PREDICTION OF ICU READMISSION USING CLINICAL NOTES

An Undergraduate Research Scholars Thesis

by

JUSTIN LOVELACE

Submitted to the Undergraduate Research Scholars program at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by Research Advisor: Dr. Bobak Mortazavi

May 2018

Major: Computer Science

TABLE OF CONTENTS

ABSTRACT

Prediction of ICU Readmission Using Clinical Notes

Justin Lovelace
Department of Computer Science
Texas A&M University

Research Advisor: Dr. Bobak Mortazavi
Department of Computer Science and Engineering
Texas A&M University

Unplanned readmissions to the ICU result in higher medical costs and an increase in the likelihood of adverse events, extended hospital stays, and mortality. Machine learning models can leverage the large amount of data stored in electronic health records to predict these cases and provide physicians with more information about patient risk at the time of ICU discharge. Most prior work in this area has focused on developing models using only the structured data found in electronic health records and neglects the large amount of unstructured information stored in clinical notes. This work applies deep learning techniques to these notes to predict ICU readmission and develops models that outperform prior work that focuses only on structured data.

NOMENCLATURE

ICU	Intensive Care Unit
EHR	Electronic Health Record
MIMIC	Medical Information Mart for Intensive Care
ICD-9	9th Revision of the International Classification of Diseases
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory Network

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION AND RELATED WORKS

The intensive care unit (ICU) provides care for hospital patients with serious illnesses or injuries. Intensive care is extremely resource intensive so decisions to step down patients from the ICU must be made in order to allocate limited resources efficiently. However, premature discharge from the ICU can result in an insufficient level of care which can then lead to readmission to the ICU. In North America and Europe, the average readmission rate is approximately 7% [1]. These readmissions come with a much higher mortality rate, longer hospital stays, and increased costs to both the hospital and patients [1] [2]. An accurate predictor for the likelihood of readmission would be useful for planning the step-down process and determining the level of monitoring that the patient would receive after discharge from the ICU.

Previous work has found success using data stored in electronic health records (EHR) to develop predictive models for ICU readmission. Pakbin et al. conducted work predicting ICU readmissions using the structured data found within the publicly available Medical Information Mart for Intensive Care (MIMIC-III) database [3]. The MIMIC-III dataset contains EHR data collected during routine clinical care for over $60,000$ ICU stays which makes it a good venue for developing clinical models [4]. The authors developed a machine learning model that predicted patient risk for ICU readmission over a number of time frames using a variety of numerical, categorical, and binary features extracted from the MIMIC-III database. This work is also performed upon MIMIC-III database to allow for direct comparison to this previous model.

The model developed by Pakbin et al. outperformed the earlier work performed by Fialho et al. who developed a model for predicting ICU readmission using the MIMIC-II database (an earlier, less comprehensive iteration of the MIMIC-III database) [5]. How-

ever, a limitation of both of these works is that the authors focused solely on the structured data found within EHRs. They ignored the large amount of unstructured information stored in free text notes. This work directly addresses this limitation.

The model developed by Pakbin et al. was also heavily reliant upon codes from the 9th Revision of the International Classification of Diseases (ICD-9) as features. ICD-9 codes are medical codes detailing diagnostic information that are assigned for billing purposes at the end of hospital stays. Because of this, they are not available to clinicians at the time that step-down decisions from the ICU are being made. Thus, reliance upon these codes prevents their model from actually being deployable in the clinical setting.

Additionally, these codes are assigned to the entire hospital stay, so some of the codes used in the prior work could pertain to future conditions that arose after the patient was discharged from the ICU, but was still in the hospital. This leak of future diagnosis information presents another problem for the practical use of their model. This work also addresses this shortcoming by leveraging the large amount of diagnostic information that exists in clinical notes to replace ICD-9 codes as features.

