Exploiting Text Data to Improve Critical Care Mortality Prediction*

Bryan Auslander
Knexus Research Corporation
National Harbor, MD, USA
bryan.auslander@knexusresearch.com

Kalyan Gupta
Knexus Research Corporation
National Harbor, MD, USA
kalyan.gupta@knexusresearch.com

Michael W. Floyd
Knexus Research Corporation
National Harbor, MD, USA
michael.floyd@knexusresearch.com

Sam Blisard
Navy Center for Applied Research in AI
Naval Research Laboratory
Washington, DC, USA
samuel.blisard@nrl.navy.mil

David W. Aha
Navy Center for Applied Research in AI
Naval Research Laboratory
Washington, DC, USA
david.aha@nrl.navy.mil

Abstract—**There has been a significant increase in the quantity, quality, and availability of unstructured clinical notes, motivating numerous machine learning approaches that leverage such data to improve predictive capabilities in medical settings. However, the question of whether patient group properties under observation influence the effectiveness of including unstructured data sources remains unanswered. The inclusion of unstructured clinical notes adds both an acquisition cost such as recording the notes by a clinician and converting records to an appropriate digital format, and a computational cost such as more complex and computationally expensive machine learning algorithms. Thus, it is important to understand the potential benefits offered by these unstructured data sources before attempting to use them. We empirically evaluate the performance impact of including unstructured clinical notes when performing mortality prediction by reproducing 29 previously published studies in this area. We use two common feature extraction methods, Word2Vec and Bag-Of-Words, with two existing machine learning models, XGBoost and Logistic Regression. Our results show that our approaches have significantly different performances depending on the properties of the patient group under study. Additionally, we identify several key findings that can be used to predict whether the inclusion of data from unstructured clinical notes will be beneficial based on properties of the patient groups.**

Index Terms—**Natural Language Processing; Feature evaluation and selection; Machine Learning; Modeling structured, textual and multimedia data**

I. Introduction

Electronic health records have increased the volume, availability, and quality of clinical patient notes [1]. These clinical patient notes contain important details about the patient from the clinicians' (i.e., physicians', practitioners', or medical staff's) perspective and may include details such as issue assessments, treatment management, and administrative information. They may contain valuable information that cannot be found in numeric and qualitative laboratory observations. As such, computational tools may require information stored in these clinical patient notes to provide the most complete and accurate diagnosis or treatment. Recent work (e.g., [2], [3]) has begun to examine how this information can be extracted and utilized by machine learning (ML) algorithms. For example, many of these efforts (e.g., [4], [5]) leverage temporal data and use recurrent neural networks to incorporate this longitudinal information.

However, to our knowledge, *the impact of patient group properties, like patient inclusion criteria, notes available per patient, or experiment duration on ML performance using clinical patient notes* has not been rigorously investigated. Using clinical patient note data adds two types of complexity: data collection cost (i.e., the difficulty in recording, collecting, or obtaining useful patient notes from clinicians) and algorithm complexity (i.e., the additional complexity of extracting and using unstructured data compared to structured laboratory results). Thus, it is crucial to understand how different patient groups, and their individual properties, effect ML performance. Having such information will allow clinicians and researchers to better understand if their patient group will potentially benefit from clinical patient note data, and that the benefits outweigh the costs.

Specifically, we empirically evaluated the effects of patient group properties on the task of *patient mortality prediction* when clinical notes are used for model training and prediction. Our investigation reproduces the results of a previous patient mortality prediction reproducibility study [6] of 38 distinct published experiments with the MIMIC III dataset [7]. Using the software and data from these original experiments, we conducted additional experiments that use both structured laboratory data (i.e., as was done in the original studies) and unstructured clinical patient notes.

The contributions of our work are as follows:

- An extension of a previous reproducibility study on published patient mortality prediction experiments [6] that includes results from using clinical patient notes to train ML algorithms.

* We thank the Office of Naval Research for supporting this research.

- An extensive comparison of several existing algorithms for textual feature extraction, all of which have been commonly used in previous medical research, on the described experiments.
- Several key findings related to patient group properties and how they impact ML performance and the benefit of including features from unstructured clinical notes. Although some of our findings seem intuitive, to the best of our knowledge, no previous work has empirically confirmed them.
- Insight for clinicians as to the predictive performance they may obtain based on their patient group and whether there is a potential benefit to using quantitative numeric data such as laboratory results together with text data such as clinical patient notes. For example, if certain patient groups have shown no positive benefits from using clinical patient notes for mortality prediction, a clinician would know that there is limited benefit to offset the cost of collecting additional clinical notes concerning their patients.

