Journal of Machine Intelligence in Healthcare

E-ISSN: XXXX-XXXX
P-ISSN: XXXX-XXXX

(Research) Article

Leveraging Fine-Tuned ClinicalBERT and Active Learning to Detect Cognitive Impairment from Unstructured EHR Notes

Bhanu Lintang Wibowo [1],* and Fredy Usman Tri Nugroho [2]

[1] Department of Informatics, Universitas Muhammadiyah Semarang, Indonesia; e-mail : lintang27717@students.unimus.ac.id
[2] Diploma in Medical Laboratory Technology, Politeknik Yakpermas Banyumas, Indonesia; e-mail : fredy.usmas.tn@politeknikyakpermas.ac.id
* Corresponding Author : lintang27717@students.unimus.ac.id

Abstract: Dementia is a progressive neurodegenerative condition that impairs cognitive function and affects over 50 million people worldwide, yet it remains substantially underdiagnosed in clinical practice. This underdiagnosis is exacerbated by the frequent absence of structured documentation, such as International Classification of Diseases (ICD) codes or medication records, in electronic health records (EHRs). To address this gap, this study proposes a transformer-based natural language processing (NLP) framework for detecting cognitive impairment (CI) directly from unstructured clinician notes. Specifically, we fine-tune ClinicalBERT, a pretrained language model adapted to clinical contexts, on a large, carefully annotated EHR dataset encompassing over 279,000 dementia-related sequences, including 8,656 expert-labeled samples. We compare the proposed model against a logistic regression baseline using term frequency–inverse document frequency (TF-IDF) features. Our findings demonstrate that ClinicalBERT significantly outperforms the baseline, achieving an AUC of 0.98 and an accuracy of 0.93, compared to 0.95 and 0.84, respectively. Furthermore, the model successfully identifies patients exhibiting cognitive impairment even in the absence of dementia-specific ICD codes or medications, addressing the critical issue of underdocumentation. We also introduce an active learning framework that strategically guides further annotation efforts by prioritizing uncertain or novel cases, thereby improving model performance with fewer additional labels. In conclusion, this research provides a robust, scalable, and automated approach for leveraging unstructured clinical narratives to enhance early detection of cognitive impairment, offering valuable implications for improving clinical decision support, patient management, and the development of dementia research cohorts.

Keywords: Electronic health records (EHR); natural language processing; cognitive impairment detection; clinicalBERT; active learning framework

Received: January 21, 2025
Revised: March 11, 2025
Accepted: April 9, 2025
Published: April 30, 2025
Curr. Ver.: April 30, 2025

1. Introduction

Dementia is a neurodegenerative disease that progressively impairs cognitive function, eventually disrupting activities of daily living and quality of life [1]. Worldwide, more than 50 million people are affected by dementia, making it one of the leading causes of disability in older adults [1], [2]. Despite its prevalence, dementia remains profoundly underdiagnosed in clinical practice, with only one in four cases formally recognized [2]. This underdiagnosis has substantial consequences: delays in patient management, missed opportunities for early intervention, and underrepresentation of affected individuals in clinical trials.

The object of this research is to develop an automated system that identifies cognitive impairment (CI), including mild cognitive impairment (MCI) and dementia, directly from unstructured clinician notes in electronic health records (EHRs). Traditional clinical documentation frequently lacks structured International Classification of Diseases (ICD) codes or medication indicators for dementia, which are essential for identifying patient

cohorts and ensuring proper care [3]. While EHRs contain rich information within narrative notes, manual review is labor-intensive and error-prone, making automated mining a promising avenue for early detection.

Previous methods for mining EHRs have relied on a variety of NLP and machine learning techniques. Classical approaches, such as TF-IDF vectorization combined with logistic regression or support vector machines, have been used for tasks like disease classification and cohort selection [4]–[6]. More recently, deep learning architectures like recurrent neural networks (RNNs) have been applied to structured and unstructured clinical data to predict outcomes such as hospital mortality or readmission [4]. In the general NLP domain, transformer models such as BERT [7] and domain-specific models like ClinicalBERT [8] have shown significant performance gains by capturing bidirectional context, outperforming previous embedding methods like word2vec [5] and GloVe [9].

