesian hyperparameter optimization | → | Evaluate on held-out test set |

Figure 4: Deep Learning Model Overview Diagram

Overview

The architecture of a pre-trained language model called ClinicalBERT was used.[28] The model was programmed using the implementation available in the Huggingface Transformers and Simpletransformers packages.[29, 30] 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 epilson, and the number of train epochs on the held-out validation set.[31, 32] An early stopping rule was used to

prevent overfitting by ensuring that training stopped if the loss did not change substantially over 3 epochs.
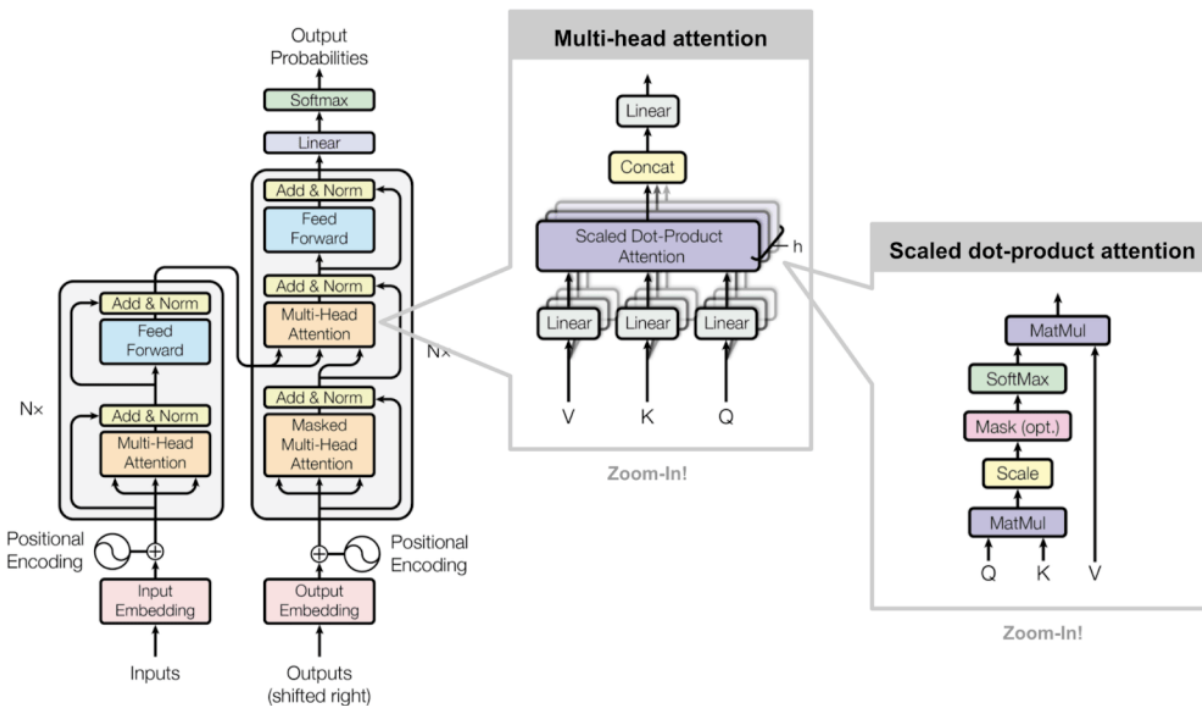
Model Procedure

Figure 5: ClincalBERT Model Architecture

Source: <u>1706.03762.pdf (arxiv.org)</u>

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.[33] This design overcomes the limits of previous models such as LSTMs, 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 allow ClinicalBERT to achieve higher performance levels than the TF-IDF approach.

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 (σ) function to obtain the weights on the values. $\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$,

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

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. $MultiHead(Q, K, V) = [head_0, head_1, head_2, ..., head_h]W_0 \ where \ head_i = attention(QW_i^Q, KW_i^K, VW_i^V), W = learnable \ parameter \ matrices$. 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.

The held-out validation set was used to tune the hyperparameters learning rate, Adam epilson, and the number of train epochs. To tune these hyperparameters, the 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 TPE sampling algorithm.[34] 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.

Results

Comparison between Results of Machine and Deep Learning Model

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

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.

