DSCI560 - Data Science Professional Practicum

Deep Learning-Based NLP Data Pipeline for EHR-Scanned Document Information Extraction

Authors: Enshuo Hsu, Ioannis Malagaris, Yong-Fang Kuo, Rizwana Sultana, Kirk Roberts

Shida Yan

Background and Objectives

Research Background

- Scanned documents in Electronic Health Records (EHR) remain a long-standing challenge
- Sources: faxed reports, paper-based documents, and external laboratory reports
- Processing requires image preprocessing, Optical Character Recognition (OCR), and Natural Language Processing (NLP)
- Limited systematic evaluation of method combinations and their interactions

Research Objectives Extract two key sleep apnea indicators from 955 scanned sleep study reports:

1. Apnea Hypopnea Index (AHI) - Gold standard for sleep apnea diagnosis

2. Oxygen Saturation (SaO2) - Additional clinical information for intervention

Data Source

- Sleep study reports from 2015-2018
- 955 reports, 2,988 scanned pages
- 83,915 numeric candidate values

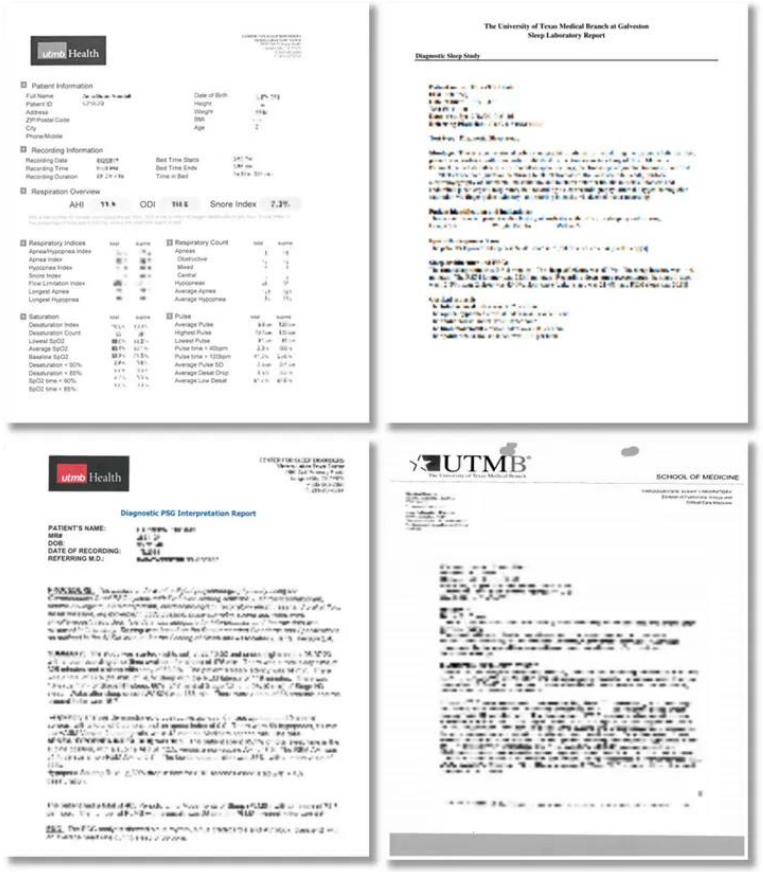

Figure 5. Collection of scanned sleep study reports. The images have been intentionally blurred, their purpose here is to provide a sense of the overall structure and consistency (and lack thereof) between scanned documents.

Methodology and Data Pipeline

Complete Data Processing Workflow:

1. Image Preprocessing (955 documents)

- Grayscale conversion
- Dilation and erosion (1 iteration)
- Contrast enhancement (20%)
- Evaluated 6 different combinations

2. Optical Character Recognition (OCR)

- Tesseract 4.0.0 for text extraction
- Word position information (pixel coordinates)
- De-identification processing

3. Text Segmentation (83,915 segments)

- Identify numeric candidate values
- Extract 21-word context window

4. NLP Model Classification

- **Traditional Methods**: 7 models including Logistic Regression, SVM, Random Forest
- **Deep Learning**: BiLSTM, BERT, ClinicalBERT
- **Innovative Features**: Incorporating position, page number, and structured information

