RespBERT: A multi-site validation of a Natural Language Processing algorithm, of Radiology Notes to Identify Acute Respiratory Distress Syndrome (ARDS)

Ashwin Pathak, Curtis Marshall, Carolyn Davis, Philip Yang, and Rishikesan Kamaleswaran

***Abstract*— Acute respiratory distress syndrome (ARDS) is a severe organ dysfunction that is associated with significant mortality and morbidity among critically ill patients admitted to the Intensive Care Unit (ICU). The etiology associated with ARDS can be highly heterogeneous, with most cases being associated with infection or trauma. ARDS is often described as a clinical syndrome associated with poor oxygenation even in the presence of mechanical ventilation. The Berlin criteria of ARDS is the current gold standard for identifying whether patients had developed ARDS, however it often requires manual adjudication of the chest radiograph, resulting in limited tools to automate the process. Since the determination of ARDS is dependent on the presence of bilateral infiltrates on radiographic images, and this information is not typically available in Electronic Medical Record (EMR). Automated determination of the presence of radiological evidence would enable robust study of the syndrome by eliminating expensive individual inspection by physicians of the images. The text of radiological reports provides an opportunity for Natural Language Processing (NLP) to determine the status of the lungs for evaluating the imaging criterion. We developed a Natural Language Processing (NLP) pipeline to analyze radiology notes of 362 patients satisfying sepsis-3 criteria from the Electronic Medical Record (EMR) to determine possible ARDS diagnosis. The radiology notes were denoised and preprocessed. They were further vectorized through the word-embedding pipeline BERT and fitted to a classification layer using transfer learning. These classification models showed F1-score of 74.5% and 64.22% for Emory and Grady dataset respectively.**

***Index Terms*— Natural Language Processing, Large Language Models, ARDS, Critical Care, Sepsis**

I. Introduction

Acute respiratory distress syndrome (ARDS) is associated with severe inflammatory lung injury that results in acute respiratory failure [1] and severe hypoxemia. ARDS is associated with a mortality rate as high as 43% [2] and a 10% prevalence period in all intensive care unit (ICU) admissions, but only 34% of cases are recognized by clinicians [3]. The time-sensitive nature of ARDS, accompanied with complexities on laboratory data, radiological data, respiratory data and disease characteristics [4], necessitates the automation of ARDS diagnosis [5]–[7] using clinical radiology notes.

Traditional methods for diagnosing ARDS require evaluation and interpretation of patients' chest imaging [8]. Based on the ARDS definition, commonly used phrases have been identified throughout clinical radiology notes of ARDS patients, such as acute respiratory distress syndrome, ARDS, bilateral infiltrates, ground glass opacities, patchy, diffuse, interstitial, multifocal, extensive, and airspace disease [9]. Algorithms based on 'sniffer' systems automate the identification of ARDS in patients from their radiology notes [10], relying on simple keywords search, such as 'edema' or 'bilateral infiltrates'. However, these methods are not generalizable and heavily rely on keywords that often vary between institutions and cohorts. Furthermore, these methods are limited in their ability to distinguish the sub-phenotypes of ARDS and are prone to misclassification [11]–[13].

Recent methods for ARDS diagnosis leverage the inherent natural language in the clinical notes to better understand the valuable information of the patients [14], [15]. Such methods use Natural Language Processing (NLP), an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with developing methods for analyzing human language, in order to extract the information contained

This study was conducted according to the principles set forward in the Declaration of Helsinki and according to Good Clinical Practice. The dataset were collected from Emory affiliated hospitals (Atlanta, GA), after Emory University institutional review board (IRB) approval with reference number IRBSTUDY00000302.

A. Pathak is with School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA 30308, USA (e-mail: apathak60@gatech.edu) and Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, USA (e-mail: apatha8@emory.edu).

C. Marshall is with Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, USA (e-mail: curtis.e.marshall@emory.edu).

C. Davis is with Department of Surgery, Emory University School of Medicine, Atlanta, GA 30322, USA (e-mail: carolyndavis@emory.edu).

P. Yang is with Division of Pulmonary, Allergy, Critical Care, and Sleep Medicine, Atlanta, GA 30322, USA (e-mail: philip.yang@emory.edu).

R. Kamaleswaran is with Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, USA (e-mail: rkamaleswaran@emory.edu) and Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30308, USA.

R. Kamaleswaran was supported by the National Institutes of Health under Award Numbers R01GM139967 and UL1TR002378.

R. Kamaleswaran and C. Marshal were supported by Surgical Critical Care Initiative, funded through the Department of Defense's Health Program—Joint Program Committee 6/Combat Casualty Care (USUHS HT9404-13-1-0032 and HU0001-15-2-0001).

C. Davis is supported by National Institutes of Health (GM095442).

in the clinical notes. With the current development of the combination of machine learning with NLP, text-based feature extraction and classification have become more efficient.

