Predicting ICU Length of Stay for Patients using Latent Categorization of Health Conditions

Sudeshna Jana, Manjira Sinha and **Tirthankar Dasgupta**
TCS Research
India
(sudeshna.jana, sinha.manjira, dasgupta.tirthankar)@tcs.com

Abstract

Predicting the duration of a patient's stay in an Intensive Care Unit (ICU) is a critical challenge for healthcare administrators, as it impacts resource allocation, staffing, and patient care strategies. Traditional approaches often rely on structured clinical data, but recent developments in language models offer significant potential to utilize unstructured text data such as nursing notes, discharge summaries, and clinical reports for ICU length-of-stay (LoS) predictions. In this study, we introduce a method for analyzing nursing notes to predict the remaining ICU stay duration of patients. Our approach leverages a joint model of latent note categorization, which identifies key health-related patterns and disease severity factors from unstructured text data. This latent categorization enables the model to derive high-level insights that influence patient care planning. We evaluate our model on the widely used MIMIC-III dataset, and our preliminary findings show that it significantly outperforms existing baselines, suggesting promising industrial applications for resource optimization and operational efficiency in healthcare settings.

1 Introduction

Intensive Care Units (ICUs) deliver critical care for severely ill patients, but due to the high costs associated with their setup and operation, hospitals face limitations on the number of available ICU beds. Efficient resource management is essential to maximize ICU capacity and avoid life-threatening shortages. Predictive planning, powered by historical patient data—such as medical history, test results, treatments, nursing notes, and previous ICU admissions—can significantly enhance the allocation of ICU resources. By leveraging advanced analytics and machine learning models, healthcare providers can optimize bed usage, streamline staffing, and improve patient outcomes, ensuring that ICU resources are deployed where they are needed most. This approach has wide industrial applications in healthcare operations, improving both efficiency and patient care while reducing operational costs.

Nursing notes contain vital information about a patient's physical and psychological condition, offering insights beyond physiological data or radiology reports. These notes also document a patient's response to treatment through behavioral descriptions, making them a rich source for predicting critical care needs. Our model leverages unstructured nursing notes, which include linguistic expressions like "extensive cardiac hx" or "slightly tachypneic," providing human assessments that numerical data alone cannot capture. These details are crucial for distinguishing between similar patients with different treatment responses. Figure 1 illustrates a sample nursing note with highlighted clinical details.

Earlier models typically process all nursing notes as input to predict a specific output, limiting their ability to predict outcomes during the ICU stay (Rocheteau et al., 2020; Gentimis et al., 2017; Harutyunyan et al., 2019; Rocheteau et al., 2020). Recent efforts have aimed at early prediction of ICU length of stay (LoS), readmission, and interventions, but their performance remains sub-optimal due to the lack of domain knowledge and the nuances of text discourse (Alghatani et al., 2021; Su et al., 2021; van Aken et al., 2021; Huang et al., 2019; Li et al., 2024).

In this paper, we present a technique for predicting ICU length-of-stay (LoS) by analyzing nursing notes, a rich source of unstructured data. By extracting health status information from these notes, our model identifies both common and unique features, leading to enhanced prediction accuracy. We introduce a joint model of latent note categorization, which recognizes critical health contexts that shape language patterns in nursing documentation. This model not only improves predictions but also offers insights that can be used for more effi-

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 422–430

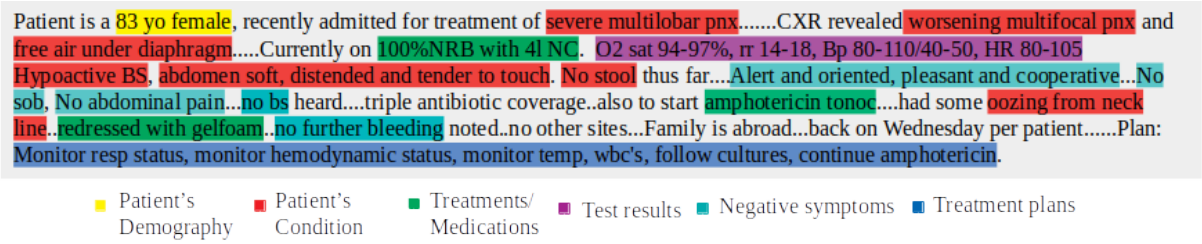

Figure 1: Illustration of a nursing notes with highlighted clinical details.

cient ICU resource management. Evaluated on the MIMIC-III dataset, our approach outperforms competitive baselines, including large language models such as LLAMA-3.1 and fine-tuned BioMistral-7B. These results demonstrate the potential of integrating unstructured text data into industrial applications like predictive healthcare analytics, optimizing ICU operations, and improving patient care strategies.