In the remainder of this paper, we describe our experimental approach and our key findings. In Section II, we describe the patient groups, taken from existing studies, that we use for our experiments. Section III describes both the structured (i.e., laboratory data) and unstructured (i.e., clinical notes) features we use, and Section IV presents the ML algorithms we use. Our results and analyses are presented in Section V, followed by a description of related work in Section VI and concluding remarks in Section VII.

II. Study Patient Groups

The patient groups we selected were based on the meta-study described by [6]. This meta-study used the publicly available Medical Information Mart for Intensive Care (MIMIC III) dataset [7], which includes approximately 60,000 ICU admissions with longitudinal records spanning over a decade. In this meta-study, the MIMIC III dataset was used to reproduce 38 distinct studies which vary in terms of data inclusion, observation time windows, outcome definitions, and inclusion criteria. Although the inclusion criteria differed between each individual study, four exclusion criteria were common to all. All studies removed records (1) for patients under 15 years old, (2) with incomplete data for admission and discharge dates, (3) for organ donors, and (4) for stays of less than 4 hours. It should be noted that we did not select these exclusion criteria ourselves, but instead used the criteria used by the existing studies.

In addition to these exclusions, we also applied one other universal exclusion: *removal of any record in which the patient was discharged or died during or within 12 hours of the observed window*. Since both of these situations provide nearly trivial predictions (e.g., someone who died during observation does not need a predictive system to determine their mortality), they were removed to make the prediction task realistic. This allowed us to emulate several studies discussed in [6] (e.g., [8], [9], and [10]) that did not include patients that died during their observed ICU admission. On this basis, we took this observation one step further and removed patients who were discharged during the observed window, since discharge is a strong negative indicator of mortality. Finally, the buffer of an additional 12 hours was added to account for unstructured features from clinical notes that tend to provide similar obvious indicators of mortality or survival. For example, in an initial pilot study we discovered certain words in clinical notes had a substantial influence on mortality prediction. From a prediction standpoint, having clinical notes that list a patient as being *dead* or having an *autopsy* are useful for identifying obvious (i.e., already deceased) cases but improbable in real-world continuous prediction tasks. Similarly, *discharge* provides a nearly trivial indication that the patient will be discharged. Having explored various window lengths, we settled on removing patients that die or are discharged within 12 hours of the observed window (i.e., they are already very near death or release during observation). This removes the majority of the unrealistic highly-predictive words (i.e., words that describe the events of death or discharge) and ensures that the quantitative features are not unfairly biased with predictive feature values as compared to the textual features. We believe this is a fair comparison as it focuses on longer-term mortality prediction rather than just predicting when mortality occurred.

III. Patient Features

We used the same set of quantitative data features (i.e., laboratory data) reported by [6] in their reproducibility study. Depending on the individual laboratory data item, the feature values consist of the first, last, minimum, maximum, and summation for the data values during the observation window (although not all data items may report all feature values). The time window used for these values was 24 or 48 hours, or all available (depending on the individual study) and may include values collected outside the ICU if the patient was not in the ICU for the full duration of observation. In summary, the quantitative data contains 93 total features, which we will refer to as the *structured features*. The original paper provides full details of the features used (omitted here for brevity).

In addition to the structured features, we extracted information from the clinical notes for each ICU stay. Our data extraction method used two common methods to extract numeric feature representations from natural language text: *Bag-Of-Words (BOW)* and *Word2Vec*. For both approaches, the clinical notes were preprocessed[1] to remove tags, punctuation, multiple characters of white space, numerics, stopwords, words less than 3 characters long, and performed stemming.

Bag-Of-Words: For the BOW approach, the entire preprocessed text corpus is analysed to create a dictionary. The dictionary contains the frequency of each word and the percentage of documents each word occurs in. Any words that appeared less than 5 times in the corpus were labelled as rare words and removed from the dictionary. Similarly, any word that appeared in 50% or more of the documents

[1] Using the Python library Gensim [11].

were labelled as non-informative common words and removed from the dictionary. Thus, the remaining words occur in a sufficient subset of documents (i.e., at least 5), but are not so common that they provide little discriminatory power. Of the remaining words in the dictionary, the 3000 most frequent words were kept[2]. For each patient record, the frequency of the 3000 dictionary words was computed, resulting in a BOW feature vector containing 3000 numeric values. Since each patient record may contain multiple clinical documents, the values are summed across all documents.

