Improving Clinical Outcome Predictions Using Convolution over Medical Entities with Multimodal Learning

Suraj Bisht and Shubhendu Bhaskar
surajb2@illinois.edu, sb59@illinois.edu

Group ID: 03
Paper ID: 211, Difficulty: Easy
Presentation link: https://youtu.be/Q7F1twmcm-k
Code link: Project Code

1 Introduction

With the advent of Electronic Health Records (EHR), there lies a wide opportunity to review the existing data and predict the outcomes for any patient especially the length of stay (LOS) in the ICU and mortality in the hospital (including ICU). As there are existing models predicting clinical problems, several of them are not utilizing the clinical notes because of the sparsity and the high dimensional nature of the data. The information that is written by the hospital staff from doctors nurses, and different departments contains exhaustive and critical patient-specific information which can be used in prediction processes.

This paper (Bardak and Tan, 2021) leverages the EHR clinical notes and uses the Named Entity Recognition (NER) task of Natural Language Processing (NLP) to extract entities from the text and categorize them into predefined classes. The paper also assesses the effect on the prediction results by using different word representations like Word2Vec, FastText, and concatenation of both.

2 Scope of reproducibility

The paper argues that the integration of structured data in EHR and medical entities improves the efficiency of mortality and LOS predictions.

The proposed model uses 1D Convolutional Neural Networks (CNN) based multimodal approach in NLP that gives higher accuracy than the averaging multimodal and baseline timeseries models.

2.1 Addressed claims from the original paper

We are the confirming following claims made from the study:

- For the given data sets study claims that Gated Recurrent Unit (GRU) has better results than Long Short Term Memory (LSTM).
- The average multimodal baseline has better performance than GRU alone.
- Proposed 1D CNN multimodal has better performance than other baseline models.

3 Methodology

We are using the original approach as specified in the paper to verify the results.

The paper proposes the model using 1D convolution layers to extract features from medical relevant entities (classes) and use them along with time-series features (Intensive Care unit signals - ICU) to enhance the accuracy of the predictions. It also applies the multimodal approach and uses extracted features in different permutations to show the effectiveness of the proposed work.

The process starts with the extraction of the first 24 hours time-series features from the raw data and run them through LSTM and GRU. Then extract the clinical notes from the MIMIC-III and pre-process them (used med7 in extracting medical entities) (Kormilitzin et al., 2021). In the baseline models Word2Vec, FastText embeddings and their concatenation were added with timeseries and multimodal.

1D CNN is applied to the extracted features from medical entity representations. In the final stages, the time-series features are concatenated with entity representations and are run through full connected layers to predict 4 different clinical tasks.

The experiments are conducted on Intel (R) Xeon (R) CPU @ 2.3GHz, memory 36GB RAM, Tesla T4 GPU.

3.1 Model descriptions

This section illustrates the baseline models and the proposed model.

