Enhance the after-discharge mortality rate prediction via learning from the medical notes

Zijiang YANG
The University of Texas at Austin
Austin, Texas, United States
zy4957@utexas.edu

Abstract

With the increase of the Electronic Health Records (EHR) data, more and more researchers are developing machine learning models to learn from the medical notes. These unstructured text data pose significant challenges on the learning process as the quality of data is low. These data are often messy, repetitive and redundant. We show the usefulness of these notes data to be informative by conducting the after-discharge mortality rate prediction task. The AUC-ROC for models using the medical note information is generally 0.1 higher than those without the medical notes. Furthermore, we propose the Deep Neural Network(DNN) model with 'pooling' mechanism to enhance the mortality prediction. Based on the experimental results, we demonstrate that the proposed model outperforms the traditional machine learning models like the tree-based models. The AUC-ROC for the proposed model is 2% to 14% higher than the traditional ones. Moreover, we can discover new knowledge through the proposed model. These knowledge are consistent with the previous findings.

Keywords

Machine learning, Neural network, Medical notes, Healthcare, Mortality prediction

1 Introduction

The Electronic Health Records (EHR) data has revolutionized the patient management, treatment planning, readmission and length-of-stay prediction. The data consist of both the basic information and detailed diagnoses and treatments, providing unprecedented information to discover the patients' health condition and study the disease progression. However, the data contains both the structured and unstructured ones.Structured data such as the gender, age, ethnicity and marital status are easy to process and analyzed. Researchers are currently focusing on harness the power from the unstructured data such as the medical notes. However, these unstructured text data pose significant challenges on the learning process as the quality of data is low. These data are often messy, repetitive and redundant.[6]

Although the unstructured data is of low quality, to show the usefulness of these notes data, we conduct the experiments on the after-discharge mortality rate prediction task. Specifically, we conduct the experiments on the patients diagnosed as 'kidney failure'. The AUC-ROC for models using the medical note information is generally 0.1 higher than those without the medical notes.

Furthermore, we propose the implementation of a Deep Neural Network (DNN) model that integrates an advanced 'pooling' mechanism to enhance the accuracy and effectiveness of mortality prediction. This pooling mechanism, which involves weighting and aggregating information across different categories of the medical notes. It is improving the model's ability to by capturing the important information in the medical notes data. By leveraging this mechanism, the DNN model is able to retain important information and ignore the redundant one, thus making the model more predictive.

Through comprehensive experiments and evaluation, we demonstrate that our proposed DNN model significantly outperforms traditional machine learning models, particularly tree-based models such as Random Forests, and Extreme Gradient Boosting model. The experimental results show that the proposed DNN model consistently achieves higher performance metrics in mortality prediction across various time span. Specifically, the Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC) for the DNN model is observed to be between 2% and 14% higher than that of the traditional tree-based models.

Moreover, we can discover new knowledge through the proposed model. These knowledge are consistent with the previous findings. For example, we discover that the medical notes under 'Discharge summary' category are most informative.

The paper is organized as follows: section 2 introduces the related work; section 3 gives the details of the methodology; section 4 presents and analyze the experimental results; section 5 shows the discovered knowledge from the proposed model; section 6 concludes this paper.

2 Related work

Previous researchers have done a lot of work on analyzing the electronic health records(EHR) and solving multiple tasks. However, we still have much difficulty in understanding and extracting the useful information from the vast amount of EHR data.

Weir et al.[6] identified difficulties in using the electronic medical documentation: 1) information overload; 2) hidden information; 3) lack of trust; These factors will be detrimental to making decisions.

Edin et al. reproduced, compared, and analyzed state-of-the-art automated medical coding machine learning models. They noted the weakness of previous work due to weak configurations, poorly sampled train-test splits, and insufficient evaluation[1].

West et al.[7] and Grover et al.[3] pointed out that the physicians would burn out and blindly copy the computer generated data to the medical notes. These will inevitably create much unnecessary and irrelevant information that may not reflect the patients' symptoms.

