2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871867

PC-LSTM: Ontology-based Long Short-Term Memory State Model for Data Incompleteness Prediction*

Muhammed Shelleh[1] and Varadraj P. Gurupur[2]

Abstract— **Medical practices are engaged and motivated by new technologies and methods to enhance patient care as efficiently as possible. These new methods and technologies give way for medical practices and clinicians to have the insight, comprehension, and projections to develop better decisions and overall levels of care. In this paper, we propose a model, PatientCentered-LSTM (or PC-LSTM), using the states of the LSTM model to produce a novel, ontology-based state system for data incompleteness. The overall architecture and system design are based around utilizing the hidden and cell states of the LSTM model to produce a network of states for each of the corresponding hierarchies in an Electronic Health Record (EHR) system. The resulting methodology allows for an accurate and precise approach to predicting data incompleteness in electronic health records.**

Clinical relevance— **The method presented uses the hierarchical nature of electronic health record systems to positively influence the analysis of its data completeness; thereby, increasing the possibility of improved healthcare outcomes.**

I. INTRODUCTION

In the healthcare industry, Electronic Health Record (EHR) systems control and help designate patient and medical data throughout modern medical practices [1]. Medical infrastructure requirements can be highly demanding, and many EHR systems have been created to solve most of the problems seen in the practices the systems are employed in [2]. The architecture of most EHR systems tend to be quite linear and generally follow a design flow, leaving little room for improvement and providing a simple, yet uninspired user experience. System predictions and design flow is a critical problem when it comes to medical records, and some methods have been proposed to add an element of prediction and improve the overall user experience of the different records applications [3].

Different quality of life improvements, such as optimizing the speed of data access and creating a user interface that is more accessible, are critical for most medical practices to keep up with the demands of modern society [4], [5]. Based on this fact, the investigators chose to utilize and improve a Long Short-Term Memory (LSTM) model to provide greater accuracy and improved data hierarchy definitions.

The specific research objectives in this work are as follows:

1) To propose a novel method of predicting data completeness in Electronic Health Record (EHR) systems while leveraging the LSTM model's states
2) To test the method on a series of EHR datasets where there is significant amounts of missing data
3) To synthesize an ontology-based state model within the LSTM network to alter the hierarchy of the EHR system and assist with accuracy when predicting incomplete data.

The authors propose a novel approach to predict the completeness of data in electronic health records using PatientCentered-LSTM (PC-LSTM). The section that follows will explain the rationale behind the methods chosen for this approach.

II. BACKGROUND

A. Recurrent Neural Networks (RNNs) for Data Incompleteness

RNNs are a deep learning technique that utilize sequential information and perform the same "recurrent" tasks on each element in a sequence of information. Traditional RNNs have useful applications in data incompleteness and EHR data [6], [7], such as the creation of domain knowledge graphs [8] and understanding patient uncertainty [9]. Other applications in the medical field include patient monitoring [10] and support [11] in modern medical practices.

B. Application of Long short-term memory networks (LSTMs) in Data Incompleteness

LSTMs are a sub-type of RNNs that have a much greater memory capability for learning dependent tasks/computations. LSTMs have several uses, and solve the issues seen in vanilla RNNs. LSTMs have been applied to Data Incompleteness in past studies such as [12] and [13], however, each of these studies were not directly measuring the completeness of data and were more focused on using the completeness measurements during pre-processing. As a result, there is a gap in the literature for the use of deep learning models on data incompleteness.

III. METHODS

A. Mathematical Formulations

In this section, we present a methodology for creating the process ontology and state functionality for an electronic health records system, with the given structure assisting in the design and modeling of Recurrent Neural Networks (RNNs). The data collected in each of the categories and subcategories of the health records systems contain a timestamp,

*This work was supported by the University of Central Florida

[1]Muhammed Shelleh is a Ph.D student of Computer Science, University of Central Florida, Orlando, Florida, 32816, USA muhammed.abdel.shelleh@knights.ucf.edu

[2]Varadraj P. Gurupur with the School of Global Health Management and Informatics, University of Central Florida, Orlando, Florida, 32816, USA varadraj.gurupur@ucf.edu