3.1.1 Baseline Models

I) Time Series Model

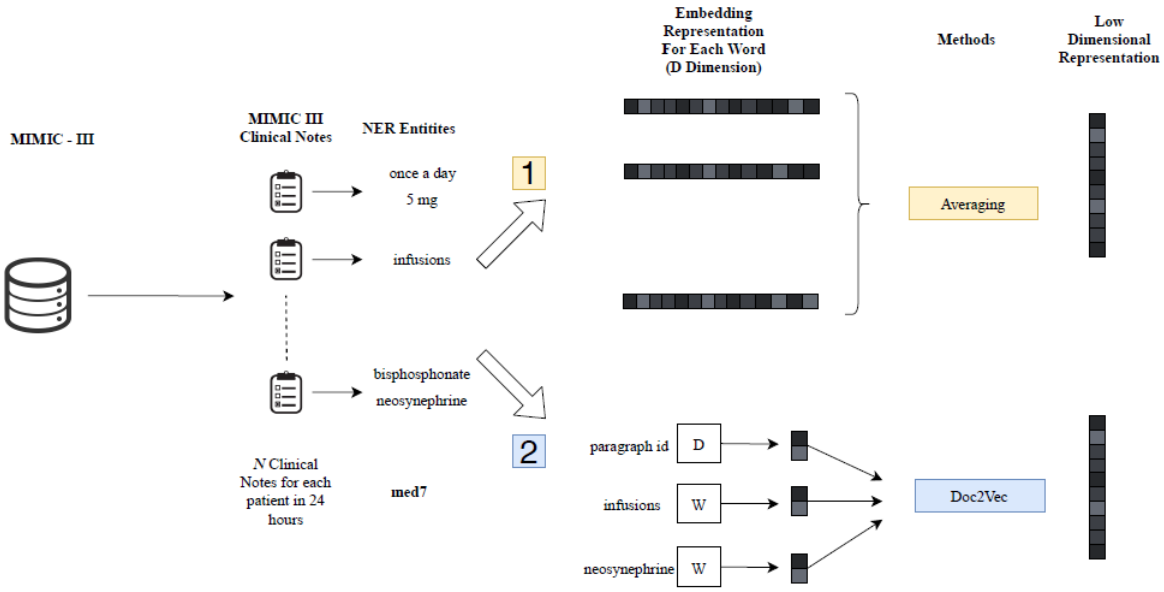

Figure 1: Baseline multimodal architecture

In this model, both LSTM and GRU are used one by one for capturing the temporal information between features. Both timeseries models uses 256 hidden units and is run under signmoid classifier for predicting mortality and length of stay. GRU is used further for multimodal architectures.

II) Multimodal approaches

In addition to time-series features, this model considers and accounts for the important information present in the clinical notes. It assesses whether the addition of clinical notes impacts the prediction performance positively. This model leverages the pre-trained clinical NER Model called med7. Applying med7 on clinical notes results in the following entities: 'Drug', 'Duration','Strength', 'Form', 'Dosage', 'Route', 'Frequency'.

Following pre-trained embeddings (Huang et al., 2019) are used in the extraction of medical entities and their performances are compared with time-series base model:

1. Word2Vec
2. FastText
3. Concatenation of Word2Vec and FastText. In case an embedding is not present for a word, zero padding is added

Figure 1 illustrates the technique used in Baseline Multimodal for learning medical entity vectors and the removal of words if they do not belong to the medical entities.

3.1.2 Proposed Model

The overview of the proposed multimodal architecture for prediction of In-Hospital Mortality, ICU mortality. LOS >3 and LOS >7 is shown in Figure 2.

The proposed model utilizes the 1D convolutional layers for the purpose of extracting features over medical entities. This technique grasps the words with the combination of their adjacent words and shows successful results related to NLP problems. The medical entities extracted say K, from the clinical notes, say N. The K Medical entities are set as a course of word embeddings using techniques such as Word2Vec, FastText, and concatenation of both. These medical entities are presented in rows (each value in a row represents a separate entity). Padding is done wherever necessary. Further, there are three back-to-back 1D CNN layers with the sizes of 32,64 and 96 keeping the same kernel size. Thenceforth max-pooling layer is applied to consecutive outcomes. Final results from Max Pooling are combined with the single-layer GRU (hidden units-256) and is lastly run in one fully connected layer consisting of 512 hidden units.

3.2 Data descriptions

The final cohort after clinical note elimination from Medical Information Mart for Intensive Care (MIMIC-III) (Johnson et al., 2016) used in this study is illustrated in Figure 3.

The MIMIC-Extract open source pipeline that transforms MIMIC-III data into a usable format is used for the experiments. The transformation includes converting raw vital signs into hourly time