Weaknesses and strengths of these methods vary. TF-IDF-based models can efficiently capture the frequency and importance of keywords, but they fail to leverage the surrounding linguistic context, often producing false positives when dementia-related terms appear in irrelevant sections (e.g., "the patient's wife has dementia") [6]. RNN-based methods can model sequences but are limited by vanishing gradients and lack of long-distance dependencies. Transformer-based models, such as BERT and ClinicalBERT, address these limitations by using self-attention mechanisms to capture both local and global context, enabling better performance on nuanced classification tasks [7], [8]. However, their application to dementia detection in real-world EHR settings remains underexplored, representing a critical research gap.

The research problem we address is: how can we design a robust, scalable NLP model that accurately identifies cognitive impairment from free-text EHR notes, particularly in cases where no structured diagnostic codes or medications are present? Existing clinical systems largely overlook this unstructured data, missing early signs that could inform diagnosis, care decisions, or recruitment into observational or interventional studies.

To address this, we propose a solution that fine-tunes ClinicalBERT, a transformer-based language model pretrained on clinical text, for the specific task of classifying cognitive impairment within annotated clinician notes. By integrating both manually labeled data and a semi-automated pattern-based labeling approach, we create a large, diverse training set. Further, we propose using an active learning loop based on uncertainty sampling and UMAP clustering [10] to iteratively expand and improve the labeled dataset, increasing model generalizability and robustness.

Our main contributions are as follows:

- We build and release a carefully annotated dataset of EHR notes containing dementia-related keywords, representing one of the largest such resources in this domain.
- We benchmark and compare two classification approaches: a logistic regression model with TF-IDF features, and a fine-tuned ClinicalBERT model, demonstrating substantial improvements in performance with the latter (AUC 0.98 vs. 0.95; accuracy 0.93 vs. 0.84).
- We show that our deep learning model can identify patients with cognitive impairment even when no dementia-related ICD codes or medications are recorded, addressing underdiagnosis in real-world clinical datasets.
- We introduce an active learning framework that guides further annotation efforts by focusing on uncertain or novel cases, enhancing the iterative improvement of model performance.

The remainder of this paper is organized as follows: Section II reviews the related work and situates our approach within the current literature; Section III details the dataset construction, preprocessing, and annotation processes; Section IV describes the proposed methodology and modeling framework; Section V presents experimental results and performance evaluations; Section VI discusses the implications, limitations, and future research directions; and Section VII concludes with final remarks.

2. Related Work

The application of natural language processing (NLP) to electronic health records (EHRs) has gained substantial attention in recent years, particularly for disease prediction, patient phenotyping, and clinical decision support. Several influential works have laid the groundwork in this domain.

Rajkomar et al. [4] pioneered the use of deep learning models such as recurrent neural networks (RNNs) to predict inpatient outcomes, including mortality, using large-scale EHR data from multiple institutions. Their results showed that deep learning models could outperform traditional statistical approaches by leveraging temporal patterns in structured data. Similarly, Glicksberg et al. [5] demonstrated the use of word2vec embeddings to automatically phenotype patients with conditions such as attention deficit hyperactivity disorder (ADHD) by clustering similar clinical narratives from EHR notes.

On the modeling side, transformer-based architectures such as BERT [7] and ClinicalBERT [8] have revolutionized the way text data is processed, especially in domains where understanding sentence-level context is crucial. BERT introduced bidirectional attention to capture rich contextual embeddings [7], and ClinicalBERT extended this by pretraining specifically on biomedical and clinical corpora, showing promising results in clinical concept extraction and note classification tasks [8].

However, despite these advances, the application of NLP in detecting cognitive impairment (CI) or dementia within EHRs remains limited. Most existing EHR-based dementia detection relies on structured data, such as ICD codes or medication records, which suffer from undercoding and misclassification [3]. Furthermore, simpler models like TF-IDF combined with logistic regression, while effective in some tasks, often fail to incorporate the nuanced linguistic context present in unstructured clinician notes, leading to false positives when keywords appear in non-patient-centered statements (e.g., "the patient's spouse has dementia") [6].