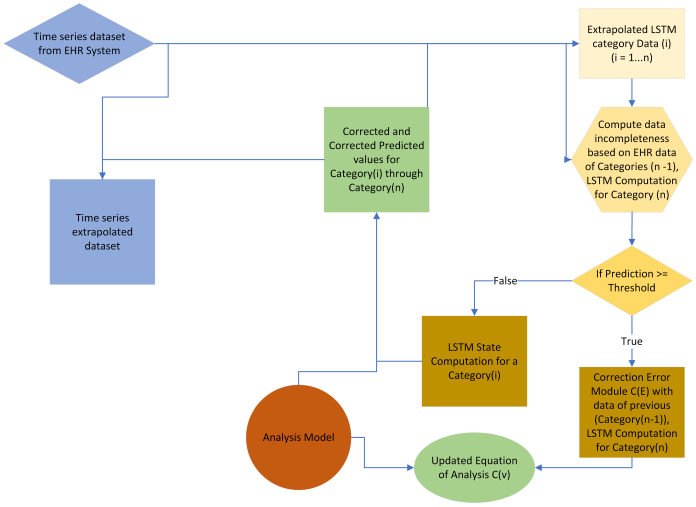


Fig. 1: Architecture using PC-LSTM on an Electronic Health Record system

Algorithm 1: PC-LSTM Model

Data: Health Records and EHR Hierarchy Data from PC(overall architecture)
Result: Forecast and Optimization of EHR Hierarchy
Let n be the output neurons;
Let X be an empty model of Neural Networks;
Let Y be the set of Health Categories in PC;
Let Z be the set of Records Sections in PC;
for *each y in Y* **do**
  X.add(layer(n)(y));
end
for *each z in Z* **do**
  X.add(layer(n)(z));
end
for *each h in Y* **do**
  Let s_y = [i is a set instance of Health Categories in PC — instance(i, h)];
  for *each j is* s_y **do**
    layer(n)(j) → layer(n)(h);
  end
  Let s_z = [z is a set instance of Records Sections in PC — ValidC(h, z)];
  for *each k is* s_z **do**
    layer(n)(k) → layer(n)(h);
  end
  if s_c = *[l is a set instance of Health Categories in PC — instance(n, l)]* **then**
    layer(n)(h) → LSTM;
  end
end
return X;

allowing the design of the ontology to be modeled after a time ordered sequence, or a time series.
For this, we define a time series for the ordered data as the following:

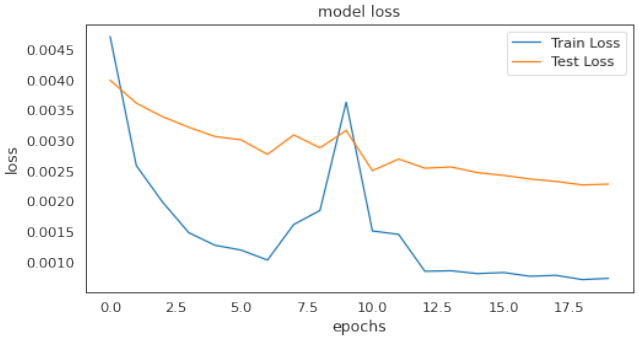
- A time-ordered sequence of arrays and data points
- The arrays and data points contain n values (where values are numbered 1, 2, ..., n)
- Each value is associated with a timestamp
- The arrays and data points of n values match with the inputs associated with the data measured at certain, periodic intervals

This definition of a time series is adopted from several works on time series networks [14], but does not encompass the definition needed within the medical records ecosystem. To reflect these needs, a new definition for an EHR system time series is as follows:

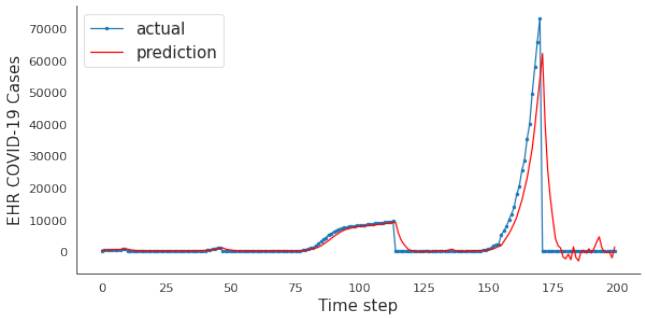
- An EHR time series, $Y = [c^1, c^2, ..., c^n]$, is a time-ordered sequence of arrays and data points
- Each point, c_n^t represents n data values for t timestamps for c categories in the hierarchy, where the array of hyperedges, $X = [c_1^t, c_2^t, ..., c_n^t]$.
- Each point, $c_n^t \in W = (X, H)$, where W is a hypergraph with x vertices to represent the sub-category data values corresponding to each record inputted into the system.
- $[h_k, h_l]$, where the ordered pair of directed hyperedges is the subset of X and contains the values of the categories measured in Y from c^n.
- Our formal definition of hyperedges $H = [(c_k, c_l)|(c_k$ precedes $c_l), c_k, c_i \subseteq X, c_k \cap c_i = \varnothing]$.

B. PC-LSTM Algorithm and Explanation

Briefly, a typical medical practice requires precise and accurate processes to deliver the highest quality patient care possible. Ontologies can be used as a state within current RNN models to better improve patient quality of care and provide a more streamlined process in how EHR systems



(a) Loss Values for data completeness



(b) Predictive data completeness

Fig. 2: Loss values and predictive data completeness using PC-LSTM on COVID-19 EHR dataset

operate. Given the hierarchy of how these systems operate, we propose a series of classes and rules to further define the ontology development and state usage of the LSTM model and provide further explanation for what each of these classes mean:

- Each EHR system includes several different categories and sub-categories where information is stored, where the output in categories depends on the input from the sub-categories, and each category has multiple sub-categories. These sub-categories contain a hierarchy of how they are stored, and their respective data.
- Measures are taken for all of the data in each of the sub-categories, include timestamps and datatypes.
- Dependencies are measured in *instances*.
- *validC* identifies categories, and *hasData* identifies completeness and availability of data in the sub-categories.
- *values* represents the data measured for each data point in the EHR system.

The methodology proposed follows a typical EHR architecture and goes through the system, checking for data completeness as:

1) Identify the categories of an EHR system and subsequent hierarchical dependencies (*instances*(***category0, category1***))
2) Identify the sub-categories of each indexed category (*validC*(***category0, subcategory1***))
3) Identify whether the sub-categories have complete datasets (*hasData*(***subcategory0, data1***))
4) Identify the completeness of the dataset in the sub-categories (*values*(***data0, data1***))

By applying these stages and steps through the methodology, the process for finding incomplete data becomes much faster and has higher predictability. PC-LSTM utilizes the states of a typical LSTM model and creates a binary representation from the time series EHR data, with representations of the incompleteness ranging from 0 to 1 (where 0 is incomplete and 1 is complete). The states of the LSTM are utilized to ”remember” and record the overall predicted score of incompleteness, which get tuned further as the model is run.

IV. RESULTS

To show the practicality and use of the PC-LSTM model, tests were done on several different datasets to measure data completeness, where the goal was to properly predict when a data field may be left incomplete and what tends to be the most incomplete within the set of electronic health records.

A. Preprocessing and Data Incompleteness Analysis

The dataset used contains over 44,000,000 unique observations from COVID-19 cases in electronic health records from across the country. This preprocessing and model extraction methodology was represented in Figure 1 and gives a high-level overview of what pre-processing steps were taken to ensure high accuracy for data incompleteness prediction when using PC-LSTM. The steps for preprocessing were the following:

1) Define the incompleteness ratio of the dataset and the distribution fitting model to be used
2) Get the fitted X and Y points from the dataset
3) Define column completeness ratios by the inverse of incompleteness for each row of a specified column
4) Plot the data points from the completeness ratios onto several different distributions to understand the fit of the data. Based off of the fit, choose the most robust incompleteness measure along with the date column to convert data points into a time series dataset
5) Measure the unit root using the Dickey-Fuller Test, then refit the dataset using a scaler and fit along the incompleteness ratio