Curto et al. explored many of the difficulties that come with working with medical notes in the MIMIC-II database [6]. There are a number of particularities specific to clinical notes that makes applying natural language processing techniques difficult. Some of these particularities include frequent use of medical terminology and abbreviations, a prevalence of numerical values associated with physiological readings, and a large number of spelling errors. They developed a number of preprocessing techniques to address these challenges and reformat the data to be better suited for applying natural language processing techniques.

Curto et al. then used the medical text notes stored in the MIMIC-II database to predict ICU readmission in later work [7]. They, however, focused their work entirely on these notes and only attempted to predict short term readmission. They treated the notes for

6

each patient as a bag-of-words which means that they completely disregard word order and focus only on the presence and frequencies of words in the notes. Thus, their models fail to capture any meaning that may pertain to the structure or order of the words included within clinical notes. They also used notes from the entire hospital stay, and this use of notes past the time of ICU discharge prevents this model from being viable in the clinical setting. The deep learning methods explored in this work are better suited to capture information conveyed by the structure and order of words in the notes. This work also focuses only on notes available before the time of discharge in order to present a clinically deployable model.

Prior work has been conducted that combines both textual and structured information from EHRs for the prediction of events other than ICU readmission. Ghassemi et al. developed models using time series data and free text notes to predict mortality [8]. They used techniques for topic modeling to extract features from the medical notes. The topic modeling technique they used also treats the notes as a bag-of-words and, as such, also suffers from an inability to capture any structural relationships between words within the notes. Despite these shortcomings, they saw the best performance when combining the features from the notes with the structured data from the MIMIC database. This provides motivation for this work to utilize both sources of information in the development of its models.

Liu et al. developed a number of deep learning models to predict the onset of chronic diseases. Their models combine textual features from medical notes with structured information pertaining to lab and demographic data [9]. They find that combining these features leads to better performance than using either set of features alone. They explored models built around both Long Short-Term Memory Networks (LSTMs) and Convolutional Neural Nets (CNNs). They also explore the interpretability of their CNN model and demonstrated that they could identify the sections of text that were the most influential

towards making their predictions. This work explored both LSTMs and CNNs in the early stages of testing and found CNNs to consistently outperform LSTMs. Because of these early results, this work focuses on using CNNs in its models. Although this work does not explore the interpretability of its predictions, this prior work demonstrates this it would be a viable avenue for future work.

A number of groups have also applied deep learning techniques to the discharge summaries in the MIMIC-III database in order to automatically assign ICD-9 codes. Baumel et al. explored different models and achieved best performance using a a bidirectional Gated Recurrent Unit model with a Hierarchical Attention mechanism [10]. Mullenbach et al. later achieves better performance on this task using convolutional attention [11]. The current state-of-the-art was then achieved for this task by Sadoughi et al. using multi-view convolutions paired with attention mechanisms [12]. These works have all shown that deep learning techniques can effectively be applied to the free text notes stored within the MIMIC-III database.

This work directly addresses two significant shortcomings of previous readmissions models. First, they fail to utilize the information stored in free text notes in the EHR. Second, they are heavily reliant upon ICD-9 codes for good performance. This work combines manual extraction of structured features with deep learning models to extract features from notes and develops a readmissions model that performs well without relying on ICD-9 codes as features. This work then demonstrates that the models developed in this work outperform prior models by a significant margin when ICD-9 codes are not used as features.

Thus the primary contribution of this work is developing a readmissions model that both performs well and only uses data that would be available to clinicians at the time of ICU discharge in the clinical setting. Because of this contribution, this work helps bridge the gap between previously published models and the development of a clinically-

8

deployable model.

The rest of this work is organized as follows. Section 2 introduces the dataset and cohort used in this work and describes the data modalities that are explored. Section 3 describes the baseline model used for comparing results, explains the methods used to develop word embeddings for the clinical text, details the architecture of the deep learning models developed, and provides details for how they were trained. Section 4 then presents the results from the models used in this work and discusses those results. Section 5 presents possible directions for future study and finally summarizes the contribution of this work.

