Glaucoma Detection from Clinical Notes: Benefits of Multimodal Fusion Across Neural Encoders

Faias Satter (C00607330)
Data Mining (CSCE 566), Fall 2025
University of Louisiana at Lafayette

1 Introduction

Glaucoma is a major cause of irreversible blindness, affecting over 70–80 million people worldwide and projected to surpass 111 million cases by 2040 [2]. Because early stages are often asymptomatic, nearly half of patients remain undiagnosed until significant vision loss has occurred. Early detection is critical, as timely treatment can delay severe impairment for many years [16].

Recent advances in clinical Natural Language Processing (NLP) highlight the strengths of different neural architectures for modeling medical text. Bidirectional Long Short-Term Memory (BiLSTM)-based methods have been shown to effectively capture bidirectional context and clinical phrasing in radiology and other healthcare narratives [5]. Gated Recurrent Unit (GRU) architectures offer a simpler and more efficient recurrent alternative, demonstrating strong performance on clinical prediction tasks involving noisy or irregular health records [3]. Convolutional neural networks (CNNs) have also proven effective for medical text classification, particularly by identifying key diagnostic phrases and local lexical patterns [9]. More recently, Transformer-based models such as ClinicalBERT have achieved state-of-the-art performance on clinical note understanding through self-attention mechanisms [1]. Moreover, Several studies highlight the importance of integrating multiple data modalities in clinical prediction tasks. Rajkomar et al.[15] demonstrate that combining structured electronic health records (EHR) fields with unstructured clinical text often improves diagnostic accuracy compared to using either modality alone.

Motivated by these findings, we demonstrate whether glaucoma can be automatically detected by integrating clinical notes with structured demographic attributes in a unified multimodal framework. Using the FairCLIP dataset[14], we compare four neural text encoders—BiLSTM[12], GRU[4], 1D CNN[11], and ClinicalBERT[1]—to quantify how multimodal fusion influences overall and race-specific detection performance.

All source code used in this study has been made publicly available to support reproducibility and future research at: `https://github.com/FaiasPromit/Data_Mining_CSCE_566_Project_Fall_2025_ULL`

2 Related Work

Early work on glaucoma detection focused primarily on imaging modalities such as fundus photography and optical coherence tomography (OCT), where CNNs have shown high performance but rely on specialized hardware and image acquisition steps [8, 7]. More recently, several studies have investigated the use of EHRs and NLP of clinical notes for glaucoma prognosis. For example, Wang et al. developed a deep-learning model using EHR structured data and clinical text to predict glaucoma progression to surgery (AUC 0.73) [17]. Jalamangala Shivananjaiah et al. combined free-text ophthalmology notes with structured features and achieved an AUC of 0.899 for predicting near-term glaucoma progression [10]. Lin et al. proposed a multimodal model integrating both structured EHR and free-text records to predict glaucoma surgical outcomes, confirming that fusion improves performance over single-modality models [13]. Compared to prior studies, our work focuses on demonstrating how multimodal fusion of routine clinical notes with simple demographic attributes from the FairCLIP dataset improves glaucoma detection beyond text-only models.

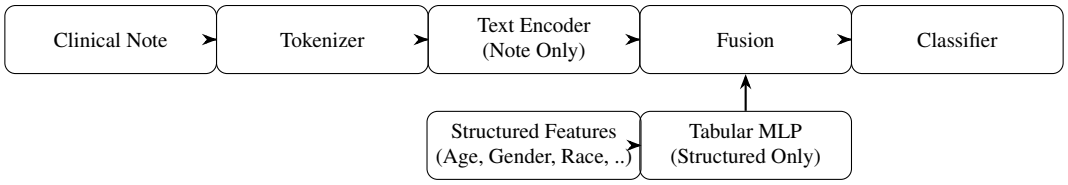
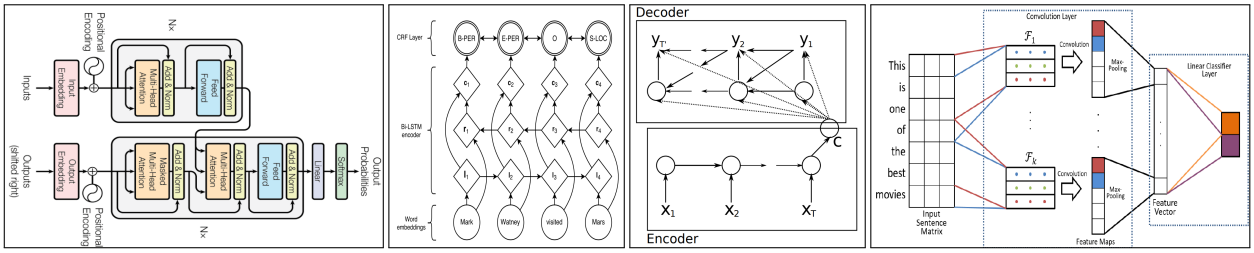