ARDS identification has been evaluated using NLP [11], [12], but these methods do not use text-based features with standardized terminology. Recent NLP-based method map the text from radiology notes to key terms from the Unified Medical Language System (UMLS) to create concept unique identifiers (CUIs) [16], which are used to train a support vector machine with labels for ARDS diagnosis. Although these methods are effective, they are not yet fully generalizable and rely on keywords in context, which limits their ability to embed the underlying information from clinical notes for better ARDS identification.

Recent advances in NLP, use transformer-based models [17] to compute text representation with context. These methods give rise to large pre-trained language models that are efficient at understanding human language. Our work aims to evaluate the ARDS diagnosis pipeline using large pre-trained language models without making use of any specific list of keywords. Building on existing models, we hypothesized that a machine learning models can be developed using word embeddings to identify ARDS based solely on the language patterns being used in the radiology notes of patients, without relying on mapping to a specific definition or metathesaurus. This makes our proposed model generic enough to be efficiently applied to different datasets. [18] uses BERT based model with Hierarchical Attention Network with Sentence Objectives but they tend to overfit and fail to generalize.

Our contribution from the proposed model is that it achieves superior performance as compared to other baselines. We compare our results with previous methods and other supervised learning methods on two different datasets and observe that our model outperforms all the previous model by a significant margin. Our proposed model achieves 74.5% F1-score to other machine learning based models achieving 46.13% on F1-score. Additionally, our model is generic and can be easily validated on different datasets with very limited training data and does not overfit on the training data. Our text-based classification model further can be easily incorporated with other modalities like X-rays and other meta-data to further make the ARDS detection pipeline robust.

II. Method

A. Derivation dataset

We consider datasets from two distinct hospital systems for our analysis: Emory University (Atlanta, GA) and Grady Memorial Hospital (Atlanta, GA). The datasets consist of radiology notes for unique radiological reports of patients. For the analysis, we included a cohort of patients admitted to Grady and Emory dataset between September 8, 2014 - October 7, 2021 and February 26, 2017 – April 1, 2018, respectively.

B. Validation dataset

A subset of randomly selected cases were adjudicated for ARDS diagnosis by a physician reviewer (PY) with board certification in critical care medicine and significant experience in ARDS research and ARDS adjudication. The adjudication process involved manual chart review of laboratory data, clinical notes, and chest radiograph images within the selected encounters. The cases were annotated as true ARDS if (1) the patient had a qualifying P/F ratio < 300 while on mechanical ventilation and a qualifying chest imaging study (chest x-ray and/or chest CT with bilateral opacities) within 24 hours of each other and worsening respiratory status within 7 days of inciting event, and (2) the patient had reasonable laboratory data and/or clinical documentation to support that the respiratory failure was not fully explained by volume overload or hydrostatic pulmonary edema.

TABLE I
DATA STATISTICS FOR EMORY AND GRADY DATASETS.

	Grady Memorial Hospital	Emory University
Unique Patients (n)	216	146
Unique Radiological Reports (n)	6557	3323
ARDS Positive Radiological Reports (n)	2546	142

C. Description of the Data

The patient and encounter description for Emory and Grady datasets are provided in Table I. These patients all satisfy the sepsis-3 criteria applied through a retrospective algorithm executed on retrospective data from the Electronic Medical Record (EMR). The patients were adjudicated by physician to either have a positive or negative diagnosis for ARDS which were used for training the model. The notes include the date at which the radiological reports are created, the patient's medical record number (Patient ID), the encounter ID (a unique value to distinguish each of the radiological reports), the document code of a chest x-ray (if available for that patient), a set of 'Findings' and a set of 'Impressions'. We consider the combination of the Findings and Impressions as our input text. Findings consist of detailed observations that are made from the chest radiograph by the interpreting radiologist. Examples of these would include 'presence of bilateral infiltrates', 'patchy opacity' and 'possible presence of edema'. Impressions contain a summary of important observations (which often re-iterate certain elements of the Findings), as well as possible medical diagnoses that are likely to result in the chest radiograph findings.

D. Pre-Processing

The raw data consists of patient notes that have been compiled into a single document. We extract the relevant dataset by selecting encounter ID, patient ID, findings, impressions, and adjudicated labels for ARDS, and filter out entries that do not contain information in any of these fields. The resulting dataset is then stratified by patient ID and split into training and test sets to prevent any leakage between them. However, we observe a high imbalance between positive and negative instances in the Emory dataset, which can make it difficult for the model to predict the presence of ARDS accurately.

To mitigate this issue, we down-sample negative instances to maintain a reasonable skewness ratio of 1:3, while we do not do such down-sampling for the Grady dataset as it does not show such imbalance. Moreover, we remove punctuation from each row and apply stemming and lemmatization using the SpaCy[1] library for statistical methods. This preprocessing step helps eliminate noise, redundant words, and accounts for different word variations by identifying their main root.