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 λ value and PCC threshold that maximized the average accuracy over the 10 CV folds. The optimal λ 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
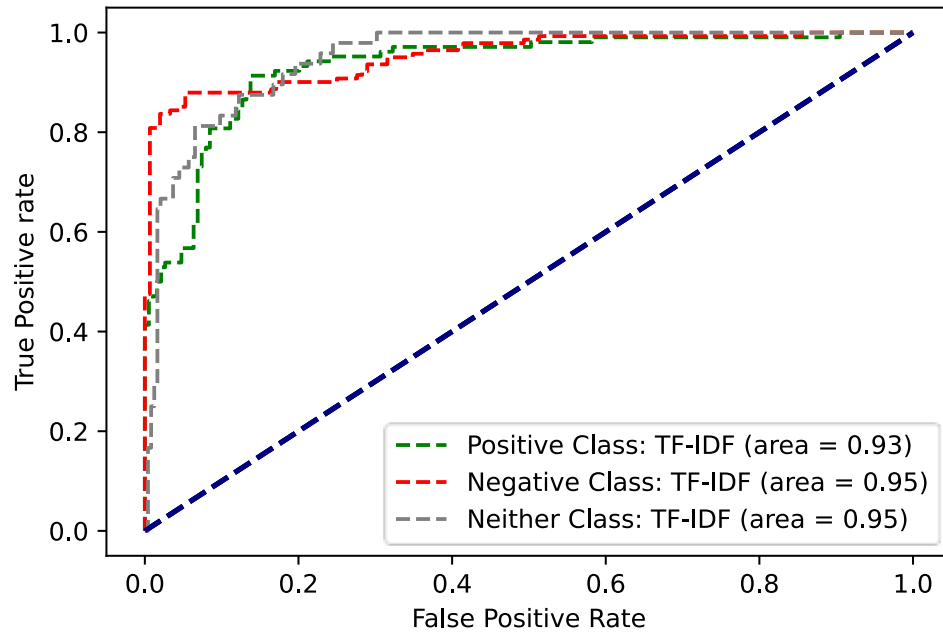
Figure 6: TF-IDF ROC Curve

While TF-IDF was able to identify the presence of a keyword or always pattern in a sequence, it was unable to the 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.
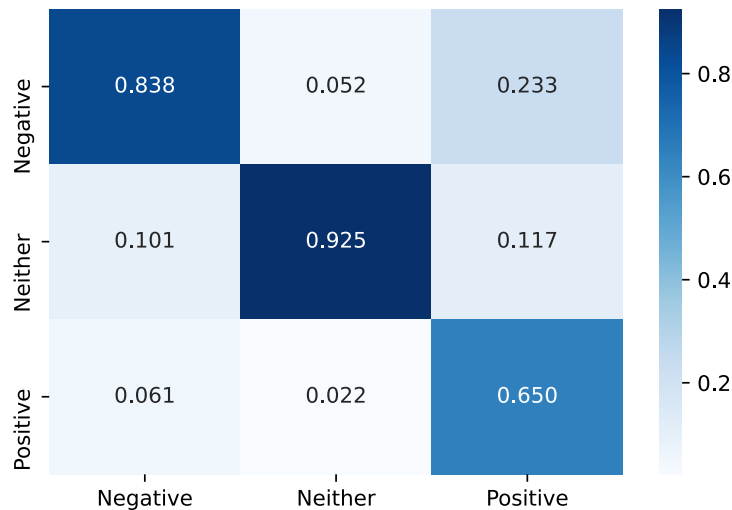

Figure 7: TF-IDF Precision Matrix

ClinicalBERT, with its more complex architecture, 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.
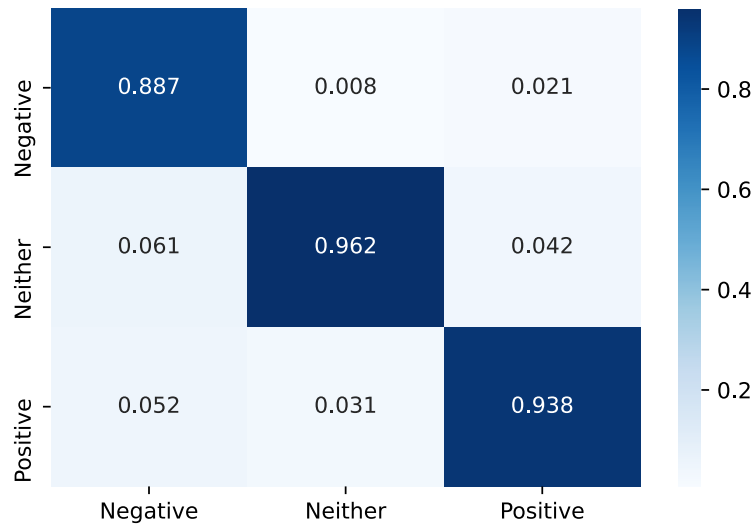
Additionally, when using a small dataset of manually annotated sequences (N =150) which did not match always pattern, ClinicalBERT was able to accurately discriminate between all three classes (see Figure 10).

