Improving EHR Data Imputation with Distribution-Based Methods

Dan Brody

Department of General Engineering, Cooper Union

Electrical Engineering Senior Projects

Professor Keene

May 16, 2022

Appendix

Abstract

This paper will be improving data imputation (filling in missing values with realistic values that match the data) in Electronic Health Records (EHRs) to improve any secondary use of EHR data. The idea to improve imputation is to use a distribution-based method. Through experimentation the method has shown to work consistently across datasets, showcasing the role that distributions have on the method of imputation.

I. Introduction

Data imputation (filling in missing values with realistic values that match the data) for Electronic Health Records (EHRs) is a pressing topic to improve any secondary use, particularly in the data science space. EHRs contain time series data for respective patients detailing their prescribed medications, results for lab tests, etc. for a given visit and day. Many machine and deep learning algorithms have used EHRs for secondary use to achieve a variety of goals including endpoint prediction (forms of personalized medicine). Now what is endpoint prediction? Consider Fig.5.

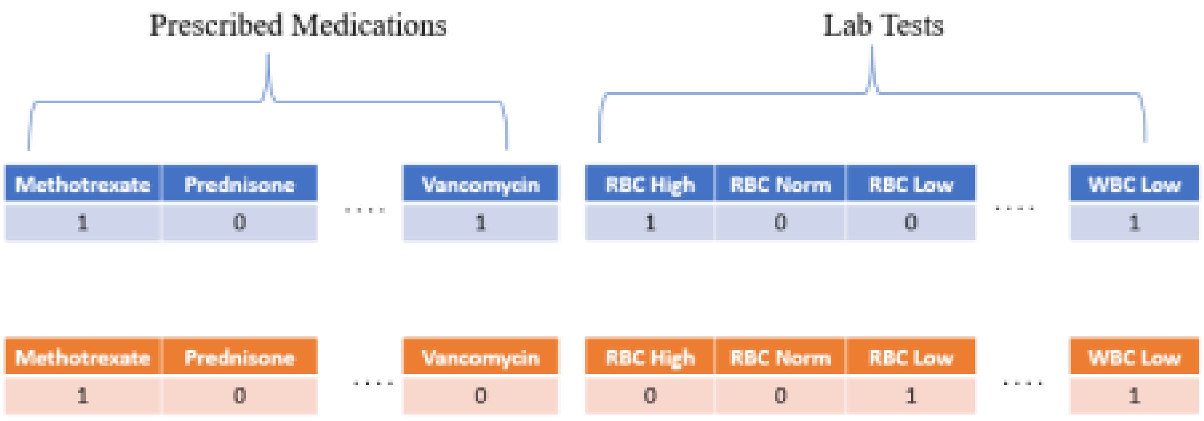


Fig.5. Illustration for Endpoint Prediction Explanation

When a patient goes into the hospital, they are prescribed medications, assigned lab tests, reporting their symptoms, etc. This can be considered as one visit which is one record in the EHR. An example of endpoint prediction would be if we want to take the features in the first visit (blue) in Fig.5. to predict whether red blood cell count (RBC) will be high in the next visit (orange). If we want to predict RBC for a visit on the next day (i.e., if the orange visit is the next day), then this is next-day endpoint prediction and if we want to predict RBC for a visit on the same day (i.e., if the orange visit is the same day) then we have same-day endpoint prediction.

As an example of some current machine and deep learning models that that do endpoint prediction on EHR data, Recurrent Neural Networks (RNNs) have been used to predict next-day as well as same-day sepsis, myocardial infarction (MI), and vancomycin administration for ICU patients using EHR data such as in [1]. Additionally, in using attention, the authors were able to make their results interpretable to a medical professional, allowing an easier time for the professional in assigning treatments.

While contributions like [1] exist, an issue with secondary EHR use is the nature of the

EHR data. EHR data is challenging to represent and model due to its high dimensionality, noise, heterogeneity, sparseness, incompleteness, random errors, and systematic biases [3,4,5]. In the ICU rates of human error have ranged from 30% to 80% [6,7,8]. A specific case of human error was discussed in [6]:

> There were 241 human errors (31%) in 161 patients, evenly distributed among planning (n = 75), execution (n = 88), and surveillance (n = 78). One error was lethal, two led to sequelae, 26 % prolonged ICU stay, and 57 % were minor and 16 % without consequence…. Human errors prolonged ICU stay by 425 patient-days, amounting to 15 % of ICU time.

