Improving EHR-based Predictions: Unifying ClinicalBERT and Variationally Regularized GNNs for Mortality Prediction Task

Yuyu (Ciel) Wang
New York University
yw7104@nyu.edu

Iktae Kim
New York University
ik798@nyu.edu

Abhilash Anand
New York University
aa9798@nyu.edu

Keigo Ando
New York University
ka2705@nyu.edu

Abstract

Electronic Health Records (EHR) are valuable sources of patient information for various predictive tasks in healthcare settings, including hospital readmission and mortality prediction. However, the sparsity and unstructured nature of EHR data make it difficult to extract useful information. Graph Neural Networks (GNN) have been proposed as a way to represent medical concepts from EHR, and have shown promising results in various predictive tasks. In this paper, we propose to enhance the Variationally Regularized Encoder-Decoder Graph Network model by incorporating the ClinicalBERT representation of patient encounters as additional input features. We find that our combined model outperforms both the fine-tuned ClinicalBERT and GNN model on the MIMIC-III mortality prediction task benchmark.

1 Introduction

Electronic Health Records (EHR) serve as a rich resource of patient data for numerous prognostic tasks within healthcare environments, such as predicting hospital readmissions and estimating patient mortality. Nonetheless, EHR data frequently exhibits sparsity and a lack of structure, which presents difficulties in deriving valuable insights (Holmes et al., 2021). Graph Neural Networks (GNN) have been proposed as a method for representing medical concepts derived from EHR, demonstrating encouraging outcomes in a range of prognostic tasks. One such model was proposed by Zhu and Razavian (2021), who developed a variationally regularized encoder-decoder graph network that achieves more robustness in graph structure learning by regularizing node representations.

ClinicalBERT (Alsentzer et al., 2019) is a Transformer encoder pre-trained on clinical notes using the BERT objective, which can be useful for predicting hospital readmission and mortality. In this paper, we propose to enhance the Variationally Regularized Encoder-Decoder Graph Network model (Zhu and Razavian, 2021) by incorporating the task specific fine-tuned ClinicalBERT representation of patient encounters as additional input features. We will then evaluate the combined model on the MIMIC-III (Pollard and Johnson III, 2016; Johnson et al., 2016; Goldberger et al., 2000) mortality prediction task.

In this paper, we propose the integration of task specific fine-tuned ClinicalBERT (Alsentzer et al., 2019) representation with Zhu and Razavian's (2021) variationally regularized formation of encoder-decoder graph neural network (VGNN) model and experiment whether this the combined model will improved the performance on MIMIC-III mortality prediction task, and compare them to the individual performance of ClinicalBERT and VGNN model.

1.1 GNN Architecture

Graph Neural Networks (GNNs) have emerged as a powerful tool for learning representations from structured data, especially graph-structured data, which is highly relevant in a healthcare setting. The basic premise of GNNs is to propagate information across the nodes of a graph, allowing each node to gather information from its neighborhood. This propagation is performed through multiple iterations, or layers, enabling each node to collect information from an increasingly larger context.

In the context of EHR data, each patient can be represented as a node, and the edges between the nodes can represent various relationships, such as similar diagnoses, treatments, or demographic characteristics. Each node carries initial features derived from the EHR, such as demographics, diagnosis codes, and treatment codes, and these features are updated through the GNN's iterative process, ultimately leading to a rich representation of each patient that captures not only their own health

information, but also the context of other similar patients.

However, a common challenge faced by traditional GNNs is the tendency to generate overly uniform node representations. These representations often collapse into tightly clustered groups, leading to uniformly-distributed attention weights and hindering the learning of meaningful connections among nodes. This is especially problematic in healthcare settings, where the objective is to capture nuanced differences between patients' health states.

1.2 VGNN Architecture

To address this challenge, the Variationally Regularized Encoder-Decoder Graph Network model (VGNN) was proposed. VGNN introduces a variational regularization that encourages node representations to be centered around the origin with moderate distances, preventing the collapse of representations into tightly clustered groups and promoting the learning of meaningful connections.

The VGNN architecture is inspired by the Variational Graph Auto-Encoder (VGAE), which improves link inference by assuming a Gaussian prior on the node representations. The VGNN extends this concept by adding a latent layer between the encoder and decoder to regularize the graph representation.

The latent variables of each observed node representation are assumed to follow a standard normal distribution. The generative encoder distribution is parameterized to ensure non-negativity. The sampled latent variables then replace the initial node representations and become the inputs to the decoder layer.

