Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 20 No. 15 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v20i15.51839

PAPER

Latent Dirichlet State Predictive Clustering Model for Disease Risk Prediction in Electronic Health Records

Prasanthi Yavanamandha(✉), D. S. Rao

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana, India

prasanthi.yavanamandha@klh.edu.in

ABSTRACT

Electronic health records (EHRs) are a valuable source of data that helps to understand patients' health conditions and generate healthcare decisions. However, modeling the longitudinal and temporal dependencies of EHRs is challenging in disease risk prediction (DRP). To overcome this problem, this study proposed a latent Dirichlet state predictive clustering (LDSPC) using medical notes for DRP in healthcare. This process includes three modules, such as posterior, prior, and likelihood. The posterior module utilized an attentive encoder for extracting data from unstructured medical notes. Additionally, the clustering approach is integrated into the similarity module to learn the patient's useful representation of the latent Dirichlet state. These states are clustered into numerous cluster centers, and a weighted average is applied for risk prediction. Moreover, the MIMIC-III and N2C2-2014 datasets contain unstructured medical notes that are preprocessed by non-English characters and stop word removal processes. The LDSPC achieves better accuracy of 0.9864 and 0.9694 for MIMIC-III and N2C2-2014 datasets, correspondingly which is better when compared to knowledge-enhanced multimodal learning for disease diagnosis generation (EHR-KnowGen).

KEYWORDS

disease risk prediction (DRP), electronic health records (EHRs), latent Dirichlet state predictive clustering (LDSPC), posterior module, unstructured medical notes

1 INTRODUCTION

The opportunity to increase healthcare quality through interoperability and healthcare information technology (IT) has gained significant attention from both the private and public sectors [1]. The e-health technologies include electronic health records (EHRs), electronic medical records (EMRs), health information exchanges, and personal health records (PHRs) [2]. As healthcare systems progressively adopt health IT, a numerous clinical data volume is gathered. These data are kept in EHRs or PERs with numerous formats, such as simple database tables or any other format [3]. These records number in millions, which is highly complex,

Yavanamandha, P., Rao, D.S. (2024). Latent Dirichlet State Predictive Clustering Model for Disease Risk Prediction in Electronic Health Records. *International Journal of Online and Biomedical Engineering (iJOE)*, 20(15), pp. 79–92. https://doi.org/10.3991/ijoe.v20i15.51839

Article submitted 2024-08-21. Revision uploaded 2024-10-04. Final acceptance 2024-10-04.

and health information systems keep data in different nature [4]. The EHR applications to efficient medicine include tasks such as disease risk prediction (DRP), mortality, survival prediction, disease diagnosis, stay duration of intensive care unit ICUs, and statistical phenotype prediction [5, 6]. Through, the effectiveness of computational techniques is limited by the capability to manage huge-dimensional, isolated, heterogeneous, irregular, and longitudinal EHRs [7]. ICUs are information-rich hospitals, and the significance of rapid responses to patient degradation is high [8].

Patient's health latent states are inferred through various types of data, which include laboratory testing results, unstructured medical notes, and signal monitoring [9]. It is critical to obtain customized healthcare to recognize every patient's health state from high-capacity data that needs numerous labor resources and domain knowledge [10]. Artificial intelligence (AI)-based techniques significantly help medical decision progress through modeling patients EHR data. The present approaches for DRP mainly focus on deep learning (DL) algorithms [11]. Convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM) are generally exploited for medical textual notes and isolated structured EHR data for the DRP process [12]. Moreover, the initiation of large language models (LLMs) and their application in healthcare domains such as BioBERT, MedBERT, and ClinicalBERT has been used to procedure textual EHR data in different healthcare tasks [13, 14]. Therefore, the integration of LLMs with other DLs has established the best prediction performance in healthcare utilizing the strength of LLM in natural language processing (NLP) [15]. Modeling the longitudinal and temporal dependencies in EHR is difficult for DRP due to patient data being irregular or high-dimensional, which makes it complex to capture meaningful patterns. To address this study, this study proposes a latent Dirichlet state predictive clustering (LDSPC) that uses unstructured medical notes to enhance DRP. By clustering patient states over time, the LDSPC captures dynamic health trends, which leading to accurate and reliable predictions for healthcare applications. The primary contributions are stated as follows:

- The LDSPC is proposed in this study using medical notes. It is considered for medical notes unstructured text data type and preserves probabilistic model characteristics to utilize neural network representation power.
- The cluster probability is employed as weights to acquire weighted embeddings of the cluster center that is utilized for DRP. The characteristic of associated clusters is interpreted to understand every latent Dirichlet state that is updated by latent Dirichlet state predictive clustering.

The literature review discusses the previous research related to DRP and identifies the research gaps. The proposed methodology discusses the detailed description of unstructured medical notes for DRP in EHR. Moreover, the LDSPC provides better accuracy by considering unstructured text data types for medical notes. The conclusion summarizes key findings and significance with potential directions for future research.

2 LITERATURE REVIEW

The existing research for DRP in the healthcare sector is analyzed below with its advantages and limitations.

Shuai Niu et al. [16] presented enhanced healthcare decision support by explainable AI for DRP. The non-parametric predictive clustering-based DRP was integrated into Dirichlet process-based predictive clustering (DirPred) through neural networks. The attention mechanism was integrated into DPMM to improve the model's interpretability, which enables the capture of local-level to cluster-level evidence through predictive clustering. The DirPred effectively captured longitudinal EHR data for disease prediction. However, DirPred requires prior specification of clusters because it relies on prior assumptions about structural data, which leads to suboptimal clustering results. Ardeshir Mansouri et al. [17] suggested a hybrid ML approach for ICU patient mortality prediction. Initially, the dataset was created and given to convolutional neural network (CNN) and its output was recorded into the probability of mortality. The temporal features are filtered through a filtering approach, and dual values are produced for each feature. The XGBoost classifier performed final classification, which helps the medical community to make accurate and immediate decisions. The XGBoost does not capture longitudinal, dynamic, and interactive patterns, which was crucial for accurate disease risk protection.