In addition to human error, some events are either not recorded in the EHR or not done because they are not billable or because financial burden needs to be considered in ordering medications, tests, surgeries, etc. Thus, many events in the EHR do not occur often and there is high missing rate and sparseness as well as invalid values. High missing rate is especially hurtful towards secondary EHR use because the accuracy of a result (i.e. tests or symptoms) decreases over time. To illustrate the need, suppose an important predictor/feature towards a model is logged for one visit but not for the next 2 or 3. Using the logged result as a placeholder for those next 2 or 3 visits may become inefficient to predict the target feature, thus decreasing model performance. Hence EHR data needs accurate data imputation for accurate secondary use. (find source)

Common data imputation methods that are used in this space are median imputation and mean imputation. Median and mean imputation impute missing values by imputing the median and mean, respectively, of the observed data of the feature for each respective feature. These imputations are quick and can be accurate under certain missing rates, however accuracy depletes as the number of samples in a feature decrease. Variability of the data is reduced and as a result standard deviation is typically underestimated [10]. As an example, consider a binary feature that is balanced where the number of ones is the same number of zeroes. If all or most of the zeroes disappear then the number one would become the median and be held as a placeholder for the missing values. If zeroes are approximated with one's using median imputation, in this case, will result in bad imputation performance.

The distribution-based algorithm will be used on datasets from the UCI machine learning repository. The paper is devised as follows. Part II will discuss related works. Part III will discuss algorithms and mathematics that enhance understanding of the model and experiments. Part IV will discuss the project in detail. Part V will discuss what has been done this semester including results. Part VI will talk about next steps.

II. Related Works

Different methods of EHR data imputation have been attempted previously. Related works include [23] and [24].

In [23] the author performs an extension of GRAPE (find source), where GRAPE represents data as a bipartite graph where sample and feature nodes are connected according to the missingness pattern of the dataset and data imputation corresponds to edge-level prediction for the missing links.

What the paper adds is a way to use a Graph Neural Network (GNN) on the bipartite graph to more accurately predict whether an event happened but was not recorded. The algorithm is specifically for EHR data because in the EHR if events are not billable then they will not be logged and hence there is high incompleteness and sparseness.

If events for the dataset are logged as 0 (i.e., did not happen) then the algorithm assumes these values as *unmeasured events with unknown outcomes* and are considered absent. There exist three different groupings. Letting k be defined as a hyperparameter there exists a set E_{neg} for k edges that are absent (events that are missing), E_{vis} for positive edges that are visible to the model (observed events), and E_{inv} for k positive edges that are unknown to the model (unmeasured events with unknown outcomes). At each iteration of the GNN these three groupings are sampled such that patient and event frequencies are preserved. This makes a powerful assumption that patient and event frequencies will not change, relying on the observed values of the data.

In [23] the authors compared their algorithm against GAIN [2], Denoising Autoencoder (DAE) [22] and k-NN imputation [25]. The dataset the authors used was the IBM Explorys dataset consisting of ICD/diagnosis codes for each patient that had values taken away until the dataset was extremely sparse (98.3% zeros).

In terms of performance the author's algorithm underperformed the other methods in terms of both sensitivity and specificity but outperformed in terms of overall accuracy by a significant amount. Accuracy for the author's algorithm was 0.79±0.09 and the second highest algorithm's accuracy for a model in [23] was 0.63±0.09. The poor performance in sensitivity indicates the algorithm has no applicability in the medical space, as sensitivity is important, but, due to its performance in accuracy, can be improved to outperform state of the art. A large difference between the author's algorithm and my algorithm is that their algorithm has data imputation of missing positive values (i.e. events that should have occurred but were not entered) whereas my algorithm has a larger emphasis on values that are known to be missing (i.e. were already empty before any preprocessing).

Another work that is relevant to EHR data imputation is [24]. [24] compared different algorithms between different levels of missingness for data imputation. For the experiments in [24] median imputation, mean imputation, SVDImpute [25], KNNImpute [25], and autoencoder with dropout were used in the MNAR case and MCAR case for ALS disease progression. Results of the paper can be found in Fig. 6 -7. In both the MNAR and MCAR case median and mean imputation performed badly in terms of root mean squared error (RMSE). The preceding result is not surprising as median and mean imputation are generally suited for MAR. In addition Fig.6-7 also shows the autoencoder with dropout outperforms in both MNAR and MCAR when spike-in ratio less than or equal to 0.5. The paper, however, makes an argument that the autoencoder and other implementations will converge to the performance of median and mean imputation in Fig. 7 for large enough spike-in ratio. As [24] was made in 2016, results may not be as relevant however in the MNAR case as in [23], where 98.3% of the values were 0, DAE was able to outperform the other methods in sensitivity.