Figure 1: Multimodal architecture with fusion of text and structured features.



(a) Transformer (BERT)[6] (b) BiLSTM[12] (c) GRU[4] (d) 1D CNN[11]

Figure 2: Model architecture used in this study. All images are snapshots taken from their respective reference papers.

3 Method

Figure 1 illustrates the overall multimodal architecture used in this work. The model contains two independent processing streams: one for the clinical note text and one for the structured demographic features. Each stream generates its own vector representation, and these are fused through simple concatenation before being passed to a final classifier. By keeping the fusion stage identical across all experiments, we ensure that performance differences arise solely from the choice of text encoder.

3.1 Text Processing for Clinical Notes

Clinical notes are first lowercased and tokenized at the word level, then truncated or padded to a fixed sequence length of 512 tokens. The resulting token IDs are transformed into trainable word embeddings, which serve as input to one of four text encoders evaluated in this study. The GPT-4 Summary feature is also processed in the same manner and concatenated with the raw clinical notes during multimodal training. Figure 2 shows the architecture of each model.

3.2 Structured Feature Processing

A separate module processes structured demographic features. Age is standardized using the dataset mean and standard deviation, while categorical variables were mapped to integer indices and converted into trainable embedding vectors of dimension 16. These embeddings are concatenated and passed through a two-layer MLP to produce a 64-dimensional tabular representation. This allows structured attributes to contribute informative signals without overshadowing the textual component. These settings were kept fixed for every architecture.

3.3 Multimodal Fusion

The outputs of the text encoder and tabular MLP are fused via concatenation. This late-fusion strategy allows each modality to be modeled independently while enabling the classifier to jointly exploit information from both sources. A single linear layer with sigmoid activation produces the final glaucoma probability.

Table 1: Training hyperparameters for all model variants.

Model	**BiLSTM**	**GRU**	**1D CNN**	**ClinicalBERT**
Epochs	5	5	5	3
Batch size	32	32	32	8
Learning rate	1e-3	1e-3	1e-3	2e-5
Optimizer	Adam	AdamW	Adam	AdamW
Max seq length	512	512	512	512
Vocab size	30000	30000	30000	–
Word embedding dim	128	128	128	–
Encoder hidden size	128×2	128×2	128 filters	768
CNN kernel sizes	–	–	{3,4,5}	–
Cat. embed dim	8	8	8	8
Tabular MLP dim	32	32	32	32
Classifier hidden dim	128	128	128	–
Tokenizer	whitespace	whitespace	whitespace	ClinicalBERT tokenizer

Table 2: Dataset demographic and label distribution of the FairCLIP Dataset.

Gender	Count	**Race**	Count	**Ethnicity**	Count	**Dataset Split**	Count
Female	5631	White	7690	Non-Hispanic	9062	Train	7000
Male	4369	Black	1491	Hispanic	398	Validation	1000
		Asian	819	Unknown	540	Test	2000
Marital Status	Count	**Glaucoma**	Count	**Language**	Count		
Married/Partnered	5739	Positive	5048	English	9250		
Single	2637	Negative	4952	Spanish	176		
Divorced	662			Other	76		
Widowed	614			Unknown	498		
Separated	97						
Unknown	251						

4 Experiments

4.1 Hardware and Implementation Details

All experiments were conducted on a local machine equipped with an NVIDIA RTX 3050 GPU and 32 GB of system memory. The models were implemented in Python using PyTorch 2.0 and the HuggingFace Transformers library for the ClinicalBERT encoder. A complete summary of model-specific hyperparameters is provided in Table 1.