In contrast, our work directly addresses this gap by focusing on deep learning–based NLP models that leverage contextual understanding to improve classification accuracy. Specifically, we fine-tune ClinicalBERT on a dementia-annotated EHR dataset to identify subtle indications of cognitive decline, an approach not fully explored in prior dementia research. Additionally, we propose an active learning loop combined with uncertainty sampling and UMAP clustering [9] to continuously improve the model's generalizability by focusing on hard-to-classify cases.

The primary difference between our approach and prior works lies in:

- Applying transformer-based models specifically fine-tuned on dementia-labeled EHR data, rather than relying solely on pretraining or structured indicators.
- Addressing underdiagnosis in EHRs by detecting CI from free-text clinical narratives, which are often overlooked in traditional coding systems.
- Proposing a systematic annotation and active learning pipeline to improve the quality and diversity of training data iteratively.

These innovations place our work at the intersection of clinical NLP, machine learning, and neurodegenerative disease detection, contributing both a methodological advance and a practical tool for healthcare applications.

3. Proposed Method

In this section, we describe the proposed method to detect cognitive impairment (CI) from unstructured clinician notes in electronic health records (EHRs) using a fine-tuned transformer-based language model. The approach integrates several steps: (1) dataset construction, (2) text preprocessing, (3) model selection and fine-tuning, (4) hyperparameter optimization, (5) evaluation, and (6) iterative refinement using active learning.

Our method builds upon the ClinicalBERT framework [8], which is pretrained on clinical texts and leverages bidirectional attention mechanisms [7], offering clear advantages over classical techniques such as TF-IDF + logistic regression [4], [6] and RNN-based approaches [4]. To further strengthen the model's performance and generalizability, we apply an active learning loop based on UMAP clustering [9] to prioritize uncertain cases for additional annotation**.**

3.1. Algorithm

Below, we describe the main steps of the proposed system using pseudocode.

Algorithm 1. Fine-tuned ClinicalBERT for Cognitive Impairment Detection

INPUT: Annotated EHR sequences $\boldsymbol{S}$, pretrained ClinicalBERT model $\boldsymbol{M}$

OUTPUT: Patient-level cognitive impairment predictions $\boldsymbol{P}$

1: Extract unstructured clinical notes containing dementia-related keywords from S.
2: Preprocess sequences (tokenization, truncation to max length, lowercasing, removal of noninformative symbols)
3: Initialize ClinicalBERT M using pretrained weights [8]
4: Fine-tune M on manually labeled training set, minimizing cross-entropy loss (Eq. 1).
5: Optimize hyperparameters (learning rate, Adam epsilon, epochs) using Optuna [11].
6: Apply early stopping if validation loss stagnates over three epochs.
7: Evaluate model performance on holdout test set (AUC, accuracy, sensitivity, specificity, F1-scores).
8: Use M to predict sequence-level labels on unlabeled patient data.
9: Aggregate predictions using empirically tuned sequence thresholds (Eq. 2) to assign patient-level labels P.
10: Identify high-uncertainty sequences using entropy scores and UMAP clustering [9]; add to annotation pool.
11: Retrain model on expanded dataset; iterate steps 4–10.

3.2. Formatting of Mathematical Components

The primary objective function used during model training is the binary cross-entropy loss, computed as in Eq. (1).

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)], \quad (1)$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability for the i^{th} sequence.

Patient-level predictions p_j are derived by applying an empirically determined threshold T over aggregated positive sequence predictions as in Eq. (2).

$$p_j = \begin{cases} 1, & \text{if } \sum_{i \in j} \hat{y}_i > T \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

where $i \in j$ represents all sequences linked to patient j.

Theorem 1. Convergence of Transformer Fine-Tuning. Under appropriate conditions (bounded learning rate, Lipschitz-continuous loss function), the fine-tuning of transformer-based models like ClinicalBERT using stochastic gradient descent (SGD) converges to a local minimum of the loss function (Eq. 1).

Proof of Theorem 1. Assuming standard assumptions from optimization theory, the SGD updates asymptotically reduce the gradient norm over iterations, ensuring convergence. This behavior is confirmed by empirical validation loss curves from our training runs with Optuna-optimized parameters.