Figure 9: ClinicalBERT ROC Curve

Aggregation to Patient Level

To make this model 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.[1] 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.
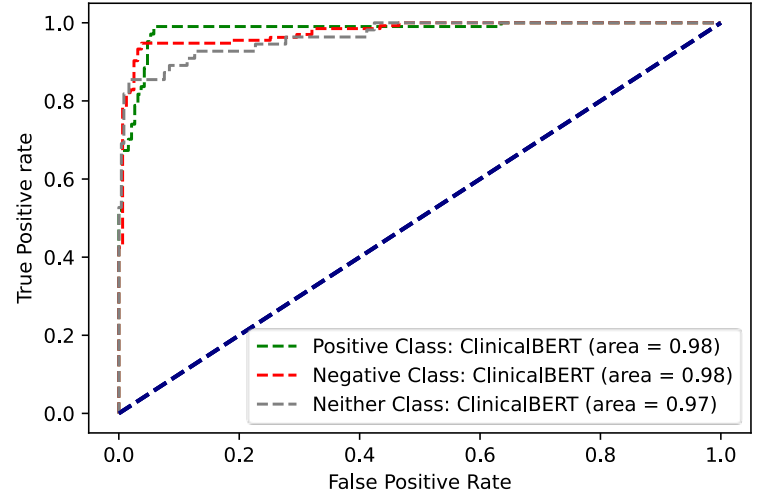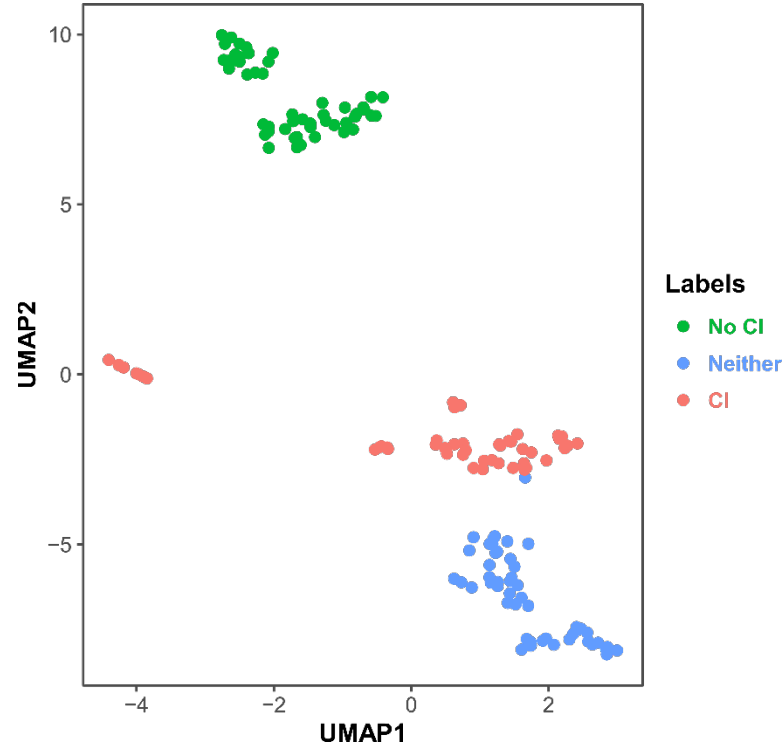

Figure 10: UMAP Clustering of ClinicalBERT Embeddings

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. loop was used. 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 λ value, a 10-fold cross validation loop was used.

When evaluated on the held-out test set, the patient level model achieved an accuracy of 0.88, AUC of 0.93, 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.

Figure 11: Patient Level Model Precision Matrix

These results are a major improvement over current clinical methods, which are only able to achieve 77% accuracy.[36] 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.
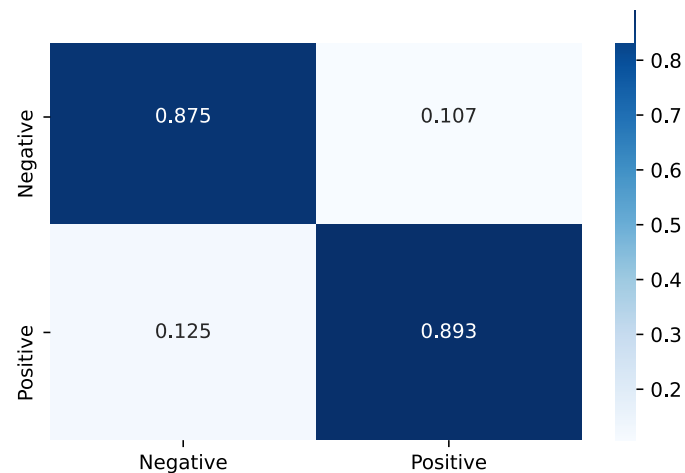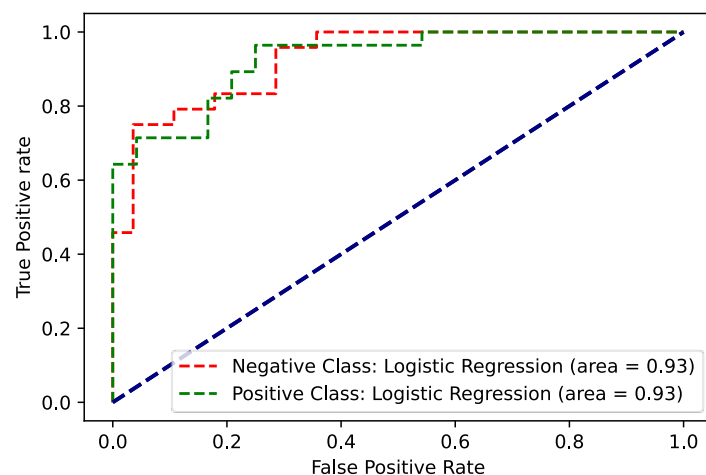

