Group3: Use of Multi-faceted Data Preprocessing and Word Embedding Methods for Deep Neural Network Models to Improve Sepsis Prediction Performance

Jihoon Kim[1], Shubham Throat[2], Aakriti Kedia[2], Manasi Agrawal[2]
[1]Health Department of Biomedical Informatics, [2]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA

Abstract

Sepsis is the body's extreme response to infection. When untreated, it damages the patient's own tissue and organs, leading to death at the rate of 30%. Hence, early recognition of sepsis is critical to improve diagnosis and timely management of this deadly disease. Previous risk scores and prediction algorithms for sepsis mostly used only structured clinical features with mixed performance and limited generalizability. In this study, we used MIMIC, a large database of unidentified patients who stayed in the critical care units of a large Academic Medical Center, dataset to build a sepsis prediction model. We added the unstructured text data from the radiology report as new features for the prediction algorithm using natural language processing and word embeddings pretrained on biomedical corpus. We also leveraged source data element which each feature was extracted from, cardinality, and missing rate to apply optimal preprocessing methods, instead of mere replacement of null values by average. We built neural network models with hyperparameter tuning and evaluated its performance on the holdout test set. We report the model performance under varied setting of model, inclusion of unstructured feature, and embedding method.

Introduction

Sepsis is the body's overwhelming and life-threatening response to infection that can lead to tissue damage, organ failure and ultimately, death [1]. It is body's overactive and toxic response to an infection and requires rapid diagnosis and treatment [2]. Sepsis can progress to severe sepsis when there are signs of organ dysfunction like lung problems, low or no urine output(kidney failure), abnormal liver tests and changes in mental status(brain). All such patients require treatment in the intensive care unit(ICU). Sepsis is responsible for the highest cost of hospitalization in US, estimated to be around $62 billion dollars annually for acute sepsis and its treatment. This accounts only for a part of the costs since there are many charges post-recovery as well. Studies investigating survival and death rates due to sepsis have reported that on average, approximately 30% of people diagnosed with severe septic shock do not survive and upto 50% of patients recovering from sepsis suffer from post-recovery side effects [3]. There is no proper cure for sepsis and until it is found, early detection and treatment of sepsis is essential for survival and limiting the side effects post-recovery. Some studies have shown that early detection of sepsis and rapid antibiotic treatment increases the chances of recovery but even a delay of 3hrs to 6hrs can increase the mortality rate by at least 8% [1]. Unfortunately, it is tough to diagnose sepsis since health deterioration along with organ failure is common with many other diseases. Also, a lot of other factors influence the symptoms of sepsis like age, gender, and comorbidities.

Artificial Intelligence and Machine Learning have been used to detect sepsis early for a long time to help increase the survival rate [4, 5, 6]. These techniques can use both structured data and unstructured data that is available in the form of electronic health records (EHR). Structured clinical data follows a fixed data model and values set like age, gender, vitals, laboratory tests, etc. In contrast, unstructured clinical data comes in different forms like physician's notes, images, nursing notes, or discharge summaries. This form of data contains a lot of inconsistencies like abbreviations, grammatical or spelling errors. Pre-processing this data is extremely important. It is a time-consuming process since natural language processing(NLP) is required to extract all relevant features. Prior studies have shown that the use of unstructured data along with structured data improves the model performance to detect or predict some diseases, including sepsis [2, 7].

The main challenge in the current research is the use of different features or physiological factors and the use of efficient machine learning algorithms for the diagnosis and prediction of sepsis [8]. In order to predict sepsis in advance, it is very crucial to choose appropriate features and design algorithms that will suit the clinical setting. The input features to the model are physiological factors like vitals (heart rate, oxygen saturation, body temperature),

biomarkers, and demographic values like age, gender, etc. The output of this machine learning algorithm would predict if the patient would suffer from sepsis after several hours.

The machine learning algorithms which have been tested for sepsis prediction include support vector machine, random forest, logistic regression, gradient boosting trees, neural networks, or a combination of these. Some researchers [9] experimented with five machine learning models on three different feature sets - biomarkers, EHR data and a combination of both. Amongst the algorithms, support vector machine and Adaboost had highest AUC scores on the data having a combination of EHR and biomarkers. Meanwhile, recurrent neural networks were tested [6] on the sepsis data provided by Medical Information Mart for Intensive Care (MIMIC) III [10, 11]. The authors compared the performance against the Insight algorithm [12], an early sepsis warning algorithm. They used specific ranges of attributes like heart rate, temperature, blood pressure, etc to train the models and showed that recurrent neural networks (RNN) outperformed the Insight algorithm.