Shuai Niu et al. [18] implemented a deep state-space model with the predictive clustering for the risk prediction (DSPCR). Initially, the prior module learns the patient's latent state transition for producing the present latent state according to the past one. The posterior estimates latent state posterior distribution. The likelihood produces prediction by predictive clustering algorithm exploitation for DRP. The DSPCR required high training to accurately capture temporal patterns and dependencies, which leads to high computational time. Sicen Liu et al. [19] developed a multi-channel fusion long short-term memory (MCF-LSTM) for DRP in EHRs. The MCF-LSTM models the correlation among various medical events through numerous channels. The task-wise fusion model was developed in that a gated network was applied to designate data and transfer among events. Moreover, irregular temporal space among nearest clinical visits is demonstrated in separate channels that were integrated by another risk. The MCF-LSTM struggled with highly variable and missing data among channels because it relies on a structural sequence of input for accurate risk prediction, which failed to capture significant patterns and dependencies.

Shuai Niu et al. [20] introduced an EHR-KnowGen, which incorporates various modalities into a combined feature space through soft prompts and influences LLMs for producing disease diagnosis. Through integrating external domain knowledge from various levels, multimodal information was extracted and combined, which outputs accurate diagnosis generation. The EHR-KnowGen has different modalities that have domain discrepancies thereby reducing overall performance. Rawan AlSaad et al. [21] suggested temporal-self attention for DRP in HER using non-linear stationary kernel approximation. The developed self-attention with non-stationary kernel approximation was applied to capture both temporal relations and contextual information among patient visits in EHR. This model does not unlock the potential to manage variable time gap data and its influence on predicted results. It was general to explore more effective alternatives to encoding that were capable of embedding non-stationary temporal patterns into higher-dimensional spaces.

From the overall analysis, the existing techniques required prior specification of clusters because they rely on prior assumptions about structural data. Those models are unable to capture dynamic and longitudinal patterns, which was crucial for accurate DRP. Required huge training to capture temporal patterns and dependencies. Struggled with high missing data among different channels due to the structural sequence of input which fails to capture significant dependencies and patterns.

To overcome this drawback, this study proposed an LDSPC for DRP. Here, the cluster probability is applied as weights to obtain a weighted embedding presentation. The associated cluster characteristic is construed to understand each latent Dirichlet state.

3 PROPOSED METHODOLOGY

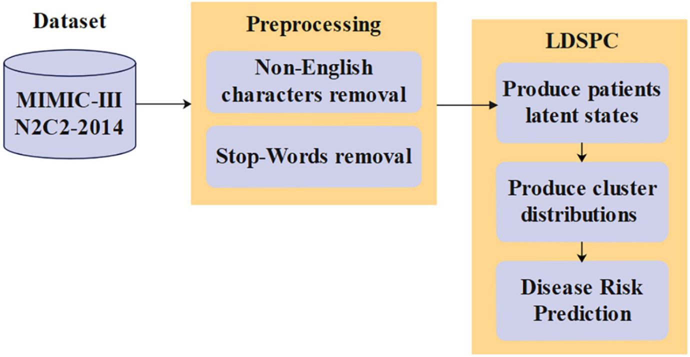
Dataset
MIMIC-III
N2C2-2014
Preprocessing
Non-English characters removal
Stop-Words removal
LDSPC
Produce patients latent states
Produce cluster distributions
Disease Risk Prediction

Fig. 1. Procedure of proposed methodology

In this study, LDSPC is proposed for DRP in EHRs. The MIMIC-III and N2C2-2014 dataset is used, which contains unstructured medical notes. The datasets are preprocessed by non-English characters and stop word removal. Then, encoding the data to acquire a latent Dirichlet state generates latent Dirichlet state cluster distribution, and DRP stages are involved. This process enables effective and efficient DRP using unstructured medical notes from healthcare. Figure 1 represents the procedure of the proposed methodology.

3.1 Dataset

The medical information mart for intensive care-III (MIMIC-III) [22] and N2C2-2014 datasets are used here, which are publicly accessible datasets. The detailed descriptions are explained in the following subsection.

MIMIC-III: This dataset contains de-identified healthy data for patients hospitalized at the ICU at Beth Israel Deaconess Medical Center in Boston, Massachusetts. A similar data-splitting strategy was employed to attain training and test datasets using a 4:1 ratio.

N2C2-2014: This dataset contains de-identified EHRs and related annotations from various hospitals, which comprise 1,304 medical notes from 296 individuals. A similar data-splitting strategy was employed to attain training and test datasets using a 4:1 ratio. Table 1 summarizes the dataset.

Table 1. Summary of dataset

Dataset	MIMIC-III	N2C2-2014
Patients	38,597	296
Avg hospital visits	2.61	4.42
EHRs	53,423	1,304
Longitudinal	9,759	1,304

3.2 Preprocessing

The medical notes preprocessing includes non-English characters and stop word removal. These preprocessing steps are explained below:

Non-English characters removal: The non-alphabetic characters such as punctuation, special symbols, and numbers are removed by using regular expressions. This process is essential for DRP to ensure data consistency and uniformity, which helps to achieve accurate prediction.