2 The Proposed LOS Prediction Model

We define the problem as follows: let X be a set of N nursing note transcripts. Each X_i is a sequence of M_i nursing notes for patient i, where P_{ij} represents the j^{th} note in X_i. Each X_i is labeled with the patient's length of stay, Y_i.

The model takes a sequence of nursing notes P_{ij} and predicts the remaining length of stay Y_i. Its success is measured by prediction accuracy and the timestamp at which the correct prediction is made. The earlier the prediction, the more valuable it is to users.

2.1 Processing of unstructured clinical notes

Clinical notes exhibit significant variability in style and content. Some document only symptoms, while others mention absences of symptoms, adverse reactions, psychological states, and appetite changes, often using non-standard terminology and abbreviations. To manage this variability, we added a processing layer that uses biomedical dictionaries to create a structured representation of clinical details. This includes extracting clinical entities such as *diseases or symptoms*, *abnormalities*, *life-style*, *mental health conditions* and *previous health histories* using GPT-4 (Waisberg et al., 2023). Along with the entities, we also identified absence indicators frequently found in clinical notes like, "*absence* of pain", or "*no* history of hypertension". Moreover, the clinical data often encompass diverse non-standard terminology, abbreviations, various formats, and coding systems to represent clinical details. For instance, "Pulmonary Edema" and "fluid in lungs" refers to the same symptom. We standardized these entities using the UMLS Metathesaurus (Schuyler et al., 1993), which assigns a "Concept Unique Identifier (CUI)" to each concept.

Once entities are extracted and represented with CUIs, each day's clinical details for a patient are consolidated using the CUIs observed on that day. Given a patient p, the clinical details at day t is defined by a vector $H_p(t) =< f(d_i) >$, $i = 1, 2, ..., |V|$, where $d_i \in V$ and the value of $f(d_i)$ is set to 1 if d_i present, -1 if it is mentioned negatively, and 0 if d_i is not mentioned in day t.

The diversity of diseases and symptoms, along with individual variability, often results in high-dimensional sparse vectors. To address high dimensionality and sparsity of vectors, we employ an autoencoder-based transformation (Wang et al., 2016) for dense, lower-dimensional representation. The encoder compresses the data to capture essential features, while the decoder reconstructs the original data, retaining key information. These compressed representations $EH_p(t)$ facilitate further processing of patient clinical details. The details of the pre-processing stages are discussed in Appendix-A.

2.2 Representing patient's health condition

A patient's health condition (HC) indicates illness severity and is assessed using various scoring systems based on data such as age, vital signs, lab results, and medical history. We used the following scores: (a) SOFA (Vincent et al., 1996), (b) APACHE (Wong and Knaus, 1991), (c) SAPS (Le Gall et al., 1993), and (d) OASIS (Johnson et al., 2013). We calculated the average of these scores to determine a unified HC for each patient. The HC scores are normalized within a range of [0,5] and are further categorized into five classes namely, $\{0 \leq HC < 1, 1 \leq HC < 2, 2 \leq HC <$

$3, 3 \leq HC < 4, 4 \leq HC < 5\}$. Lower HC score reflects better health condition.

3 Joint Latent Note Categorization for ICU-LoS Prediction

Based on the work of (Rinaldi et al., 2020), we have adopted a similar network architecture for predicting the ICU length of stay(LoS) for an individual patient. We modified the above architecture by categorizing the daily nursing notes for a patient (p) for the day (t) along with the encoded clinical details (H_p^t) of the patient. Thus, we propose an nursing note categorization model that jointly learns to predict the ICU LoS of a patient from the nursing note transcripts and encoded clinical details while grouping the information into their respective health condition (HC) classes. The rational behind the joint categorization is the fact that ICU stay for a patient will largely depend upon patients' progressive health condition.

A detailed overview of the model architecture is depicted in Figure 2. The model is composed of the following components namely,

- Input representation,
- Health condition inference layer
- Latent health condition membership layer
- Health condition aware note aggregation layer
- Decision layer

The details of each of the components are discussed in the following subsections.