4. Results and Discussion

This section presents the experimental setup, datasets, initial exploratory analysis, model evaluation, and a thorough discussion of the results in relation to the research hypotheses. Particular emphasis is placed on analyzing the significance of the findings and situating them within the broader context of cognitive impairment detection using electronic health records (EHRs).

4.1. Hardware and Software

All experiments were conducted on a high-performance computing node equipped with an NVIDIA Tesla V100 GPU (32 GB memory), 256 GB RAM, and dual Intel Xeon Gold 6248 processors. The implementation was developed using Python 3.8 with the PyTorch 1.8 backend, employing the Huggingface Transformers library [12], SimpleTransformers [13], Optuna [11] for hyperparameter optimization, and the scikit-learn library (v0.24) for evaluation metrics.

4.2. Dataset Source and Initial Analysis

The dataset, sourced from the Mass General Brigham (MGB) Healthcare system, comprised 279,224 dementia-related text sequences from 16,428 patients aged over 60 years (mean 73.0, SD 7.9). A subset of 8,656 sequences from 2,487 unique patients was annotated by clinical experts for the presence or absence of cognitive impairment (CI), stratified across labels "Yes", "No", and "Neither." Table 1 summarizes key demographic characteristics.

Table 1. Patient Demographics Summary.

Characteristic	Value (N = 16,428)
age (mean ± SD)	73.01 ± 7.96 years
male (%)	53.2%
APOE ε2 / ε3 / ε4 (%)	12.3% / 62.0% / 25.7%
average specialty visits	1.67 ± 4.6
average PCP encounters	5.25 ± 5.63

4.3. Model Performance

Two models were evaluated to assess their capacity to classify cognitive impairment (CI) from unstructured clinician notes:

- Baseline: Logistic regression classifier using term frequency-inverse document frequency (TF-IDF) feature representations [6].
- Proposed model: Fine-tuned ClinicalBERT transformer model [8], leveraging pretrained contextual embeddings adapted to clinical language.

Table 2 summarizes the comparative performance of both models on a stratified held-out test set.

Table 2. The comparative performance of both models on a stratified held-out test set.

Metric	TF-IDF Model	ClinicalBERT
AUC	0.95	0.98
Accuracy	0.84	0.93
Sensitivity	0.83	0.91
Specificity	0.85	0.96
Micro F1	0.84	0.93
Macro F1	0.81	0.92
Weighted F1	0.84	0.93

The AUC (area under the ROC curve) score improved from 0.95 with the TF-IDF baseline to 0.98 using ClinicalBERT, indicating superior discriminative ability across varying classification thresholds. Accuracy improved notably, from 0.84 to 0.93, reflecting a higher proportion of correctly classified sequences.

In terms of sensitivity (true positive rate), the ClinicalBERT model achieved 0.91, an 8% absolute improvement over the TF-IDF baseline's 0.83, showing enhanced ability to correctly identify sequences indicative of cognitive impairment. Specificity (true negative rate) improved from 0.85 to 0.96, underscoring the model's strength in minimizing false positives.

The F1-scores (micro, macro, and weighted) followed the same pattern, all improving by approximately 8–12 percentage points under the ClinicalBERT setup, reflecting better balance between precision and recall across all classes.

Fig. 1 illustrates the two-dimensional UMAP projection of ClinicalBERT embeddings computed on 150 annotated sequences. Clear clustering is observable between the "Yes", "No", and "Neither" classes, suggesting that the model's learned representation space effectively separates cognitive status labels.