Stop-words removal: The stop-words have less semantic meaning, and they were removed to focus on important words in text such as “is”, “and”, and “the”. This process enables us to focus on informative and meaningful terms related to disease risk which enhances accuracy and efficiency by eliminating noise. Table 2 summarizes the preprocessing steps.

Table 2. Summary of preprocessing steps

Preprocessing Steps	Description	Example Before the Step	Example After the Step	Purpose
Non-English characters removal	Uses a regular expression to remove punctuation, special symbols, and numbers	“Patient Age: 50.”	“PatientAge”	Ensures better prediction results
Stop-Words removal	Removes the stop words such as “was”, “the”, “and”	“The patient’s condition is critical”	“Patient critical”	Reduces the noise

3.3 Latent Dirichlet state predictive clustering

Assume every patient n is considered as an EHR sequence gathered from numerous hospital visits where every visit data sample t is signified as z_t^n. The N_t^n is several words of the patient n at visit t in medical notes. T_n is the total number of visitors. Equation (1) denotes the occurrence of various disease risks observed at numerous visits where every vector y_t^n has 0 and 1 values. In the DRP task of patient n, x^n is applied to produce the prediction value of y^n. The LDSPC implements the sequential Bayesian updating approach by prior from alteration state and likelihood model defined through the newest examination to update the present latent Dirichlet state via calculating Bayes rule-based posterior distribution. In this study, this approach is implemented to gather patient’s latent spate x_t^n at every hospital visit t.

$$y^n = \left\{ y_1^n, \ldots, y_t^n, \ldots, x_{T_n}^n \right\} \tag{1}$$

Figure 2 shows the outline of the proposed approach. The prior module produces x_t^n of prior distribution from past latent Dirichlet states. The posterior module estimates x_t^n of posterior distribution through encoding data involved in z_t^n. The likelihood module implements a clustering algorithm to produce y_t^n observations. The μ_t^n and σ_t^n are the mean and standard deviation of latent Dirichlet states, here (p) and (q) are prior and posterior subscripts. $\hat{x}_t^n$ is a latent Dirichlet state sampled vector, $c_{1:K}$ includes cluster center embeddings K, o_t^n and its s_t^n standardized form denotes similarity among $\hat{x}_t^n$ and c_k for every $k \in \{1, \ldots, K\}$. The u_t^n is a $c_{1:K}$ the weighted average, here weight is specified through s_t^n and predicted risk vector is $\hat{y}_t^n$.

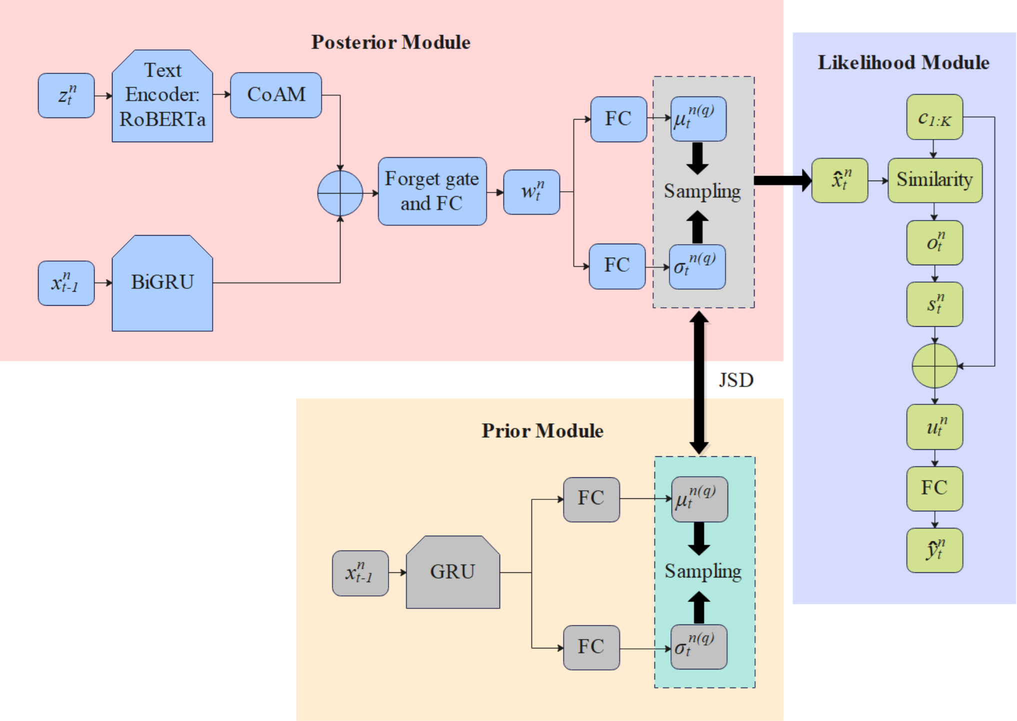

Fig. 2. Outline of proposed approach

Posterior module. In the posterior module, the variational approximation is $q_{\varnothing}\left(x_t^n \middle| x_{t-1}^n, z_t^n\right)$ where, $x_{t-1}^n = \left[x_1^n, \ldots, x_{t-1}^n\right]$ which contains the latent Dirichlet state of every previous visitor. Particularly, the posterior is recursive through the attentive encoder network and FC network by x_{t-1}^n and Z_t. For the embedding process, robustly optimized bidirectional encoder representations from the transformer approach (RoBERTa) and co-attention mechanism (CoAM) are applied to embed medical notes x_t^n into latent representation. The RoBERTa is applied for the encoding process which can enhance clinical text interpretation and patient records, which leads to accurate risk prediction. The CoAM enables concentration on relevant parts of input text by attending to clinical features and disease-related data simultaneously, which enhances the capability to extract and integrate significant features from various sources. The embedding data is signified as $E_t^n \in R^{D \times N_t^n}$, where D is a size of embedding. For the integrating step, the CoAM is adopted, which assists in capturing data involved in sequential words. Initially, the scaled-dot similarity matrix $G_t^n \in R^{N_t^n \times N_t^n}$ is applied to the present similarity among every token from E_t^n as equation (2),

$$G_t^n = \frac{\left(FC_1\left(E_t^n\right)\right)^T FC_2\left(E_t^n\right)}{\sqrt{D}} \tag{2}$$