5. Model Evaluation (286 test documents)

- Segment-level: AUROC, Recall, Precision
- Document-level: Accuracy

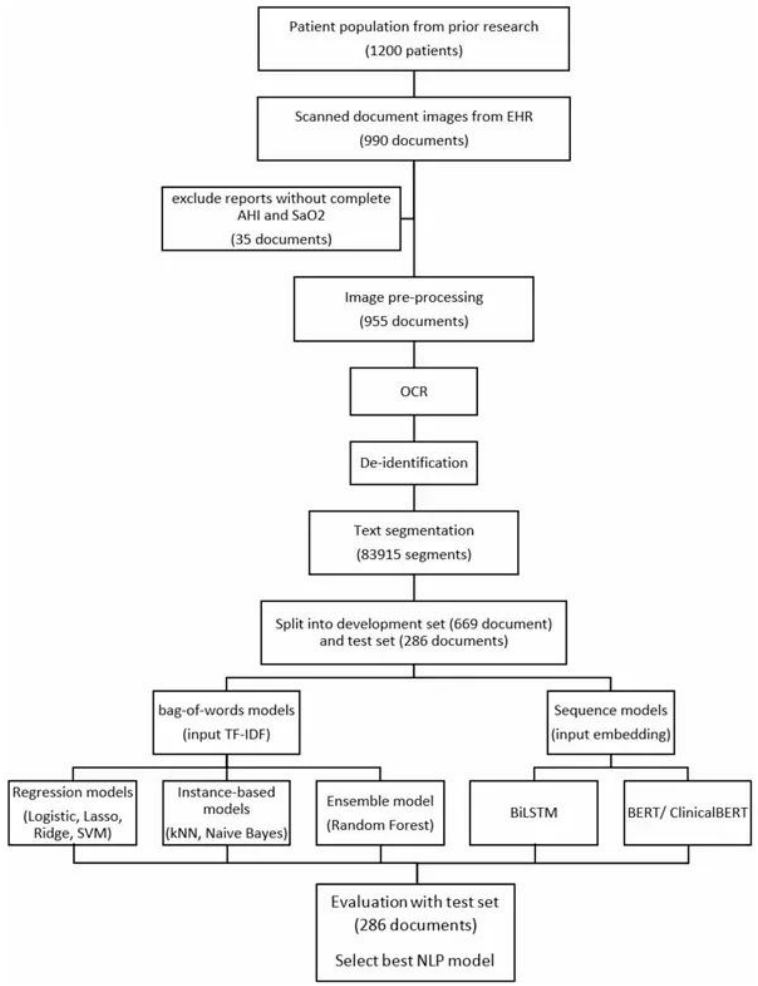

Figure 4. Data pipeline flowchart.

Main Results

Best Model Performance (ClinicalBERT + Structured Input)

AHI Extraction:

- AUROC: 0.9743
- Document Accuracy: 94.76%
- Recall/Precision: 0.73 / 0.91

SaO2 Extraction:

- AUROC: 0.9523
- Document Accuracy: 91.61%
- Recall/Precision: 0.68 / 0.89

Key Findings:

1. Deep learning models significantly outperforr traditional methods
2. ClinicalBERT achieves the best performance
3. Only 50 reports needed to reach ~90% accura
4. Document layout information significantly improves performance

Table 2. Summary of data and labels

	PDF documents		OCR outputs			
	Reports	Pages	Numeric values	Instances of AHI	Instances of SaO_2	Instances of other
Entire data set	955	2988	83 915	1904	1698	80 313
development set	669	2031	56 839	1323	1146	54 370
Test set	286	957	27 076	581	552	25 943