2. DATASET AND COHORT SELECTION

2.1 MIMIC-III

We conduct our work using the Medical Information Mart for Intensive Care (MIMIC-III) database [4]. The MIMIC-III dataset is comprised of deidentified information detailing over $60,000$ ICU stays from the Beth Israel Deaconess Medical Center in Boston, Massachusetts. This information was collected as part of routine clinical care and, as such, is representative of the information that would be available to clinicians in real-time. This dataset is made available to the research community which makes it a good venue for reproducible research. The large amount of prior work conducted with this data makes the evaluation of our results straightforward.

The database contains a wide range of numerical data such as lab results and patients' vitals which has been extracted to predict readmission in previous work. It also contains a large amount of unstructured information in the form of free text notes. These notes are taken down by medical professionals during patients' stays and contain a large amount of both qualitative and quantitative information regarding the patients' hospital stays.

These notes, however, come with many challenges that make accessing this information difficult. The primary difficulty is the unstructured format of clinical notes, which makes it very difficult to extract relevant information as structured features that could augment a predictive model. These difficulties have led to the exclusion of this information from many of the prior models developed upon this database. This work overcomes these challenges by using deep learning models which directly learn to extract features from the unstructured text, alleviating the need for manual feature engineering.

2.2 Cohort

This work explores the predictive power of the textual information stored in clinical notes as an alternative to relying on ICD-9 codes for the task of ICU readmission. This motivates the selection of this work's cohort. All patients who have at least three notes recorded before the time of ICU discharge are extracted from the MIMIC-III dataset. Patients are dropped from this cohort if they are missing a time for either ICU admission or discharge as this makes it impossible to calculate readmission times.

These criteria result in a patient cohort of $55,362$ patients. Of these patients, there have been $5,833$ total readmissions that have occured within 30 days of ICU discharge. This work examines patient discharges over a variety of time frames. Specifically, this work looked at readmissions that occured within 24 hours, 48 hours, 72 hours, 7 days, and 30 days. We also look at bounceback readmissions, readmissions that occur before being discharged from the hospital, as well as readmissions that occur between $24 - 72$ hours. Table 2.1 provides the breakdown for these readmissions within the selected cohort.

This cohort was split into training, validation, and testing splits using a $80/5/15$ split for this work. This work divides this cohort by subject ID rather than ICU stay ID to prevent leaking data from patients who have had multiple ICU stays. The breakdown of the training, validation, and testing splits is provided in detail in table 2.1.

Table 2.1: ICU Readmission Cohort

Group	Total	24 Hr	48 Hr	24-72 Hr	72 Hr	7 day	30 days	Bounceback
Training	44393	778	1310	921	1699	2718	4709	2735
Validation	2742	56	97	66	122	187	291	177
Testing	8227	130	227	178	308	491	833	522
Total	55362	964	1634	1165	2129	3396	5833	3434

11

2.3 Data Modalities

This work focuses on two different data modalities: the structured tables and the unstructured clinical notes stored within the MIMIC-III database.

2.3.1 Structured Data

This work uses the structured feature extraction performed by Pakbin et al. in their work on predicting ICU readmission in MIMIC-III [3]. They extracted a series of binary features, categorical features, and continuous features from the MIMIC-III database. These features come from 6 tables within the MIMIC-III database: Admissions, ICUStays, Patients, ProcedureEvents, Chartevents, and Labevents. Among these features are demographics data, lab values, and vitals values. More details about this structured feature extraction can be found in their work [3].

One notable difference between this work and the prior work by Pakbin et al. is that this work does not extract the ICD-9 codes stored within the MIMIC-III database. Because these codes are not available in the clinical setting until after the time of ICU discharge, this work decides to exclude them, and instead extracts clinical notes which also contain diagnostic information, although in an unstructured manner.