Word2Vec: Word2Vec is an alternative feature representation approach for textual data and can improve prediction task performance compared to BOW. Word2Vec analyzes text to leverage context by estimating and using the semantic similarity among words. The specific Word2Vec model we use is Continuous Bag-Of-Words, which learns a model to predict a target word based on the contextual words that surround it in text. For data extraction, we use the five words surrounding the target word on both sides (i.e., 10 total words, possibly with padding used if the word is at the start or end of a document) as input and the target word as output. The underlying autoencoder used to train the Word2Vec model contains 300 nodes in the hidden layers[3]. After training, the Word2Vec model was extracted by maintaining the input layer and the 300 hidden nodes representing the dense encoding which we used as the features for our models. For patient records containing multiple clinical documents, the Word2Vec representation is a vector which is a mean of all the Word2Vec document representations.

Combined Feature Vectors: The structured and unstructured features were concatenated together to create combined feature vectors. When BOW is used, there are 3093 total features (93 structured features + 3000 unstructured features). When Word2Vec is used, there are 393 total features (93 structured features + 300 unstructured features).

IV. Machine Learning Methods

The motivation and primary contribution of this work is not to develop a novel ML algorithm but instead to (1) examine the performance of ML algorithms on mortality prediction when using both quantitative laboratory data and text data in clinical notes, and (2) determine the effect of patient group selection on mortality prediction performance. We use existing ML algorithms for this task. To more accurately compare our findings to the baseline results presented by [6], we used the ML models used in prior studies: XGBoost [12] and Logistic Regression (from Scikit-learn [13]). XGBoost is a gradient boost algorithm that produces an ensemble of weak prediction models, in our case trees. The gradient boosting model attempts to fit its model by minimizing a loss function, using techniques like gradient decent. Logistic Regression is a linear model that assumes all features are independent predictors of the target variable.

These ML models are trained to perform a binary classification based on patient mortality (i.e., the patient will die or will not die within a fixed time interval). The inputs to the algorithms are the extracted feature vectors, as described previously, representing the structured and unstructured clinical data. A subset of training instances that contain both the input features and mortality labels are used to train the ML models. Given a feature vector representing a patient with an unknown (or hidden) mortality, the ML models predict whether the patient will die within a specified time interval.

V. Results

In this section, we describe our study's design, evaluation approach, the empirical results, and key takeaways.

A. Evaluation Approach/Study Design

Our evaluation is based on 29 of the 38 reproduced studies reported in [6]. We could not reproduce some of the studies. One study, *Che2016recurrent_a*, uses a 48 hour window and a death indicator of 48 hours post-admittance (i.e., predicting if death occurs during hospitalization). Another, *Luo2016predicting*, uses a unique start window of 12 hours after ICU admittance. In addition, 7 other studies used non-standard observation windows. For these reasons, we did not include these studies in our experiments. A summary of the key properties of the 29 studies we did use in our experiments is shown in Table III. Although we do not provide full details on each, some of the key differences include the length of observation (*W* of 24 hours, 48 hours, or all available time), the number of patients in the patient group (*patients*), the number of patient records (*records*), the mean number of patient records per patient ($\#notes$), and the mortality prediction window (*flag*). The types of mortality prediction windows include: death in hospital (*hospital_expire*), death 30 days after ICU discharge (*30dy_post_icu_disch*), death 30 days after hospital discharge (*30dy_post_hos_disch*), death 6 months after hospital discharge (*6mo_post_hos_disch*), death 1 year after hospital discharge (*1yr_post_hos_disch*), and death 2 years after hospital discharge (*2yr_post_hos_disch*). For each of the 29 studies, we conducted experiments using a 5-fold cross validation procedure and calculated the AURUC (Area Under the Receiver Operating Characteristics) curve to measure performance.