Table 3. Evaluation of different classifiers

	Classifier	Segment-level				Document-level
		Recall	Precision	F1	AUROC (95% CI)	Accuracy (95% CI)
AHI						
Bag-of-word models	LR	0.4819	0.8383	0.612	0.9093 (0.8932–0.9254)	87.41 (83.57–91.26)
	LASSO (L1)	0.4819	0.8889	0.625	0.9169 (0.9014–0.9325)	89.16 (85.56–92.76)
	Ridge (L2)	0.4802	0.8429	0.6118	0.9176 (0.9021–0.9331)	87.41 (83.57–91.26)
	SVM	0.6093	0.9752	0.75	0.9050 (0.8886–0.9215)	93.01 (90.05–95.96)
	kNN	0.6713	0.8534	0.7514	0.8644 (0.8454–0.8834)	93.57 (90.36–96.78)
	NaiveBayes	0.5577	0.4367	0.4898	0.9179 (0.9024–0.9334)	75.87 (70.92–80.83)
	Random Forest	0.6299	0.9865	0.7689	0.9476 (0.9350–0.9603)	93.71 (90.89–96.52)
Sequence models	BiLSTM	0.6454	0.9843	0.7796	0.9637 (0.9530–0.9743)	94.06 (91.32–96.80)
	BERT	0.747	0.8803	0.8082	0.9705 (0.9609–0.9802)	**95.10 (92.60–97.61)**
	ClinicalBERT	0.7315	0.914	**0.8126**	**0.9743 (0.9652–0.9833)**	94.76 (92.17–97.34)
SaO_2						
Bag-of-word models	LR	0.567	0.4914	0.5265	0.9153 (0.8992–0.9314)	82.87 (78.50–87.23)
	LASSO (L1)	0.538	0.5103	0.5238	0.9151 (0.8990–0.9312)	84.62 (80.43–88.80)
	Ridge (L2)	0.5543	0.4904	0.5204	0.9143 (0.8981–0.9305)	83.22 (78.89–87.55)
	SVM	0.6105	0.9133	0.7318	0.8860 (0.8678–0.9042)	87.76 (83.96–91.56)
	kNN	0.587	0.8663	0.6998	0.8429 (0.8223–0.8634)	87.86 (83.84–91.88)
	NaiveBayes	0.6322	0.2705	0.3789	0.9082 (0.8915–0.9248)	51.75 (45.96–57.54)
	Random Forest	0.6087	0.9307	0.736	0.9264 (0.9113–0.9415)	89.51 (85.96–93.06)
Sequence models	BiLSTM	0.6739	0.9051	0.7726	0.9274 (0.9123–0.9424)	**91.61 (88.40–94.82)**
	BERT	0.7319	0.8651	**0.7929**	0.9358 (0.9215–0.9500)	**91.61 (88.40–94.82)**
	ClinicalBERT	0.683	0.8871	0.7718	**0.9523 (0.9398–0.9647)**	**91.61 (88.40–94.82)**

Note: Logistic Regression does not apply penalty; Lasso regression has L1 penalty ($\lambda = 0.01$); Ridge has L2 penalty ($\lambda = 0.01$). Support Vector Machine uses a polynomial kernel. kNN uses $k = 3$. NaiveBayes classifier uses alpha = 0.5. BiLSTM uses Word2Vec model for embedding pretrained on the training set with CBOW, input vector of 100 dimensions. BERT and ClinicalBERT are fine-tuned for 100 epochs with sequence length 32, and batch size 64. We highlight the highest F1, AUROC, and Accuracy in bold.

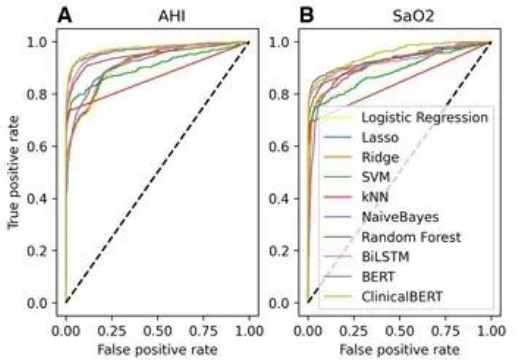

Figure 6. ROC curve for each classifier.

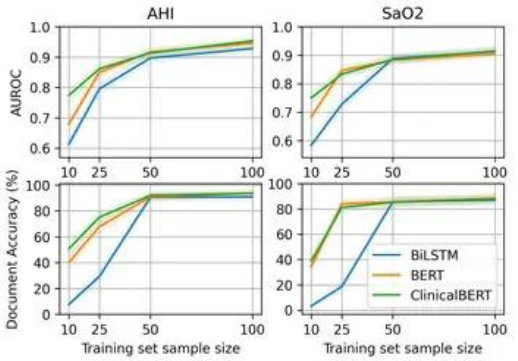

Figure 7. Evaluation of effects of training set size.