Where, FC_1 and FC_2 are Fully Connected (FC) layers, $(\cdot)^T$ is transposition function. Max-pooling with *SoftMax* activation is applied to produce a focused medical note embedding vector $e_t^n \in R^D$ as equation (3–4),

$$g_t^n = SoftMax\left(MaxPool\left(G_t^n\right)\right) \tag{3}$$

$$e_t^n = \sum_{i=1}^{N_t^n} g_{t,i}^n FC_3\left(E_{t,i}^n\right) \tag{4}$$

Here, FC_3 is FC layer, $g_t^n \in R^{N_t^n}$ is a score vector of CoAM, $E_{t,i}^n$ is a column i of E_t^n, $g_{t,i}^n$ is an element i of g_t^n. The e_t^n contains weighted data from z_t^n, the next stage is to integrate e_t^n with x_{t-1}^n for producing w_t^n as equation (5),

$$w_t^n = FC_4\left(g\left(e_t^n \oplus BiGRU\left(x_{t-1}^n\right)\right)\right) \tag{5}$$

Here, FC_4 is FC layer, g is a forget gate implemented from LSTM, ⊕ is a concatenation operator. The w_t^n is a weighted representation fed into the FC network of FC_5 and FC_6 for producing mean and standard deviations such as $\mu_t^{n(q)}$ and $\sigma_t^{n(q)}$. The sample state vector is given in equation (6),

$$\hat{x}_t^n = \mu_t^{n(q)} + \varepsilon \cdot \sigma_t^{n(q)} \tag{6}$$

where ε is a random noise.

Prior module. In the Bayesian sequential inference framework, the state transition model is applied to produce a prior circulation of the present latent Dirichlet state from the past state. Therefore, prior circulation of the patient n at time t is in equation (7–9),

$$p_\vartheta = \left(x_t^n | x_{t-1}^n\right) \sim \mathcal{N}\left(\mu_t^{n(p)}, \sigma_t^{n(p)}\right) \tag{7}$$

$$\mu_t^{n(p)} = FC_5\left(GRU\left(x_{t-1}^n\right)\right) \tag{8}$$

$$\sigma_t^{n(p)} = FC_6\left(GRU\left(x_{t-1}^n\right)\right) \tag{9}$$

Here, $\mu_t^{n(p)}$ and $\sigma_t^{n(p)}$ are mean and standard deviation of prior modules which are recursive through GRU and dual FC layers such as FC_5 and FC_6. The FC_5 is applied to produce the mean vector, and FC_6 is applied to produce the standard deviation.

Likelihood module. In the likelihood module, predictive clustering is integrated into the LDSPC model. Every latent Dirichlet is clustered into K groups in which embedding centers are as $c_{1:K} = [c_1, \ldots, c_k, \ldots, c_K]$. Every sampled latent Dirichlet state $\hat{x}_t^n$ is approached as the weighted average of $c_{1:K}$. Here, s_t^n is established through similarity among latent Dirichlet states to every cluster embedding. The center embeddings weighted average u_t^n is applied for DRP. The initial step is for the cluster's Latent Dirichlet state detection and derive embeddings of cluster centers. The possibilities of allocating $\hat{x}_t^n$ to kth cluster is estimated through calculating similarity among $\hat{x}_t^n$ and c_k based on distribution t as equation (10),

$$o_t^n = \frac{\left(1 + \left\|\hat{x}_t^n - c_k\right\|_2^2 / \alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{k'=1}^{K}\left(1 + \left\|\hat{x}_t^n - c_{k'}\right\|_2^2 / \alpha\right)^{-\frac{\alpha+1}{2}}} \tag{10}$$

Where α is a degree of freedom in distribution t, $\hat{x}_t^n$ is the latent representation of x_t^n produced through the posterior module. Then, the $SoftMax$ layer is applied to normalize $o_t^n = \left[o_t^{n1};\ldots;o_t^{nK}\right]$ as equation (10), The cluster center embedding weighted average as equation (11–12),

$$s_t^n = SoftMax\left(o_t^n\right) \tag{11}$$

$$u_t^n = (c_{1:K})^T s_t^n \tag{12}$$

Here, $c_{1:K} \in R^{K \times D}$, K is a truncation parameter, $u_t^n \in R^D$ and $(\cdot)^T$ is a transposition function, then u_t^n is fed into FC layer FC_7 to get risk prediction results as equation (13),

$$\hat{y}_t^n = FC_7\left(u_t^n\right) \tag{13}$$

The log-likelihood of perceiving every element of y_t^n by specified latent Dirichlet state z_t^n are calculated as equation (14),

$$logp_\theta\left(y_{t,j}^n | z_t^n\right) = y_{t,j}^n \log\left(\hat{y}_{t,j}^n\right) + \left(1 - y_{t,j}^n\right)\log\left(1 - \hat{y}_{t,j}^n\right) \tag{14}$$

