International Journal of Multidisciplinary Research and Growth Evaluation
ISSN: 2582-7138
Received: 28-01-2020; Accepted: 23-02-2020
www.allmultidisciplinaryjournal.com
Volume 1; Issue 1; January-February 2020; Page No. 120-124

Feature Engineering for Healthcare Big Data: Approaches to Missing Data Imputation, Dimensionality Reduction, and Time-Series Analysis

Simran Sethi
Independent Researcher, USA

Corresponding Author: **Simran Sethi**

DOI: https://doi.org/10.54660/.IJMRGE.2020.1.1.120-124

Abstract
In the domain of healthcare analytics, especially regarding electronic health records (EHRs), feature engineering plays an integrative role in clinical understanding of Big Data. The scale of EHR data has a wealth of temporal and heterogeneous information, but it is extremely challenging because of the high dimensionality, irregular sampling, and high volumes of missing data. This paper presents an attempt to capture some of the most important methods on missing data imputation, dimensionality reduction and time-series modeling. In addition, we construct an empirically grounded feature engineering pipeline based on real world experiences, which include significant deduplication projects within clinical data. The review and framework not only provide insights but also serve as a practical guide for researchers and practitioners for enhancing the use of EHR data in predictive modeling, patient stratification, and population health analytics.

Keywords: Feature Engineering, Healthcare Big Data, Electronic Health Records (EHR), Missing Data Imputation, Dimensionality Reduction, Time-Series Analysis, Machine Learning

1. Introduction
The implementation of Electronic Health Record (EHR) systems has resulted in the availability of massive amounts of clinical information, which can be utilized for these wide scales analytics and sophisticated predictive modeling. This information is important for making clinical decisions, managing the health of the populations, and applying precision medicine. Unfortunately, the EHR data comes with numerous difficulties that can obstruct machine learning processes such as (1) high dimensionality due to many possible variable that include lab tests, medications, and diagnoses, (2) missing data from the irregular documentation patterns, and (3) time series complexities because of irregular and multi-resolution measurements [1, 2]. As a result, EHR data requires significant transformation, in some cases even feature extraction, to provide machine learning models with information adequate to provide insights.

The use of deep learning, probabilistic modeling, and representation learning has shown the significance of data features. Techniques like auto encoders and generative adversarial networks (GANs) and other types of unsupervised learning have become efficient means for missing values imputation, compact patient representation learning, and irregular time interval management [3-6]. Together, these methods have the potential to improve accuracy in downstream tasks, including disease risk stratification, length-of-stay predictions, and patient subtyping when applied within an end-to-end pipeline framework [7].

This paper aims to (1) review literature for existing features engineering processes for EHR data focusing missing data imputation, data features aggregation, time-series data analysis, (2) share lessons and a suggested pipeline from real life work that covers extensive EHR de-duplication project, and (3) propose the boundaries of this research focusing what the author deems methodological gaps and unresolved problems.

2. Background and literature review
A. Missing data imputation
Some missingness in the healthcare data can be attributed to poor documentation practices, missed appointments by patients, or symptoms driven selective documentation of investigations done [8]. Traditional methods like mean/median imputation tend to be used but do not capture the rich relationships within clinical variables, or the temporal aspects of the measurements [9]. MICE is well accepted in the medical research context but it is still based on a number of regression equations, which may not account

for sophisticated interactions between variables [10].

Methods based on deep learning approaches have transformed missing data imputation for the better. The denoising autoencoders can have representations that capture latent features and reconstruct missing values from multidimensional data [3]. There are Generative Adversarial Imputation Networks (GAIN), for example, where the task and the impute missing data is presented as a contest between a generator and a discriminator and achieve state-of-the-art results on benchmark datasets [11]. Moreover, models like GRU-D aim at missing data pattern directly in the architecture of recurrent neural networks where the steps of masking and the time decay are incorporated into the structure [12]. These methods are often combined with other data processing steps considering the clinical background context such as being the informative missingness (e.g., a lab test that was not done is not clinically warranted).