3 Methodology

3.1 Basics of the data

The study uses MIMIC III data to learn medical notes. MIMIC-III is a large and freely available database comprising deidentified

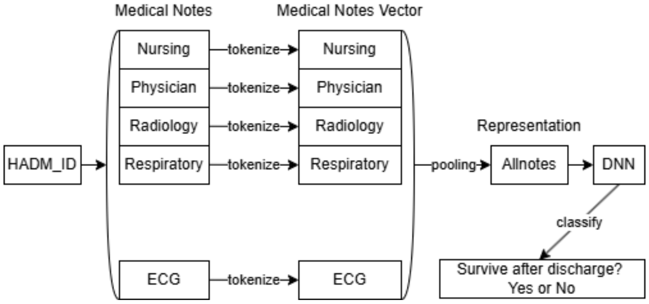

Figure 1: Flow chart for medical notes representation

health-related data associated with more than 40,000 patients who remained in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The MIMIC-III Clinical Database is available on PhysioNet. https://physionet.org/content/mimiciii/1.4/

In particular, we used the following tables from the dataset:

- *patient* This records the basic information of the patient such as the age, ethnicity, marital status and so on.
- *admission* This records the ICU admission and discharge information for the patients.
- *diagnoses* This records the diagnoses for the patients.
- *ICD diagnosis* This records the ICD code of the diagnosis for the patients.
- *noteevents* This records the medical notes for the patients.

In our experiments, for illustration, we focus on the patients that are diagnosed as 'kidney failure'. We filter out the outliers of patients with age over 120 and the patients with 'DEAD/EXPIRED' discharge location. We are left with 6365 legitimate patients. The target label we want to prediction is the mortality after discharge from the ICU. Specifically, based on the discharge date and the death date, we compute the label 'Survive 30 days after discharge' for each patient: 1 for survival and 0 for death. The same process is repeated for 15-days, 60-days, 90-days and 365-days.

3.2 Pre-processing of the data

3.2.1 Statistics of the medical notes. We first gather all medical notes corresponding to 6365 patients and apply the tokenization technique from the *nltk* package to tokenize each medical note. The basic statistics for the medical notes are displayed in the table.1. Majority of the medical notes are Nursing/other notes followed by the Radiology reports. These two consist of more than half of the medical notes. The discharge summaries has the longest token length among all the categories as they are typically very long.

After the tokenization, the *CountVectorizer* in *sklearn* package is used to count the frequency for the top 400 most frequant tokens. The counts of the tokens form a 400-dimension feature vector. For traditional methods mention below, we count the number of tokens of all the medical note for each hospital admission ID(HADM_ID).

3.3 Traditional models

This mortality prediction task is basically a binary classification task. For comparison, we set up the traditional machine learning models to perform the classification. Followings are the model we consider:

Category	Count	Percentage	Average token length
Nursing/other	80644	0.324526	153.2393
Radiology	55368	0.222811	185.6273
Nursing	39717	0.159828	264.2057
Physician	26705	0.107466	751.9017
ECG	23053	0.092769	30.2879
Discharge summary	7396	0.029763	1623.0770
Respiratory	5917	0.023811	145.7208
Echo	4983	0.020052	320.1120
Nutrition	1794	0.007219	268.7352
General	1487	0.005984	204.1432
Rehab Services	806	0.003243	403.7506
Social Work	342	0.001376	306.4766
Case Management	239	0.000962	144.0084
Pharmacy	27	0.000109	229.4815
Consult	20	0.000080	880.5000

Table 1: Categories of the medical notes

- *Logistic regression* The logistic regression model predicts probabilities using a sigmoid function (also called the logistic function), which maps any input value (a linear combination of input features) to a value between 0 and 1.
- *Random forest* This is an ensemble learning method used for both classification and regression tasks. It combines multiple decision trees to improve the overall model's accuracy and robustness.
- *XGBoost* This model, Extreme Gradient Boosting, is an optimized, scalable, and efficient implementation of Gradient Boosting. It is widely used in machine learning tasks, especially for structured/tabular data.

The above models achieves decent results in the traditional downstream tasks like readmission rate prediction and length-of-stay prediction.[1]

3.4 Proposed method

The proposed method uses the 'pooling' mechanism to gather all the feature vectors. For each category, we count the tokens and form a 400-dimension feature vector f_i. The representation for all notes, denoted as $f_{allnotes}$, is computed as follows:

$$f_{allnotes} = \sum_{i=1}^{N} f_i w_i \qquad (1)$$

where w_i is weight for i-th category and N is number of categories. Previous researchers treat all text documents equally, where w_i is the same for all i. Its inherently the idea of the Deep Average Network (DAN) [5]. However, medical notes often contain repetitive content. For example, current Nursing notes often repeats the content of the previous Nursing note. In addition, different types of medical notes convey different messages. Treating them equally will down-weigh the important ones and up-weigh the repetitive ones.