The evidence lower bound (ELB) through the adoption of Bayesian variational inference as equation (15),

$$\begin{aligned} L_{ELB} = \frac{1}{N}\sum_{n=1}^{N}\Bigg(&-E_{q\varnothing\left(x_1^n | z_1^n\right)}\left[log\, p_\theta\left(y_1^n \mid x_1^n\right)\right] + JSD\left(q_\varnothing\left(x_1^n | z_1^n\right) || p_\theta\left(x_1^n\right)\right) \\ &-\sum_{t=2}^{T^n}\left[E_{q\varnothing\left(x_t^n | x_{t-1}^n, z_t^n\right)}\left[log\, p_\theta\left(y_t^n | x_t^n\right)\right] + \sum_{t=2}^{T^n} JSD\left[q\varnothing\left(x_t^n | x_{t-1}^n, z_t^n\right) || p_\theta\left(x_t^n | x_{t-1}^n\right)\right]\right]\Bigg) \end{aligned} \tag{15}$$

Where, $JSD(\cdot)$ calculates the Jensen-Shannon divergence among dual distribution. The $p_\theta\left(y_t^n | x_t^n\right)$ is the likelihood of perceiving y_t^n provided latent Dirichlet state x_t^n. When, $t > 1$, $q_\varnothing\left(x_t^n | x_{t-1}^n, z_t^n\right)$ and $p_\theta\left(x_t^n | x_{t-1}^n\right)$ are posterior and prior of x_t^n. The $p_\theta(z_1)$ and $q_\varnothing\left(z_1^n | x_1^n\right)$ are prior and posterior for x_1^n. The $\varnothing$ and θ are neural network parameters for distribution. Figure 3 explains the flowchart of the proposed approach.

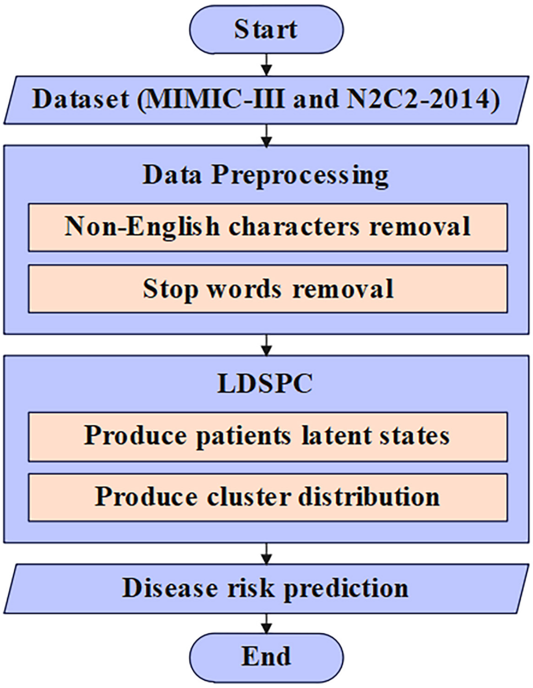
Start
Dataset (MIMIC-III and N2C2-2014)
Data Preprocessing
Non-English characters removal
Stop words removal
LDSPC
Produce patients latent states
Produce cluster distribution
Disease risk prediction
End

Fig. 3. Flowchart of proposed approach

4 EXPERIMENTAL RESULTS

The LDSPC performance is simulated in the environment of Python with system requirements of i5 processor, 16GB RAM, and Windows 10 OS. The results are crucial for healthcare and the broader community as described by the capability of LDSPC to enhance the DRP from EHR. Particularly in addressing challenges in handling unstructured medical nodes and longitudinal data modeling. This enhances the accurate and reliable predictive healthcare system in personalized involvements that leads to enhanced patient outcomes and reduced healthcare costs in medical settings. The performance is estimated with accuracy, micro precision, micro recall, micro f1-score, macro precision, macro recall, and macro f1-score which are given in equation (16–22).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

$$Micro\ Precision = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i} \tag{17}$$

$$Micro\ Recall = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i} \tag{18}$$

$$Micro\ F1 - score = \frac{2 \times Micro\ Precision \times Micro\ Recall}{Micro\ Precision + Micro\ Recall} \tag{19}$$

$$Macro\,Precision = \frac{\dfrac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i}}{L} \tag{20}$$

$$Macro\ Recall = \frac{\dfrac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i}}{L} \tag{21}$$

$$Macro\ F1 - score = \frac{2 \times Macro\ Precision \times Macro\ Recall}{Macro\ Precision + Macro\ Recall} \tag{22}$$

Where, *True Positive* (*TP*), *True Negative* (*TN*), *False Positive* (*FP*) and *False Negative* (*FN*) are *TP*, *TN*, *FP* and *FN*, *i* and *L* are class index and number of classes.

4.1 Quantitative and qualitative analysis

The LDSPC performance is calculated through accuracy, micro precision, micro recall, micro f1-score, macro precision, macro recall, macro f1-score, and computation time. Tables 3 and 4 denote the various clustering results for MIMIC-III and N2C2-2014 datasets.