For the two ML algorithms, XGBoost and Logistic Regression, we performed experiments using the following five feature vector representations:

- **Original Only (O)**: The only features used were the original quantitative/numeric features (93 features).
- **Bag-Of-Words Only (BOW)**: The only features used were the ones generated by the Bag-Of-Words method from textual data sources (3000 features).
- **Word2Vec Only (W2V)**: The only features used were the ones generated by the Word2Vec models from textual data sources (300 features).

[2] If n words remained in the dictionary and $n < 3000$, then n words would be retained. However, in practice, there were always far more than 3000 words remaining in the dictionary

[3] We performed experiments varying the number of nodes from 100 to 3000 but found minimal improvement with more than 300 nodes.

- **Original and Bag-Of-Words (O+BOW)**: The original features and the BOW features appended together (3093 features).
- **Original and Word2Vec (O+W2V)**: The original features and the Word2Vec features appended together (393 features).

B. Experiment Results

The average Area Under the Receiver Operating Characteristic (AUROC) performances of the two ML algorithms, averaged over all 29 experimental studies, are shown in Table I. The key conclusion from these results is that, when examining the results of all studies, using the textual features on their own results in a performance decrease compared to numeric features, and using a combination of structured and unstructured features generally yields a small but often statistically significant improvement compared to using only numeric features.

TABLE I
AUROC Results for each ML Algorithm

ML Algorithm	0	BOW	W2V	O+BOW	O+W2V
XGBoost	0.86	0.79	0.74	0.87	0.86
Linear Regression	0.84	0.69	0.76	0.74	0.84

These results appear to indicate that there is only a small benefit to using unstructured data sources for mortality prediction using our data extraction methods. However, as we discussed previously, each of these studies used different patient groups and/or varied the prediction task (e.g., observation window and predicted time of death). We hypothesize that these differences between studies also effect the relative performance as a result of including unstructured features.

Table II displays the performance of the algorithms based on observation period duration (24 hours, 48 hours, or all data for the entire hospitalization time). These results are the increase (positive values) or decrease (negative values) in AUROC versus using the original features only as well as the reduction (positive values) or increase (negative values) in error percentage. Statistically significant (using a paired t-test with $p < 0.05$) improvements are displayed in bold. The mixture of Bag-of-Words features derived from text data and the original numeric features yield statistically significant improvements across all categories when using XGBoost, and decreases across all categories when using Logistic Regression. The results using a mixture of Word2Vec and original features were less conclusive, and did not yield statistically significant differences. These Word2Vec findings are similar to those reported in [2], who also found limited improvement using Word2Vec for extracting features from text data. One notable finding is that, when textual features do increase an ML algorithm's performance, as the observation window increases in length, the inclusion of unstructured data decreases error (i.e., significantly increases prediction accuracy).

TABLE II
Empirical Results Partitioned by Observation Time Length

	BOW + Original		Word2Vec + Original	
Experiment Type	AUROC Increase	% Error Reduction	AUROC Increase	% Error Reduction
24-XGB	**0.011**	**6.930**	0.003	1.905
24-LogReg	-0.091	-50.648	0.005	2.809
48-XGB	**0.011**	**6.988**	-0.000	-0.130
48-LogReg	-0.122	-73.057	-0.002	-1.034
All-XGB	**0.010**	**10.348**	0.004	4.506
All-LogReg	-0.082	-78.791	0.006	5.550

Table III displays a performance summary and the details for all 29 studies using the XGBoost algorithm with a combination of the Bag-Of-Words and original features. The rows are sorted by the % error reduction that resulted when using those features compared to using only the original numeric features. Additionally, we performed experiments examining the performance of both XGBoost and Logistic Regression using all five feature combinations (Original, Bag-Of-Words, Word2Vec, Original+Bag-Of-Words, and Original+Word2Vec). Figures 1 to 3 show more detailed results from these experiments using the XGBoost classifier (the Logistic Regression results were omitted for space). In these figures, the experiment number corresponds to the experiment number (*exp #*) shown in Table III.

A noteworthy conclusion is that the patient group selection can significantly impact the ML algorithms' AUROC performance when using clinical notes. The inclusion of clinical note features resulted in performance changes from approximately -3% to +18%, with only one study increasing the error. Looking at the ordering, we see that the 9 studies with the largest error reduction percentage were all for longer-term mortality prediction (i.e., predicting the patient will expire between 30 days and 2 years after leaving the hospital). This may imply that clinical notes provide valuable information that is useful in predicting longer-term health outcomes, although additional research is necessary to validate this hypothesis.