4.2 Dataset Description

We use the FairCLIP dataset[14], which contains de-identified clinical notes paired with structured demographic attributes and a binary glaucoma label. Each example includes one free-text clinical note and several patient-level features: age, gender, race, ethnicity, language, and marital status. Table 2 summarizes the dataset statistics.

4.3 Results

4.3.1 Multimodal Performance

Table 3 presents the overall test results. The 1D CNN achieved the highest AUC (0.8801), followed closely by ClinicalBERT (0.8705). Both models also showed strong sensitivity–specificity balance. These results suggest that shallow convolutional architectures and pretrained Transformers are more effective at modeling long clinical notes than RNN-based encoders.

All models demonstrated consistent performance across racial groups, with the best subgroup AUCs again obtained by the 1D CNN and ClinicalBERT. These results indicate that the multimodal fusion setup does not introduce performance collapse across demographic subgroups.

Table 3: Test performance of all models using multimodal input (clinical notes + structured features).

Model	AUC	Sensitivity	Specificity	AUC (Asian)	AUC (Black)	AUC (White)
BiLSTM	0.7606	0.7879	0.5599	0.7714	0.7821	0.7495
GRU	0.8021	0.6891	0.7584	0.8332	0.7943	0.7963
1D CNN	**0.8801**	**0.7898**	0.7932	**0.9261**	0.8963	**0.8698**
ClinicalBERT	0.8705	0.7019	**0.8260**	0.9229	**0.8974**	0.8590

Table 4: Comparison of text-only vs multimodal training across all models.

Model	Text Only AUC	Multimodal AUC	Difference
BiLSTM	0.7301	0.7606	+0.0305
GRU	0.5000	0.8021	**+0.3021**
1D CNN	0.8684	0.8801	+0.0117
ClinicalBERT	**0.8821**	0.8705	−0.0116

4.3.2 Performance Comparison: Text-Only(Clinical Notes) vs. Multimodal(All features)

To evaluate the contribution of structured demographic information, all models were retrained using clinical notes alone. Figure 4 summarizes the difference between the two training paradigms. In spite of GRU failing in the text-only evaluation (AUC: 0.5000), GRU benefited the most from multimodal fusion, improving from 0.5000 to 0.8021 (+0.3021 AUC). BiLSTM also showed a moderate gain (+0.0305). Overall, multimodal integration yields consistent improvements for recurrent models but is less critical for CNNs and Transformers.

5 Conclusion

This work demonstrates that glaucoma can be effectively detected from routine clinical notes by integrating textual information with structured demographic attributes. Among the four evaluated architectures, 1D CNN and ClinicalBERT achieved the highest overall performance, highlighting the advantage of strong text encoders for long and noisy medical narratives. Our multimodal fusion approach consistently improved the performance of recurrent models such as BiLSTM and GRU, while offering marginal gain for CNNs, suggesting that the value of structured features depends strongly on the capacity of the text encoder. Moreover, the degradation of results for the multimodal approach in the case of ClinicalBERT suggests that it is already pretrained on massive clinical text and is highly sophisticated. The additional structured features might:

1. Add noise: Simple demographics may not add signal for this task. 2. Dilute powerful text representations: The 768-dim BERT embeddings get concatenated with only 32-dim tabular features. 3. Overfitting: The extra parameters from tabular MLP (with only 1000 validation samples) may cause overfitting.

Furthermore, the probable reasons behind GRU failing in the text-only evaluation are:

1. Insufficient capacity: GRU alone cannot extract meaningful patterns from raw clinical notes. 2. Gradient issues: Bidirectional GRU may suffer from vanishing gradients on long sequences. 3. Model initialization: Could have gotten stuck in a poor local minimum. 4. Learning dynamics: BCELoss + sigmoid output may have caused instability.