B. Evaluation of PC-LSTM

The task utilized in the case of testing PC-LSTM is classification-based, and we can map missing data values to 0, while the existing, complete data is mapped to 1. As a result, we can check this using the Dickey-Fuller Test for testing the null hypothesis and finding differences in a time series set [15]. The results of the Dickey-Fuller Test can be seen in Table I. Further representation of the analysis can be seen in Figure 3. Figure 3 shows the predicted data completeness values from the EHR usage dataset, following the predictions from PC-LSTM. Figure 4 shows the state representation of the ontology hierarchy in the EHR system.

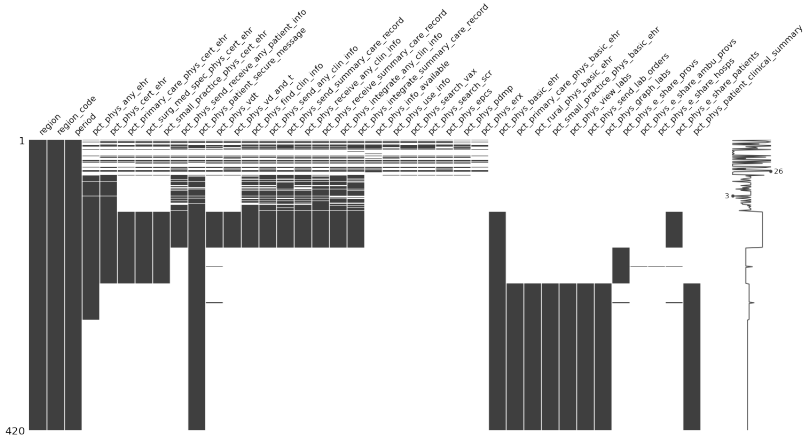


Fig. 3: Predicted data completeness processed by PC-LSTM

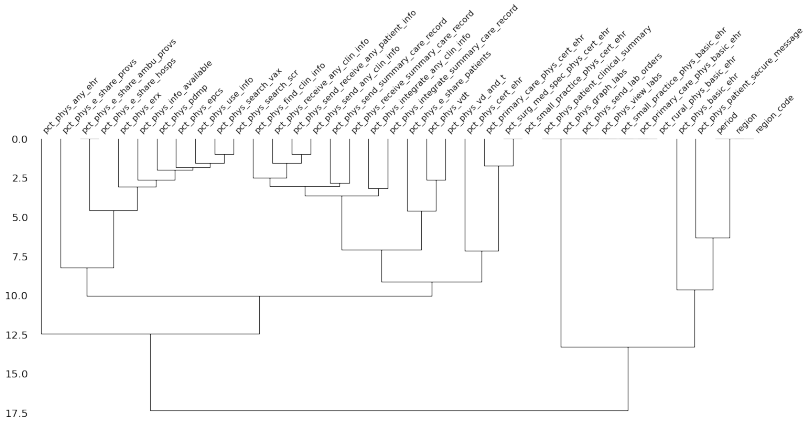


Fig. 4: Ontology representation of EHR hierarchy processed by PC-LSTM

TABLE I: Measures and Results of Dickey-Fuller Test on COVID-19 EHR Dataset

Results of Dickey-Fuller Test	
Test Statistic	**-2.5842**
p-value	**0.0963**
#Lags Used	**7.0000**
Number of Observations Used	**4972.0000**
Critical Value (1%)	**-3.4709**
Critical Value (5%)	**-2.8793**
Critical Value (10%)	**-2.5763**

V. DISCUSSION

A. PC-LSTM versus MADI

MADI, a previous work done by the authors, was proposed by Gurupur et. al as a method to analyze data incompleteness in EHR systems [6]. PC-LSTM, however, not only analyzes and builds off the existing work, but predicts where data may be incomplete and allows for the user of the EHR system to curb fields that may otherwise be left incomplete. PC-LSTM also uses a different algorithm, where the main focus is in RNNs and state models in the LSTM network. As a result, PC-LSTM improves on the previous model in many different ways while also leaving a similar foundation.