In this study, we use a publicly available ICU dataset to build and evaluate neural network model performance under different setting of carefully designed data prepocessing based on the feature charactersitics, word embedding, and different input feature types (structured vs. unstructured) for sepsis prediction. The goal is to figure out the best combination of setting that prediction model would perform best on the holdout dataset.

Methods

The workflow of this study is illustrated in Figure 1. Below we explain how the final output of area under the receiver operating characteristic (AUC) curve plot was obtained from the raw input MIMIC database.

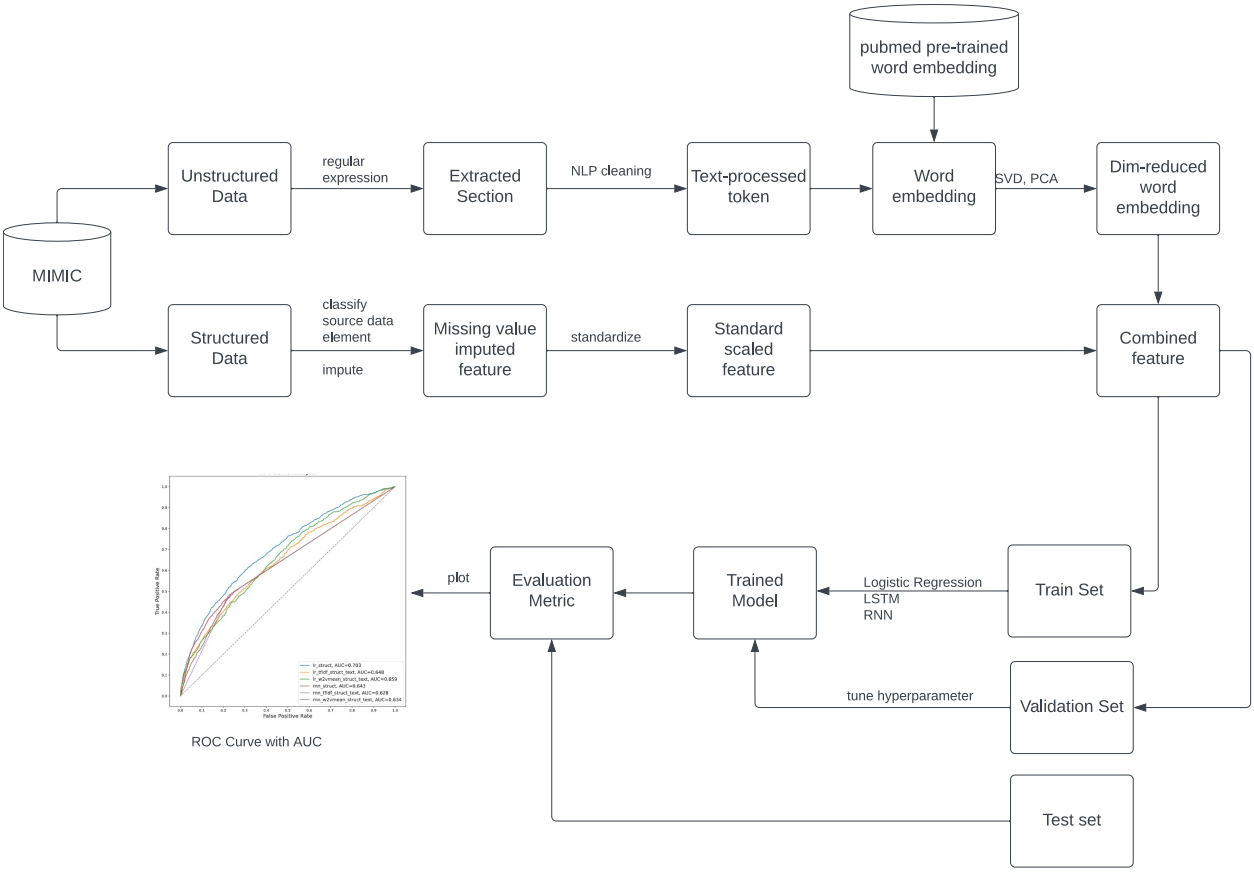

Figure 1: Schematic diagram of study workflow

We used MIMIC, a large database of unidentified patients who stayed in the critical care units of a large Academic Medical Center, Beth Israel Deaconess Medical Center (BIDMC) affiliated with Harvard Medical School, between 2001 and 2012 [10, 11]. The database has multiple data elements such as demographics, vital sign measurements made at the bedside, laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality. In addition to the structured clinical data, we also used free-text radiology reports that summarize the findings from the chest X-ray images in Digital Imaging and Communications in Medicine (DICOM) format during routine clinical care [11].