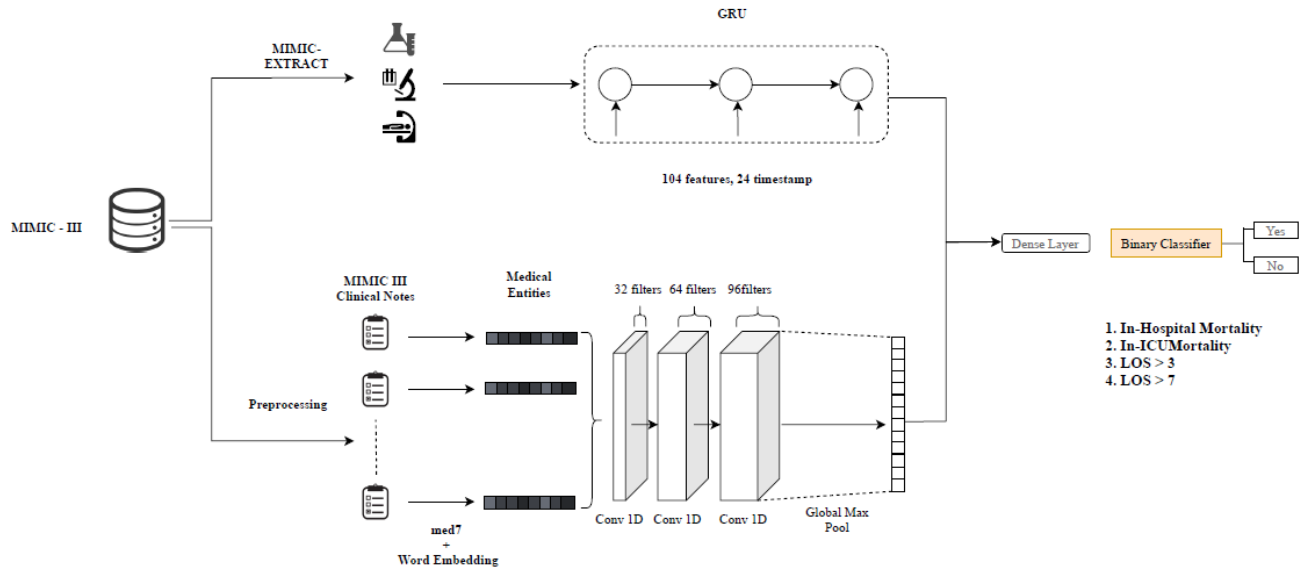

Figure 2: Proposed multimodal architecture

series and applying unit conversion, imputing missing data, and handling outliers. To process medical entities from clinical notes Med7 model is used. This data, eliminates the data of patients younger than 15 years old and LOS, not between 12 hours and 10 days. Only the first 24 hours of patient's ICU data and only patients with a minimum of 30 hours of data. Discharge summaries are removed before applying the clinical NER model on notes. Finally, all patients were dropped with no notes in 24 hours.

The experiment results are illustrated into the following class distributions:

1. **In-hospital mortality:** - Patients succumbed to death in the hospital (Significantly imbalanced %10.5).

2. **In-ICU mortality:** - Patients succumbed to death in ICU stay after admitted in ICU (Significantly imbalanced %7).

3. **Length-of-stay >3:** - Patient with stay longer than 3 days in the ICU (Slightly imbalanced %43.2)

4. **Length-of-stay >7:** - Patient with stay longer than 7 days in the ICU (Significantly imbalanced %7.9)

Additional Experiment: LOS >5 : - Patient with stay longer than 5 days in the ICU (Significantly imbalanced %18.7)

3.3 Hyperparameters

Below listed are hyperparameters used in this research:

- Hidden Layers: 3 1D-CNN Layers
- Hidden unit: 128 and 256
- Filter Sizes - 32, 64, 96
- Dropout - 0.2 (used in average multimodal and proposed model)
- L2 Regularizer factor - 0.01
- Learning Rate - 0.001

3.4 Implementation

The author's code has been updated with the latest libraries (for e.g. tensorflow libraries). The obsolete methods are replaced with the current library's methods. The updated code can be found at https://github.com/surajbisht1809/CS598_DL4H_Project_Team03_2022Spring.git

3.5 Computational requirements

All the experiments are conducted on Intel (R) Xeon (R) CPU @ 2.3GHz, memory 36GB RAM, Tesla T4 GPU.

In our run, Google Colab with TPU/GPU and high memory based virtual machine is used for all executions. Restricted Google drive is used for storing pre-processed (Khadanga et al., 2019), intermediary data and results in files.