Figure 12: Patient Level Model ROC Curve

Model Deployment

Input: File with notes, 1 note per line

Output:

1. Overall yes/no patient level prediction, probability
2. # sequences created from note txt, len sequence, predictions for each individual sequence and probability
3. Highlight keyword matches that allowed for creation of sequence

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. This work can help address the underdiagnosis of dementia and alert primary care physcians to do a formal cognitive evaluation or refer to specialists. Such a tool can be used 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.

We anticipate that our tool will more efficiently and accurately detect signs of cognitive impairment, and hence can be readily used within our institution—or adapted by others. Improved dementia detection will enhance the identification of individuals for recruitment into clinical trials and studies, as well as facilitate real-world data research ranging from in-silico trials for drug repurposing to evaluating how policies and programs meet the needs of patients and caregivers at each stage of illness. In the future, the proposed tool may also be useful in clinical settings to flag patients with cognitive concerns who could benefit from an evaluation or be referred to specialist care. Cutting-edge deep learning algorithms have been applied to many real-world tasks (e.g. spam detection) but in a limited manner to healthcare problems—and never before to dementia.

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. Our work illustrates the need of more complex, expressive language models for the nuanced task of detecting dementia in electronic health records.

We used the sequence level class assignments of the deep learning model to generate patient level classes. For a few patients, we manually verified that our model can successfully identify patients with cognitive impairment who lack dementia-related ICD codes or medications in their records.

[achievement: using always patterns you can enhance dataset, train bert,. And now bert is accurately generalizing on non-always ppaterns sequences, and non-manually annotated sequences are extremely hard to come by, this project shows that we were able to work successfully without manually annotated sequences. All done without costly manual annotation. Shows very high potential]

Future Plans

1. Hand Label 6000 sequences
    a. In order to improve the generalizability of our models by labeling more sequences that do not match an always pattern, we plan to hand label 6000 sequences.
2. Active Deep Learning Loop
    a. An active learning loop 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

b. ***Active learning:*** *We will compute the classification probabilities of cognitive concern for all ~45K Medicare ACO patients who are alive, have at least 5 visit notes in the outpatient setting in the period of interest, and aged 65 years or older as of 12/31/2016. We will use both uncertainty (based on the entropy of classification probabilities) and diversity (based on similarity with labeled cases) measures for selecting the cases for labeling. Clinical adjudication will be performed by a team of 10 neurologists. We will iterate though this active learning process, oversampling minorities, until we have a total of 5000 gold-standard labels. In the next rounds of labeling, we will oversample minorities: Black and Hispanic individuals. To improve the efficiency of the adjudication process, we will compile and present digested information that highlights dementia-related phrases and important encounters such as a PCP or specialist visit with a web-based annotation tool.*

3. Upsample sequences that do not have keyword matches

4. Investigate different types of patient level models (NN) - maybe try incorporating structured features as well

Appendix A

SLAT PICS Appendix A

References

[1] Robinson, L., Tang, E., & Taylor, J. P. (2015). Dementia: timely diagnosis and early intervention. Bmj, 350.

[2] 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.

[3] 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.

[4] 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/.

[5] 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.

[6] 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.

[7] Association As. 2019 Alzheimer's disease facts and figures. Alzheimers Dement. 2019;15(3):321-87.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Glicksberg, 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).

[14] 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).

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

[16] 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).

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Reuben, D. B., Hackbarth, 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.

[23] 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.

[24] 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).

[25] Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In Noise reduction in speech processing (pp. 1-4). Springer, Berlin, Heidelberg.

[26] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

[27] Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. Encyclopedia of database systems, 5, 532-538.

[28] 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.

[29] 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.

[30] Rajapakse, T. (2019). Simple transformers.

[31] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[32] 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).

[33] 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).

[34] Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. Advances in neural information processing systems, 24.