We represent every day nursing note of a patient as contextual embeddings $N_t \in \mathbb{R}^E$. Along with this we extract the specific clinical details of the patient $H_p(t)$ from the notes as discussed in section A.5. We concatenate these two representations together and form a patient-centric contextual embedding $P_t \in \mathbb{R}^{E+V}$. Where V is the dimension of the clinical detail vector. We hypothesize that each note can be grouped into K latent categories such that similar category of patient will exhibit unique, useful patterns. We have used the *health condition (HC)* of each patient per day, corresponding to each note as the latent categories. To perform a soft assignment of the notes to the HC classes, for each note, our model computes a category membership vector $h_j = [h_j^1, ..., h_j^K]$. Here, h_j represents the probability distribution for the j^{th} note of the patient over each of K latent categories for the patient's health condition. h_j is computed as a function ϕ of P_j and trainable parameters θ_{CI}. This is depicted as the Category Inference layer:

$$h_i = \phi(P_i, \theta_{CI})$$

Based on these category memberships for each nursing note, the model then analyze the corresponding health categories so that unique patterns can be learned for each category. Specifically, we form K category-aware note aggregations ($\bar{P}_t^k$). Each of these aggregations, $(\bar{P}_t^k) \in \mathbb{R}^{\mathbb{E}}$, is a category-aware representation of all the nursing notes till the t^{th} timestamp with respect to the k^{th} category.

$$\bar{P}_t^k = \frac{1}{Z_t^k} \sum_{t=1}^{M_t} h_t^k P_t \; ; \; Z_i^k = \sum_{j=1}^{M_i} h_{ij}^k$$

Here, h_t^k is the k^{th} scalar component of the latent category distribution vector h_t. Z_t^k is the normalizer added to prevent varying signal strength, which interferes with training. We then compute the output class probability vector y_i as a function ψ of the note aggregations $[\bar{P}_t^1, ..., \bar{P}_t^K]$ and trainable parameters θ_D (illustrated as the Decision Layer in Figure 2). The predicted label Y_i is selected as the class with the highest probability based on y_i.

3.1 The Category Inference Layer

We compute the latent category membership for all notes for a patient X using a feed-forward layer with K outputs and softmax activation:

$$\phi(P_t, \theta_{CI}) = \sigma(row_j(P_t W_{CI} + B_{CI})) \quad (1)$$

As shown in Equation 1, as $\phi(.)$ is computed using a softmax, it generate a probability distribution. Thus, $\phi(.)$ produces the desired category membership vector h_j over latent categories for the j^{th} nursing note of X. $(P_t W_{CI} + B_{CI})$ computes a matrix where row j is a vector of the latent category distribution for the j^{th} note, and σ denotes the softmax function. $(W_{CI}) \in \mathbb{R}^{E \times K}$ and $(B_{CI}) \in \mathbb{R}^K$ are the trainable parameters for this layer:

$$\theta_{CI} = \{W_{CI}, B_{CI}\} \quad (2)$$

3.2 The Decision Layer

The decision layer models the probabilities for remaining length of stay using a regression model.

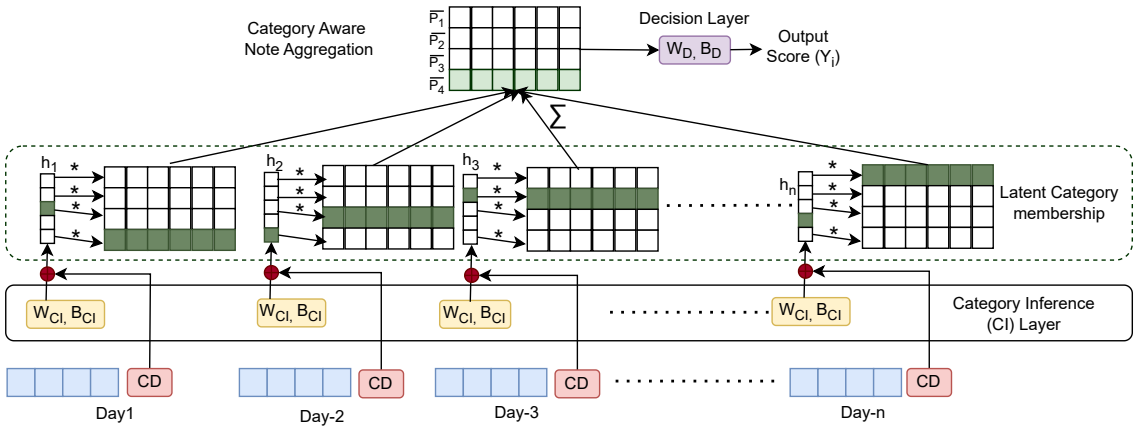

Figure 2: Overview of the joint nursing note categorization model for forecasting ICU LoS outcome.