Conclusions

1. Methodological Innovation

- First systematic evaluation of deep learning for scanned medical document processing
- Validated the importance of image preprocessing and document layout information

2. Best Practices

- Image preprocessing: Grayscale + Dilation/Erosion + 20% contrast
- NLP model: ClinicalBERT + Structured features
- Training data: 50 reports sufficient for good performance

3. Practical Value

- Generalizable to other clinical document types
- Reduces manual annotation costs
- Provides solutions for legacy scanned documents

Table 4. Comparing ClinicalBERT with BERT, BiLSTM, and Random Forest

Adjusted *P*-value	ClinicalBERT vs BERT		ClinicalBERT vs BiLSTM		ClinicalBERT vs Random Forest	
	AHI	SaO_2	AHI	SaO_2	AHI	SaO_2
AUROC	0.4528	**0.0029**	**0.0008**	**<0.0001**	**<0.0001**	**<0.0001**
Document accuracy	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Note: AUROC was pair-wisely compared with DeLong's test. Document accuracy was pair-wisely compared with the chi-squared test. All *P*-values are corrected with the Bonferroni procedure. We highlight statistically significant *P*-values in bold.

Table 5. Comparison of different image preprocessing methods

	Image preprocessing	Segment-level				Document-level
		Recall	Precision	F1	AUROC (95% CI)	Accuracy (95% CI)
AHI	Gray scale	0.7187	0.9249	0.8089	0.9699 (0.9601–0.9796)	95.45 (93.04–97.87)
	Gray scale+dilate and erode	0.6961	0.9687	0.81	0.9679 (0.9573–0.9784)	94.41 (91.74–97.07)
	Gray scale+contrast 20%	0.7126	0.9324	0.8078	0.9705 (0.9609–0.9802)	94.06 (91.32–96.80)
	Gray scale+contrast 60%	0.7268	0.9216	0.8127	0.9692 (0.9593–0.9790)	95.45 (93.04–97.87)
	Gray scale+dilate and erode+contrast 20%	0.7315	0.914	0.8126	**0.9743 (0.9652–0.9833)**	94.76 (92.17–97.34)
	Gray scale+dilate and erode+contrast 60%	0.7268	0.9216	0.8172	0.9715 (0.9620–0.9810)	**95.80 (93.48–98.13)**
SaO_2	Gray scale	0.7258	0.8617	0.7879	0.9334 (0.9190–0.9478)	91.61 (88.40–94.82)
	Gray scale+dilate and erode	0.7427	0.8819	0.8063	**0.9620 (0.9504–0.9736)**	90.21 (86.77–93.65)
	Gray scale+contrast 20%	0.6957	0.8889	0.7805	0.9431 (0.9296–0.9566)	91.26 (87.99–94.53)
	Gray scale+contrast 60%	0.6863	0.8671	0.7662	0.9495 (0.9366–0.9623)	91.61 (88.40–94.82)
	Gray scale+dilate and erode+contrast 20%	0.683	0.8871	0.7718	0.9523 (0.9398–0.9647)	91.61 (88.40–94.82)
	Gray scale+dilate and erode+contrast 60%	0.6863	0.8671	0.7684	0.9486 (0.9356–0.9616)	**91.96 (88.81–95.11)**

Note: Each image preprocessing method was followed by fine-tuning a downstream ClinicalBERT. We highlighted the highest AUROC and Accuracy in bold.

Table 6. Comparison of different sequence model architectures

	Model architecture	Segment-level					Document-level	
		Recall	Precision	F1	AUROC (95% CI)	*P*-value	Accuracy (95% CI)	*P*-value
AHI	Sequence input	0.7522	0.8723	0.8078	0.9703 (0.9606–0.9800)	**.0092**	94.41 (91.74–97.07)	1.0000
	Sequence input+structured input	0.7315	0.914	0.8126	**0.9743 (0.9652–0.9833)**		**94.76 (92.17–97.34)**	
SaO_2	Sequence input	0.692	0.8761	0.7733	0.9430 (0.9295–0.9565)	**.0123**	90.91 (87.58–94.24)	.8823
	Sequence input+structured input	0.683	0.8871	0.7718	**0.9523 (0.9398–0.9647)**		**91.61 (88.40–94.82)**	

Note: We highlighted the highest AUROC and Accuracy, and statistically significant *P*-value in bold.

Question

Why does ClinicalBERT significantly outperform BiLSTM with a small training set (25 reports), but the three deep learning models converge to similar performance when the training set increases to 50 reports?

Answer

ClinicalBERT was pre-trained on large-scale clinical text, so it already has medical domain knowledge and can generalize well even with limited task-specific data (25 reports). BiLSTM learns from scratch and requires more data to build effective representations. When training data reaches 50 reports, all models have sufficient task-specific information to learn the patterns effectively, eliminating the pre-training advantage.

Thank You