The raw training data set was annotated to assign data elements to predefined categories manually by authors. For instance, while aspartate aminotransferase (AST) and bilirubin were from the lab test category, heart rate and blood pressure were from the physical exam category. To avoid data leakage, an 80%/20% split of the train/validation set of the raw training set was conducted at patient level. Also, to conserve the proportion of positive outcomes, stratified sampling was applied to perform random split within each label stratum (positive or negative). Then, for each feature, its cardinality, missing rate, and unique values were assessed. There are many ways to fill in the missing values like imputation or mean filling methods. However, the optimal data preprocessing methods need to be chosen according to the characteristics of the data set in use. For example, consider the feature **on_insulin**, which came from the medication category, has possible values of 1 if the patient received insulin but otherwise 0. A naive application of replacing null values with mean values or numbers other than 0 or 1 would be adding unnecessary noise and bias, negatively impacting classifier performance. Standard scaling was applied to have a mean of 0 and a standard deviation of 1 for all numeric features after preprocessing. The hold-out test set followed the same preprocessing method as the training set.

In the raw data, the radiology reports were stored as a single column called **TEXT**. Since this is a whole note, two sections, Findings and Impression, were identified and extracted using a regular expression. A pre-trained word2vec model trained on PubMed data set is used as a word embedding. Various experiments were conducted to get the best results from word embeddings. First, the embeddings of each word present in the text were flattened. They were either truncated or padded with extra values to bring all the vectors to the same size. In the second experiment, each note was divided into several tokens and the mean of all these individual tokens was taken to generate a final vector of size 300. In the third experiment, the term frequency inverse document frequency (TFIDF) method was used to construct text-based features.

We used two different models for the sepsis prediction task. First, we used the sklearn implementation of Logistic Regression. It is a supervised classification algorithm and performs well over binary output prediction tasks. As our dataset was skewed with a lot of training samples having no sepsis results and very few samples showing sepsis, we applied class_weight="balanced" to ensure that equal weight was assigned to both label classes (0 for no sepsis, 1 for sepsis). We employed optimization techniques to train within an acceptable amount of time as the number of training samples along with the structured and unstructured features generates a huge training set. We limited the number of iterations to 1000 to reduce the training time as well as allow the model to learn sufficient information to come up with correct predictions. We also adopted dimensionality reduction techniques like Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) for restricting the number of structured dimensions, appending the processed reports column and using these columns as features in the model. Through the experimentation, we found that the original model without dimensionality reduction techniques yielded the best results.

We also trained both RNN and Long Short Term Memory (LSTM) deep learning models as as they can handle sequence data while memorizing historical data. A sequential keras model with three hidden layers along with an output layer was used. The layers of our network model are described as follows.

- Input layer - The number of neurons was set to 512. It takes an input of 1 * the embedding size (the number of features = 465). Dropouts were used to prevent overfitting.
- Hidden layers - We generated two hidden layers. The first layer with 1024 neurons and the next layer with 512 neurons. In both of these layers, we employed dropouts. Different experiments gave the most optimal dropout value of 0.2
- Flatten layer - We used this layer to reshape our input into a single column for our next Dense layer.

- Dense layer - This layer took the input from the Flatten layer and yielded a single binary output stating if a patient has sepsis or no. The activation function used was 'ReLU'. Reasons for choosing ReLU are as follows.
  - Reduced likelihood of the gradient to vanish
  - Constant gradient of ReLUs resulted in faster learning
  - ReLUs generates sparse representations which are more beneficial in training

The final architectures of neural network models are illustrated in Figure 2 for RNN and Figure 3 for LSTM.

Layer (type)	Output Shape	Param #
simple_rnn_18 (SimpleRNN)	(None, 1, 512)	500736
dropout_18 (Dropout)	(None, 1, 512)	0
simple_rnn_19 (SimpleRNN)	(None, 1, 1024)	1573888
dropout_19 (Dropout)	(None, 1, 1024)	0
simple_rnn_20 (SimpleRNN)	(None, 1, 512)	786944
dropout_20 (Dropout)	(None, 1, 512)	0
flatten_6 (Flatten)	(None, 512)	0
dense_6 (Dense)	(None, 1)	513

Figure 2: Model architecture of recursive neural network (RNN)

Layer (type)	Output Shape	Param #
lstm_3 (LSTM)	(None, 1, 512)	2002944
dropout_3 (Dropout)	(None, 1, 512)	0
lstm_4 (LSTM)	(None, 1, 1024)	6295552
dropout_4 (Dropout)	(None, 1, 1024)	0
lstm_5 (LSTM)	(None, 1, 512)	3147776
dropout_5 (Dropout)	(None, 1, 512)	0
flatten_1 (Flatten)	(None, 512)	0
dense_1 (Dense)	(None, 1)	513

Figure 3: Model architecture of Long Short Term Memory (LSTM) network

Binary cross-entropy loss function and Root Mean Squared Propagation (RMSprop) optimizer were used to measure the accuracy of the model. The model was trained for 20 epochs and monitor the training and the validation loss. The model performance was evaluated with Area Under the Receiver Operating Characteristic (AUROC) values and its curve in figures.