We have used a feed-forward layer over the concatenation of the daily nursing note aggregations $[\bar{P}_t^1, ..., \bar{P}_t^K,]$ also denoted as $[L_t^1, ...L_t^{2K}]$. This allows each note aggregations to contribute to the final regression parameters through a separate set of trainable parameters.

$$\psi(L_t^1, ...L_t^{2K}, \theta_D) = \sigma(\bar{L}_t^T W_D + B_D) \quad (3)$$

As shown in Equation 3, $\psi(L_t^1, ...L_t^{2K}, \theta_D)$ produces the output class probability vector yi. $W_D \in \mathbb{R}^{(EK)\times C}$ and $B_D \in \mathbb{R}^C$ are the trainable parameters for the decision layer: $\theta_D = \{W_D, B_D\}$. We then compute the cross entropy loss $L(Y, Y')$ between ground truth labels and y_i.

4 Evaluation

Experiments: We investigate the performance of the proposed model in terms of the following criteria: a) *Efficacy of the joint model* with respect to the other base lines. b) *Prediction accuracy* of the network architectures, and c) The *timeliness* of the prediction. Accordingly, we propose baseline models that considers only the nursing notes as input (NotesOnly), Clinical Details ($H_p(t)$) only (CD), and taking both the inputs into account but without considering the joint categorisation (Notes+CD).

In terms of the *neural network architectures*, we have used the ClinicalBERT and Blue-BERT models (Devlin et al., 2018) fine-tuned on our dataset as baselines. We also present our experimental results on fine-tuned open-source LLMs such as LLAMA-3.1 (He et al., 2024) and BioMistral-7B (Labrak et al., 2024). First, we have evaluated the LoS prediction ability of LLAMA-3.1 using zero-shot (Labrak et al., 2023) and few-shot prompt techniques. Here, we have used the few-shot technique demonstrated by (Labrak et al., 2023) and given examples of series of notes for two patients as prompt. We have also fine-tuned the pre-trained BioMistral-7B Model with the MIMIC-III Dataset to compare its ability to perform LoS prediction. Details of the fine-tuning process is discussed in Appendix A.1.

Evaluation Metrics: Prediction accuracy of the models are computed in terms of evaluation matrices such as R^2 score for accuracy, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). We have also performed evaluation with Area Under the ROC Curve (AUC-ROC) and Cohen Kappa Scores. The *ROC curve* shows the trade-off between true positive rate (TPR) and false positive rate (FPR) and provides the ability of a classifier in distinguishing between classes. The closer an AUC-ROC curve is to the upper left corner, the more efficient in distinguishing the classes. *Cohen's Kappa* score measures the agreement between model predictions and actual class values and it is defined by, $\kappa = \frac{p_0 - p_e}{1 - p_e}$ where p_0 is the observed agreement of the model and p_e is the chance agreement.

Since the model aims to predict ICU LoS, it is important to evaluate how early it provides predictions. Early warnings enable hospital administrators to adjust strategies effectively. To measure this, we calculate the time between the model's initial warning and the end of the patient's ICU stay. We introduce a *time-coupled prediction score*, which modifies the existing evaluation parameters by combining the prediction accuracy with the elapsed time from the model's warning to the patient's ICU

NN Model	Data input	Accuracy R^2	MAE	RMSE	AUC-ROC	Kappa	MAE'	RMSE'
LLAMA-3.1	zero-shot	0.341	0.61	0.63	0.818	0.482	0.683	0.694
LLAMA-3.1	few-shot	0.441	0.55	0.61	0.818	0.492	0.676	0.644
BioMistral	zero-shot	0.319	0.65	0.67	0.818	0.451	0.673	0.691
BioMistral	few-shot	0.449	0.45	0.53	0.818	0.462	0.511	0.633
BioMistral	fine-tune	0.641	0.43	0.46	0.818	0.521	0.472	0.577
ClinicalBioBERT	NoteOnly	0.680	0.49	0.47	0.571	0.559	0.571	0.594
ClinicalBioBERT	CD	0.578	0.58	0.57	0.557	0.556	0.573	0.694
ClinicalBioBERT	Note+CD	0.690	0.45	0.43	0.664	0.642	0.471	0.569
ClinicalBioBERT	Note+CD Joint	0.761	0.41	0.43	0.818	0.682	0.488	0.54
BlueBERT	NoteOnly	0.717	0.23	0.39	0.871	0.594	0.29	0.44
BlueBERT	CD	0.692	0.28	0.4	0.873	0.573	0.371	0.494
BlueBERT	Note+CD	0.749	0.21	0.28	0.872	0.678	0.287	0.294
BlueBERT	Note+CD Joint	**0.826**	**0.19**	**0.26**	**0.833**	**0.693**	**0.271**	**0.284**