2.3.2 Clinical Notes

This work focuses on utilizing deep learning models for the extraction of features from clinical notes. This involves extracting all notes from each patient's textual history up to the point of ICU discharge and concatenating them to produce a single textual narrative for each patient. The text is then preprocessed by removing stop words, normalizing punctuation, and replacing numerical text and de-identified information with generic tokens. This work then uses pre-trained word embeddings to represent the text. Specifically, this work experiments with embeddings trained on the MIMIC-III dataset using the word2vec

algorithm developed by Mikolov et al. [13] and the starspace algorithm developed by Wu et al. [14]. More details about the embeddings used in this work and how they are trained are presented in our methods section.

3. METHODS

In this work we focus on both the free text data and the structured data stored within the MIMIC-III database. In the following sections we detail the methods that we used to develop predictive models from both the textual and structured data.

3.1   Baseline

For a baseline, we use the model previously developed by Pakbin et. al [3] for comparison. Because this work intends to do away with the dependency on ICD-9 codes, this work reports the results from their model both with and without ICD-9 codes.

They used XGBoost [15] to develop a model from a number of numerical, binary, and categorical features that were extracted from the structured tables within the MIMIC-III database. This work follows their training setup and uses internal cross-validation to tune hyperparameters and then trains and evaluates their model using this work's training and testing cohorts for the sake of direct comparison.

3.2   Dense Word Embeddings

This work explored two different methods for developing word embeddings. The first method involves applying the word2vec algorithm to all notes in the MIMIC-III database associated with subjects who are not in our validation or testing set. This allows for training on a greater selection of notes than if training had been limited to the training set. This training is done using the continuous bag of words implementation and it generates embeddings for all words that appear in at least 10 notes in our corpus. This results in a vocabulary of $101,374$ words. Both 100 and 300 dimensional word embeddings were explored and early testing showed that 100 dimensional word embeddings led to better performance.

The other embedding algorithm explored in this work is the starspace embedding algorithm which allows for the training of word embeddings based on document labels. This work leverages the ICD-9 codes in the MIMIC-III database to label all the notes associated with a given hospital stay with their assigned ICD-9 codes. This algorithm treats notes in the same stay as a bag-of-words and embeds both words and documents into the same feature space. It then maximizes the cosine similarity between words and the labels that occur together. This results in words related to the same diseases and procedures being placed close to each other in the embedding space. This helps our model capture some of the information represented by the ICD-9 codes without relying on them as an explicit feature to our model. Words that appear in over 80% of patient narratives are ignored and then embeddings are generated for a vocabulary of the $50,000$ most frequent words. Both 100 and 300 dimensional word embeddings were explored for this embedding scheme as well and the 300 dimensional embeddings were found to perform better.

3.3 Convolutional Neural Net

The Convolutional Neural Net (CNN) architecture used in this work is based on that developed by Lee et. al [16] for the task of sentence classification and is presented in figure 3.1. This architecture was later used by Liu et al. [9] for the predictiaon of the onset of chronic diseases. This architecture consists of three different kernels of width 1, 2, and 3. These kernels convolve over the textual input with a stride of 1, applying learned filters for every 1-gram, 2-gram. and 3-gram in the input. Through learning these filters, this model learns to extract features from the unstructured clinical text. The ability of this model to extract features from short windows of text provides an advantage over the bag-of-words models used in prior works that failed to capture any information pertaining to local context or word order.