Table 3. Performance on MIMIC-III dataset

Method	Micro			Macro			Accuracy
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
DSPC	0.8866	0.8952	0.8908	0.8846	0.8919	0.8882	0.9273
DSSPC	0.9035	0.9162	0.9098	0.9012	0.9128	0.9069	0.9467
HDSPC	0.9251	0.9316	0.9283	0.9164	0.9186	0.9174	0.9571
LSPC	0.9364	0.9437	0.9405	0.9253	0.9349	0.9302	0.9638
Proposed LDSPC	**0.9538**	**0.9829**	**0.9681**	**0.9462**	**0.9816**	**0.9635**	**0.9864**

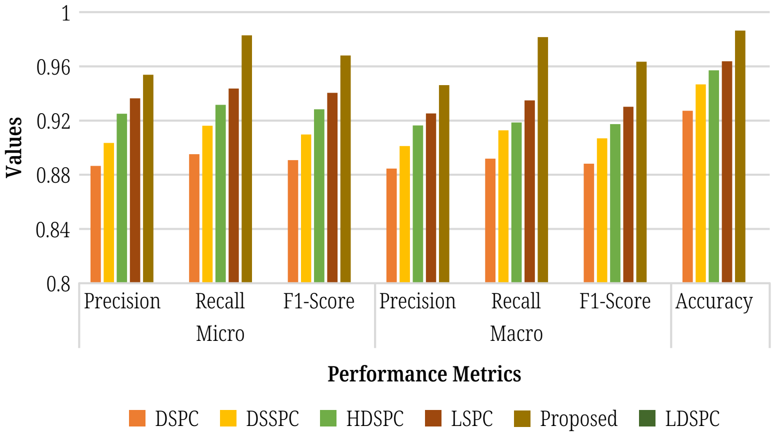

Fig. 4. Performance of LDSPC on MIMIC-III dataset

In Table 3 and Figure 4, for the MIMIC-III dataset, LDSPC performance is calculated and compared with other techniques such as density-based spatial predictive clustering (DSPC), deep state space predictive clustering (DSSPC), hierarchical Dirichlet state predictive clustering (HDSPC), and latent state predictive clustering (LSPC). The LDSPC performance is estimated through seven performance metrics. The LDSPC achieves 0.9864 accuracy, 0.9538 micro precision, 0.9829 micro recall, 0.9681 micro f1-score, 0.9462 micro precision, 0.9816 macro recall, and 0.9635 macro f1-score for the MIMIC-III dataset, which is better than existing techniques.

Table 4. Performance on N2C2-2014 dataset

Method	Micro			Macro			Accuracy
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
DSPC	0.8952	0.8912	0.8931	0.8826	0.8816	0.8821	0.9047
DSSPC	0.9138	0.9167	0.9152	0.9064	0.9018	0.9040	0.9238
HDSPC	0.9364	0.9352	0.9358	0.9251	0.9239	0.9245	0.9486
LSPC	0.9522	0.9646	0.9583	0.9487	0.9491	0.9489	0.9527
Proposed LDSPC	**0.9667**	**0.9683**	**0.9674**	**0.9536**	**0.9542**	**0.9539**	**0.9694**

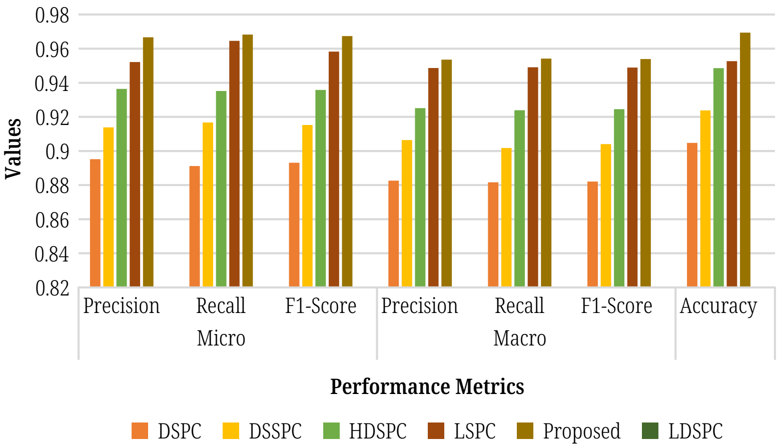

Fig. 5. Performance of LDSPC on N2C2-2014 dataset

In Table 4 and Figure 5, for the N2C2-2014 dataset, LDSPC performance is calculated and compared with other techniques such as DSPC, DSSPC, HDSPC, and LSPC. The LDSPC performance is estimated through seven performance metrics. The proposed LDSPC achieves 0.9694 accuracy, 0.9667 micro precision, 0.9683 micro recall, 0.9674 micro f1-score, 0.9536, micro precision 0.9542 macro recall, and 0.9539 macro f1-score for the N2C2-2014 dataset, which is better than existing techniques.

In Figure 6, LDSPC performance is calculated and compared with other techniques such as DSPC, DSSPC, HDSPC, and LSPC. The LDSPC performance is estimated through computation time. The LDSPC achieves less computation time of 1.76s and 2.13s for MIMIC-III and N2C2-2014 datasets, which is better than existing techniques.

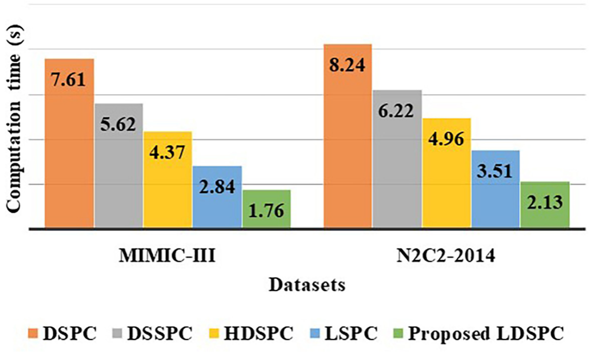

Fig. 6. Computation time (s) of the proposed approach

4.2 Comparative analysis

The LDSPC performance is estimated and compared with existing techniques such as DirPred [16], DSPCR [18], and EHR-KnowGen [20]. All seven performance metrics are considered for comparison. The LDSPC achieves 0.9864 accuracy, 0.9538 micro precision, 0.9829 micro recall, 0.9681 micro f1-score, 0.9462 micro precision 0.9816 macro recall, and 0.9635 macro f1-score for the MIMIC-III dataset, which is better as signified in Table 5. The LDSPC achieves 0.9694 accuracy, 0.9667 micro precision, 0.9683 micro recall, 0.9674 micro f1-score, 0.9536 micro precision, 0.9542 macro recall, and 0.9539 macro f1-score for the N2C2-2014 dataset, which is better as signified in Table 6.