E. Feature Extraction

Our proposed architecture utilizes BERT-based large language models, which are designed to learn the representation of a sentence by using attention. Specifically, BERT focuses on the most relevant information in the input and disregards irrelevant parts of the sentence. This is accomplished by assigning weights to each input token, which are computed by an alignment model that scores the match between the tokens. Additionally, BERT creates a context vector for each target word, which is then weighted based on its alignment with the input tokens. By using multiple stacks of attention networks, known as Transformers, BERT is able to better capture the nuances and context of the sentence, resulting in more accurate and meaningful representations.

To extract features and learn word embeddings from text entries, we utilize a pre-trained BERT model. Initially, we use a word-piece tokenizer to break words into sub-words, based on the pre-defined vocabulary for the BERT base uncased model. Next, each set of tokens is used as input to the BERT model to extract important features from the tokenized text. The BERT model's primary objective is to generate vector-representations of word usage in text-based data, known as word embeddings. These embeddings enable the model to provide a rich representation of the text data, allowing for accurate analysis and prediction.

Let X denote a text-based instance space and $Y = \{0, 1\}$ denote a label space. The goal is to learn function $h : X \to Y$ using a dataset $D = \{(x_i, y_i)\}_{i=1}^{N} \subset X \times Y$. A pretrained classifier M parameterized as $f(.;\theta)$ is fine-tuned on D. We define M as a function that takes as input a text sequence x and outputs a sequence of hidden states h:

$$M(x) = [h_1, h_2, ..., h_n]$$

where h_i is the hidden state corresponding to the i^{th} token in the input sequence. During fine-tuning, the model takes the input text sequence and passes it through a series of transformer layers to generate a sequence of hidden representations. The final hidden state corresponding to the [CLS] token is used as input to a fully connected layer, which maps the hidden state to the desired number of output classes. We take the final hidden state h_{cls} corresponding to the [CLS] token in the input sequence as a representation of the sequence embedding :

$$h_{cls} = h[CLS]$$

[1]https://spacy.io/

F. Classification

Next, we define the task-specific output layer as a function g that takes as input the final hidden state h_{cls} and outputs a vector of class probabilities p:

$$p = g(h_{cls})$$

where p is a vector of length C, where C is the number of output classes.

The output layer g is a fully connected layer with weights W and biases b, followed by a softmax activation function:

$$p = softmax(W * h_{cls} + b)$$

where

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

The parameters of the output layer (W and b) are trained using a cross-entropy loss function L to minimize the difference between the predicted class probabilities p and the ground truth labels y:

$$L(p, y) = -\sum_{i} y_i * log(p_i)$$

where y is a one-hot vector representing the ground truth label.

To regularize the parameters, we use a dropout layer before the linear layer and introduce ReLU as non-linearity. Our proposed architecture is shown in Figure 1.

III. Results

A. Patient and data characteristics

The Grady corpus includes clinical notes from 216 patients and 6557 unique encounters. Of these, 70 patients were ARDS positive, and we consider only notes recorded after ARDS detection as positive examples for classification. The remaining notes for ARDS positive patients and those for ARDS negative patients serve as negative examples. The Grady dataset has 2546 positive examples and 4011 negative examples for classification. ARDS patients in this dataset had a mean age of 48.98, compared to 42.67 for non-ARDS patients. Deaths among ARDS patients were more than twice as frequent as those among non-ARDS patients, and ARDS was more prevalent in male patients.

The Emory dataset includes clinical notes from 146 patients and 3323 unique encounters, of which 22 patients were ARDS positive. We use only notes recorded after ARDS adjudication as positive examples, resulting in 142 positive examples and 426 negative examples for classification. This small number of positive examples makes the Emory dataset highly skewed and difficult for the model to learn.