Table 1: Performance of baseline models in terms of R^2, MSE, RMSE, AUC-ROC, Kappa and modified MAE' and RMSE' scores.

discharge. Accordingly, we modify the MAE, and RMSE scores of the proposed model as follows:

1. $M\bar{A}E' = \frac{\tau}{N} * (\sum_{i=1}^{N} |y - y'| + \epsilon)$
2. $R\bar{M}SE' = \sqrt{\frac{\tau}{N} * (\sum_{j=1}^{N} (y_i - y_j)^2 + \epsilon)}$

Where, τ is the elapsed time from the model's warning to the patient's ICU discharge and ϵ is a constant set to 0.0001.

All the models have used sentence embeddings from either the pre-trained $BlueBERT$ or the pre-trained $ClinicalBERT$ model. The models are trained using the Adam optimizer. Mean validation performance was used to select hyper-parameter values. We trained the models with 10 epochs, and the learning rate of 5×10^{-4}.

4.1 Results

We computed the accuracy scores of the predicted LoS averaged over the 10 test sets. Table 1 summarizes our results. The *NoteOnly* model performs better than the *Clinical details(CD) only*, indicating the nursing notes are useful. The Note+CD baseline improves over the *NoteOnly* baseline indicating that the combination of notes and the CD information is more informative. The proposed model outperform all the above baselines by achieving a statistically significant improvement ($p < 0.05$) over them. This indicates the utility of our notes-category aware analysis of the clinical texts.

In terms of network architectures, We observe that BlueBERT performs better than the Clinical BioBERT model in this task, as expected. It is also observed that, compared to NoteOnly data input, adding clinical details with the joint model gives better accuracy, which assures that the latent categorization of the health condition does a better job for this classification and can effectively learn important health characteristics from the notes that are indicative of severity or lack of it. Incorporating the Joint model of the health condition has further increased classifier accuracy by providing more information to the network about the distinguishing phrases of the output scores. Further, the CD features contains more information about organ dysfunction, physiological decompensation from different physiological and disease-related variables. In addition to this, there are phrases like *"HR dropping"*,*"requiring mask ventilation for resp failure"*, *"couldn't breathe"* that are indicative of high risk patients who usually need longer ICU stays, whereas *"good effect from Ativan"*, *"comfortable breathing"*, *"hemodynamically stable"* are indicative of healing since these talk of signs of improvement of a patient's condition.

Detailed analysis of results show that including the joint modeling of Note+CD improves the performance of the prediction model by improving the predictions for certain categories patients namely those suffering from *Myocardial Infarction*, *Coronary Artery Disease*, *Sepsis*, *Congestive heart failure*. This also indicates that better CD measures, if available, can possibly improve the performance of other categories also. This is identified as one of our future endeavours.

Overall we have observed that our proposed approach outperforms the state of the art for all evaluation metrics. However, we would like to point out that since each reported state of the art chose different features and different points during the stay of a patient to predict the length of ICU stay, the set of patient data used for the tasks reported are not always identical. For example, some patient records did not have nursing notes. These were not used

for our experiments. Similarly, the work reported by (Su et al., 2021) used the data for Sepsis patients only, and not the entire dataset. Accordingly, Appendix B provides a summary of performance reported by other work discussed earlier.

Analyzing erroneous predictions revealed that many misclassifications were for patients who died within a day or two of ICU admission, despite the model predicting a longer stay. Although the features suggested a longer stay, the early deaths altered the outcomes. This highlights the importance of nursing notes in reflecting a patient's true condition, suggesting the need for separate accommodation in the prediction model, possibly by incorporating additional outputs. Another challenge faced by our model is due to multiple non-standard abbreviations, spelling mistakes etc. all of which were declared as unknown tokens by the language models. Some examples of such tokens are *" .....GI: Abd soft, hypoactive bs. OGT to LCS, clear drainage....."*. The language model thus needs to be enhanced to accommodate these.