Results

The raw training set had 97,512 examples from 1,972 unique patients with 287 structured features, 1 unstructured feature in text format, and 1 binary outcome (sepsis). The number of examples per patient ranged from 1 to 402 and with mean 49.4 and standard deviation 43.6. The raw train set had 1,972 unique patients and 629 or 31.9% experienced at least one event of sepsis. While 588 had 8 records of sepsis, 1,343 had no sepsis record. To prevent data leakage, train/validation (80%/20%) data split was conducted at patient level. Also, stratified sampling was applied to preserve the same proportion of positive labels. This resulted in 5% proportion of positive label, sepsis, in both train and validation data sets. In order to characterize the feature and make an informed decision of missing value imputation method, the category of data elements where each features was derived from was assigned. In total, 287 structured features came from nine categories (Figure 4). As expected, the 50.2% (144 out of 287) features came from lab test category, followed by comorbidity (19.9%) and physical exam (19.5%). Notably, the risk score category had 11 risk scores of either of Sequential Organ Failure Assessment (SOFA) or Systemic inflammatory response syndrome (SIRS) that was derived from other features [13, 14].

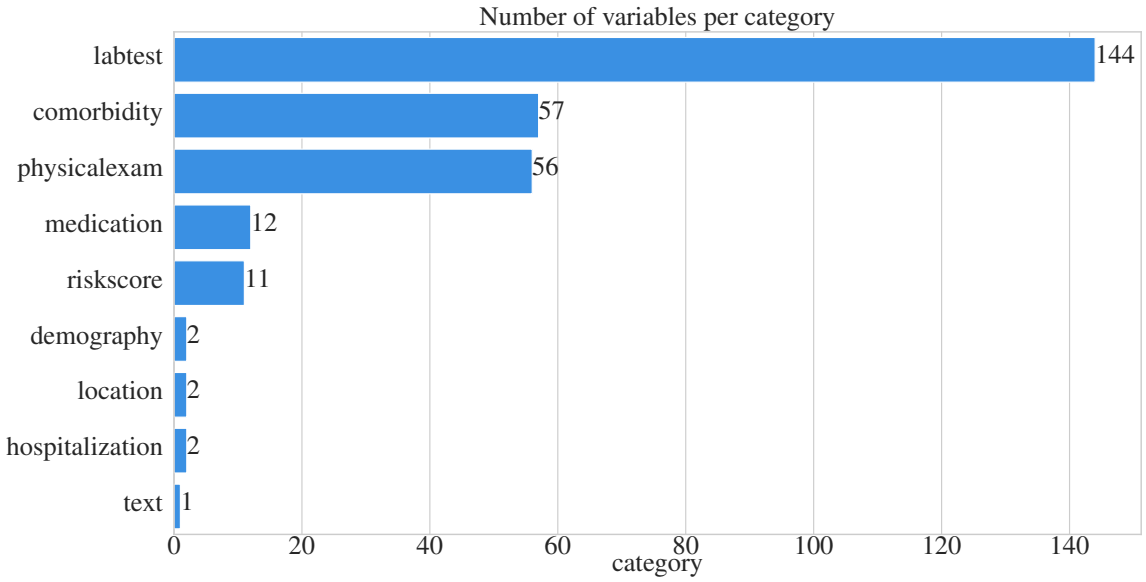

Figure 4: Bar plot of source data element category of features

Based on the feature characteristics of data element category, data type, cardinality and missingness of features, four types of missing value preprocessing were defined and corresponding number of features in each type is shown in Figure 5. The largest type was one with features having cardinality greater or equal to 2, largely consisted of lab test features, and imputed with mean value of non-missing values in each feature.