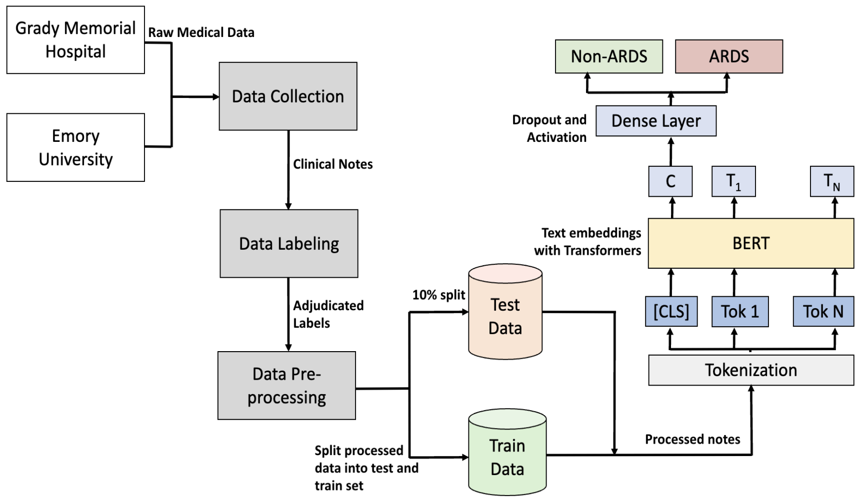

Fig. 1. Outline of our RespBERT architecture. The raw medical data is collected and clinical notes are selected. The selected notes are sent for adjudication to the clinicians for acquiring the true labels. The notes are then processed to remove punctuations, tags and converted to lower case. The data is further split into train and test splits. The processed splits are tokenized and passed to the BERT model for getting text embeddings from the notes using transformers. We use the dense layer with activations and dropouts on these embeddings to get the predicted probabilities from the model. ARDS presence is then predicted with these probabilities. For training, the loss is calculated and back propagated to adjust the training parameters of the model. For evaluation, the metrics are computed using the gold labels for the test data.

B. Discrimination and Calibration of NLP and Machine Learning Models

We compare our proposed method to several baseline machine learning algorithms :

- **SVM:** Support Vector Machines [19] is a machine learning algorithm that finds the best hyperplane to separate classes in a dataset by maximizing the margin.
- **GNB:** Gaussian Naïve Bayes [20] is a probabilistic classification algorithm that assumes features are independent and normally distributed.
- **RFC:** Random Forest Classifier [21] is an ensemble learning method that constructs multiple decision trees and combines their predictions to make a final classification. It uses a subset of features and data samples to build each tree and applies bagging and random feature selection to reduce overfitting.
- **XGBoost:** XGBoost [22] is a powerful ensemble learning method that uses gradient boosting to build a predictive model by iteratively adding decision trees to minimize a loss function.

These algorithms were implemented using scikit-learn's[2] package. The training data for these machine learning models consisted of the clusters that were generated through the PCA [23] in feature extraction, and the labels for the training data consisted of the adjudications for those data (provided by the physicians). The results are validated by evaluating these models on a separate held-out test data. We train machine learning models by using stratified 5-fold validation technique to calculate the confidence intervals.

To generate word embeddings from clinical notes, we utilized the skip-gram model of pre-trained word2vec [24]. We first removed stop words, except for common negation words such as no', nor', didn't', doesn't', and not'. Negation words were kept to learn negation patterns separately from positive ones, as they appear in close proximity to the current word in most notes. For instance, the phrases bilateral infiltrates', no bilateral infiltrates', doesn't contain bilateral infiltrates', and bilateral infiltrates are not present' all contain the term bilateral infiltrate'. The extracted features were then used as input to machine learning models in a binary classification task to predict the presence of ARDS.

We compared our results to the CUI-ARDS baseline [16] too, which uses UMLS named entity mentions to standardize language variations between radiologists. Each named entity mentioned was mapped to a UMLS concept unique identifier (CUI), and the CUIs vs. n-grams were input to SVM. To ensure a fair comparison between our method and the baseline, we used their pre-trained models on SVM and evaluated the performance. We also compared our results with HANSO [18] which uses BERT to obtain the embeddings and uses sentence objectives to design a hierarchical attention network. For fairness and generalizability, we use the same hyper-parameters as mentioned in the research work for our comparison.

C. Comparison between traditional model and NLP model

We evaluate our model's performance on a held-out test dataset that maintains the same positive-to-negative examples ratio as in the original dataset to ensure a representation of the real-world distribution of data. We find that large language models are better able to capture the implicit information from clinical notes to predict ARDS. Given the skewed nature of our dataset, we primarily evaluate our model using sensitivity and F1-score. Accuracy is not an appropriate evaluation metric because it does not give importance to false negatives and false positives.

[2]https://scikit-learn.org/

TABLE II
RESULTS FOR EMORY DATASET.

Algorithm	Sensitivity (%)	Positive Predictive Value (%)	F1-Score (%)
SVM	60 (43.1 – 76.9)	38.31 (30.9 – 45.7)	46.13 (36.9 – 55.3)
GNB	58.57 (46.5 – 70.6)	35.01 (25.3 – 44.7)	43.66 (32.9 – 54.4)
XGBoost	12.86 (4.7 – 21)	54 (22.6 – 85.4)	20.13 (8.1 – 32.1)
RFC	4.29 (-1.3 – 9.9)	30 (-9.2 – 69.2)	7.5 (-2.3 – 17.3)
CUI-ARDS	27.14 (20.3 – 34)	35.81 (25.5 – 46.1)	30.38 (23.4 – 37.3)
HANSO	80 (40.8 – 119.2)	20 (10.2 – 29.8)	32 (16.3 – 46.7)
RespBERT	75.14 (69.4 – 80.9)	75.25 (70.4 – 80)	74.5 (69.3 – 79.7)

TABLE III
RESULTS FOR GRADY DATASET.