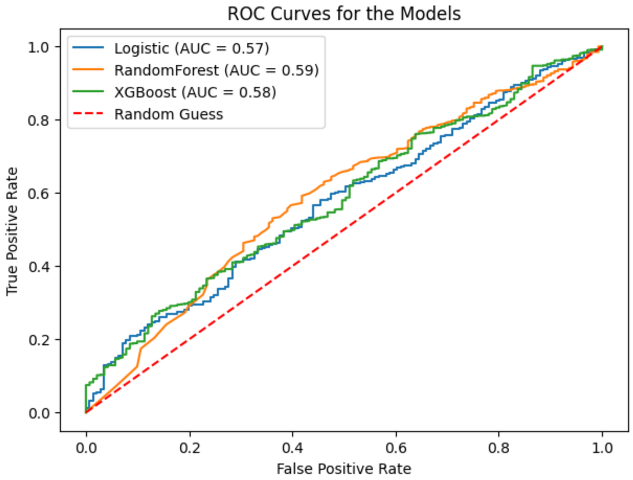

Figure 2: AUC-ROC curve on 30-days mortality prediction using multiple machine learning classifiers without medical notes

We build up our method by constructing the pooling weight and passing it into a Deep Neural Network(DNN) as a learnable parameter. In our experimental settings, we configure the DNN as 4-layer neural network with 2 hidden layers. The hidden layer size is set to be 70. All the hidden layers have short-cut links interconnected.

To train the model, we carefully initialize the weights, use the Adam optimizer and set the learning rate to be as small as 0.0005. As the over-fitting issues are fairly common in medical tasks, we use the validation set to early-stop the training process.

4 Experimental results

We first start with the simplest mortality prediction based on the basic patient information only. The survival 30 days after discharged is predicted using the traditional methods mentioned above. The samples are randomly split into the training, validation and test sets with ratio 7 : 1.5 : 1.5. The hyper-parameters are learned via the validation set. The experimental results are reported as the AUC-ROC graph in the figure.2.

4.1 Baseline model using patient basic information only

It is reported that these 3 models have very similar model accuracy. Based on the result, we can see that these models cannot deliver a good result. The AUC is just around 0.58, which barely outperforms the random guessing.

4.2 Baseline model using medical notes information

By tokenizing the medical notes and constructing them as input feature vectors, we repeat the experiments using the same models. We are able to observe a significant improvement: all models perform better than basic information only; the AUC increase by 0.09 from 0.59 to 0.68 for the Random forest method. This results demonstrate the effectiveness of the medical notes in mortality prediction, which contradicts to the Ghassemi et al.[2] as they used a limited set of structured variables.

We then apply our method on the mortality prediction task for 15-days, 30-days, 60-days and 365-days time spans. It is reported that this DNN based method achieves best performance across all the time spans. It outperforms the traditional machine learning methods by 2% to 14% based on the AUC-ROC criterion. It is worth noting that the DNN gives a better performance than the prevailing tree based methods for the medical prediction task.

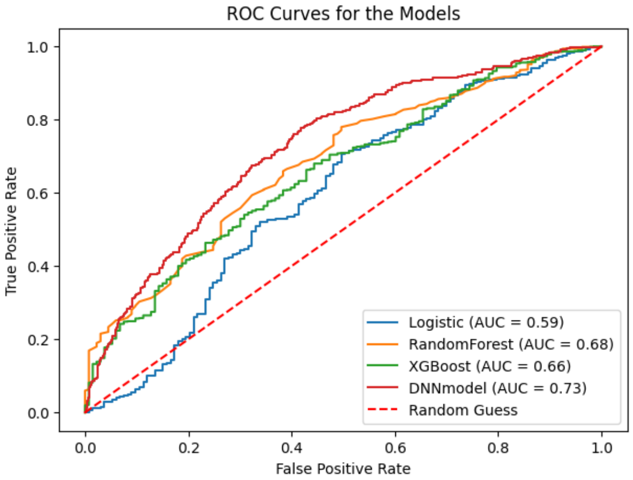

Figure 3: AUC-ROC curve on 30-days mortality prediction using multiple machine learning classifier with medical notes