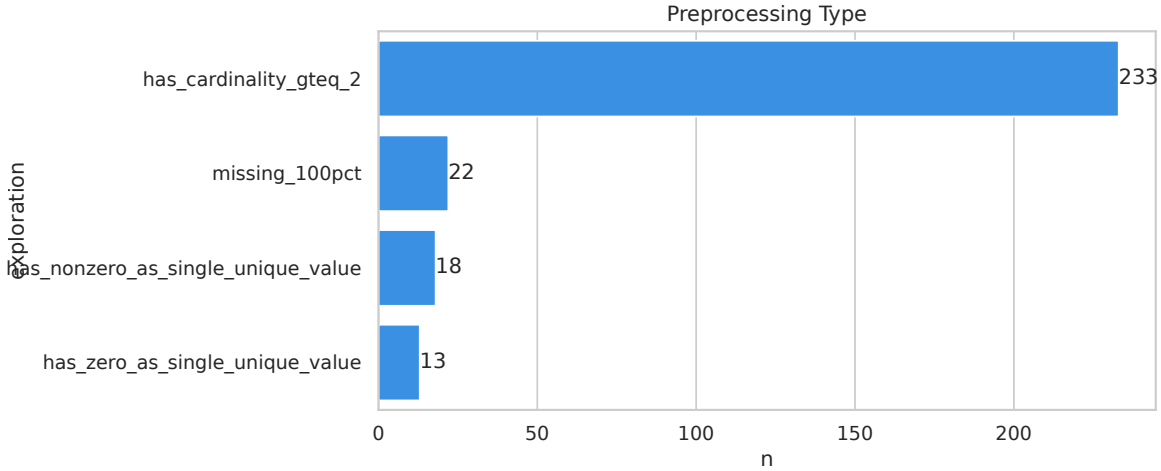

Figure 5: Bar plot of features counts by preprocessing type

The 22 features with 100% missing were discarded (Figure 6). The 18 features with a single non-zero unique value were imputed with Gaussian noise around that unique value as a mean. The 13 features with a single value was imputed with Gaussian noise $N(0, 0.02)$.

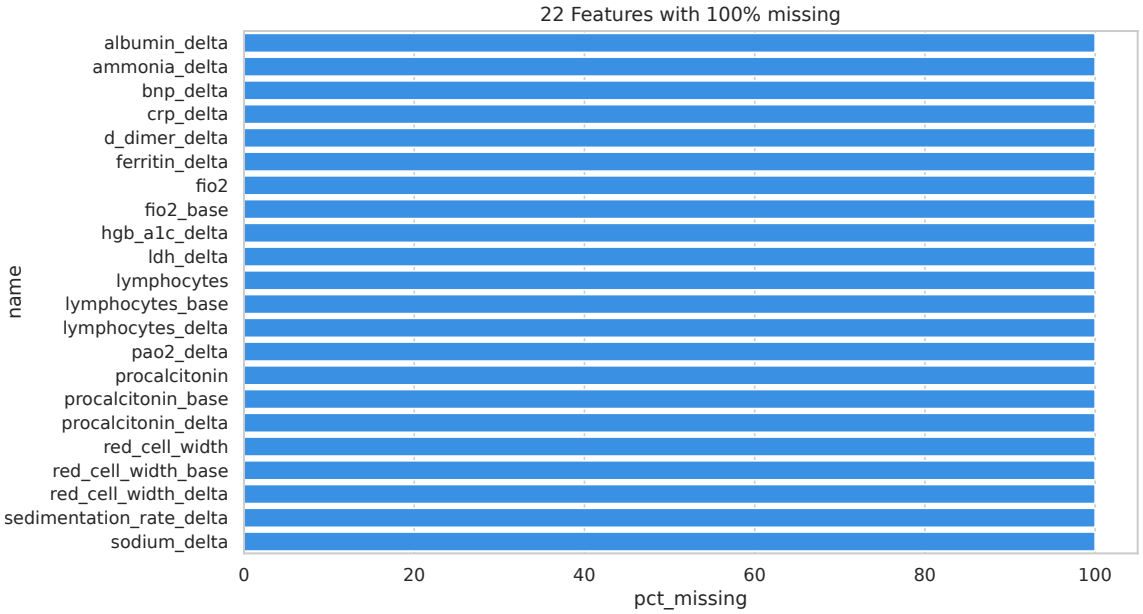

Figure 6: Bar plot of 22 features with 100 % missing values

Figure 7 shows the forest plot of 60 single-feature logistic regression results with p-value less than 0.0005 and the absolute value of log odds ratio less than 10, sorted by log odds ratio. Shock, pneumonia, bacteria, and high temperature were associated with positive log odds in line with previous findings. Of the 11 risk scored features, only perfusion score and cardiovascular (CV) SOFA score were included in the forest plot. Although not shown in the forest plot (due to its large magnitude), use of medication Prostacyclin, a class of vasodilator drugs, had largest positive log odds, 821.5 in association with sepsis, which is in line with literate findings [15].