When we examine the results in terms of the net change in AUROC performance, we see a range of -0.005 to +0.020. Like our error reduction results, only one study showed a decrease in AUROC when using the combination of Bag-Of-Words and original features. The best performing study, *ghassemi2015multivarite_b*, used a patient selection method that is well suited for our approach. During patient selection, the study excludes patients who have fewer than 6 clinical notes, fewer than 100 non-stop words in the notes, were in the hospital for less than 24 hours, or did not have a SAPS value. On the other side of the performance spectrum, *joshi2012prognostic* and *hug2009icu* show the worst performance. Both of these studies use the same patient group and differ in the mortality prediction task being performed (death in hospital vs. death 30 days post-ICU discharge). Thus, we conclude that the patient group has a greater effect on performance than the prediction task. Similarly, *hug2009icu* performs significantly worse than the other longer-term predictive studies, providing further evidence to support the importance of patient group selection.

Based on the high-performance on the *ghas-*

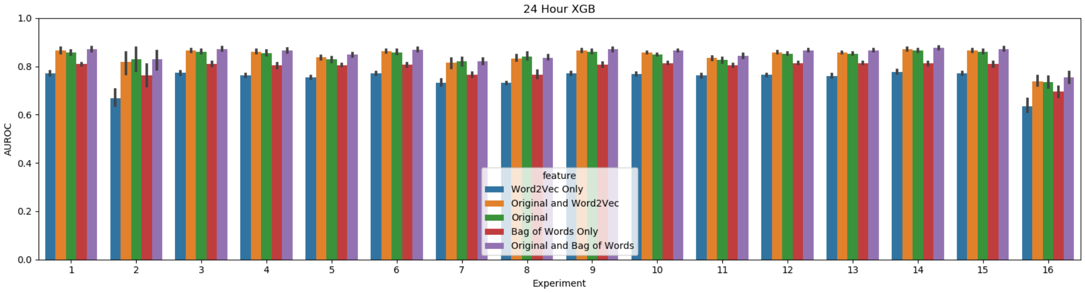

Fig. 1. The average AUROC performance of the XGBoost algorithm across experimental studies with a 24 hour observation window. The results show how the performance changed depending on the features used: original (structured) features only, Word2Vec (unstructured) features only, Bag-Of-Words (unstructured) features only, Original and Word2Vec (structured and unstructured) features, and Original and Bag-Of-Words (structured and unstructured) features.

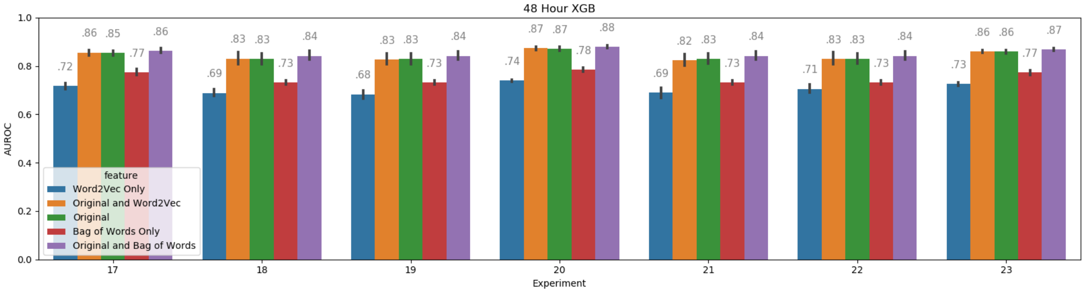

Fig. 2. The average AUROC performance of the XGBoost algorithm across experimental studies with a 48 hour observation window. The results show how the performance changed depending on the features used: original (structured) features only, Word2Vec (unstructured) features only, Bag-Of-Words (unstructured) features only, Original and Word2Vec (structured and unstructured) features, and Original and Bag-Of-Words (structured and unstructured) features.