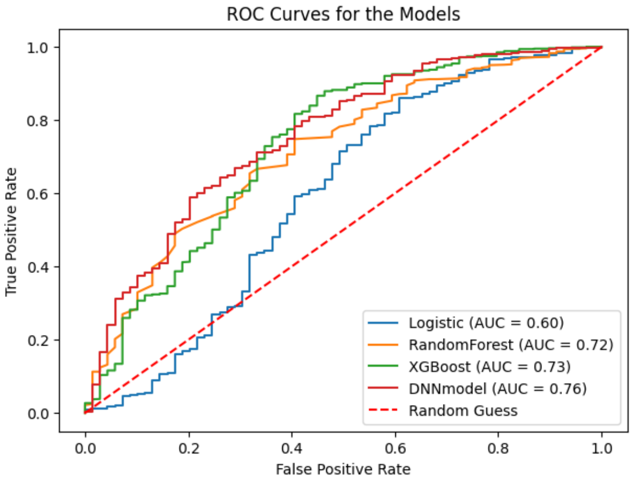

Figure 4: AUC-ROC curve on 15-days mortality prediction using multiple machine learning classifier with medical notes

5 Knowledge discovery

Based on the DNN model constructed, we are able to perform the analysis on the medical notes. Firstly, we can analyze the weights of

Category	Weight	Sensitivity	Normalized Sensitivity	Token length correlation with survival
Nursing/other	**0.255617**	0.002279	0.349164	-0.02560
Radiology	0.036422	-0.000036	-0.006710	**-0.11285**
Nursing	-0.163920	0.001789	0.472738	-0.07814
Physician	0.001094	0.000506	0.380522	-0.08529
ECG	0.020193	-0.002660	-0.080600	-0.03462
Discharge summary	**0.243545**	-0.001050	**-1.697190**	**-0.22884**
Respiratory	**0.230256**	-0.000340	-0.049690	-0.05800
Echo	0.082943	-0.002190	-0.699590	-0.07606
Nutrition	-0.185750	-0.001080	-0.290890	-0.06459
General	0.150531	0.000165	0.033651	-0.04432
Rehab Services	-0.082580	0.004298	1.735236	-0.04160
Social Work	0.058556	0.001955	0.599275	-0.03227
Case Management	-0.192240	-0.006180	-0.889630	-0.05788
Pharmacy	-0.089710	0.000913	0.209617	-0.01427
Consult	-0.121180	-0.000430	-0.374790	-0.00873

Table 2

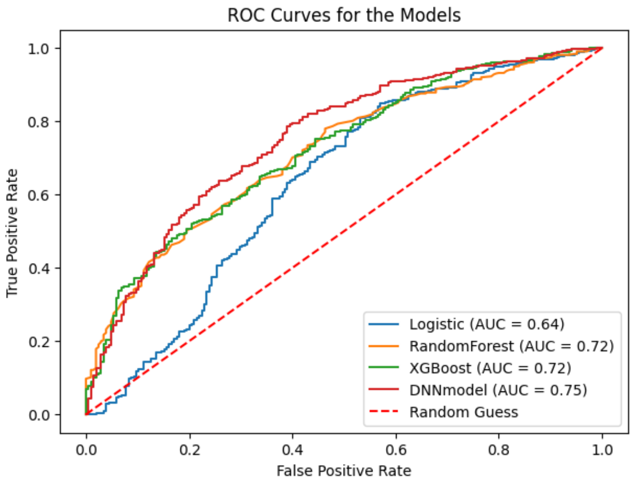

Figure 5: AUC-ROC curve on 60-days mortality prediction using multiple machine learning classifier with medical notes

each category passed into the DNN model. The table.2 tabulates the weights. As is shown in the table, the categories 'Nursing/other', 'Nursing', 'Discharge summary' and 'Respiratory' contributes most to the prediction result. This is consistent with the preliminary results from Hsu et al.[4]. In their experiments on readmission prediction, the discharge summary is the most informative category among the medical notes.

We then compute the sensitivity of the prediction score y_{pred} with respect to each token via:

$$\frac{\partial y_{pred}}{\partial countoftoken_i} \approx \frac{\Delta y_{pred}}{\Delta countoftoken_i} = y_{pred}|(token_i + 1) - y_{pred}|token_i \quad (2)$$

We artificially alter the count of token i and see the change in the prediction score y_{pred}. The results are tabulated in the table.2.