B. Dimensionality reduction and representation learning

EHR data is made up of multiple elements that can be classified as structured (ICD codes, lab results, and vitals) or unstructured (clinical notes). Typical feature extraction techniques (e.g., principal component analysis, PCA) are limited by how well they can model the complex interactions associated with clinical coding [13]. As in many other contexts, PPCA and autoencoder-based methods have surfaced as more viable options for dimensionality reduction, most of the time also decreasing the data reconstruction error [14].

Med2Vec and other neural embedding techniques translate clinical codes into dense lower dimensional vectors, simultaneously translating vast amounts of medical data [3], [15]. Some other hierarchical attention models augment concept representation learning by incorporating domain ontologies, like ICD or SNOMED hierarchies [16]. One of the greatest merits of these approaches is that they can capture local code-level co-occurrences (e.g., laboratories, diagnoses, medications) as well as group-level relatedness (e.g., a cluster of diagnoses suggesting a more entrenched single chronic condition).

C. Time-series analysis in EHRS

Due to the constantly evolving clinical measures, healthcare professionals face the challenge of unevenly distributed time-series data sampling. This reality severely hampers the standard methods that rely on fixed time intervals. Besides, many standard methods require a complete dataset for each time step which is highly problematic when you are dealing with delays or missing information [2].

In recent years, sequence analysis in EHR systems has greatly benefited from architectures based on RNNs Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) [2, 4]. Sometime-aware models like T-LSTM and GRU-D use the time interval between two successive events to modulate the hidden state decay rates [12, 17]. These models improve the measurement gap problem, using the clinical importance of time, i.e. how certain conditions change slowly over hours versus weeks, to improving the clinical data capture. Convolutional neural networks (CNNs) have also seen applications in physiological signals, however, they usually depend on data being sampled uniformly or on special preprocessing where the data is aligned to a template [18].

3. Proposed feature engineering framework

Based on existing works and my own experience on utilizing big data in healthcare, I suggest a framework that does feature engineering in three stages to mitigate missing data, dimensionality, and temporal issues. This framework is discussed in more detail in Fig. 1.

A. Data cleaning and integration

- **Deduplication of EHR records:** Unifying patient records within extensive healthcare data repositories is challenging during the merging process. Issues stem from name identifiers, name spellings, and missing demographic details. In my previous experience with an enterprise healthcare technology firm, we used a very specific toolkit (Duke) designed for matching and merging records. Integration of such tools is necessary to take the first steps toward attaining a single patient record to mitigate data merges and feature set incompleteness.
- **Metadata Standardization:** Translation of local codes into sets of standard terminologies (ICD-10 or LOINC) as well as standardization of unit measures. This step is crucial because if left unattended will render elimination of dimensional reduction approaches in later steps impossible.

B. Missing data handling

a) Missingness Assessment: Evaluate missingness patterns including Missing at Random, Missing Completely at Random, or Missing Value Not at Random. Analyze the amount and allocation of missing values for a proportion of variables to help direct imputation methods.

b) Imputation Approach:

- **Unstructured or low-dimensional:** For basic laboratory measurements (a single variable), try out the 3D-MICE algorithm for time-series data if longitudinal data is available [9].
- **High-Dimensional:** For data containing complex multi-clinical parameters, use autoencoder-based [11] and GAIN [13] techniques to facilitate imputation.
- **Temporal Modeling:** GRU-D accounts for the last observation's time interval and patient state change, which helps integrate the evolving patient state into the model's imputation structure [12].

3. Dimensionality reduction and temporal feature extraction

- **Representation Learning:** Neural embedding models such Med2Vec or stacked denoising autoencoders should be trained to build compact representations capturing the latent clinical context Med2Vec or vai. Where possible, incorporate domain knowledge via ontological hierarchies to improve interpretability.
- **Sequence Modeling:** Extract temporal features from sequences of clinical events using RNN variations (LSTM, GRU, T-LSTM, GRU-D). In the other case, if data can be aligned into evenly spaced intervals, then 1D convolutions or temporal CNNs can be used.
- **Hybrid Approaches:** It is also possible to combine learned embeddings with time-sensitive models. For example, embed medical codes to be represented as a discrete metric in continuous space, then incorporate, in addition to continuous labs/vitals, these embedded representations into a T-LSTM model.