	# of Patient	# of hospital admission	# of ICU admission
MIMIC-III (> 15 years old)	38,597	49,785	53,423
MIMIC-Extract	34,472	34,472	34,472
MIMIC-Extract (at least 24+6 (gap) hours patient)	23,937	23,937	23,937
Final Cohort (After clinical note elimination)	21,080	21,080	21,080

Figure 3: Data Description

The experiment used 40 hours of GPU with 100 epochs, model patience 3, using 4 different seeds. It is noted that the execution stops at 6 epochs as there was no performance improvement.

In summary for the entire experiment, following is the breakdown of execution time:

- Pre-processing - Requires around 18 hours
- Baseline timeseries model took 2 hours
- Baseline timeseries multimodal took 5 hours
- Proposed multimodal execution expected to take approximately 6-8 hours

The average runtime for each epoch is listed below:

- Timeseries- 6 seconds per epoch
- Timeseries with Multimodal - 5 seconds per epoch
- Proposed model - 6 seconds per epoch

The consolidated run of BaseLine Model (LTSM and GRU), Averaged Multimodal GRU (Word2Vec,fastext Concat) is depicted in Figure 4.

The Proposed Model GRU (Word2Vec,fastext Concat) along with best Baseline Multimodal is shown in Figure 5.

4 Result

After running all the experiments, Baseline models performance figures are better than the original baseline models. The proposed model results are quite similar to the proposed model in the original paper. Due to high performance of baseline models, it outperforms the proposed model which is not inline with original paper claim. Details of the statistics are present Figure 4 and Figure 5.

4.1 Result 1

GRU on an average shows 1.12% improvement over LSTM across all categories auroc, auprc and F1

4.2 Result 2

For most of the tasks, averaged multimodal performs better than the timeseries model as shown in the Figure 4.

4.3 Result 3

Baseline Average Multimodal performs 0.5-1% better than proposed 1D CNN model as shown in Figure 5.

4.4 Additional results not present in the original paper

Apart from the above results we have consolidated other observations in our experiment as below:

- In our results, all the model's performance is higher than the model's performance in the original paper. For instance, our Baseline model performs at least 2% higher than baseline of the original paper
- It is also observed that Word2Vec embedding has a higher performance than the other embeddings.

See Figure 4 and 5 for the statistics.

Experiments beyond original paper: Since the original paper cited the patient data with LOS>3 and LOS>7, we extracted the data for LOS>5 when compared with Baseline and Averaged Multimodal. In our results, we found that Proposed Model has the best outcomes for LOS>5. Please refer to Figure 5 for proposed model results.

In the original paper, the baseline model is run with only 256 hidden units. In our experiment, we also ran the baseline models for 128 hidden units along with 256 units. Please see Figure 4 for results metrics.