The representations that come from applying the learned filters to the text are then
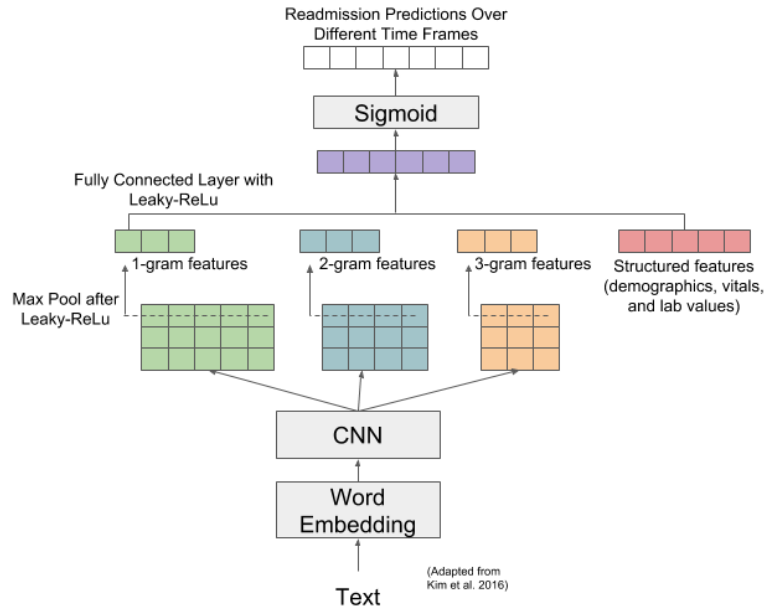
Figure 3.1: Convolutional Neural Network Architecture

concatenated and max-pooled over time. In doing so, the model extracts the most relevant windows of text for the prediction task. This aids in the interpretability of the model which is demonstrated by Liu et. al [9]. The outputs of the max-pooling layer are then concatenated and fed to a dense hidden layer to allow for the model to capture interactions between the textual features. In this work, structured features are integrated with the textual features by concatenating them with the output of the CNN before the dense hidden layer. The output of the dense layer is then passed through a fully connected layer with a sigmoid to make the final predictions.

This work does make two modification to the CNN models used in previous work. For the first one, this work replaces the ReLu activation function used by Lee et. al [16] and Liu et. al [9] with the Leaky ReLu. The ReLu function completely drops the negative part of the activation function while the Leaky Relu assigns a slight, non-zero slope to it. A

16

comparison between these two functions can be seen in figure 3.2 [1]. The Leaky ReLu has been demonstrated to outperform the ReLu function with CNN architectures in previous work [17] and early testing found that this was also the case for this application.
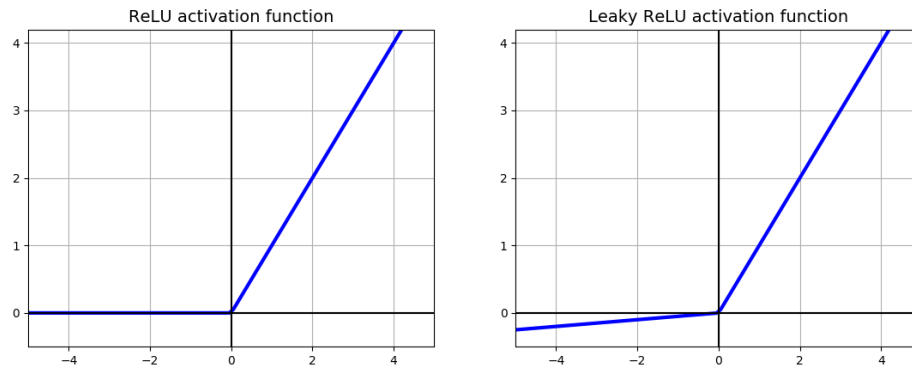

Figure 3.2: Comparison Between ReLu and Leaky ReLu Activation Functions

The second change involves ensembling two different sets of dense word embeddings. The architecture for this is presented in figure 3.3. To do this, the input text is passed to two different embedding layers. The two different embedding layers are the word2vec and starspace embeddings discussed previously. The embedded text is then passed to two separate CNNs that are trained in parallel. These outputs are then concatenated along with the structured features as was done in the previous model. The rest of the model follows the architecture described previously.