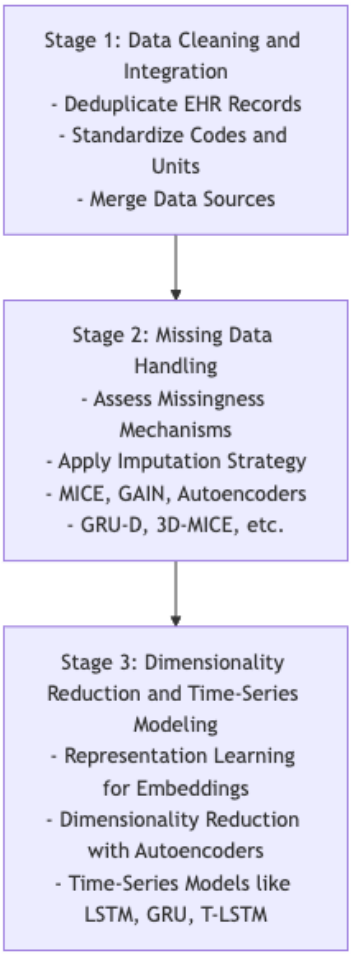

Fig 1: Flowchart illustrating different stages of feature engineering

4. Methodology

A. Dataset Description

In practice, the above framework can be illustrated using an EHR dataset from multiple hospital systems that have been de-identified. Usual data repositories may include the following categories:

- **Demographics:** Age, gender, ethnicity, and insurance coverage.
- **Diagnoses:** ICD codes for chronic and acute conditions.
- **Medications:** Pharmacy orders with medication names, doses, and administration dates.
- **Lab Tests:** Wide range of laboratory results with a set of comprehensive, and often irregular, sampling frequency.
- **Vital Signs:** Heart rate, respiratory rate, blood pressure, temperature and other vital signs usually recorded at patient visits.

B. Experimental Setup

a) Data Pre-processing:

- Combine exact matching using patient ids along with probabilistic matching using name, DOB, and address to remove duplicate patient records.
- Normalise clinical codes and consolidate lab measurement units.
- Stratified sampling for patients' dataset with training, validation, and test subsets of 70%, 10%, and 20%, respectively.

b) Imputation Evaluation:

- For evaluating imputation quality through RMSE or MAE, simulate missingness in a subset of the variables that remain complete.
- Evaluate different imputation methods including MICE, PPCA, denoising autoencoders, GAIN, and GRU-D [9, 12, 14].

c) Dimensionality Reduction:

- Fit autoencoder based decoders to the training set to learn low dimensional representations for each patient visit, then validate using reconstruction loss on the validation set to ensure no overfitting.
- Alternatively, represent medical codes for vectorized or sequence-based analysis using Med2Vec or hierarchical attention models [3, 16].

d) Predictive Modeling:

- An LSTM or GRU classifier is implemented to predict a selected clinical outcome (e.g., 30-day readmission).
- The features are the learned embeddings from the reduction step, with additional clinical factors or alone.
- AUC, precision-recall, or F1 score will be used as measures to evaluate the predictive performance.

5. Discussion

The earlier constructs highlight the aspects of data quality (for example, deduplication) alongside an effort for missing data imputation, dimensionality reduction, and temporal modeling. According to my understanding, data cleaning and deduplication, in particular, can mitigate the level of fragmentation and feature noise caused by the patient identity

mistreatment, which in turn is biased. Additionally, advanced imputation techniques have proven to be far better compared to naive ones when it comes to capturing variable correlations and time-varying states [9, 10, 12].

A. Challenges and Limitations

- **Computational Costs:** Deep imputation models and large-scale representation learning tend to be expensive. For hospitals or research institutions with limited access to ample GPU cores, training on large datasets can prove to be difficult.
- **Interpretability:** At the same time, autoencoders and deep neural networks, while offering performance increases, may not be straightforward to interpret. The introduction of attention mechanisms, or knowledge embedding from other domains (for instance, ontologies), can solve part of this problem [16].
- **Data Heterogeneity:** EHR datasets deduplicated and standardized usually differ across various hospital systems in reference to terminology, lab ranges, and documentation. This makes model generalization difficult.
- **Regulatory Considerations:** Ensuring patient privacy and compliance to regulations like HIPAA and GDPR is always paramount. Processes of de-identification must be fully robust, restricting certain analyses (having timestamps to the exact date and time) in the process.