B. Limitations

Given the novelty and scope of the methodology, several limitations were found during design and implementation. The first limitation is that there were no similar methods to utilize as a baseline for comparison, given the lack of use of LSTM networks for EHRs and data incompleteness. The second limitation the authors found is that the model is not always able to process data that is mostly complete and works best with a dataset that has many missing values.

VI. CONCLUSION

In this paper, we presented a methodology and proposed a model to utilize the states of an LSTM network for predicting where data may be incomplete in an EHR system. We introduced our model, PC-LSTM, as a method to predict and classify missing data. As a result, we demonstrated that such a concept has the potential to assist medical practices in predicting where their data may be incomplete, allowing them to better target incomplete or missing medical data. Our main contributions were the implementation of a novel LSTM model and a methodology for solving the data incompleteness problem. In the future, we are interested in looking more into the topic of deep learning and seeing how it can be applied to further predicting data incompleteness.

REFERENCES

[1] A. Nasir, V. Gurupur, and X. Liu, "A new paradigm to analyze data completeness of patient data," *Applied clinical informatics*, vol. 7, no. 03, pp. 745–764, 2016.

[2] V. Gurupur, A. Nasir, and X. Liu, "Method and system for managing health care patient record data," Sep. 29 2020, uS Patent 10,790,049.

[3] E. Jamoom, *Physician experience with electronic health record systems that meet meaningful use criteria: NAMCS Physician Workflow Survey, 2011*. US Department of Health and Human Services, Centers for Disease Control and ..., 2013, no. 129.

[4] K. V. Bolgova, S. V. Kovalchuk, M. A. Balakhontceva, N. E. Zvartau, and O. G. Metsker, "Human computer interaction during clinical decision support with electronic health records improvement," in *Research Anthology on Decision Support Systems and Decision Management in Healthcare, Business, and Engineering*. IGI Global, 2021, pp. 1316–1330.

[5] L. Rundo, R. Pirrone, S. Vitabile, E. Sala, and O. Gambino, "Recent advances of hci in decision-making tasks for optimized clinical workflows and precision medicine," *Journal of biomedical informatics*, vol. 108, p. 103479, 2020.

[6] V. P. Gurupur and M. Shelleh, "Machine learning analysis for data incompleteness (madi): Analyzing the data completeness of patient records using a random variable approach to predict the incompleteness of electronic health records," *IEEE Access*, vol. 9, pp. 95 994–96 001, 2021.

[7] A. Nasir, X. Liu, V. Gurupur, and Z. Qureshi, "Disparities in patient record completeness with respect to the health care utilization project," *Health informatics journal*, vol. 25, no. 2, pp. 401–416, 2019.

[8] C. Yin, R. Zhao, B. Qian, X. Lv, and P. Zhang, "Domain knowledge guided deep learning with electronic health records," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 738–747.

[9] M. W. Dusenberry, D. Tran, E. Choi, J. Kemp, J. Nixon, G. Jerfel, K. Heller, and A. M. Dai, "Analyzing the role of model uncertainty for electronic health records," in *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020, pp. 204–213.

[10] A. Procházka, O. Dostál, P. Cejnar, H. I. Mohamed, Z. Pavelek, M. Vališ, and O. Vyšata, "Deep learning for accelerometric data assessment and ataxic gait monitoring," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 360–367, 2021.

[11] L. De Vree and R. Carloni, "Deep reinforcement learning for physics-based musculoskeletal simulations of healthy subjects and transfemoral prostheses' users during normal walking," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 607–618, 2021.

[12] F. Biessmann, D. Salinas, S. Schelter, P. Schmidt, and D. Lange, "" deep" learning for missing value imputationin tables with non-numerical data," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 2017–2025.

[13] N. Handa, A. Sharma, and A. Gupta, "Framework for prediction and classification of non functional requirements: a novel vision," *Cluster Computing*, pp. 1–19, 2022.

[14] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[15] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American statistical association*, vol. 74, no. 366a, pp. 427–431, 1979.