Table 5. Comparison on MIMIC-III dataset

Method	Micro			Macro			Accuracy
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
DirPred [16]	0.8722	0.9340	0.9022	0.8714	0.9307	0.8997	0.6041
DSPCR [18]	0.8272	0.9797	0.8971	0.8262	0.9768	0.8952	0.5651
EHR-KnowGen [20]	0.2742	0.3228	0.2965	0.2354	0.2537	0.2376	0.3813
Proposed LDSPC	**0.9538**	**0.9829**	**0.9681**	**0.9462**	**0.9816**	**0.9635**	**0.9864**

Table 6. Comparison on N2C2-2014 dataset

Method	Micro			Macro			Accuracy
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
DirPred [16]	0.9270	0.9379	0.9323	0.9209	0.9360	0.9278	0.7050
DSPCR [18]	0.9093	0.9330	0.9210	0.9042	0.9260	0.9140	0.6432
EHR-KnowGen [20]	0.9449	0.9534	0.9492	0.9389	0.9459	0.9422	0.7826
Proposed LDSPC	**0.9667**	**0.9683**	**0.9674**	**0.9536**	**0.9542**	**0.9539**	**0.9694**

4.3 Discussion

The results of LDSPC are analyzed with different clustering approaches for MIMIC-III and N2C2-2014 datasets. The drawbacks of existing research are DirPred [16] required prior requirement of clusters due to its prior assumption about structural data that leads to suboptimal clustering performance. The DSPCR [18] required extensive training to capture temporal patterns and dependencies accurately, which led to huge computation time. The EHR-KnowGen [20] has various modalities that have domain discrepancies, thereby reducing overall performance. In this work, LDSPC is proposed for DRP in healthcare. The RoBERTa is applied for the encoding process, which can enhance clinical text interpretation and patient records, which leads to accurate risk prediction. The CoAM enables concentration on relevant parts of input text by attending to clinical features and disease-related data simultaneously, which enhances the capability to extract and integrate significant features from various sources. The LDSPC achieves better accuracy of 0.9864 and 0.9694 for MIMIC-III and N2C2-2014 datasets, respectively, when compared to DirPred [16], DSPCR [18], and EHR-KnowGen [20].

5 CONCLUSION

The LDSPC is proposed in this study for DRP in healthcare using medical notes. This process includes three modules, such as posterior, prior, and likelihood. The posterior module utilized an attentive encoder for extracting data from unstructured medical notes. Moreover, the clustering approach is integrated into the similarity module to learn the patient's useful representation of the latent Dirichlet state. These latent states are gathered into numerous cluster center weighted average, which are applied for DRP. The encoder with CoAM is used for encoding raw medical notes from actual language space to latent presentation. The datasets contain unstructured medical notes that are preprocessed by non-English characters and stop word removal processes. The LDSPC achieves better accuracy of 0.9864 and 0.9694 for MIMIC-III and N2C2-2014 datasets, respectively. Exploring divergence and hybrid clustering approaches for DRP leads to an accurate and computationally efficient model that enhances the scalability and predictive performance in managing complex healthcare data. In the future, this study suggests to exploring divergence and hybrid clustering approaches, which further enhance the accuracy and computation efficiency.

6 REFERENCES

[1] M. Ayaz, M. F. Pasha, T. Y. Le, T. J. Alahmadi, N. N. B. Abdullah, and Z. A. Alhababi, "A framework for automatic clustering of EHR messages using a spatial clustering approach," *Healthcare*, vol. 11, no. 3, p. 390, 2023. https://doi.org/10.3390/healthcare11030390

[2] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *npj Digital Medicine*, vol. 4, 2021. https://doi.org/10.1038/s41746-021-00455-y

[3] I. Kaur and T. Ahmad, "A cluster-based ensemble approach for congenital heart disease prediction," *Computer Methods and Programs in Biomedicine*, vol. 243, p. 107922, 2024. https://doi.org/10.1016/j.cmpb.2023.107922

[4] E. Werner *et al.*, "Explainable hierarchical clustering for patient subtyping and risk prediction," *Experimental Biology and Medicine*, vol. 248, no. 24, pp. 2547–2559, 2023. https://doi.org/10.1177/15353702231214253

[5] L. A. Carrasco-Ribelles, J. R. Pardo-Mas, S. Tortajada, C. Sáez, B. Valdivieso, and J. M. García-Gómez, "Predicting morbidity by local similarities in multi-scale patient trajectories," *Journal of Biomedical Informatics*, vol. 120, p. 103837, 2021. https://doi.org/10.1016/j.jbi.2021.103837

[6] H. Aguiar, M. Santos, P. Watkinson, and T. Zhu, "Learning of cluster-based feature importance for electronic health record time-series," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, 2022, pp. 161–179. [Online]. Available: https://proceedings.mlr.press/v162/aguiar22a.html