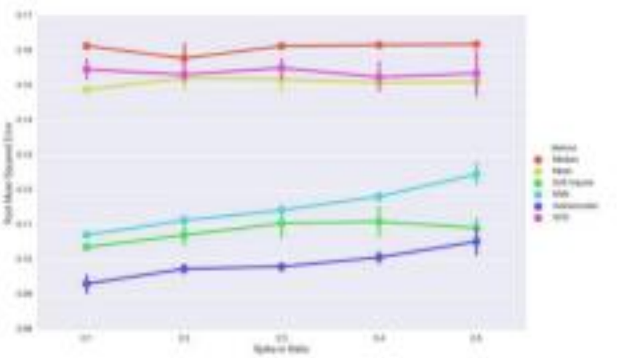


Fig. 6. Effect of the amount of spiked-in missing data on imputation for the MCAR case. Error bars indicate 5-fold cross validation score ranges. [24]

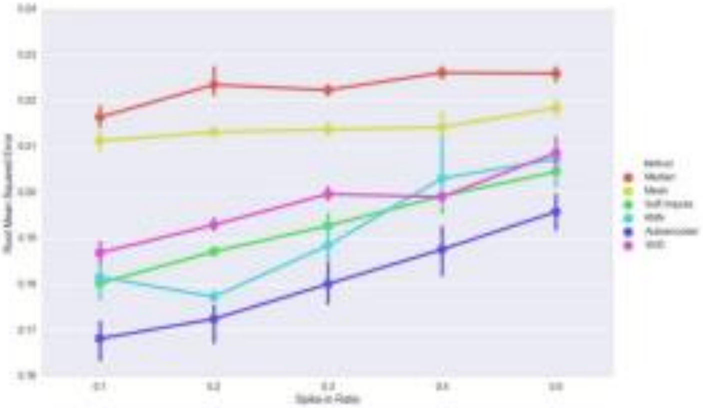


Fig. 7. Effect of the amount of spiked-in non-random missing data on imputation for the MNAR case. Error bars indicate 5-fold cross validation score ranges. [24]

II. Background

For a general background in data imputation part A will discuss the theory of missingness and part B will discuss state of the art methods in data imputation. MIM-based GAIN is an algorithm that was researched for one semester but ended up failing due to poor post-imputation performance. To describe the algorithm Part C will discuss why Message Importance Measure (MIM) is better than other information measures such as Shannon entropy [14] and Renyi entropy [15, 16].

A. Missingness

The concept of missingness is unknown to many people. The question is, why should we care about what values are missing or how they are missing? The reason is that understanding the missingness of a dataset can help us better understand how to tailor an algorithm towards that dataset. Consider multiple imputation methods used in data imputation [add citation]. Multiple imputation imputes values with a divide and conquer method by adding placeholders for all missing values in other features except one then does a regression to figure out the missing values on the held-out feature (this is one cycle). Typically, the placeholders are created using either of median or mean imputation which uses the mean or median of the observed values of a feature as the placeholder for that feature. No matter which imputation is used, however, it is assumed that each feature in the dataset can be predicted from the observed features in the dataset as a regression with a target and dataset is done to fill in missing values. The type of missingness that multiple imputation would call for is Missing at Random (MAR) where missing data depends only on the observations observed [18]. This makes sense as a large issue with multiple imputation is that if multiple features have many missing values, no matter how many cycles there is high uncertainty in the imputations. Hence observed data MUST be highly predictive of the feature and multiple imputation can only be used in the case of a dataset that has MAR missingness.

Other types of missingness include Missing Not at Random (MNAR) and Missing Completely at Random (MCAR). MNAR is when the missing data of a variable is dependent both on the observed and missing values of the variable [18]. Consider a drug survey. Applicants are allowed to leave areas blank where they don’t want to answer. The missing values here are not random but on purpose. Suppose all females did not answer a question but some females did not denote their sex in the survey. The missingness then becomes related to both observed and unobserved values. Hence the situation is MNAR. MCAR is completely different from both MNAR and MAR. MCAR is where the missing data of a variable is not related to other variables or itself [18]. As an example getting rid of 50% of the data randomly is an example of the generation of an MCAR dataset.