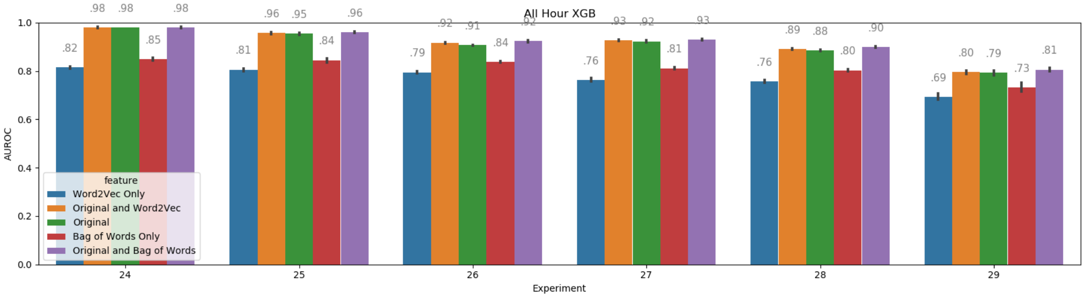

Fig. 3. The average AUROC performance of the XGBoost algorithm across experimental studies with an observation window of all available data. The results how the performance changed depending on the features used: original (structured) features only, Word2Vec (unstructured) features only, Bag-Of-Words (unstructured) features only, Original and Word2Vec (structured and unstructured) features, and Original and Bag-Of-Words (structured and unstructured) features.

semi2015multivarite_b study, we examined in greater detail the impact of clinical note availability on prediction performance. When comparing error reduction percentage to the number of patients in the group, we found a moderate positive correlation between them (a Pearson correlation coefficient of 0.45). Similarly, comparing the error reduction percentage to the number of clinical records in the patient group shows a moderate positive correlation (a Pearson correlation coefficient of 0.44). Finally, comparing the increase in AUROC to the mean number of records per patient, we also found a moderate positive correlation (a Pearson correlation coefficient of 0.49). All of these results, considered together, provide evidence that the inclusion of unstructured features is more beneficial for larger patient groups with a higher number of patient records per patient. These findings support our conjecture that text data features are more beneficial when more text data is available.

C. Summary of Key Findings

The following are our key findings for the mortality prediction task using the MIMIC III dataset:

- Using only structured numeric or quantitative laboratory data outperformed using only unstructured text data.

- Using a combination of structured and unstructured data outperformed using only structured data across most test conditions.
- The benefit from a Bag-Of-Words representation for unstructured features is strongly related to the specific ML algorithm used.
- There was no noticeable benefit of using a Word2Vec representation for unstructured features, regardless of ML algorithm.
- Patient group selection had a substantial effect on increased AUROC and decrease in percentage prediction error.
- When including text features, error reduction significantly increased with observation window duration.
- Inclusion of text features appeared to provide more benefit for longer-term mortality prediction tasks.
- The highest-performing experiment used explicit exclusion criteria that removed patients with few or less-dense clinical notes.
- Error reduction was positively correlated to the number of patients in the patient group and the total number of records contained in the patient group.
- AUROC improvement was positively correlated to the mean number of records per patient in the patient group.

VI. Discussion and Related Work

Recently, several papers have examined the use of textual clinical data for ML tasks. Similar to our work, [2] evaluated how the clinical note representations impact various classification and prediction tasks. Their work compares Bag-Of-Words, Word2Vec, and Recurrent Neural Network encodings on MIMIC III data. Their findings showed that the Bag-Of-Words representations tended to outperform Word2Vec and Recurrent Neural Network representations, which is similar to our results. [14] took a similar approach for a baseline mortality prediction task from the MIMIC III dataset. They use the feature extraction/construction pipeline from [15] to obtain 17 structured baseline features, and then use clinical notes to generate several representations including Doc2Vec, applying medical text tokenization with Doc2Vec, and using a neural network for named entity extraction/negation [5]. These features were then evaluated using a Long-Short Term Memory (LSTM) and Multimodal approach, resulting in a small improvement in AUROC (approximately 2%, similar to our findings). [4] made use of an LSTM for in-hospital mortality estimates using a Word2Vec embedding combined with a Convolutional Neural Network for feature extraction with clinical notes. They also found that exclusively using features extracted from unstructured sources were inferior to exclusively structured features. They reported improvements only when the data were examined as a time series.

Our results are consistent with those reported by other researchers on the overall, global applicability of features extracted from unstructured clinical notes. Although they often examined different individual patient groups, different feature extraction methods, and different ML algorithms, their results are generally in agreement with our findings and show that there is benefit to using unstructured features. However, the primary difference between what we present in this paper and these prior reports is that we focused more locally on how patient group selection impacts ML performance. Thus, the combined results that we've summarized (i.e., ours and those reported in related work) provide more compelling evidence on the value of using features from both structured and text data sources. Additionally, our work also provides preliminary evidence on how properties of a patient group may impact how valuable features from unstructured sources may be toward ML performance.