In the VGNN, the Evidence Lower Bound (ELBO) is maximized, which consists of two terms: the loss for reconstructing the input and the Kullback-Leibler (KL) divergence between the prior distribution and the likelihood of the latent space. The KL-term regularizes the node representations to center around the origin, while the reconstruction term ensures sufficient distance between the node representations to prevent mode collapse and retain expressiveness.

In practice, the VGNN combines the KL-term with a cross-entropy loss as the loss function to minimize. The result is a model that provides a balance between regularization and expressiveness, offering a promising approach for learning meaningful medical concepts from complex healthcare data (Zhu and Razavian, 2021).

2 Related Work

In a survey paper, Xu et al. (2022) provide an introduction to existing deep learning-based methods on EHR and the most recognized EHR datasets.

Zhu and Razavian (2021) proposed a Variationally Regularized Encoder-Decoder Graph Network model (VGNN) that enhances graph structure learning in EHR data by regularizing node representation.

Alsentzer et al. (2019) introduce a pre-trained language model called "ClinicalBERT" that is specifically designed for clinical text. The model is trained on a large corpus of electronic health records and can generate high-quality representations of medical concepts. The authors also release publicly available embeddings generated by the model that can be used for various clinical natural language processing tasks.

Huang et al. (2019) utilized the pre-trained ClinicalBERT, and demonstrated its effectiveness for predicting hospital readmission.

Golmaei and Luo (2021) proposed a GNN model for predicting hospital readmission based on clinical notes and patient networks. It demonstrates the effectiveness of GNNs for predicting hospital readmission and served as an additional guideline in developing our model.

Shang et al. (2019) proposed pre-training Graph Augmented Transformers for medication recommendation, which demonstrates the potential of combining graph-based methods with transformer models for healthcare applications.

Lyu et al. (2022) explored the fusion of clinical notes with structured EHR data using a multimodal transformer for interpretable in-hospital mortality prediction, and demonstrated the potential benefits of combining textual data with structured data, such as the output of ClinicalBERT and GNN models, to improve the performance of predictive tasks in healthcare.

3 Methodology

In this section, we describe our methodology for data preprocessing and the integration of ClinicalBERT representation with VGNN. We evaluate the proposed methods by predicting mortality within 24 hours after admission

based on the MIMIC-III cohort. The code implementation and resources used for this study can be found in our GitHub repository at https://github.com/kimit0310/Unified-ClinicalBERT-VGNN.

3.1 Dataset Description

MIMIC-III (Medical Information Mart for Intensive Care III, (Pollard and Johnson III, 2016; Johnson et al., 2016; Goldberger et al., 2000)) is a large, freely-available dataset of de-identified electronic health records (EHRs) of patients admitted to the intensive care units (ICUs) at Beth Israel Deaconess Medical Center between 2001 and 2012. The dataset includes information such as demographics, diagnoses, medications, laboratory results, and survival outcomes for over 50,000 patients. In the context of the ClinicalBERT and VGNN training and testing sets, we leveraged a critical index known as 'HADM_ID'. This identifier, unique to each hospital admission, facilitated the integration of clinical notes for individual patients, providing a holistic perspective of their medical histories. This approach enabled us to align the datasets effectively and maintain consistency between the training and testing stages.

Clinical Notes Clinical notes are free-text descriptions written by healthcare providers during a patient's hospital stay. These notes may include nursing notes, physician notes, and discharge summaries, among others. They typically contain detailed information about a patient's medical history, symptoms, physical examination findings, medical tests, treatments, and healthcare providers' observations and plans.

3.2 ClinicalBERT Preprocessing

To handle the large volumes of clinical note data in the MIMIC-III dataset, we adopt a chunk-based processing approach. This approach allows us to work with manageable portions of data, reducing memory requirements for scalability.

Details on the preprocessing can be found in Appendix A

By following this preprocessing methodology, we ensure that the clinical notes data is properly cleaned, normalized, and aggregated for subsequent analysis.

3.3 VGNN Preprocessing

For the preprocessing of MIMIC-III data for VGNN, the process is nearly identical to Zhu and Razavian's 2021, in processing the patient, diagnosis, treatment data, and laboratory results. The only difference is that our model requires the concatenation of 'HADM_ID's to the input data in order to match the inputs with the CLS features extracted from our ClinicalBERT model. Details on the preprocessing can be found in Appendix B

3.4 ClinicalBERT Fine-tuning

We employed the ClinicalBERT model, which underwent a process of training using our designated training dataset, and subsequent evaluation on the validation dataset. The allocation of data into training, validation, and test sets for ClinicalBERT was done in accordance with the HADM_ID, ensuring consistency with the datasets used for the VGNN.