B. Data Imputation Methods

Data imputation is the task of imputing missing values with “real” values. Some data imputation algorithms that have been used consistently include median imputation, mean imputation, MICE [20], MissForest [21], and GAIN[2].

Median and mean imputation of a feature can be defined as using the median or mean of the observed values of the feature as a placeholder respectively. As mean and median imputation both rely on the fact that the missing data is related to the observed values median and mean imputation are both best for the MAR case.

Both MissForest and MICE use multiple imputation which, as discussed in A., is only useful for the MAR case of missingness.

GAIN leverages a Generative Adversarial Network (GAN). GANs are structured as follows: there exists a generator, which generates fake data, and a discriminator, which determines whether the data from the generator is real. The optimal case is when the generator starts outputting real values and the discriminator registers all the values as real. GAIN works by having the discriminator predict the mask matrix from the output of the generator where the mask matrix has zeroes in the places that data is missing (pictured in Fig.2. as the mask matrix of the original data). If the discriminator can determine which values correspond to 0 in the mask matrix, then it can learn the distribution of the missing values. The process is summarized in Fig. 2.

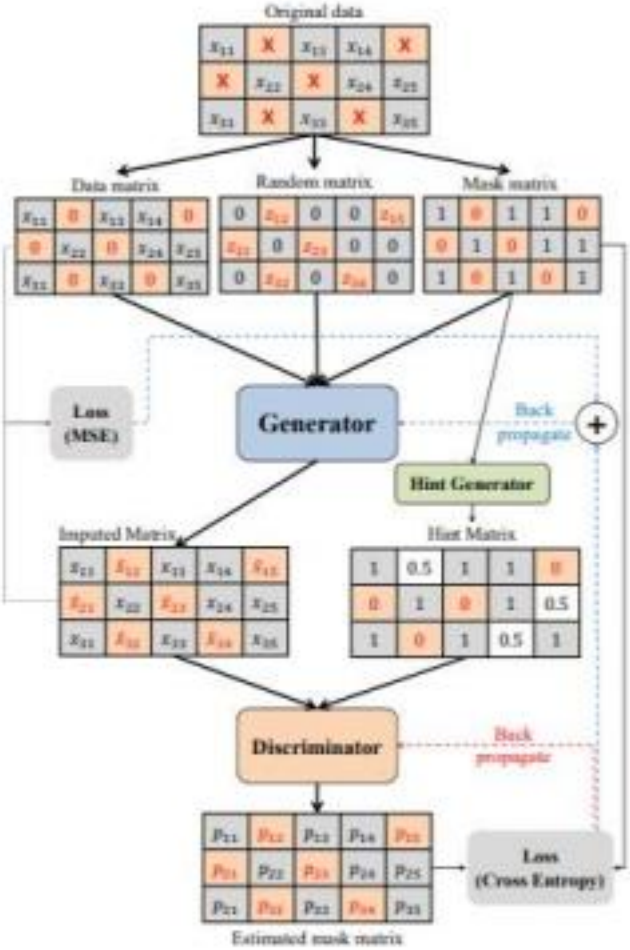


Fig. 2. Architecture of GAIN [?]

The mask matrix **m** can be defined as the original data with observed and missing values (where 'X' in Fig.2. corresponds to missing values and (x_{11}, x_{13}, ….x_{35}) are the observed values) being separated into a matrix with 1s where values are observed and 0s where values are missing. From Fig.2. we can see the generator G takes realizations of $\tilde{\mathbf{X}}$ (realization is the data matrix $\tilde{x}$), **M** (realization is the mask matrix **m)**, and a noise matrix **Z** (realization is random matrix **z)**, and outputs a matrix $\hat{x}$ as well as a hint matrix **h**. The matrix $\tilde{x}$ can be defined as

$$\hat{x} = \begin{cases} \tilde{x}, \tilde{x} \text{ observed} \\ G(\tilde{x}, m, (1 - m) \odot \quad z)\hat{x}, else \end{cases} \qquad (12)$$

and the hint matrix **h** can be considered a little bit of a cheat to nudge the discriminator to estimate the mask matrix **m**. “The hint mechanism **H** is a random variable from a space $\mathcal{H}$ dependent on M such that for an imputed sample ($\tilde{x}$, m) the hint matrix h is drawn from H|M=m “. In varying H, we can vary how much of a hint the matrix can give to the discriminator [2]. The reason why a hint matrix is needed is proven in [2] by Theorem 1, which asserts that if the hint matrix does not contain enough information about the mask matrix an optimal state is not guaranteed. Although GAIN does use a bit of a cheat to get a result that cheat is *necessary*.