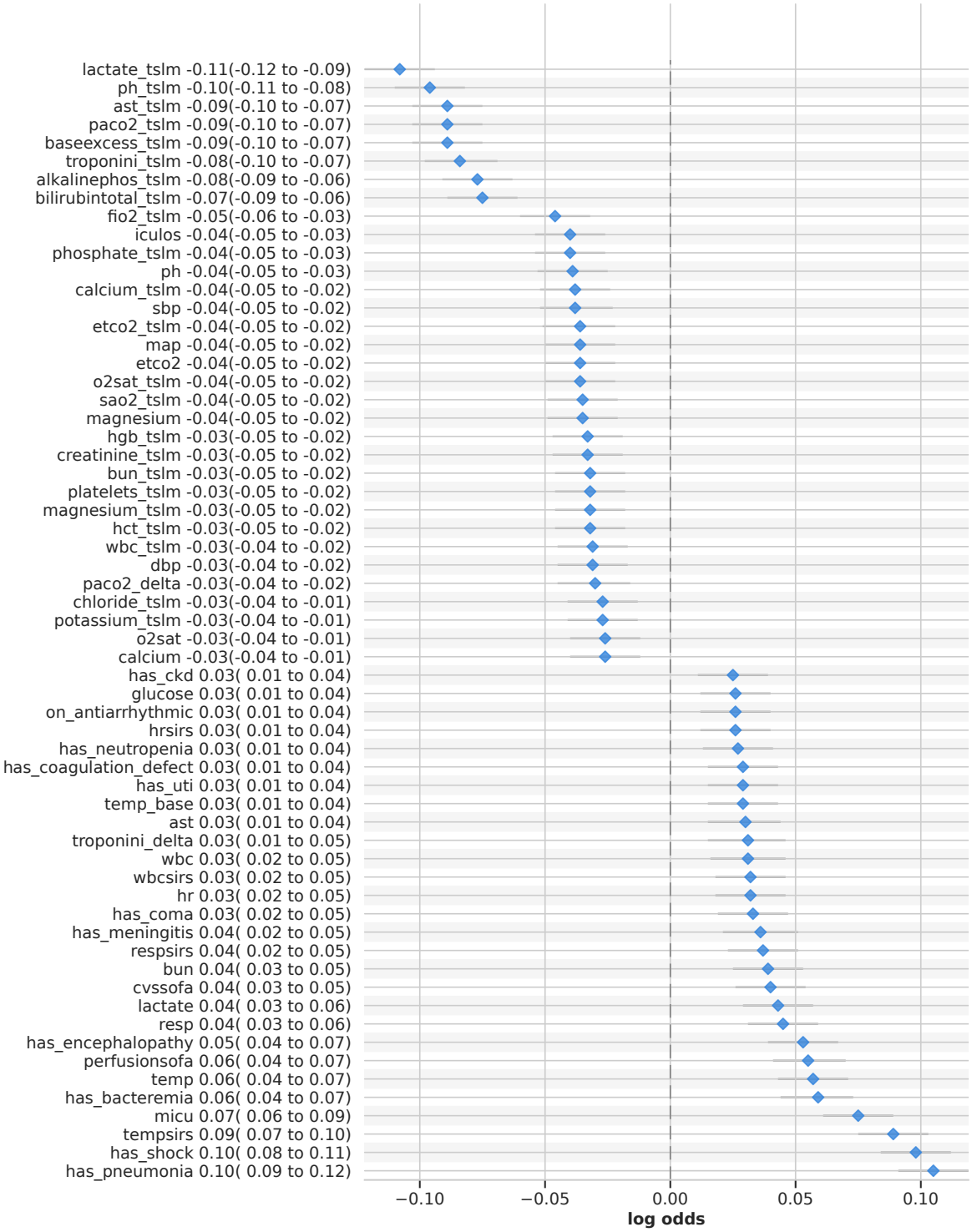

Figure 7: Forest plot of single-featured logistic regression models

Table 1 shows the results of model performance under different setting of model type, input feature type, and embedding method. The best performance was achieved by logistic regression (LR) using only structured input features. Among neural network models, Long Short Term Memory (LSTM) outperformed Recursive Neural Network (RNN) under all settings. In embedding methods, use of TFIDF and word2vec achieved higher AUC Values in reference fo mean word2vec. At predefined cutoff of prediction probability 0.5, LSTM with TFIDF had highest accuracy. Recall was highest in RNN with no embedding using only structured input features.