4.2 Comparison with LLMs

We compare the performance of the proposed model with LLMs such as LLAMA-3.1 and BioMistral-7B with zero-shot, few shot and fine-tuned strategies. We observe the performance of both LLAMA-3.1 and BioMistral-7B using both zero-shot and few-shot approach was notably limited. This limitation stemmed from the complexity of defining clinical concepts, which necessitates a comprehensive representation beyond the provided examples as prompt. While LLAMA-3.1 achieved a high precision score, its recall and F1 scores were significantly lower, primarily due to its tendency to classify the majority of the clinical notes towards a longer ICU stay. We also observe LLMs limitations while processing sequence of notes with larger contexts.

We have also fine-tuned the BioMistral-7B model with the proposed dataset. Out of the test sentences, the trained BioMistral Model provided a distinct classification for only 25% cases, while out of the remaining 75% cases resulted in a rather confusing answer. Among those, a manual verification reveals that it categorized correctly for 22% cases. Therefore, we concluded that while training the large language model on a specific domain can improve its classification capacity, however, the inherent hallucination properties can still pose a challenge.

4.3 Analyzing the timeliness of prediction

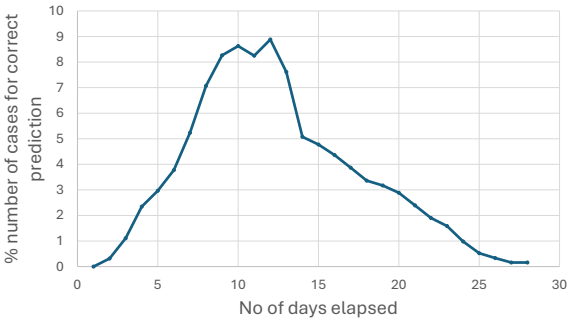

Figure 3: Distribution of number of days elapsed between the proposed model's warning till end of the patient is discharged/deceased from ICU.

A detailed comparison of the original MAE and RMSE scores with the modified *time-coupled scores* reveals that while most models exhibit low MAE and RMSE scores, indicating strong performance, the *time-coupled score* shows that many models predict the ICU LoS too late, diminishing the utility of early predictions. Models relying solely on NotesOnly or clinical details (CD) are particularly disadvantaged in making early predictions. In contrast, the joint model demonstrates greater stability in predicting LoS earlier. Empirical analysis indicates that baseline models typically require around 50% of the total elapsed time to make a prediction, whereas the joint latent categorization model achieves comparable predictions within the first 25-30% of the elapsed time, thereby preserving the benefits of early warning. Figure 3 depicts the distribution of these counts across the test set.

5 Conclusion

In this paper, we develop a neural network architecture that uses the nursing notes, prepared at the time of admission to ICU, to predict ICU LoS. The novelty of the model lies in the fact that it processes the the notes during the development of the patient's ICU stay. We proposed a joint model of latent categorization of patient's health status for the task. We have demonstrated that the proposed approach allows the model to identify high-level health status that influence the prediction. Results showed that the proposed joint model outperforms the baseline systems that uses individual clinical notes or health status representations.

References

Khalid Alghatani, Nariman Ammar, Abdelmounaam Rezgui, Arash Shaban-Nejad, et al. 2021. Predicting intensive care unit length of stay and mortality using patient vital signs: Machine learning model development and validation. *JMIR Medical Informatics*, 9(5):e21347.

Alan R Aronson. 2006. Metamap: Mapping text to the umls metathesaurus. *Bethesda, MD: NLM, NIH, DHHS*, 1:26.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Thanos Gentimis, Alnaser Ala'J, Alex Durante, Kyle Cook, and Robert Steele. 2017. Predicting hospital length of stay using neural networks on mimic iii data. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pages 1194–1201. IEEE.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. 2024. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*.

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

Alistair EW Johnson, Andrew A Kramer, and Gari D Clifford. 2013. A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. *Critical care medicine*, 41(7):1711–1718.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Yanis Labrak, Mickael Rouvier, and Richard Dufour. 2023. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. *arXiv preprint arXiv:2307.12114*.

Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. 1993. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963.

Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. 2024. Frozen language model helps ecg zero-shot learning. In *Medical Imaging with Deep Learning*, pages 402–415. PMLR.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Alex Rinaldi, Jean E Fox Tree, and Snigdha Chaturvedi. 2020. Predicting depression in screening interviews from latent categorization of interview prompts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7–18.