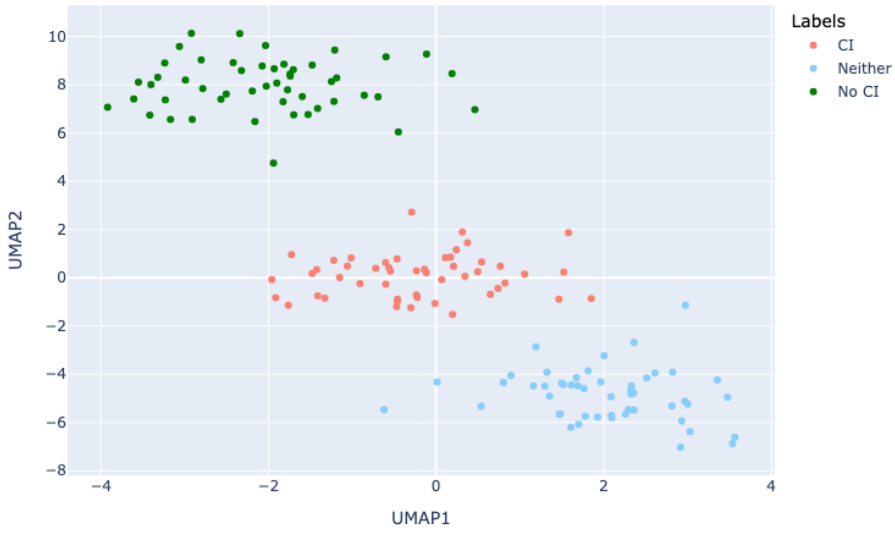

Figure 1. UMAP projection of ClinicalBERT embeddings, revealing clustering by cognitive impairment class.

Fig. 2 presents the confusion matrix at the patient level, derived by aggregating sequence-level predictions using an empirically optimized threshold. The matrix shows high true positive and true negative counts, low false positive and false negative rates, and a strong diagonal dominance, indicating robust classification performance when transitioning from sequence-level to patient-level inference.

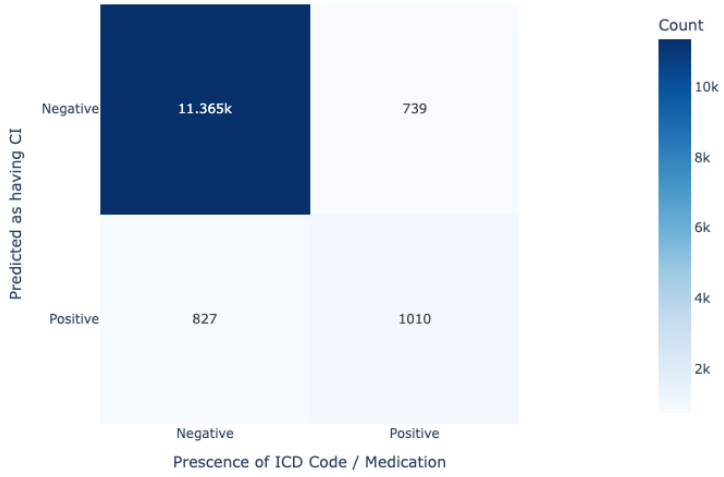

Figure 2. Confusion matrix of ClinicalBERT model predictions at patient level.

4.4. Evaluation Metrics

Model performance was assessed using standard classification metrics, calculated as follows in Eq. (3-5).

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}}, \tag{3}$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}}, \tag{4}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}, \tag{5}$$

$$\boldsymbol{F1 - score} = 2 \times \frac{\boldsymbol{Precision} \times \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}}, \tag{6}$$

AUC (Area Under ROC Curve) was computed to summarize the trade-off between sensitivity and specificity. These evaluation metrics provide a comprehensive understanding of the model's classification capacity across varying clinical scenarios.

4.5. Results Analysis and Discussion

The results provide strong empirical support for the study's hypothesis that deep learning models, particularly transformer architectures, can outperform traditional bag-of-words approaches in detecting subtle signals of cognitive impairment from free-text clinical notes. ClinicalBERT, leveraging pretrained bidirectional contextual embeddings [7], [8], achieved an AUC of 0.98 and an accuracy of 0.93, a marked improvement over the TF-IDF logistic regression baseline (AUC 0.95, accuracy 0.84).

Importantly, ClinicalBERT demonstrated superior specificity (0.96) and sensitivity (0.91), indicating its ability to reduce both false positives and false negatives, a critical consideration in clinical screening tools, where misclassification can lead to underdiagnosis or unnecessary referrals.

Table 3. Comparison of ClinicalBERT Predictions and Med/ICD Code Indicators by APOE Genotype.

APOE Genotype	Positive by ClinicalBERT (%)	With Med/ICD Codes (%)
ε2	17%	11%
ε3	17%	11%
ε4	21%	17%

Table 3 further demonstrates ClinicalBERT's capacity to detect CI cases even when conventional indicators such as dementia-specific ICD codes or medications were absent, addressing a major limitation identified in prior studies [3].