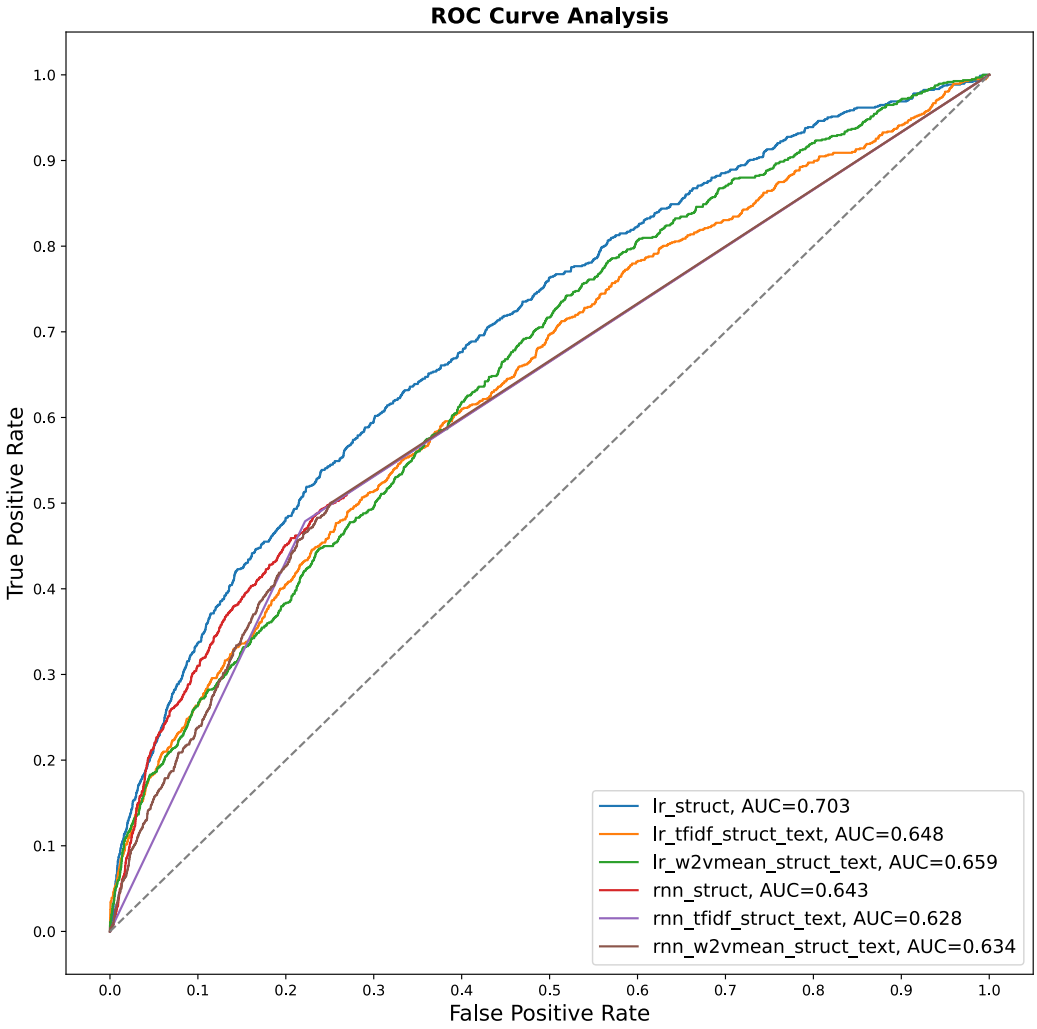

Figure 8: ROC Curve of model evaluated on the holdout test set. AUC: Area Under the Receiver Operating Characteristic Curve (AUC), LR: Logistic Regression, RNN: Recursive Neural Network, struct: Structured features, TFIDF: Term Frequency Inverse Document Frequency, W2V: Word-To-Vector

Model	Structured Input Feature	Unstructured Input Feature	Embedding	AUC	ACC	Recall
LR	Y	N	none	**0.703**	0.744	0.531
LSTM	Y	N	none	0.631	0.829	0.320
RNN	Y	N	none	0.643	0.722	**0.506**
LR	Y	Y	Mean w2v	0.659	0.724	0.461
LSTM	Y	N	Mean w2v	0.586	0.836	0.261
RNN	Y	Y	Mean w2v	0.634	0.740	0.492
LR	Y	Y	TFIDF x w2v	0.648	0.722	0.481
LSTM	Y	N	TFIDF x w2v	0.585	**0.868**	0.243
RNN	Y	Y	TFIDF x w2v	0.634	0.740	0.479

Table 1: Model Performance. ACC: Accuracy, AUC: Area Under the Receiver Operating Characteristic Curve, N: not used as input feature: N/A: Not Applicable, LR: Logistic Regression, TFIDF: Term Frequency Inverse Document Frequency, RNN: Recursive Neural Network, W2V: Word-to-Vector, Y: used as input feature

Discussion and Concluding Remarks

Our models were evaluated on the basis of 3 parameters AUC score, Accuracy, and Recall. The MIMIC dataset we used is imbalanced in the favor of cases having non-sepsis, which makes accuracy an unreliable metric. The recall was directly proportional to the number of actual cases that have sepsis and are correctly predicted which was a good indicator of model performance. The Logistic Regression approach worked better on a structured part of the dataset whereas a deep learning-based approach called RNN was used to handle the sequential part (radiology reports) of the dataset. This approach worked better on this dataset even after incorporating the unstructured dataset as the length of reports is too short for the RNN model to capture enough information. The RNN was developed to work with sequential datasets, we utilized this model to make predictions based on the radiology reports. We faced some limitations due to the way the reports are structured which will be explained in later paragraphs.