Although the models exhibit robust subgroup performance across racial categories, future work could explore more expressive fusion mechanisms, richer structured inputs (e.g., EHR timelines), and specialized medical foundation models to further enhance diagnostic accuracy and fairness.

References

[1] Emily Alsentzer, John Murphy, Willie Boag, Wen-Haw Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019.

[2] Mona Ashtari-Majlan, Mohammad Mahdi Dehshibi, and David Masip. Glaucoma diagnosis in the era of deep learning: A survey. *Expert Systems with Applications*, 256:124888, 2024.

[3] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

[4] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[5] Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana. Modelling radiological language with bidirectional long short-term memory networks. In Cyril Grouin, Thierry Hamon, Aurélie Névéol, and Pierre Zweigenbaum, editors, *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 17–27, Auxtin, TX, November 2016. Association for Computational Linguistics.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Osama N Hassan, Serhat Sahin, Vahid Mohammadzadeh, Xiaohe Yang, Navid Amini, Apoorva Mylavarapu, Jack Martinyan, Tae Hong, Golnoush Mahmoudinezhad, Daniel Rueckert, Kouros Nouri–Mahdavi, and Fabien Scalzo. Conditional gan for prediction of glaucoma progression with macular optical coherence tomography. *arXiv preprint arXiv:2010.04552*, 2020.

[8] Ruben Hemelings, Bart Elen, João Barbosa–Breda, Matthew B Blaschko, Patrick De Boever, and Ingeborg Stalmans. Deep learning on fundus images detects glaucoma beyond the optic disc. *Scientific Reports*, 11:1–9, 2021.

[9] Michael C. Hughes, Yichao Li, Spyros Kotoulas, and Toyotaro Suzumura. Medical text classification using convolutional neural networks. *Journal of the American Medical Informatics Association*, 24(3):555–560, 2017.

[10] Sunil K Jalamangala Shivananjaiah, Sneha Kumari, Iyad Majid, and Sophia Y Wang. Predicting near-term glaucoma progression: An artificial intelligence approach using clinical free-text notes and data from electronic health records. *Frontiers in Medicine*, 10:1157016, 2023.

[11] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

[12] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[13] W-C Lin and ... (others). Prediction of multiclass surgical outcomes in glaucoma using multimodal deep learning of ehr data and free-text notes. *Journal of the American Medical Informatics Association*, 31(2):456–467, 2024.

[14] Yan Luo, Min Shi, Muhammad Osama Khan, Shuaihang Yuan Yu Tian Song Luo Muneeb Afzal, Hao Huang, Ava Kouhana, Tobias Elze, Yi Fang, and Mengyu Wang. Harvard-fairvlmed (10,000 patient multimodal glaucoma dataset). `https://github.com/Harvard-Ophthalmology-AI-Lab/FairCLIP`, 2024. Dataset includes 10,000 clinical notes and demographic attributes for glaucoma diagnosis; used under CC BY-NC-ND 4.0.

[15] Alvin Rajkomar, Eran Oren, Kai Chen, Andrew M Dai, Noa Hajaj, Peter J Liu, Xuefeng Liu, Min Sun, Patrik Sundberg, Henry Yee, Po-Han Zhang, Kun Zhang, Gisela Flores, Gary Duggan, Lawrence Horton, Michael D Howell, Bobak J Mortazavi, Elizabeth Murray, David McLean, Aditya Rao, David Odell, Gordon Schiff, and Nigam H Shah. Deep learning in electronic health records. *NPJ Digital Medicine*, 1(1):1–10, 2018.

[16] Tanay Saha, Suman Sedai, Benny Antony, Sampath Abeysinghe, Han Ta, Michael Garvey, Stephen Daniel, David Hwang, Christopher A. Girkin, Steven Harper, Sandra Smith, and Vinay Prabhu. A fast and fully automated system for glaucoma detection using color fundus photographs. *Scientific Reports*, 13(1):18171, 2023.

[17] Sophia Y Wang, Brian Tseng, and Tina Hernandez–Boussard. Deep learning approaches for predicting glaucoma progression using electronic health records and natural language processing. *Ophthalmology Science*, 2(2):100127, 2022.