In order to address the class imbalance in our dataset, originally characterized by a distribution of 10.21% for the positive class and 89.79% for the negative class, we implemented the weighted cross-entropy loss function during training and increased the positive class representation by twofold. This approach is in line with the preprocessing steps outlined in the original VGNN study, effectively addressing the challenge posed by the class imbalance.

3.5 VGNN Fine-tuning with Additional Input from ClinicalBERT

The Variational Graph Neural Network (VGNN) is trained on preprocessed data, augmented with the CLS token (EHR representation) sourced from ClinicalBERT. The CLS token serves as an additional set of input features (N = 768) for the VGNN model, enriching its original structure. Aside from this addition, the training procedure remains in strict accordance with the methodology proposed by Zhu and Razavian (2021)

4 Results

Table 1 shown in this section compares the performance of different models on the MIMIC-III test dataset for mortality prediction. We preferred the model which showed the highest Area Under the Precision-Recall Curve (AUPRC) in the validation set as our primary metric. This decision was based on the inherent class imbalance in the dataset.

Model	AUPRC	Accuracy
ClinicalBERT	0.5340	0.8746
VGNN	0.6864	0.9157
ClinicalBERT+VGNN	**0.7387**	**0.9332**

Table 1: Performance Comparison of Different Models on Prediction of Mortality

The models compared were ClinicalBERT, VGNN, and a combination of ClinicalBERT+VGNN. The combined model showed the best results with an AUPRC of 0.7387 and an accuracy of 0.9332. This comparison demonstrates that our approach of integrating ClinicalBERT and VGNN outperforms using them individually.

It should be noted that despite our efforts to address the class imbalance during training by upsampling the positive class, and using a weighted loss, the imbalance is inherent and was not adjusted for the validation set. This validates our decision to prioritize the AUPRC as a key performance metric to capture the performance of our models accurately in light of the prevailing class imbalance.

5 Discussion and Conclusion

We aimed to develop an effective approach for mortality prediction using Electronic Health Records in this study. Our approach combined ClinicalBERT, with the Variationally Regularized Encoder-Decoder Graph Network (VGNN), which has shown promise in graph structure learning for EHR data.

Our combined model showed better performance than the standalone ClinicalBERT and VGNN models on the MIMIC-III mortality prediction benchmark. This indicates the potential of combining structured EHR data with unstructured clinical notes to improve predictive performance.

We faced a significant challenge with the inherent class imbalance in the MIMIC-III dataset. Although we mitigated this imbalance during the training phase using upsampling strategies and a weighted loss function, it remained a substantial issue.

Our model has the potential to integrate diverse data types and machine learning models for healthcare predictive tasks. Future work could explore other ways to combine ClinicalBERT and VGNN, such as integrating the VGNN to represent the internal hierarchical structures of medical codes, or investigate the integration of other types of EHR data. Our methodology could also be extended to other predictive tasks in healthcare.

Zhu and Razavian's model on which our research is based has proven to be inherently versatile. It has demonstrated its robustness and effectiveness on a variety of tasks and datasets. For our model as well, it may be worthwhile in the future to evaluate the generalizability of this model by, for example, applying it to the task of long-term readmission prediction using the eICU dataset and evaluating its performance. This would provide valuable insight into the consistency and reliability of the model across different time scales and tasks.

Future work can explore other ways to combine ClinicalBERT and VGNN, such as integrating the VGNN to represent the internal hierarchical structures of medical codes and integrating into a transformer-based visit encoder (ClinicalBERT), similar to the work of Shang et al. (2019), or investigate the integration of other types of EHR data, such as imaging or genomic data. Furthermore, the methodology developed in this work could be extended to other predictive tasks in healthcare, such as disease progression or treatment response prediction.

In conclusion, our findings suggest that the integration of structured EHR data and unstructured clinical notes, represented through models such as ClinicalBERT and VGNN, can significantly enhance the performance of predictive tasks in healthcare.

6 Limitations

Despite the promising results, our study has several limitations. The first is that our model was trained and validated solely on the MIMIC-III dataset, which includes data from only one medical center. This may limit the generalizability of our model to other settings, as EHR systems and patient demographics may vary across different healthcare providers. Furthermore, our model primarily relies on the assumption that EHR data is complete and accurately recorded, which may not always be the case in real-world settings. The quality and completeness of EHR data can be influenced by various factors, such as the healthcare provider's data entry practices and the specific EHR system being used. Lastly, the inherent class imbalance in the MIMIC-III dataset was

a significant challenge, and while we attempted to address it during training, it is still a limitation that could affect the model's performance.