Emma Rocheteau, Pietro Liò, and Stephanie Hyland. 2020. Predicting length of stay in the intensive care unit with temporal pointwise convolutional networks. *arXiv preprint arXiv:2006.16109*.

Peri L Schuyler, William T Hole, Mark S Tuttle, and David D Sherertz. 1993. The umls metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217.

Longxiang Su, Zheng Xu, Fengxiang Chang, Yingying Ma, Shengjun Liu, Huizhen Jiang, Hao Wang, Dongkai Li, Huan Chen, Xiang Zhou, et al. 2021. Early prediction of mortality, severity, and length of stay in the intensive care unit of sepsis patients based on sepsis 3.0 by machine learning models. *Frontiers in Medicine*, 8:883.

Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A Gers, and Alexander Löser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. *arXiv preprint arXiv:2102.04110*.

J-L Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PeterM Suter, and Lambertius G Thijs. 1996. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure.

Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Sharif Amit Kamran, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. 2023. Gpt-4: a new era of artificial intelligence in medicine. *Irish Journal of Medical Science (1971-)*, 192(6):3197–3200.

Yasi Wang, Hongxun Yao, and Sicheng Zhao. 2016. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242.

David T Wong and William A Knaus. 1991. Predicting outcome in critical care: the current status of the apache prognostic scoring system. *Canadian journal of anaesthesia*, 38(3):374–383.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.

A Pre-processing the data: Extraction of clinical details

Clinical details in Nursing notes vary greatly in style and content. Some document only symptoms, while others detail absences of symptoms, adverse reactions, psychological states, and appetite changes, often using non-standard terminology and abbreviations. To manage this variability, we added a processing layer that uses biomedical dictionaries to create a structured representation of clinical details, as shown in Figure 4. Details of this processing pipeline are presented below.

A.1 Entity Extraction

We employed two BioNER tools, ScispaCy (Neumann et al., 2019) and Metamap (Aronson, 2006), for the extraction of patients' health conditions from clinical notes. The pre-trained ScispaCy model, was utilized for recognizing "disease" names. We use Metamap to identify eight medical entities, including "Sign or Symptom", "Disease or Syndrome", "Acquired Abnormality", "Anatomical Abnormality", "Congenital Abnormality", "Injury or Poisoning", "Mental Process", and "Mental or Behavioral Dysfunction" within these notes.

A.2 Detecting Negations

Subsequently, the Negex algorithm (Chapman et al., 2001), designed to identify negative modifiers such as "no", "not", etc., is employed to detect negative mentions of entities within the text. The initial list was expanded to encompass commonly occurring negation concepts like 'deny", "refuse", "absent", "decline", etc., frequently encountered in clinical notes. For instance, in a sentence like "The patient has shortness of breath but denies any chest pain", the two symptoms identified would be "shortness of breath" and "neg chest pain." These negative symptoms play a crucial role in providing a comprehensive understanding of individual patients.

A.3 Clinical Entity Normalization

Clinical notes use varied terminology, abbreviations, formats, and coding systems. For example, "Hemorrhage" might be called "Bleeding," "Blood Loss," or "oozing of blood" by different professionals. To standardize these terms, we used the UMLS Metathesaurus(Schuyler et al., 1993), which assigns a Concept Unique Identifier (CUI) to each term. When exact UMLS matches were unavailable, we applied an approximate string-matching algorithm based on Levenshtein distance (Yujian and Bo, 2007) to find the closest CUI. For unmatched entities, we created unique identifiers to ensure no conditions were missed, referring to these as CUIs.

Thus, each clinical note is represented by the presence or absence of CUIs. We use a comprehensive vocabulary of CUIs, denoted as V, to describe relevant diseases and symptoms, allowing us to express a patient's condition at any time using these CUIs.

A.4 Handling Missing Data

Our EHR analysis revealed two main issues: missing medical records for certain hospital days and incomplete clinical notes. For example, information about a disease might be recorded on Day_{n-1} and Day_{n+1} but not on Day_n, creating uncertainty about the disease's presence. To address these problems and maintain a continuous understanding of the patient's condition, we have established the following rules:

1. If a disease or symptom d is present in Day_{n-1} and Day_{n+1}, we consider it to be present in Day_n as well.

2. If a disease or symptom d is noted as negative in Day_{n-1} and Day_{n+1}, we assume it is also negative in Day_n.

3. If a disease or symptom d is present in Day_{n-1} and negative in Day_{n+1}, we assume it is positive in Day_n.

4. If a disease or symptom d is noted as negative in Day_{n-1} and never occurred in the future, we consider it to be negative in all future days.