Algorithm	Sensitivity (%)	Positive Predictive Value (%)	F1-Score (%)
SVM	53.18 (49.5 – 56.8)	59.38 (56.6 – 62.1)	56.07 (53 – 59.1)
GNB	54.7 (52.3 – 57.1)	59 (55.9 – 62.1)	56.76 (54.1 – 59.4)
XGBoost	44.86 (42.2 – 47.5)	62.42 (59.1 – 65.8)	52.18 (49.4 – 54.9)
RFC	48.68 (44.8 – 52.6)	64.26 (60.6 – 67.9)	55.34 (51.7 – 58.9)
CUI-ARDS	6.72 (4.5 – 8.9)	69.21 (60.2 – 78.2)	12.2 (8.5 – 16)
HANSO	99.84 (99.5 – 100.2)	38.38 (37.3 – 39.5)	55.43 (54.3 – 56.5)
RespBERT	61.24 (53.7 – 68.8)	70.33 (65.3 – 75.3)	64.22 (57.9 – 70.5)

To evaluate the efficacy of our proposed architecture, RespBERT, we compared it with existing baselines. We utilized a publicly available model called ClinicalBERT - Bio + Clinical BERT, which is pre-trained on electronic health records from ICU patients. We kept the BERT model parameters learnable and fine-tuned the embeddings using our train dataset. The resulting model was evaluated on test dataset. By leveraging ClinicalBERT and fine-tuning its parameters with our dataset, we were able to demonstrate the effectiveness of RespBERT in accurately identifying ARDS patients.

We perform our evaluation on both the Emory (Table II) and Grady (Table III) datasets. To test the generalizability of our model, we train it on the Grady dataset and test it on the Emory dataset. However, due to the differences in definitions and writing styles of clinical notes, it is challenging for the model to perform well in a highly limited data setting. To address this problem, we add an additional 30 examples from the Emory dataset to our training data to help the model learn to adapt better to the Emory dataset (Table IV). Our proposed model outperforms all other models by a significant margin on the training set.

Large language models perform well, especially in limited data and cross-dataset validation settings, highlighting the generalizability of the model and its better adaptation using very few annotated examples.

TABLE IV
RESULTS FOR MODEL TRAINED ON GRADY DATASET WITH ADDITIONAL 30 EMORY EXAMPLES AND TESTED ON EMORY DATASET.

Algorithm	Sensitivity (%)	Positive Predictive Value (%)	F1-Score (%)
SVM	0 (0 – 0)	0 (0 – 0)	0 (0 – 0)
GNB	2.86 (-0.6 – 6.3)	30 (-9.2 – 69.2)	5.2 (-1 – 11.4)
XGBoost	2.86 (-2.7 – 8.5)	13.33 (-13 – 39.5)	4.7 (-4.5 – 13.9)
RFC	4.29 (0.9 – 7.7)	40 (3.3 – 76.7)	7.7 (1.5 – 13.8)
HANSO	75.71 (38.3 – 113.2)	19.43 (9.9 – 29)	30.92 (15.7 – 46.1)
RespBERT	62.86 (46.1 – 79.7)	36.12 (33.6 – 38.7)	44.95 (39.1 – 50.8)

Figure 2 shows the AUROC curve and the Precision-Recall curve for the Grady and Emory datasets, comparing different models.

IV. DISCUSSION

Diagnosing ARDS is a complex task that requires consideration of multiple data points and disease characteristics to either rule in or rule out patients. The time-sensitive nature of ARDS makes it crucial to automate its diagnosis using clinical notes. However, the diverse range of definitions and practices for clinical note preparation necessitates a generalizable machine learning-based model for ARDS identification. In this study, we propose a scalable and generalizable NLP model that utilizes large language models. Our proposed model outperforms existing models with an improved F1 score from 46.13% to 74.5%. This achievement demonstrates the potential for using NLP models to aid in ARDS diagnosis and highlights the importance of developing generalizable models for clinical practice.

Our study utilized datasets curated from two distinct hospital systems to identify sepsis patients. Adjudication for ARDS was based on multiple criteria, including PaO_2/FiO_2 ratio, radiological and clinical reports. We also included patients with multiple etiologies for respiratory failure to increase the dataset size and incorporate real-world settings. To further enhance the dataset, we included encounters of potential ARDS patients provided by clinicians. We considered encounters ARDS-positive for our machine learning model only if the adjudication for ARDS was made before the clinical notes were recorded; otherwise, the encounter was considered ARDS-negative. Our approach ensured that our dataset was robust and representative of real-world ARDS patients, enabling us to develop a machine learning model that accurately identifies and diagnoses ARDS in clinical settings.