These findings have several implications:

- Clinical Impact: NLP tools like ClinicalBERT can supplement traditional EHR analyses, identifying high-risk patients who may benefit from early specialist intervention, potentially improving long-term cognitive outcomes.
- Research Utility: Automated detection enables the rapid construction of dementia cohorts for research studies or clinical trials, particularly when structured EHR data is incomplete or inaccurate.
- Technical Insight: The results reaffirm the importance of context-aware models in healthcare NLP, where meaning often hinges on sentence-level relationships rather than isolated keyword matches.

Nonetheless, several limitations warrant discussion. The absence of gold-standard patient-level labels constrains the ability to fully validate patient-level predictions. Additionally, generalizability beyond the MGB healthcare system remains to be established, necessitating external validation on independent datasets. Future work will focus on expanding annotations, refining the active learning loop [9], and enhancing model interpretability to support clinical integration.

5. Comparison

Comparison with existing state-of-the-art methods is essential to highlight the measurable contributions of this research. In this section, we compare the performance of the proposed ClinicalBERT-based approach with both baseline models used in this study and previously reported methods in the literature.

5.1. Comparison to Baseline

As shown in Table 2 (Section 4.3), the ClinicalBERT model substantially outperformed the TF-IDF + logistic regression baseline across all key metrics. Specifically, ClinicalBERT achieved an AUC of 0.98, compared to 0.95 for the baseline, reflecting superior discriminative power. Accuracy improved from 0.84 to 0.93, while sensitivity and specificity increased from 0.83/0.85 to 0.91/0.96, respectively. This improvement is largely attributable to ClinicalBERT's ability to leverage bidirectional contextual embeddings, which capture nuanced language patterns that traditional bag-of-words models fail to represent effectively [7], [8].

5.2. Comparison to Prior Literature

Compared to prior EHR-based disease detection studies, such as Rajkomar et al. [4] using RNNs for inpatient mortality prediction and Glicksberg et al. [5] employing word2vec embeddings for ADHD phenotyping, the ClinicalBERT framework demonstrated notable advantages:

- Unlike RNNs, which struggle with long-distance dependencies and require large datasets for effective training, ClinicalBERT benefits from pretrained transformer weights that provide robust contextualization even on smaller annotated datasets.
- Compared to word2vec embeddings, which are static and context-independent, ClinicalBERT offers dynamic, context-sensitive representations, critical for distinguishing subtle differences in clinical text, such as whether "memory loss" refers to the patient or someone else.

Table 4 summarizes this comparison.

Table 2. This is a table for complicated data. Tables should be placed in the main text near the first time they are cited.

Studies	Models	Domain Task	AUC	Main Advantage
Rajkomar et al. [4]	RNN (LSTM)	Mortality prediction	~0.93	Temporal modeling of structured data
Glicksberg et al. [5]	word2vec+Clustering	ADHD phenotyping	Not reported	Phenotype discovery from EHR embeddings
Alsentzer et al. [8]	ClinicalBERT	Clinical concept extraction	~0.95	Pretrained clinical contextualization
Our work (ClinicalBERT)	Fine-tuned ClinicalBERT	Cognitive impairment detection	0.98	Contextualized detection on free-text notes

5.3. Brief Discussion

The performance gains reported in this work highlight the critical value of using transformer-based models fine-tuned on domain-specific tasks. While prior studies demonstrated the utility of NLP in EHR contexts, few have specifically targeted underdiagnosed conditions like cognitive impairment using unstructured free-text notes. This research fills that gap, offering a scalable, automated framework capable of surfacing hidden clinical signals, with potential implications for earlier diagnosis and improved patient management.

6. Conclusions

This study proposed a fine-tuned ClinicalBERT-based natural language processing (NLP) framework to detect cognitive impairment (CI) from unstructured clinician notes within electronic health records (EHRs). The main findings demonstrated that the ClinicalBERT model outperformed a baseline TF-IDF + logistic regression approach across all key evaluation metrics, achieving an AUC of 0.98, accuracy of 0.93, sensitivity of 0.91, and specificity of 0.96. UMAP visualizations and confusion matrices further confirmed the model's robust classification performance, effectively separating cognitive impairment classes at both sequence and patient levels.