---

[1] The negative portion of the Leaky ReLu used in this visualization is assigned a slope of 0.05 while the actual slope used in this work is 0.01. This is done for visualization purposes.
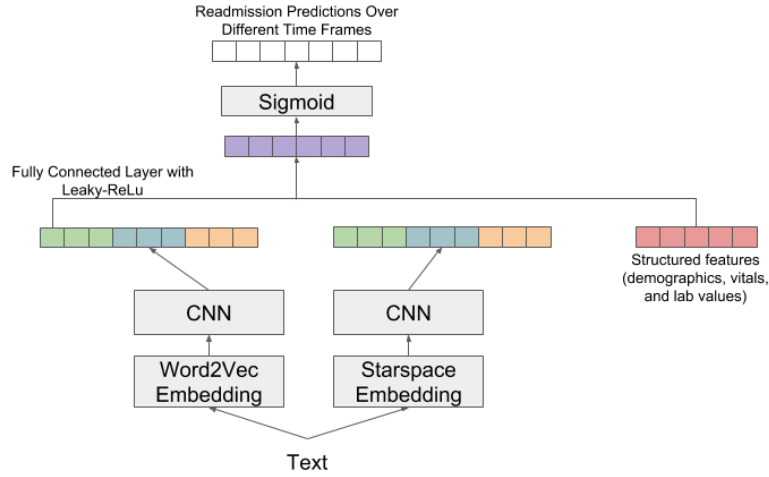
Figure 3.3: Convolutional Neural Network Architecture with Two Embedding Layers

3.4 Training Setup and Parameter Tuning

The training procedure used in this work minimizes the binary cross entropy loss,

$$L_{BCE}(X, y) = -\sum_{\ell=1}^{\mathcal{L}} y_\ell log(\hat{y}_\ell)$$

using the Adam optimizer [18]. Hyperparameter tuning was conducted using a grid search on the validation set. All models are trained upon the training set and performance on the validation set is used for early stopping. Training is terminated after the average performance on the validation set across all of the time intervals does not improve for 10 epochs; the model with the greatest average performance is then evaluated using the test set.

The results presented in this work are achieved using a learning rate of .001 and a batch size of 16. The CNNs used in this work have 64 filters and dropout with a ratio of 0.5 is applied along with batch normalization at all levels of the network. The textual narratives are padded or truncated to a sequence length of 8000 before being fed to the embedding layers. This configuration was selected based on the grid search mentioned above.

18

4. RESULTS

4.1 Results

In this work we look at predicting readmission over a variety of time time frames. This work found that the deep learning models performed better when trained to predict readmission jointly over every time frame than when trained to predict each time frame individually. Thus, all results in this work, except for the baseline model, are reported from training in that setting. The baseline model is trained to predict each time frame individually as was done in previous work.

Results using the baseline XGBoost model are reported both with and without ICD-9 codes as features. Results using the CNN models explored in the work are reported when using both embedding schemes individually, as well as when they are ensembled as described in the methods section. This work also reports the results from integrating the structured features without ICD-9 codes with the ensembled CNN model. The results for all of the different methods explored in this work are presented in table 4.1.

Due to the unbalanced nature of the dataset, this work reports performance using the Area Under the Curve (AUC) of the Receiver Operating Curve. This metric directly measures how well a model is able to distinguish between positive and negative classes, with an AUC of $.5$ corresponding to a random baseline and an AUC of 1.0 corresponding to a perfect model. This is much easier to interpret than accuracy because poor models can still achieve high accuracies when working with unbalanced data. For example, a model that predicts that no patients are readmitted would achieve an accuracy of over 98% for 24 hour readmissions, but such a model would not actually be predictive of readmissions.

Table 4.1: Model Performance (AUC) by Readmission Type

Model	24 Hr	48 Hr	24-72 Hr	72 Hr	7 day	30 days	Bounceback
Structured Data Baseline	.57	.59	.57	.59	.61	.64	.65
CNN model + W2V	.62	.64	.67	.65	.66	.70	.71
CNN model + Starspace	.61	.65	.69	.67	.67	.70	.70
CNN model + both embeddings	.63	**.67**	**.72**	**.69**	**.69**	**.71**	**.72**
CNN model + both embeddings + structured features	**.65**	**.67**	**.72**	**.69**	**.69**	**.71**	**.72**
Structured Data Baseline + ICD Codes	.68	.72	.72	.73	.74	.72	.80

4.2   Discussion

The testing performed finds that all of the CNN models using the clinical notes outperform the structured baseline by a large margin. The best CNN model improves results across every time frame by at least .07 AUC and improves one time frame by as much as .15 AUC. On average, the CNN with starspace embeddings seems to outperform the CNN model using word2vec embeddings, although this is not true across every timeframe. The model using word2vec embeddings outperforms the model using starspace embeddings by a small margin when predicting 24 hour and bounceback readmission.