[7] X. Xiao *et al.*, "Treatment initiation prediction by EHR mapped PPD tensor based convolutional neural networks boosting algorithm," *Journal of Biomedical Informatics*, vol. 120, p. 103840, 2021. https://doi.org/10.1016/j.jbi.2021.103840

[8] Y. Z. Abd Elgawad *et al.*, "New method to implement and analysis of medical system in real time," *Healthcare*, vol. 10, no. 7, p. 1357, 2022. https://doi.org/10.3390/healthcare10071357

[9] F. Yan, H. Huang, W. Pedrycz, and K. Hirota, "A disease diagnosis system for smart healthcare based on fuzzy clustering and battle royale optimization," *Applied Soft Computing*, vol. 151, p. 111123, 2024. https://doi.org/10.1016/j.asoc.2023.111123

[10] M. Bampa, I. Miliou, B. Jovanovic, and P. Papapetrou, "M-ClustEHR: A multimodal clustering approach for electronic health records," *Artificial Intelligence in Medicine*, vol. 154, p. 102905, 2024. https://doi.org/10.1016/j.artmed.2024.102905

[11] M. A. Mohammed, O. Bismark, S. Alornyo, M. Asante, and B. O. Essah, "ResFCNET: A skin lesion segmentation method based on a deep residual fully convolutional neural network," *IETI Transactions on Data Analysis and Forecasting (iTDAF)*, vol. 1, no. 1, pp. 4–19, 2023. https://doi.org/10.3991/itdaf.v1i1.35723

[12] Y. Meng, W. Speier, M. K. Ong, and C. W. Arnold, "Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 3121–3129, 2021. https://doi.org/10.1109/JBHI.2021.3063721

[13] W. Shao *et al.*, "Application of unsupervised deep learning algorithms for identification of specific clusters of chronic cough patients from EMR data," *BMC Bioinformatics*, vol. 23, no. S3, 2022. https://doi.org/10.1186/s12859-022-04680-4

[14] G. Harerimana, G. Il Kim, J. W. Kim, and B. Jang, "HSGA: A hybrid LSTM-CNN self-guided attention to predict the future diagnosis from discharge narratives," *IEEE Access*, vol. 11, pp. 106334–106346, 2023. https://doi.org/10.1109/ACCESS.2023.3320179

[15] X. Chen and U. Comite, "Knowledge inference combining convolutional feature extraction and path semantics integration," *IETI Transactions on Data Analysis and Forecasting (iTDAF)*, vol. 2, no. 1, pp. 48–62, 2024. https://doi.org/10.3991/itdaf.v2i1.40095

[16] S. Niu *et al.*, "Enhancing healthcare decision support through explainable AI models for risk prediction," *Decision Support Systems*, vol. 181, p. 114228, 2024. https://doi.org/10.1016/j.dss.2024.114228

[17] A. Mansouri, M. Noei, and M. S. Abadeh, "A hybrid machine learning approach for early mortality prediction of ICU patients," *Progress in Artificial Intelligence*, vol. 11, pp. 333–347, 2022. https://doi.org/10.1007/s13748-022-00288-0

[18] S. Niu, J. Ma, Q. Yin, L. Bai, C. Li, and X. Yang, "A deep clustering-based state-space model for improved disease risk prediction in personalized healthcare," *Annals of Operations Research*, vol. 341, pp. 647–672, 2024. https://doi.org/10.1007/s10479-023-05817-1

[19] S. Liu, X. Wang, Y. Xiang, H. Xu, H. Wang, and B. Tang, "Multi-channel fusion LSTM for medical event prediction using EHRs," *Journal of Biomedical Informatics*, vol. 127, p. 104011, 2022. https://doi.org/10.1016/j.jbi.2022.104011

[20] S. Niu, J. Ma, L. Bai, Z. Wang, L. Guo, and X. Yang, "EHR-KnowGen: Knowledge-enhanced multimodal learning for disease diagnosis generation," *Information Fusion*, vol. 102, p. 102069, 2024. https://doi.org/10.1016/j.inffus.2023.102069

[21] R. AlSaad, Q. Malluhi, A. Abd-alrazaq, and S. Boughorbel, "Temporal self-attention for risk prediction from electronic health records using non-stationary kernel approximation," *Artificial Intelligence in Medicine*, vol. 149, p. 102802, 2024. https://doi.org/10.1016/j.artmed.2024.102802

[22] MIMIC-III Dataset link: https://physionet.org/content/mimiciii/1.4/ [Accessed on November 2024].

7 AUTHORS

Prasanthi Yavanamandha is a research scholar at the Department of Computer Science and Engineering at Koneru Lakshmaiah Deemed to be University, Hyderabad, India. She earned her Bachelor's degree in Information Technology in 2007 and her Master's degree in Software Engineering in 2011. Currently, she is pursuing Ph.D. in Computer Science, with a focus on machine learning and deep learning. Her research interests include deep learning, data science, machine learning. Prasanthi has contributed significantly to her field by publishing numerous research articles in esteemed journals and conferences (E-mail: prasanthi.yavanamandha@klh.edu.in).

Dr. D. S. Rao, PhD (NIT Rourkela), is a Professor at the department of Computer Science Engineering, at the KLEF (KL Deemed to be University) Hyderabad Campus. He is also an associate member of the INCE USA. He has published more than 20 peer review research articles indexed in SCI and Scopus databases, 7 patents and 7 book chapters. His work has been published in soft computing; noise prediction; noise induced hearing loss and lung cancer prediction. He is a reviewer of 3 SCI indexed journals. He is also an editorial board member of two academic journals. For his work, Dr. D.S. Rao has received awards from many organizations including Noise Pollution and Healthcare (E-mail: dsrao@klh.edu.in).