The identification of ARDS is challenging due to the highly imbalanced nature of the classification problem. As such, accuracy is not an accurate metric for evaluating performance. Instead, F1-score is the most crucial metric because a false negative result can have severe consequences given the potentially lethal nature of ARDS, but at the same time, predicting

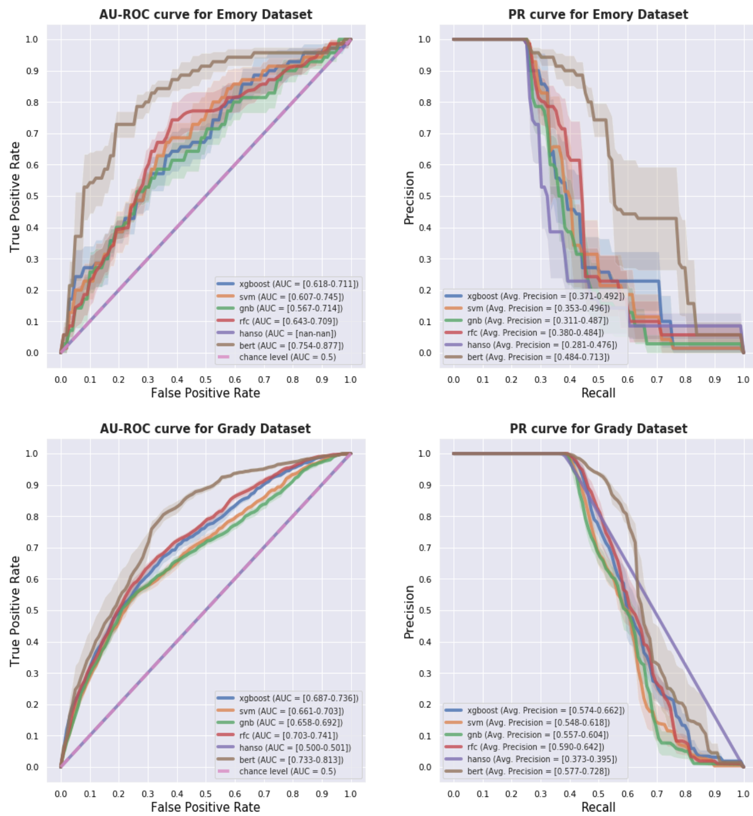

Fig. 2. We observe that BERT outperforms all other machine learning baselines for Emory dataset by a significant margin, which highlights the importance of using large pre-trained language models for highly skewed and limited datasets. In contrast, for Grady dataset, which is comparatively balanced and has more examples, the gap between BERT's performance and other machine learning models is smaller. However, BERT still outperforms other baselines by a significant margin, which demonstrates the effectiveness of our proposed model.

a lot of patients as ARDS obviates developing robust ARDS detection models. We compared our NLP-based approach with other machine learning models using two datasets (Table II and Table III). We also compared our model with a recent state-of-the-art NLP-based model. We found that computable phenotype-based baselines that utilize radiology notes tend to overfit and lack generalizability across a broad range of settings and datasets. In contrast, our proposed model achieved significantly better results than other baselines. Our approach demonstrates the potential for using NLP-based models to improve ARDS diagnosis, highlighting the importance of developing models that are both accurate and generalizable across diverse clinical settings. We also consider a Hierarchical Attention Network with Sentence Objectives (HANSO). We observe better sensitivity score with HANSO at the cost of positive predictive value. This shows the superiority of BERT based large language models for ARDS detection. However, HANSO ends up classifying a lot of notes as positive which undermines the very purpose of developing a robust ARDS detection model.

Our experiments showed that RespBERT outperforms all other methods by a significant margin, confirming the importance of using language models trained in the NLP domain for accurate ARDS diagnosis. The precision-recall curve and AUROC in Figure 2 further support the effectiveness of our proposed model.

Of all the machine learning models, SVM performs the best for both Emory and Grady datasets, as it is better able to learn from the available features. GNB also performs well, despite the imbalanced nature of the datasets. In contrast, other machine learning models tend to overfit and are not able to learn effectively in the limited data setting. The Emory dataset presents a more challenging learning environment due to its greater imbalance, leading to worse performance for all models compared to the Grady dataset. We also observed that RespBERT outperforms other models more significantly on the Emory dataset than the Grady dataset.

We found that CUI-ARDS did not perform well on either dataset, likely because it relies on specific keywords that may not be standard across clinicians. In contrast, the RespBERT model demonstrated robust performance and was not affected by overfitting, even in the limited data and imbalanced data scenarios.

HANSO performs well on sensitivity but performs poorly on F1-score which makes HANSO non-deployable in a real world setting.