Our work has shown results exclusively on English language datasets for evaluating textual features. Dashtipour et al. [16] provides a summary of a number of language representation techniques used for multilingual sentiment analysis. The textual features we make use of fall under their corpus-based approaches and are compiled from the actual subject matter. As such, we believe the results shown in this paper are likely to be similar when data from other languages are used once a Bag-Of-Words dictionary or Word2Vec model is built. A limiting factor for evaluating the use of clinical notes in other languages is that they are not as readily available. If we identify a medical dataset in another language we plan to evaluate the effectiveness of this approach versus other methods such as normalizing the data via translation.

VII. Conclusion

We explored the effect of patient group selection on the performance of ML algorithms that use text data from clinical documents for mortality prediction. Our work builds on [6], who reproduced a series of mortality prediction studies with the MIMIC III dataset, by extending their results to include unstructured features rather than purely structured numeric features. We found a small but significant improvement in ML performance when including structured features and, more importantly, that performance varied significantly depending on the properties on the patient group. We identified several key factors that appear to be useful indicators of whether considering text data features will benefit a specific patient group, thereby allowing clinicians to better understand the potential data requirements and expected performance of ML systems for mortality prediction. However, many of our findings are preliminary and will require a larger study on whether these patient group properties are accurate predictors of ML performance and whether our findings hold beyond the mortality prediction task. Additionally, we plan to explore the relative explainability and interpretibility of models, and whether more interpretable models are more readily usable by clinicians [17].

References

[1] H. B. Burke, L. L. Sessums, A. Hoang, D. A. Becher, P. Fontelo, F. Liu, M. Stephens, L. N. Pangaro, P. G. O'Malley, N. S. Baxi *et al.*, "Electronic health records improve clinical note quality," *Journal of the American Medical Informatics Association*, vol. 22, no. 1, pp. 199–205, 2015.

TABLE III

Detailed Experiment Result Summary for XGBoost Algorithm with Bag-Of-Words + Original Features. *W* = window size, *auroc inc* = increase in AUROC, *error reduction* = percent error reduction, and $\overline{\#notes}$ = mean notes per patient

exp	exp #	W	vocab	auroc	auroc inc	error red	$\overline{\#notes}$	flag	patients	records
joshi2012prognostic	8	24	28797	0.836	-0.005	-3.135	3.672	hospital_expire	8765	32188
hug2009icu	7	24	28797	0.821	0.001	0.280	3.672	30dy_post_icu_disch	8765	32188
celi2012database_b	2	24	11927	0.830	0.001	0.822	7.012	hospital_expire	843	5911
grnarova2016neural_a	24	all	58510	0.981	0.000	1.763	5.023	hospital_expire	21615	108564
joshi2016identifiable	23	48	79302	0.867	0.008	5.854	8.410	hospital_expire	25006	210300
caballero2015dynamically_b	17	48	47643	0.863	0.009	6.176	6.949	hospital_expire	11569	80396
wojtusiak2017c	29	all	56595	0.805	0.013	6.274	6.660	30dy_post_hos_disch	17186	114466
harutyunyan2017multitask	20	48	61045	0.880	0.008	6.464	5.267	hospital_expire	21169	111489
lee2015customization_a	9	24	59871	0.871	0.010	7.085	5.977	hospital_expire	20593	123087
ripoll2014sepsis	16	24	22531	0.755	0.019	7.304	10.803	hospital_expire	2010	21715
hoogendoorn2016prediction	6	24	54988	0.869	0.010	7.346	6.203	hospital_expire	17275	107165
johnson2012patient	21	48	27674	0.842	0.013	7.487	6.864	hospital_expire	3974	27277
johnson2014data	22	48	27674	0.842	0.013	7.496	6.864	hospital_expire	3974	27277
ding2016mortality	19	48	27674	0.842	0.013	7.496	6.864	hospital_expire	3974	27277
che2016recurrent_b	18	48	27674	0.842	0.013	7.496	6.864	hospital_expire	3974	27277
ghassemi2015multivariate_a	4	24	63460	0.866	0.011	7.538	6.496	hospital_expire	21670	140770
ghassemi2014unfolding_a	3	24	64413	0.872	0.011	7.764	6.215	hospital_expire	23017	143048
pirracchio2015mortality	15	24	64547	0.873	0.011	8.143	6.218	hospital_expire	23070	143443
lehman2012risk	14	24	52021	0.877	0.011	8.444	4.173	hospital_expire	21317	88954
caballero2015dynamically_a	1	24	45227	0.871	0.013	9.405	6.224	hospital_expire	11573	72035
lee2015customization_c	11	24	59871	0.844	0.017	9.595	5.977	2yr_post_hos_disch	20593	123087
lee2015personalized	12	24	64413	0.867	0.015	10.154	6.215	30dy_post_hos_disch	23018	143049
lee2017patient	13	24	64413	0.867	0.015	10.154	6.215	30dy_post_hos_disch	23018	143049
luo2016interpretable_a	27	all	63819	0.931	0.008	10.407	6.262	30dy_post_hos_disch	22542	141148
lee2015customization_b	10	24	59871	0.866	0.016	10.588	5.977	30dy_post_hos_disch	20593	123087
ghassemi2015multivariate_b	5	24	63460	0.848	0.020	11.400	6.496	1yr_post_hos_disch	21670	140770
grnarova2016neural_b	25	all	58510	0.960	0.006	12.200	5.023	30dy_post_hos_disch	21615	108564
luo2016interpretable_b	28	all	63819	0.899	0.014	12.396	6.262	6mo_post_hos_disch	22542	141148
grnarova2016neural_c	26	all	58510	0.923	0.017	17.726	5.023	1yr_post_hos_disch	21615	108564