Task	Timeseries Model	Modal	Embedding	AUROC	AUPRC	ACC	F1
In-Hospital Mortality	LSTM	Baseline	-	87.58 ± 0.003	55.46 ± 0.007	91.14 ± 0.001	42.64 ± 0.024
	GRU	Baseline	-	87.89 ± 0.002	56.57 ± 0.008	91.23 ± 0.001	44.58 ± 0.022
		Averaged Multimodal	Word2Vec	87.33 ± 0.004	57.63 ± 0.007	91.51 ± 0.001	45.57 ± 0.033
			FastText	87.32 ± 0.003	57.28 ± 0.008	91.44 ± 0.002	45.18 ± 0.022
			Concat	87.17 ± 0.003	57.06 ± 0.012	91.43 ± 0.002	44.55 ± 0.031
In-ICU Mortality	LSTM	Baseline	-	88.45 ± 0.003	49.94 ± 0.011	93.86 ± 0.001	41.55 ± 0.031
	GRU	Baseline	-	89.06 ± 0.002	52.41 ± 0.009	94.02 ± 0.001	43.8 ± 0.023
		Averaged Multimodal	Word2Vec	88.13 ± 0.006	51.99 ± 0.011	94.15 ± 0.001	42.12 ± 0.049
			FastText	88.04 ± 0.004	51.53 ± 0.01	94.12 ± 0.001	42.81 ± 0.032
			Concat	88.22 ± 0.006	51.9 ± 0.009	94.18 ± 0.001	44.24 ± 0.022
LOS > 3 Days	LSTM	Baseline	-	69.11 ± 0.003	63.22 ± 0.004	65.87 ± 0.004	54.53 ± 0.012
	GRU	Baseline	-	69.82 ± 0.001	64.14 ± 0.003	66.12 ± 0.004	55.33 ± 0.012
		Averaged Multimodal	Word2Vec	70.35 ± 0.004	64.69 ± 0.004	66.26 ± 0.004	56.42 ± 0.019
			FastText	70.24 ± 0.004	64.61 ± 0.004	66.21 ± 0.004	55.9 ± 0.025
			Concat	70.1 ± 0.004	64.44 ± 0.005	66.08 ± 0.005	55.31 ± 0.021
LOS > 5 Days	LSTM	Baseline	-	70.2 ± 0.004	36 ± 0.008	82.29 ± 0.003	20.65 ± 0.031
	GRU	Baseline	-	71.21 ± 0.004	37.72 ± 0.003	82.52 ± 0.001	21.61 ± 0.017
		Averaged Multimodal	Word2Vec	71.89 ± 0.006	37.83 ± 0.01	82.61 ± 0.003	20.8 ± 0.04
			FastText	71.78 ± 0.005	37.9 ± 0.008	82.52 ± 0.002	22.45 ± 0.043
			Concat	71.81 ± 0.005	37.64 ± 0.008	82.54 ± 0.002	20.16 ± 0.05
LOS > 7 Days	LSTM	Baseline	-	72.76 ± 0.007	20.42 ± 0.007	91.78 ± 0	2 ± 0.016
	GRU	Baseline	-	73.21 ± 0.005	21.33 ± 0.006	91.74 ± 0.001	4.73 ± 0.02
		Averaged Multimodal	Word2Vec	72.31 ± 0.01	20.82 ± 0.009	91.8 ± 0	2.37 ± 0.017
			FastText	72.68 ± 0.006	20.95 ± 0.009	91.8 ± 0	2.74 ± 0.03
			Concat	72.51 ± 0.008	20.82 ± 0.009	91.81 ± 0	2.96 ± 0.023

Figure 4: Baseline and Average Multimodal

Task	Timeseries Model	Modal	Embedding	AUROC	AUPRC	ACC	F1
In-Hospital Mortality	GRU	Best Baseline		87.89 ± 0.002	57.63 ± 0.007	91.51 ± 0.001	45.57 ± 0.033
		Proposed Model	Word2Vec	87.51 ± 0.004	56.85 ± 0.009	91.37 ± 0.002	44.02 ± 0.033
			FastText	87.11 ± 0.005	56.2 ± 0.007	91.31 ± 0.002	44.84 ± 0.022
			Concat	87.31 ± 0.006	56.01 ± 0.014	91.23 ± 0.002	44.24 ± 0.033
In-ICU Mortality	GRU	Best Baseline		89.06 ± 0.002	52.41 ± 0.009	94.18 ± 0.001	44.24 ± 0.022
		Proposed Model	Word2Vec	87.88 ± 0.005	50.66 ± 0.011	93.99 ± 0.001	39.76 ± 0.057
			FastText	88.01 ± 0.005	50.44 ± 0.01	93.96 ± 0.001	40.27 ± 0.03
			Concat	87.79 ± 0.005	50.06 ± 0.013	93.89 ± 0.001	40.99 ± 0.036
LOS > 3 Days	GRU	Best Baseline		70.35 ± 0.004	64.69 ± 0.004	66.12 ± 0.004	56.42 ± 0.019
		Proposed Model	Word2Vec	70.16 ± 0.004	64.6 ± 0.005	66.4 ± 0.003	55.76 ± 0.018
			FastText	69.51 ± 0.004	64.17 ± 0.004	65.87 ± 0.004	55.32 ± 0.021
			Concat	69.72 ± 0.004	64.4 ± 0.005	66.12 ± 0.004	54.63 ± 0.023
LOS > 5 Days	GRU	Best Baseline		71.89 ± 0.006	37.9 ± 0.008	82.61 ± 0.003	22.45 ± 0.043
		Proposed Model	Word2Vec	72.16 ± 0.005	38.86 ± 0.005	82.68 ± 0.002	19.84 ± 0.05
			FastText	71.34 ± 0.006	38.28 ± 0.007	82.71 ± 0.002	20.54 ± 0.043
			Concat	71.45 ± 0.006	38.06 ± 0.006	82.46 ± 0.004	23.17 ± 0.047
LOS > 7 Days	GRU	Best Baseline	-	73.21 ± 0.005	21.33 ± 0.006	91.81 ± 0	4.73 ± 0.02
		Proposed Model	Word2Vec	72.45 ± 0.009	20.95 ± 0.01	91.8 ± 0	2.47 ± 0.02
			FastText	71.7 ± 0.008	20.94 ± 0.011	91.81 ± 0	2.14 ± 0.02
			Concat	71.73 ± 0.011	20.9 ± 0.012	91.8 ± 0	2.44 ± 0.026