The objective function of GAIN with discriminator D and generator G can be formulated as follows:

$$V(D, G) = E\hat{X}, M, H[M^{T} \log D(\hat{X}, H) + (1 - M)^{T} \log (1 - D(\hat{X}, H))] \qquad (13)$$

$$\min_{G} \min_{D} V(D, G) \qquad (14)$$

Additionally, we can also describe the loss function, which is formulated as

$$L(a, b) = \sum a_i \log(b_i) + (1\text{-}a_i)\log(1\text{-}b_i) \qquad (15)$$

And as D($\hat{X}$, H), the output of the discriminator, is the estimation of M, we can then rewrite (15) as

$$\min_{G} \min_{D} \mathbb{E}(L(M, \widehat{M})) \qquad (16)$$

A reason to use GAIN in MIM-GAIN is that GAIN performs very well, outperforming state of the art. Let us consider Fig.3.

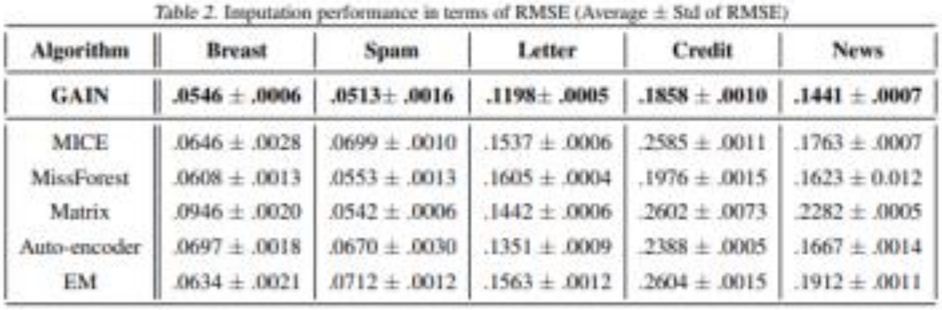


Table 2. Imputation performance in terms of RMSE (Average ± Std of RMSE)

Algorithm	Breast	Spam	Letter	Credit	News
GAIN	**.0546 ± .0006**	**.0513± .0016**	**.1198± .0005**	**.1858 ± .0010**	**.1441 ± .0007**
MICE	.0646 ± .0028	.0699 ± .0010	.1537 ± .0006	.2585 ± .0011	.1763 ± .0007
MissForest	.0608 ± .0013	.0553 ± .0013	.1605 ± .0004	.1976 ± .0015	.1623 ± 0.012
Matrix	.0946 ± .0020	.0542 ± .0006	.1442 ± .0006	.2602 ± .0073	.2282 ± .0005
Auto-encoder	.0697 ± .0018	.0670 ± .0030	.1351 ± .0009	.2388 ± .0005	.1667 ± .0014
EM	.0634 ± .0021	.0712 ± .0012	.1563 ± .0012	.2604 ± .0015	.1912 ± .0011

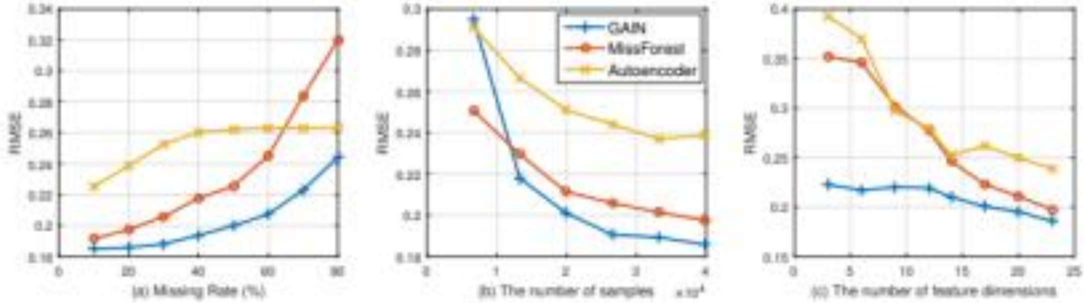


Fig.3. RMSE Performance on Different Settings (a) Various missing rates (b) Various numbers of samples (c) Various Feature Dimensions