[2] W. Boag, D. Doss, T. Naumann, and P. Szolovits, "What's in a note? unpacking predictive value in clinical note representations," *AMIA Summits on Translational Science Proceedings*, vol. 2018, p. 26, 2018.

[3] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits, "Unfolding physiological state: Mortality modelling in intensive care units," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 75–84.

[4] S. Khadanga, K. Aggarwal, S. R. Joty, and J. Srivastava, "Using clinical notes with time series data for ICU management," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019, pp. 6431–6436. [Online]. Available: https://doi.org/10.18653/v1/D19-1678

[5] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar, "Deep active learning for named entity recognition," *CoRR*, vol. abs/1707.05928, 2017. [Online]. Available: http://arxiv.org/abs/1707.05928

[6] A. E. Johnson, T. J. Pollard, and R. G. Mark, "Reproducibility in critical care: a mortality prediction case study," in *Machine Learning for Healthcare Conference*, 2017, pp. 361–376.

[7] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[8] P. Grnarova, F. Schmidt, S. L. Hyland, and C. Eickhoff, "Neural document embeddings for intensive care patient mortality prediction," *arXiv preprint arXiv:1612.00467*, 2016.

[9] M. Ghassemi, M. A. Pimentel, T. Naumann, T. Brennan, D. A. Clifton, P. Szolovits, and M. Feng, "A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[10] Y.-F. Luo and A. Rumshisky, "Interpretable topic features for post-icu mortality prediction," in *AMIA Annual Symposium Proceedings*, vol. 2016. American Medical Informatics Association, 2016, p. 827.

[11] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[12] 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*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[14] M. Jin, M. T. Bahadori, A. Colak, P. Bhatia, B. Celikkaya, R. Bhakta, S. Senthivel, M. Khalilia, D. Navarro, B. Zhang, T. Doman, A. Ravi, M. Liger, and T. A. Kass-Hout, "Improving hospital mortality prediction with medical named entities and multimodal learning," *CoRR*, vol. abs/1811.12276, 2018. [Online]. Available: http://arxiv.org/abs/1811.12276

[15] H. Harutyunyan, H. Khachatrian, D. C. Kale, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *CoRR*, vol. abs/1703.07771, 2017. [Online]. Available: http://arxiv.org/abs/1703.07771

[16] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Y. Hawalah, A. Gelbukh, and Q. Zhou, "Multilingual sentiment analysis: state of the art and independent comparison of techniques," *Cognitive computation*, vol. 8, no. 4, pp. 757–771, 2016.

[17] R. Davoodi and M. H. Moradi, "Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier," *Journal of biomedical informatics*, vol. 79, pp. 48–59, 2018.