A consistent improvement, however, is seen when using a model that combines these embedding schemes. Using two CNNs trained in parallel with each of the two embedding schemes results in a consistent increase in performance ranging from $.01 - .03$ AUC across every time frame. Although the boost in performance is relativly small, it is very consistent, with the ensembled model achieving better performance across every prediction window than either of the models using only one embedding scheme alone.

Experimentation with augmenting the text based models with structured features finds that it makes a relatively minor difference. It does increase performance by $.02$ AUC for the 24 hour readmission case, but performance over every other time frame remains the same. It is possible that structured information such as vitals is more predictive of short term readmission, but is not as helpful for predicting readmission over longer time frames.

Despite the large improvement of the CNN models over the structured baseline without any ICD codes, it is still unable to match the performance of the XGBoost model that utilizes ICD codes across most of the time frames. Although it is actually able to match the XGBoost model for the $24 - 72$ hour time frame and only lags by $.01$ AUC for the 30 day time frame, it does not compare as favorably over the other time frames.

However, it is important to keep two things in mind when comparing the models explored in this work to the model using ICD-9 codes. First, the ICD-9 codes are not available at the time that discharge decisions are made, so the use of ICD-9 codes for readmission models is not actually possible in the clinical setting. Second and more importantly, the ICD-9 codes can represent diagnostic information from after the time of ICU discharge. So not only are the codes themselves unavailable at ICU discharge, but they actually leak information from the future as well. The predictive power of this future information likely accounts for some of the large performance boost that comes with using these codes.

This leak of future data is especially significant for bounceback readmissions because both ICU stays are contained within one hospital visit and thus share a single set of ICD-

9 codes. So using ICD-9 codes for bounceback readmissions allows the model to have access to diagnostic information for the future ICU stay as well as any conditions that arose leading up to that readmission. This may explain why the gap in performance between the models explored in this work and the model using ICD codes is the largest for bounceback readmissions.

Because of these problems with using ICD-9 codes, any models that incorporate them as features are unusable in the clinical setting and will have an inflated performance from their access to future information. Despite that, the above problems with models that use ICD-9 codes also make them a very strong baseline for the purposes of comparison. The models developed in this work compare much more favorably to this strong baseline than the previous work using only structured data. The most successful CNN model is even able to achieve comparable performance across a couple of the readmission time frames. Because of these promising results, this work represents a large step towards developing a more accurate, clinically deployable model.

5. FUTURE WORK AND CONCLUSION

5.1 Further Study

While this work did outperform previous work that focused on the structured within the MIMIC-III database, it does have its limitations. Prior work has demonstrated the interpretability of the models used in this work, but that was not directly explored in this work. Further work looking at what contributed to the predictions made by these models could provide valuable insight into what information from these notes is important for readmission.

Given the success of ensembling two different embedding schemes, it would also be interesting to observe whether these two CNNs were able to extract different features from the text. If each embedding scheme was more successful at extracting different subsets of features, then this could explain the boost in performance found when using them together.

The structured extraction used in this work is another area for improvement. The methods used in this work for structured extraction were borrowed directly from previous work and the time-series work conducted was relatively simplistic. This could possibly account for the lack of improvement when incorporating the structured data. More sophisticated time-series work utilizing LSTMs could lead to greater improvement. These LSTMs could be jointly trained with the CNNs to allow for the model to directly account for interactions between the textual and time series features while training.