By applying these rules, we aim to alleviate the impact of missing or incomplete data, providing a more comprehensive understanding of the patient's medical history and progression.

	Dataset	Feature used	Method	Best Result
Alghatani et al., 2021	44,000 ICU stays from MIMIC	patient's vital signs like, heart rate, BP, temp., resp. etc	Random Forest	65% accuracy
Su et al., 2021	2224 Sepsis patients PICMISD	Age, P(v-a)CO2 /C(a-v)O, SO, wbc etc.	XG-Boost model	F1: 0.69, AUC-ROC:0.76
Rocheteau, Liò, et al., 2020	eICU critical care dataset	medical features, Gender, Age, Ethnicity, etc.	Temporal convolution	Kappa score = 0.58
Harutyunyan et al., 2019	42276 ICU stays of 33798 unique patients from mimic database	17 clinical variables like, Capillary refill rate, Diastolic blood pressure etc. from first 24 hours of admission.	LSTM	AUC-ROC : 0.84
van Aken et al., 2021	38013 admission notes from MIMIC III	Created admission notes from discharge summaries	Pretrained CORe + BioBERT	AUC-ROC : 0.72%

Table 2: Performance of different SOTA prediction models as reviewed in the present paper. Note that different works have used different set of data, and evaluation parameters. As a result of this, the results could not be compared with that of the present task.

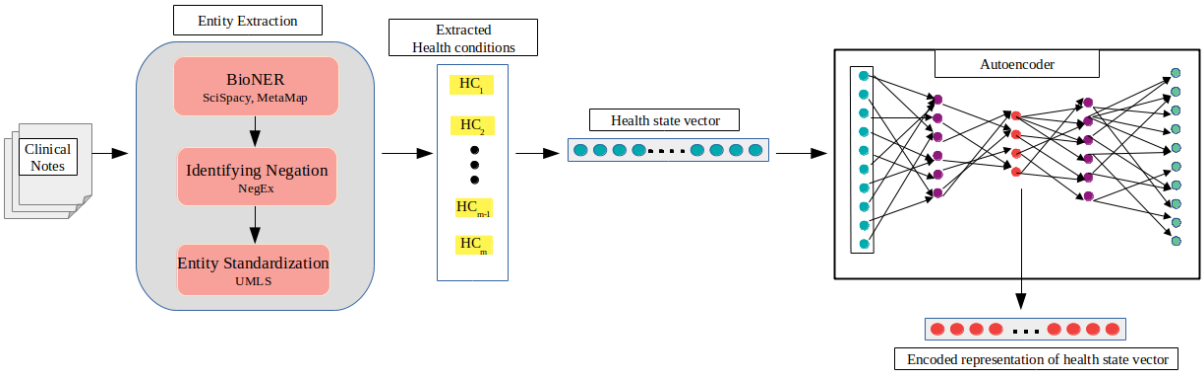

Figure 4: Overview of the process for extraction and representation of patient health conditions from clinical notes.

A.5 Encoding the clinical details

Once entities are extracted and represented with CUIs, each day's clinical details for a patient are consolidated using the CUIs observed on that day.

Given a patient p, the clinical detils at day t is defined by a vector $H_p(t) =< d_i >, i = 1, 2, ..., |V|$, where $d_i \in V$ and

$$d_i = \begin{cases} 1 & \text{if } d_i \text{ present in day } t \text{ for } p \\ -1 & \text{if } d_i \text{ negative in day } t \text{ for } p \\ 0 & \text{if } d_i \text{ not mentioned in day } t \text{ for } p \end{cases}$$

Due to the high dimensionality and sparsity of vectors from numerous diseases and symptoms, we use an autoencoder-based transformation (Wang et al., 2016) to achieve a dense, lower-dimensional representation. The autoencoder's encoder compresses the data, capturing essential features, while the decoder reconstructs the original data from this compressed form, preserving key information. These compressed representations are then used for further processing of patient clinical details.

B Performance of different SOTA Length of Stay (LoS) prediction models as reviewed in the present paper

Table 2 reports the performance of different SOTA prediction models as reviewed in the present paper. Note that different works have used different set of data, and evaluation parameters. As a result of this, the results could not be compared with that of the present task.