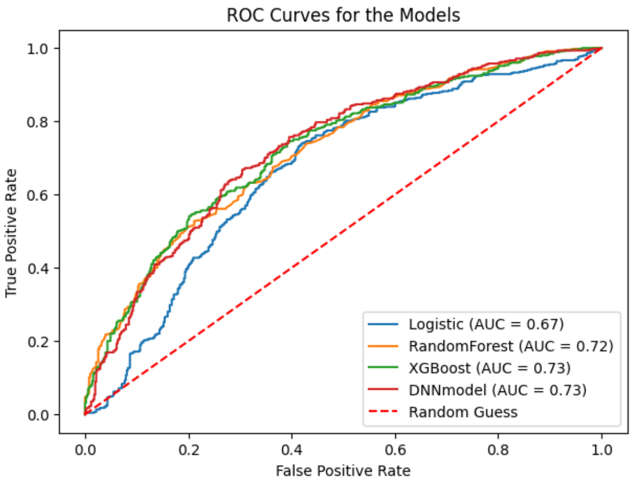

Figure 6: AUC-ROC curve on 365-days mortality prediction using multiple machine learning classifier with medical notes

As different medical notes have different lengths, we account for this by multiplying the average token length in the table.1. The normalized sensitivities are generated. The meaning of this quantity is to measure how the prediction score will change when a new medical note is presented. Again, it is reported that the discharge summary is the most influential medical note for predicting the mortality. (Ignore the last few rows as they contains very limited samples.)

Finally, we compute the correlation of the token length and the patients' survival in the table.2. So, we are investigating how the token length of the medical notes will after survival. Generally, they are negatively correlated with survival, which means longer the notes, more likely the patient will not survive. The discharge summary is most negative one, suggesting that it is a very strong predictor for the survival.

There are also few interesting observations from the table. It is shown that the radiology and ECG reports are not informative for prediction based on the weight and sensitivity analysis. Physician notes help with the prediction but not as strong as the nursing notes. Lastly, the Echo notes information is also helpful for this prediction task.

6 Conclusion

Based on the experiments, our results confirm the value of medical notes for the after-discharge mortality prediction. The AUC-ROC for the models using medical note information is generally 0.1 higher than that without using the medical notes.

While Natural Language Processing techniques are shown to be effective in extracting out the useful information, we further enhance the mortality prediction by introducing the 'pooling' mechanism. Through assigning the weights to different categories and learning the weights through the training process, we are able to discover the significance of each medical notes. The performance of our DNN model achieves the best prediction performance for 15-days, 30-days, 60-days and 365-days after-discharge mortality. The proposed model outperforms the traditional models by 2% to 14% based on the AUC-ROC criterion. In addition, it is found that 'Discharge summary' and 'Nursing' are the most informative categories for predicting the mortality among the categories of the notes.

References

[1] Joakim Edin, Alexander Junge, Jakob D. Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. 2023. Automated Medical Coding on MIMIC-III and MIMIC-IV: A Critical Review and Replicability Study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2572–2582. doi:10.1145/3539618.3591918

[2] Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. 2014. Unfolding physiological state: mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, New York, USA) *(KDD '14)*. Association for Computing Machinery, New York, NY, USA, 75–84. doi:10.1145/2623330.2623742

[3] Sandeep Grover, Himani Adarsh, Chandrima Naskar, and Natarajan Varadharajan. 2018. Physician burnout: A review. *Journal of Mental Health and Human Behaviour* 23 (01 2018), 78. doi:10.4103/jmhhb.jmhhb_47_19

[4] Chao-Chun Hsu, Shantanu Karnwal, Sendhil Mullainathan, Ziad Obermeyer, and Chenhao Tan. 2020. Characterizing the Value of Information in Medical Notes. doi:10.48550/arXiv.2010.03574

[5] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong and Michael Strube (Eds.). Association for Computational Linguistics, Beijing, China, 1681–1691. doi:10.3115/v1/P15-1162

[6] Nebeker JR. Weir CR. 2007. Critical issues in an electronic documentation system.. In *AMIA Annu Symp Proc.*, Vol. 3. 786–790. doi:99.9999/woot07-S422

[7] Shanafelt TD. West CP, Dyrbye LN. 2018. Physician burnout: contributors, consequences and solutions. *Journal of Internal Medicine* 283 (06 2018), 516–529. doi:10.1111/joim.12752