Additionally, prior work has demonstrated that negation tagging can improve the performance of predictive models applied to clinical notes. This suggests that negation tagging could likely be applied to this work to improve performance. Prior work utilized a pre-existing regex-based negation tagging tool, but deep learning driven negation tagging could potentially lead to an even greater improvement in performance than was found

using the existing tool.

5.2 Conclusion

This work helps address the failure of recent readmissions models developed using the MIMIC-III database to utilize the large amount of information stored in free-text clinical notes. It also attempts to do away with the previous readmissions models' reliance upon ICD-9 codes. Although the model using the clinical text is not able to match the performance of the previous model that uses ICD-9 codes along with structured data, it outperforms that model by a large margin when the ICD-9 codes are removed. It even compares favorably to the model using ICD-9 codes across some of the readmission time frames. Because of these strong results, this work achieves it's purpose of developing a more accurate model that could potentially be viable in a clinical setting.

The success of the models explored in this work demonstrates that there is a large amount of information stored in clinical notes that is predictive of readmission. Additionally, this information appears to be much more predictive than the structured data that has been the primary focus of previous work. It also demonstrates the viability of deep learning models to utilize this textual information for developing predictive models. This work presents a promising avenue for future work to be conducted upon clinical text for both readmissions models and other applications within the medical domain.

REFERENCES

[1] A. L. Rosenberg and C. Watts, "Patients readmitted to icus: a systematic review of risk factors and outcomes," *Chest*, vol. 118, no. 2, pp. 492–502, 2000.

[2] A. A. Kramer, T. L. Higgins, and J. E. Zimmerman, "The association between icu readmission rate and patient outcomes," *Critical Care Medicine*, vol. 41, no. 1, 2013.

[3] A. Pakbin, P. Rafi, N. Hurley, W. Schulz, M. Harlan Krumholz, and J. Bobak Mortazavi, "Prediction of icu readmissions using data at patient discharge," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4932–4935, July 2018.

[4] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, pp. 160035 EP –, May 2016. Data Descriptor.

[5] A. Fialho, F. Cismondi, S. Vieira, S. Reti, J. Sousa, and S. Finkelstein, "Data mining using clinical physiology at discharge to predict icu readmissions," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13158 – 13165, 2012.

[6] J. P. Carvalho and S. Curto, "Fuzzy preprocessing of medical text annotations of intensive care units patients," *2014 IEEE Conference on Norbert Wiener in the 21st Century (21CW)*, pp. 1–7, 2014.

[7] S. Curto, J. P. Carvalho, C. Salgado, S. M. Vieira, and J. M. C. Sousa, "Predicting icu readmissions based on bedside medical text notes," in *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 2144–a–2151–h, July 2016.

[8] 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," *KDD: Proc Int Con Knowl Discov Data Mining.*, vol. 2014, pp. 75–84, 08 2014.

[9] J. Liu, Z. Zhang, and N. Razavian, "Deep ehr: Chronic disease prediction using medical notes," in *Proceedings of the 3rd Machine Learning for Healthcare Conference* (F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, eds.), vol. 85 of *Proceedings of Machine Learning Research*, (Palo Alto, California), pp. 440–464, PMLR, 17–18 Aug 2018.

[10] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, and N. Elhadad, "Multi-label classification of patient notes: Case study on icd code assignment," 2018.

[11] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1101–1111, Association for Computational Linguistics, 2018.

[12] N. Sadoughi, G. P. Finley, J. Fone, V. Murali, M. Korenevski, S. Baryshnikov, N. Axtmann, M. Miller, and D. Suendermann-Oeft, "Medical code prediction with multi-view convolution and description-regularized label-dependent attention," 11 2018.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *CoRR*, vol. abs/1310.4546, 2013.

[14] L. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, and J. Weston, "Starspace: Embed all the things!," *arXiv preprint arXiv:1709.03856*, 2017.

[15] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *CoRR*, vol. abs/1603.02754, 2016.

[16] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Association for Computational Linguistics, 2014.

[17] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *CoRR*, vol. abs/1505.00853, 2015.

[18] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.