To evaluate the generalizability of our proposed model, we conducted an additional experiment where we trained our model on the Grady dataset and only included 30 examples from the Emory dataset for training, and then tested the model on the remaining Emory dataset. The results of this experiment are presented in Table IV, and we found that our proposed model outperformed all of the baselines by a significant margin, providing further evidence of its generalizability. Unfortunately, we could not compare our method with CUI-ARDS because we did not have a trained model available. The poor performance of the baselines on the Emory test set can be attributed to the highly imbalanced nature of the dataset, as well as the fact that the clinical notes in Emory differ significantly from those in Grady. In contrast, RespBERT was able to achieve good results even in this challenging setting.

Figure 3 depicts the most frequent unigrams and bigrams that are associated with ARDS in the dataset. Traditional feature-based models without appropriate embeddings may not adequately represent the complex information contained in the clinical notes. This is due to the presence of many common n-grams in both ARDS and non-ARDS clinical notes, which makes it challenging for skip-gram models to capture the relevant information. In contrast, sub-word level embedding models such as BERT are more effective in incorporating information through transformers and thus perform better than traditional models for classification tasks. This is because BERT is capable of embedding information over longer sequences, allowing it to capture the intricate relationships between words and phrases in the clinical notes.

Figure 4 presents the most important features for Emory and Grady dataset obtained from our trained model. We use Captum[3] for fetching the most important features for predictions. We observe an overlapping of the important features for positive and negative predictions too which highlights the importance of more adjudicated samples and strategical selection of notes for adjudication to further improve the model's learning.

In this study, we have demonstrated the efficacy of RespBERT in predicting the presence of ARDS using clinical notes. With the increasing amount of data being collected in ICUs, machine learning is becoming an essential tool for research and clinical practice. Machine learning provides powerful methods for identifying patterns in data that can predict outcomes such as ARDS, particularly when these patterns are complex and nonlinear. Using BERT for ARDS prediction has the advantage of better generalizability, allowing the model to perform well even with limited data, which can be helpful in reducing adjudication costs.

For future work, we plan to improve the model's performance by intelligently selecting encounters to be adjudicated, which can be most beneficial for the model to learn from challenging input clinical notes. We also plan to incorporate active learning into our proposed method to achieve more reliable results for ARDS identification. This can help to reduce the need for manual adjudication while achieving better performance in ARDS prediction. By continuing to explore and develop machine learning methods for ARDS identification, we can help ICU clinicians to make better-informed decisions and improve patient outcomes.

[3]https://captum.ai

Fig. 3. To gain insight into the most commonly used language in clinical notes related to ARDS, we analyzed the uni-grams and bi-grams present in both positive and negative instances of the Grady and Emory datasets. We observed that many uni-grams and bi-grams occur in both positive and negative instances, suggesting that the presence of these terms alone may not be enough to accurately predict ARDS. Additionally, we found notable differences in the n-grams used between the Grady and Emory datasets, which may be attributed to variations in clinical note-taking practices across different hospital systems and environments.

V. References

[1] L. B. Ware and M. A. Matthay, "The acute respiratory distress syndrome," *New England Journal of Medicine*, vol. 342, no. 18, pp. 1334–1349, 2000.

[2] M. Diamond, H. L. Peniston, D. Sanghavi, and S. Mahapatra, "Acute respiratory distress syndrome," in *StatPearls [Internet]*. StatPearls Publishing, 2022.

[3] G. Bellani, J. G. Laffey, T. Pham, E. Fan, L. Brochard, A. Esteban, L. Gattinoni, F. Van Haren, A. Larsson, D. F. McAuley *et al.*, "Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries," *Jama*, vol. 315, no. 8, pp. 788–800, 2016.

[4] A. Agrifoglio, J. Lopez, J. Figueira, M. Hernández, L. Fernández, M. Irazabal, S. Yus, and M. Jiménez, "Acute respiratory distress syndrome-the berlin definition: impact on an icu of a university hospital," *Critical Care*, vol. 17, pp. 1–200, 2013.