Fig.3. shows GAIN's imputation performance on five different real-world datasets from UCI Machine Learning Repository [19]: Breast, Spam, Letter, and Credit. Concerning RMSE in Fig.3. Table 2, GAIN outperformed the other algorithms in all datasets and showed stability with standard deviation consistently less than or equal to 0.0016. There is however a case shown in Fig.3. in which GAIN did not outperform. This is in Fig.3. (b) where the number of samples was cut by 75% leaving only 25% of samples. In this case the autoencoder implementation outperformed. Overall, however, GAIN has shown itself to outperform state of the art and so has become state of the art.

C. MIM

Many algorithms that are involved in rare event detection are efficient for their respective applications, however the underlying issue in these algorithms is that they are based on information measures and frameworks which were originally designed for the processing of typical events. Commonly used information measures in this space that fit this criteria are Shannon [14] and Renyi [15,16] entropy. As an indicator of the commonality several measures stemming from these measures includes Kullback-Liebler (KL) divergence, f-divergence, Renyi divergence, and Fisher information [16,17]. The popular loss, cross entropy, stems from KL divergence.

Consider the following formulations of Shannon and Renyi entropy:

For a given probability distribution $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ the Shannon entropy *H(**p**)* can be defined as:

$$\mathrm{H}(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log(p_i) \qquad (1)$$

Which measures the uncertainty or the information that the distribution contains. On the other hand, for a given probability distribution $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ the Renyi entropy $\mathrm{H}_\alpha(\mathbf{p})$ can be defined as:

$$\mathrm{H}_\alpha(\mathbf{p}) = \frac{1}{1-\alpha} \sum_{i=1}^{n} p_i^\alpha \qquad (2)$$

for $0 < \alpha < \infty,\ \alpha \neq 1$ where, as α approaches 1, $\mathrm{H}_\alpha(\mathbf{p})$ converges to Shannon entropy H(**p**). Additionally, some other interesting properties of Renyi entropy are as follows: if we take the limits as $\alpha \rightarrow 0$ and $\alpha \rightarrow$ infinity for $\mathrm{H}_\alpha(\mathbf{p})$, letting n denote the length of the probability distribution **p**, and letting the uniform distribution **u** = $(\frac{\mathbf{1}}{\boldsymbol{n}}, \frac{\mathbf{1}}{\boldsymbol{n}}, \ldots. \frac{\mathbf{1}}{\boldsymbol{n}})$ we receive:

$$\lim_{\alpha \to 0} \mathrm{H}_\alpha(\mathbf{p}) = \log \mathrm{n} = \mathrm{H}(\mathbf{u}) \qquad (3)$$

$$\lim_{\alpha \to \infty} \mathrm{H}_\alpha(\mathbf{p}) = \min_i(-\log(p_i)) = -(\max_i \log(p_i)) = -\log(\max_i p_i) \qquad (4)$$

Both (3) and (4) show that Renyi entropy is bounded above by H(**u**) and that even when α is as low as possible the measure still does not put high importance on less typical events. Either all events have equal probability as α approaches 0 or the measure is leaning towards more typical events to have more importance as α increases. Similarly, as $\mathrm{H}_\alpha(\mathbf{p})$ converges to Shannon entropy when α = 1 Shannon entropy also puts more importance on typical events and places an upper bound on $\mathrm{H}(\mathbf{u})$. This can be summarized as follows:

$$\mathrm{H}(\mathbf{p}) \leq \mathrm{H}(\mathbf{u}) \tag{5}$$

$$\mathrm{H}_\alpha(\mathbf{p}) \leq \mathrm{H}_\alpha(\mathbf{u}) \tag{6}$$

MIM, on the other hand, has a special property that allows the measure to supersede the uniform distribution. Consider the same probability distribution $\boldsymbol{p} = (p_1, p_2, \dots, p_n)$. MIM with importance coefficient ϖ can be formulated as:

$$L(\boldsymbol{p}, \varpi) = L_\varpi(\boldsymbol{p}) = \log \sum_{i=1}^{n} p_i \exp\,(\varpi(1 - p_i)) \tag{7}$$