7 Ethical Considerations

Several ethical considerations arise in the development and application of predictive models in healthcare, especially when using patient data. One such issue is patient privacy. Even though the MIMIC-III dataset is de-identified, it is important to ensure that any model developed does not inadvertently re-identify individuals or expose sensitive information. Furthermore, there are potential risks of algorithmic bias, where the model might perpetuate existing disparities in healthcare due to biases in the training data. For instance, if the EHR data used to train the model disproportionately represents certain demographic groups, the model's predictions may be less accurate for underrepresented groups. Therefore, it is crucial to ensure fairness in the model's predictions across different patient populations. Lastly, the use of such models in clinical decision-making should be carefully considered, as over-reliance on algorithmic predictions could potentially devalue the role of clinical judgment and patient autonomy. It is essential that predictive models are used as supportive tools in healthcare, complementing rather than replacing the expertise of healthcare professionals.

8 Contributions

In this work, Keigo, Iktae, Abhilash, and Ciel have contributed equally to the literature review, proposal write-up, paper drafting, fine-tuning of ClinicalBERT, and integration of the CLS token from ClinicalBERT into the VGNN. Iktae, Abhilash, and Ciel shared equal responsibilities in preprocessing the MIMIC-III data. Keigo was responsible for training the VGNN.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, pages 72–78. https://doi.org/10.18653/v1/W19-1909.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation* 101(23):e215–e220.

S. N. Golmaei and X. Luo. 2021. Deepnote-gnn: predicting hospital readmission using clinical notes and patient network. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. pages 1–9.

John H Holmes, James Beinlich, Mary R Boland, Kathryn H Bowles, Yong Chen, Tessa S Cook, George Demiris, Michael Draugelis, Laura Fluharty, Peter E Gabriel, et al. 2021. Why is the electronic health record so challenging for research and clinical care? *Methods of information in medicine* 60(01/02):032–048.

K. Huang, J. Altosaar, and R. Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint 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.

W. Lyu, X. Dong, R. Wong, S. Zheng, K. Abell-Hart, F. Wang, and C. Chen. 2022. A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction. *arXiv preprint arXiv:2208.10240* .

Tom J Pollard and AEW Johnson III. 2016. The mimic iii clinical database, version 1.4. *The MIMIC-III Clinical Database. PhysioNet* .

J. Shang, T. Ma, C. Xiao, and J. Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346* Https://github.com/jshang123/G-Bert.

J. Xu, X. Xi, J. Chen, V. S. Sheng, J. Ma, and Z. Cui. 2022. A survey of deep learning for electronic health records. *Applied Sciences* 12(22):11709.

W. Zhu and N. Razavian. 2021. Variationally regularized graph-based representation learning for electronic health records. In *Proceedings of the Conference on Health, Inference, and Learning*. pages 1–13. Https://github.com/NYUMedML/GNN_for_EHR.

A Appendix A

Text Cleaning: We employ a combination of regular expressions and string operations to clean the text data. This includes removing de-identified

brackets, numeric sequences, abbreviations, and special characters, newline and carriage return characters, and eliminating excess spaces. Although the original ClinicalBERT did not clean their pre-training data exactly as our steps, our steps replicate the text cleaning processes from Huang et al. (2019)

Aggregation and Grouping: In order to facilitate analysis at the patient level, we aggregate the preprocessed text based on the Hospital Admission ID ('HADM_ID') which is a unique ID given for each admissions to a hospital. This allows us to consolidate the clinical notes for each patient, enabling a comprehensive view of their medical history.

B Appendix B

Processing Patient Data: We extract relevant information from the patient data, including the patient ID, encounter ID, and encounter timestamp. This data is organized into a dictionary structure, sorted chronologically by encounter timestamp.

Processing Diagnosis Data: We extract the encounter ID and corresponding diagnosis codes from the diagnosis data. These codes are appended to the respective encounters, linking them to the specific medical conditions identified.

Processing Treatment Data: Similar to the diagnosis data, we process the treatment information associated with each encounter. The encounter ID and treatment codes are extracted and appended to their respective encounters, ensuring accurate association of treatment information with each patient encounter.

Calculating Laboratory Mean and Standard Deviation: To assess the significance and variability of laboratory results, we calculate the mean and standard deviation of the laboratory values. This statistical information serves as a basis for subsequent analysis.

Processing Laboratory Data: The laboratory data is processed by extracting the encounter ID, laboratory ID, and laboratory timestamp. Each laboratory result is compared to the calculated mean and standard deviation. Based on this comparison, an appropriate suffix is added to the laboratory ID, indicating the range within which the result falls.