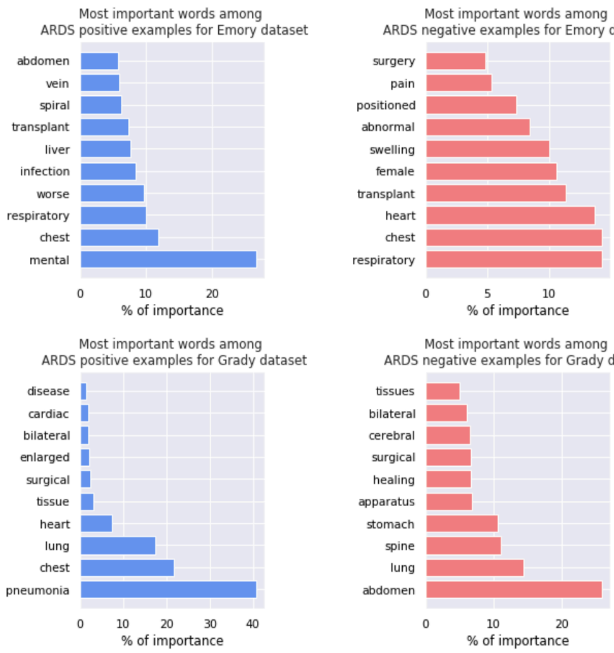

Fig. 4. We analyzed the most important tokens or words (features) learnt by the model to predict the presence of ARDS. While there are notable differences in Grady and Emory dataset, we can observe few overlaps in the most important features for positive and negative predictions.

[5] V. Herasevich, M. Yilmaz, H. Khan, R. D. Hubmayr, and O. O. Gajic, "Validation of an electronic surveillance system for acute lung injury," *Intensive Care Medicine*, vol. 35, pp. 1018–1023, 2009.

[6] H. C. Koenig, B. B. Finkel, S. S. Khalsa, P. N. Lanken, M. Prasad, R. Urbani, and B. D. Fuchs, "Performance of an automated electronic acute lung injury screening system in intensive care unit patients," *Critical care medicine*, vol. 39, no. 1, pp. 98–104, 2011.

[7] H. C. Azzam, S. S. Khalsa, R. Urbani, C. V. Shah, J. D. Christie, P. N. Lanken, and B. D. Fuchs, "Validation study of an automated electronic acute lung injury screening tool," *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 503–508, 2009.

[8] X. Su, A. Mayampurath, M. Churpek, S. Shah, B. Patel, D. Dligach, and M. Afshar, "External validation of an acute respiratory distress syndrome prediction model using clinical text," in *A25. ARDS: NEW APPROACHES TO DIAGNOSIS AND TREATMENT*. American Thoracic Society, 2020, pp. A1129–A1129.

[9] V. Fanelli, A. Vlachou, S. Ghannadian, U. Simonetti, A. S. Slutsky, and H. Zhang, "Acute respiratory distress syndrome: new definition, current and future therapeutic options," *Journal of thoracic disease*, vol. 5, no. 3, p. 326, 2013.

[10] A. C. McKown, R. M. Brown, L. B. Ware, and J. P. Wanderer, "External validity of electronic sniffers for automated recognition of acute respiratory distress syndrome," *Journal of intensive care medicine*, vol. 34, no. 11-12, pp. 946–954, 2019.

[11] I. Solti, C. R. Cooke, F. Xia, and M. M. Wurfel, "Automated classification of radiology reports for acute lung injury: comparison of keyword and machine learning based natural language processing approaches," in *2009 IEEE international conference on bioinformatics and biomedicine workshop*. IEEE, 2009, pp. 314–319.

[12] M. Yetisgen-Yildiz, C. A. Bejan, and M. Wurfel, "Identification of patients with acute lung injury from free-text chest x-ray reports," in *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, 2013, pp. 10–17.

[13] C. S. Calfee, D. R. Janz, G. R. Bernard, A. K. May, K. N. Kangelaris, M. A. Matthay, and L. B. Ware, "Distinct molecular phenotypes of direct vs indirect ards in single-center and multicenter studies," *Chest*, vol. 147, no. 6, pp. 1539–1548, 2015.

[14] V. M. Castro, D. Dligach, S. Finan, S. Yu, A. Can, M. Abd-El-Barr, V. Gainer, N. A. Shadick, S. Murphy, T. Cai *et al.*, "Large-scale identification of patients with cerebral aneurysms using natural language processing," *Neurology*, vol. 88, no. 2, pp. 164–168, 2017.

[15] E. Joffe, E. J. Pettigrew, J. R. Herskovic, C. F. Bearden, and E. V. Bernstam, "Expert guided natural language processing using one-class classification," *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 962–966, 2015.

[16] M. Afshar, C. Joyce, A. Oakey, P. Formanek, P. Yang, M. M. Churpek, R. S. Cooper, S. Zelisko, R. Price, and D. Dligach, "A computable phenotype for acute respiratory distress syndrome using natural language processing and machine learning," in *AMIA Annual Symposium Proceedings*, vol. 2018. American Medical Informatics Association, 2018, p. 157.

[17] 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.

[18] K. Lybarger, L. Mabrey, M. Thau, P. K. Bhatraju, M. Wurfel, and M. Yetisgen, "Identifying ards using the hierarchical attention network with sentence objectives framework," in *AMIA Annual Symposium Proceedings*, vol. 2021. American Medical Informatics Association, 2021, p. 823.

[19] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[20] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[22] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[23] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.