The first technique used was the flattening of tf-idf which comes with its own limitation as it introduced a huge amount of features that make the model easily overfit. If the final vector size is too small then it can result in a feature loss and hence underfit the data. The dimensionality reduction technique could be used to get a final vector of reasonable length that can capture a lot of information through this method without exploding the number of features. The two variants of word2vec used in this paper (mean w2v and tfidf) kept the number of features uniform for all the text in the corpus without too much of data loss. utilizing the tf-idf along with word2vec doesn't affect the AUC much but gives some improvement in the recall, as it is able to predict the sepsis cases more accurately.

The drawback of this approach is that the tf-idf based approach gives more importance to the words that are seen only in the training set making it biased towards the words that are new in the test dataset. whereas if the training set has rich coverage of different data points it can give the best results. With a reasonably extensive training set, it can capture tf-idf values of all the words effectively. As we can see from the results the word embedding used to train the models plays a very important role in getting good results. To further reduce the noise and capture the importance of the words in radiology a different word embedding can be used like a pre-trained clinical Bert. There is also scope for improvement by using the doc2vec method to learn the vectorized form directly from the reports rather than learning it from individual words.

Acknowledgments We thank MED 277 course instructors, Drs. Mike Hogarth and Shamim Nemati, for their clear teaching of clinical NLP and deep learning, helpful guidance and constructive comments. We also appreciate course TAs, Jonathan Lam and Fatemeh Amrollahi, for data preparation and positive feedback on the project.

References

1. Yan MY, Gustad LT, Nytrø Ø. Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review. Journal of the American Medical Informatics Association. 2022;29(3):559-75.

2. Amrollahi F, Shashikumar SP, Razmi F, Nemati S. Contextual embeddings from clinical notes improves prediction of sepsis. In: AMIA Annual Symposium Proceedings. vol. 2020. American Medical Informatics Association; 2020. p. 197.
3. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). Jama. 2016;315(8):801-10.
4. Nemati Zargaran F, Akya A, Ghadiri K, Ranjbarian P, Rostamian M. Detecting the Dominant T and B Epitopes of Klebsiella pneumoniae Ferric Enterobactin Protein (FepA) and Introducing a Single Epitopic Peptide as Vaccine Candidate. International journal of peptide research and therapeutics. 2021;27(4):2209-21.
5. Fagerström J, Bång M, Wilhelms D, Chew MS. LiSep LSTM: a machine learning algorithm for early detection of septic shock. Scientific reports. 2019;9(1):1-8.
6. Scherpf M, Gräßer F, Malberg H, Zaunseder S. Predicting sepsis with a recurrent neural network using the MIMIC III database. Computers in biology and medicine. 2019;113:103395.
7. Liu R, Greenstein JL, Sarma SV, Winslow RL. Natural language processing of clinical notes for improved early prediction of septic shock in the ICU. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2019. p. 6103-8.
8. Schinkel M, Paranjape K, Panday RN, Skyttberg N, Nanayakkara PW. Clinical applications of artificial intelligence in sepsis: a narrative review. Computers in biology and medicine. 2019;115:103488.
9. Taneja I, Reddy B, Damhorst G, Dave Zhao S, Hassan U, Price Z, et al. Combining biomarkers with EMR data to identify patients in different phases of sepsis. Scientific reports. 2017;7(1):1-12.
10. Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3(1):1-9.
11. Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng Cy, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data. 2019;6(1):1-8.
12. Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, et al. A computational approach to early sepsis detection. Computers in biology and medicine. 2016;74:69-73.
13. Li W, Wang M, Zhu B, Zhu Y, Xi X. Prediction of median survival time in sepsis patients by the SOFA score combined with different predictors. Burns & Trauma. 2020;8.
14. Usman OA, Usman AA, Ward MA. Comparison of SIRS, qSOFA, and NEWS for the early identification of sepsis in the Emergency Department. The American journal of emergency medicine. 2019;37(8):1490-7.
15. Zardi EM, Zardi D, Dobrina A, Afeltra A. Prostacyclin in sepsis: a systematic review. Prostaglandins & other lipid mediators. 2007;83(1-2):1-24.