B. Future Directions

- **Federated Learning:** In an effort to overcome privacy and data silo issues, federated learning can be used to build collaborative models within organizations that do not require the exchange of raw patient data. This should increase the effectiveness of feature engineering pipelines.
- **Causality-informed feature engineering:** The introduction of causal reasoning into the process of feature extraction and construction enables the selection and modification of features that are truly reflective of the underlying phenomena as opposed to being superficial proxies.
- **Multi-modal data integration:** Later versions should merge the imaging and genomic information with data from sensors and EHR event streams to form a comprehensive representation of the patient. This creates a need for novel feature engineering approaches that differ in their modality and their data structure.
- **Real-time EHR analytics:** Applying streaming or online learning algorithms could be helpful to enhance patient monitoring systems and early warning scores by providing real-time updates of features based on new incoming data.

6. Conclusion

The understanding of EHR Big Data would not be possible without feature engineering. Data gaps filling, shrinking the number of features, and time series analysis are key pillars in unlocking the potential of these large, complex datasets. The development of deep learning–based imputation and sequence models has achieved significant gains in predictive accuracy and the representation of patients. However, problems such as data repetition and non-data base-shard standardization need to be solved at the beginning and not the end of the pipeline to provide strong analytics at the end.

This paper is built on existing literature and the author's practical experience, which helps build a comprehensive feature engineering framework. In the future, researchers may build on those strategies by employing federated learning, causality, multi-modal data fusion, and others to improve clinical perspective even more. Ultimately, precision medicine and better healthcare outcomes will be supported through EHR data transformed into actionable information by well-crafted feature engineering workflows.

7. References

1. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific Reports. 2016;6:1-10.
2. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to Diagnose with LSTM Recurrent Neural Networks. In: Proceedings of the International Conference on Learning Representations (ICLR); 2016.
3. Choi E, Bahadori MT, Sun J, *et al*. Multi-layer Representation Learning for Medical Concepts (Med2Vec). In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD); 2016.
4. Choi E, Bahadori MT, Song L, Stewart WF, Sun J. GRAM: Graph-based Attention Model for Healthcare Representation Learning. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD); 2017.
5. Baytas IM, Xiao C, Zhang X, Wang F, Jain A, Zhou J. Patient Subtyping via Time-Aware LSTM Networks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD); 2017.
6. Beaulieu-Jones BK, Moore JH. Missing Data Imputation in the Electronic Health Record Using Deeply Learned Autoencoders. In: Proceedings of the Pacific Symposium on Biocomputing (PSB); 2017.
7. Rajkomar A, Oren E, Chen K, *et al*. Scalable and Accurate Deep Learning with Electronic Health Records. npj Digital Medicine. 2018;1(18):1-10.
8. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE Journal of Biomedical and Health Informatics. 2018;22(5):1589-1604.
9. Luo Y, Szolovits P, Dighe AS, Baron JM. 3D-MICE: Integration of Cross-Sectional and Longitudinal Imputation for Multi-Analyte Longitudinal Clinical Data. Journal of the American Medical Informatics Association. 2018;25(6):645-653.
10. Yoon J, Jordon J, van der Schaar M. GAIN: Missing Data Imputation Using Generative Adversarial Nets. In: Proceedings of the International Conference on Machine Learning (ICML); 2018.
11. Beaulieu-Jones BK, Lavage DR, *et al*. Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis. JMIR Medical Informatics. 2018;6(1):e11.
12. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Scientific Reports. 2018;8(1):1-13.
13. Venugopalan J, Chanani N, Maher K, Wang MD. Novel Data Imputation for Multiple Types of Missing Data in

Intensive Care Units. IEEE Journal of Biomedical and Health Informatics. 2019;23(3):1255-1262.
14. Hegde H, Shimpi N, Panny A, *et al*. MICE vs PPCA: Missing Data Imputation in Healthcare. Informatics in Medicine Unlocked. 2019;15:100178.
15. Xu D, Hu PJ, Huang T-S, Fang X, Hsu C-C. A Deep Learning–Based Unsupervised Method to Impute Missing Values in Electronic Health Records. Journal of Biomedical Informatics. 2020;102:103370.