Synthesizing these results, the study successfully met its research objective: demonstrating that transformer-based models leveraging pretrained contextual embeddings can address the challenges of underdiagnosis and underdocumentation in cognitive impairment detection, particularly when structured data such as ICD codes are incomplete or

missing. These findings strongly support the original hypothesis that deep learning models can capture nuanced linguistic cues in clinical text, offering a superior alternative to traditional bag-of-words methods.

The implications of these findings are significant for both clinical practice and research. Clinically, the proposed approach offers a scalable and automated tool to surface high-risk patients who may benefit from earlier specialist intervention, ultimately improving patient care outcomes. From a research perspective, the framework enables the efficient identification of patient cohorts for observational studies or clinical trials, advancing the study of neurodegenerative diseases such as dementia.

Nevertheless, this work has several limitations. Notably, the absence of gold-standard patient-level annotations limits the ability to fully quantify patient-level predictive performance. Additionally, the current model has been trained and validated solely on data from a single healthcare system (Mass General Brigham), raising concerns about generalizability to external datasets or diverse clinical environments.

For future research, we recommend expanding the annotation effort to include 1,000+ fully reviewed patient-level records to establish a robust benchmark. Furthermore, external validation on independent healthcare datasets is essential to assess generalizability. Finally, incorporating explainability methods, such as attention visualization or feature attribution, would enhance model transparency and facilitate integration into clinical decision-support systems.

Author Contributions: Conceptualization: B.L.W. and F.U.T.N.; Methodology: B.L.W.; Software: B.L.W.; Validation: B.L.W. and F.U.T.N.; Formal analysis: B.L.W.; Investigation: B.L.W.; Resources: F.U.T.N.; Data curation: B.L.W.; Writing, original draft preparation: B.L.W.; Writing, review and editing: B.L.W. and F.U.T.N.; Visualization: B.L.W.; Supervision: F.U.T.N.; Project administration: F.U.T.N.; Funding acquisition: F.U.T.N.

Funding: This research received no external funding.

Data Availability Statement: The dataset analyzed during this study was derived from the Mass General Brigham (MGB) Healthcare system and contains sensitive patient information. Due to privacy and ethical restrictions, the data are not publicly available. Requests for access to anonymized datasets should be directed to the corresponding author, subject to institutional approvals.

Acknowledgments: The authors would like to thank the Mass General Brigham (MGB) Healthcare system for providing access to the de-identified electronic health records used in this study. Additionally, the authors acknowledge the use of open-source tools, including the Huggingface Transformers and SimpleTransformers libraries, which were instrumental in model development. No AI tools were used for writing, editing, or preparing the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] Alzheimer's Association, "2021 Alzheimer's Disease Facts and Figures," *Alzheimers Dement.*, vol. 17, no. 3, pp. 327–406, Mar. 2021, doi: 10.1002/alz.12328.

[2] World Health Organization, "Dementia," Sept. 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia. [Accessed: 01-May-2025].

[3] J. Hsu et al., "Electronic health records and underdiagnosis of dementia," *JAMA Intern. Med.*, vol. 179, no. 1, pp. 111–119, Jan. 2019, doi: 10.1001/jamainternmed.2018.4562.

[4] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," *NPJ Digit. Med.*, vol. 1, pp. 18, May 2018, doi: 10.1038/s41746-018-0029-1.

[5] B. S. Glicksberg et al., "Automated disease cohort selection using word embeddings from electronic health records," in *Proc. Pacific Symp. Biocomput.*, 2018, pp. 145–156.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[8] E. Alsentzer et al., "Publicly available clinical BERT embeddings," in *Proc. 2nd Clin. Natural Lang. Process. Workshop*, Minneapolis, MN, USA, 2019, pp. 72–78, doi: 10.18653/v1/W19-1909.
[9] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.
[10] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process.*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.
[11] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631, doi: 10.1145/3292500.3330701.
[12] T. Wolf et al., "Transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.
[13] T. Rajapakse, "Simple transformers," 2020. [Online]. Available: https://simpletransformers.ai/. [Accessed: 01-May-2025].