Figure 5: Proposed Model v/s Best Baseline

5 Discussion

After updating the obsolete references to the latest python libraries and making code changes to replace non-supported methods, experiments were conducted successfully as the original paper. It is found the performance of all the models in our result metrics are higher than the respective models in the original paper.

However, the proposed model fails to outperform other baseline models as claimed in the original paper. The same results were observed for additional experiments with 128 hidden units.

Also, in our additional experiments with LOS >5, the proposed model performs better than the best of the baseline model.

As per the original author Batuhan Bardak, one of the reasons could be that the original experiments were run using different seeds and the mean of the scores was evaluated and then the comparison was performed. The paper does not specify what specific seeds the author used in running their experiments.

As per the original author Batuhan Bardan, one of the reasons for the

5.1 What was easy

In running our experiment we found the below steps were comparatively easy:

- Getting access to the MIMIC-III data used in the experiment
- The data was already processed and did not require any preparation to make it fit to run for the experiment
- Mounting the data to drives and consuming them over colab for code execution
- Since the code was publicly available in github, we did not need to request further access to get to the code repository

5.2 What was difficult

While working on the code we encountered the following difficulties:

- Since the code was old, identifying the necessary current replacements was challenging
- When the latest changes were incorporated into the code, it triggered issues in other components which we had to debug and fix them manually
- Processing of clinical notes was highly time consuming, each iteration of the experiment ran for almost 16-20 hours

5.3 Recommendations for reproducibility

Please use the latest tensor versions and make appropriate code changes for maintaining compatibility. Updated executable code with detailed documentation is available in public repository at GitHub: Link. This code is compatible with the latest python and tensorflow versions.
Below pointers will help future researcher in smooth experiments run:

- Pre-requisite, environment and execution sequence details with output and expected time duration are also added in the same repository.
- Computation requirements and the suggested platform is provided with code documentation.
- MIMIC-III Dataset and pre-processed data can be accessed with approved physionet credentials.

6 Communication with original authors

We were able to connect to the original author Batuhan Bardak quite early. We had multiple email threads communication to discuss various aspects of the paper, covering computational requirements, execution environment, code details and observed results. The author was quite responsive and provided the required guidance. We informed him regarding the results, the proposed model has worked best only in the case of LOS >5 days and somehow the baseline model is performing better for rest of the experiments. The author is keen to review our results and has mentioned that he will perform the experiment on his end. The email conversation is attached to the code repository in pdf format. The communication email can be found at this link

References

Batuhan Bardak and Mehmet Tan. 2021. Improving clinical outcome predictions using convolution over medical entities with multimodal learning. *Artificial Intelligence in Medicine*, 117:102112.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. 2019. Using clinical notes with time series data for icu management. *arXiv preprint arXiv:1909.09702*.

Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, and Alejo Nevado-Holgado. 2021. Med7: a transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, 118:102086.