In "Message Importance Measure and Its Application to Minority Subset Detection in Big Data" by Pingyi Fan er al. [11] one of the key findings is Lemma 3, which will now be restated:

Lemma 3 : For a probability distribution $\boldsymbol{p} = (p_1, p_2, \dots, p_n)$ with 0 < $\min\limits_i p_i$, there exists a ϖ_0> 0 such that for $\varpi \geq \varpi_0$, the parametric message importance measure satisfies

$$\mathrm{L}(\mathrm{p}\,\varpi) = \log \sum_{i=1}^{n} p_i \exp\,(\,\varpi\,(1 - p_i)) > \mathrm{L}(\mathrm{u}, \varpi) \tag{8}$$

where **u** = $\left(\frac{\mathbf{1}}{\boldsymbol{n}}, \frac{\mathbf{1}}{\boldsymbol{n}}, \dots .\frac{\mathbf{1}}{\boldsymbol{n}}\right)$ is the uniform distribution. When the importance coefficient is sufficient large, we further have

$$\mathrm{L}(\mathrm{p}, \varpi)\ .\!= \varpi\,(1 - \min_i p_i) + \log(\min_i p_i) \tag{9}$$

Based on Lemma 3 with a sufficiently large importance coefficient we can have the measure giving highest importance to the rarest events (those with lowest probability) as shown by (9). This brings MIM to be the exact opposite of Renyi entropy where if, for Renyi entropy, $\alpha \to \infty$ the events with maximum probability will be given the highest importance. The property given by Lemma 3 allows MIM to place more importance on rarer events than typical events such that there exists importance coefficient ϖ that strikes a balance between importance on rare events and typical events. For an understanding of this idea consider weighted cross entropy loss. Cross entropy does not perform well in the case of class imbalance, so each class has to be weighted differently based on their probability. That idea is not too far from what we are trying to achieve here.

Another point that the paper talks about is the lower bound of MIM such that (8) applies. In restating theorem 1:

Theorem 1 : Given a probability distribution $\boldsymbol{p} = (p_1, p_2, \dots, p_n)$ without zero elements, if ϖ satisfies

$$\varpi \geq \frac{\log(\min_i p_i)}{\frac{1}{n} - \min_i p_i)} \tag{10}$$

then we have

$$\mathrm{L}(\mathbf{p}_{,}\varpi) > \mathrm{L}(\mathbf{u}, \varpi) \tag{11}$$

Where $\mathbf{u} = (\frac{\mathbf{1}}{\boldsymbol{n}}, \frac{\mathbf{1}}{\boldsymbol{n}}, \dots \frac{\mathbf{1}}{\boldsymbol{n}})$ is the uniform distribution where n >2.

One successful application of MIM has been towards anomaly detection where the goal is to detect outliers in the data. In particular MIM-GAN (Generative Adversarial Network) [12] has shown state of the art performance as shown in Fig. 1.

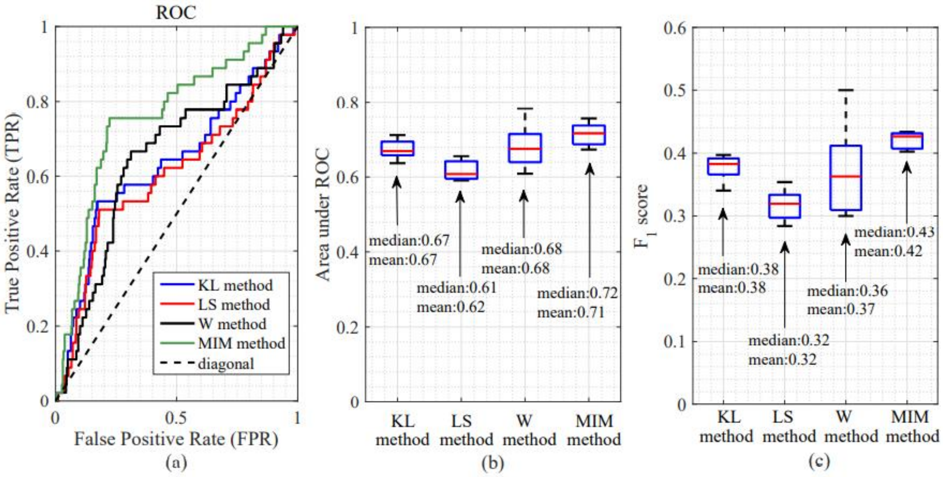
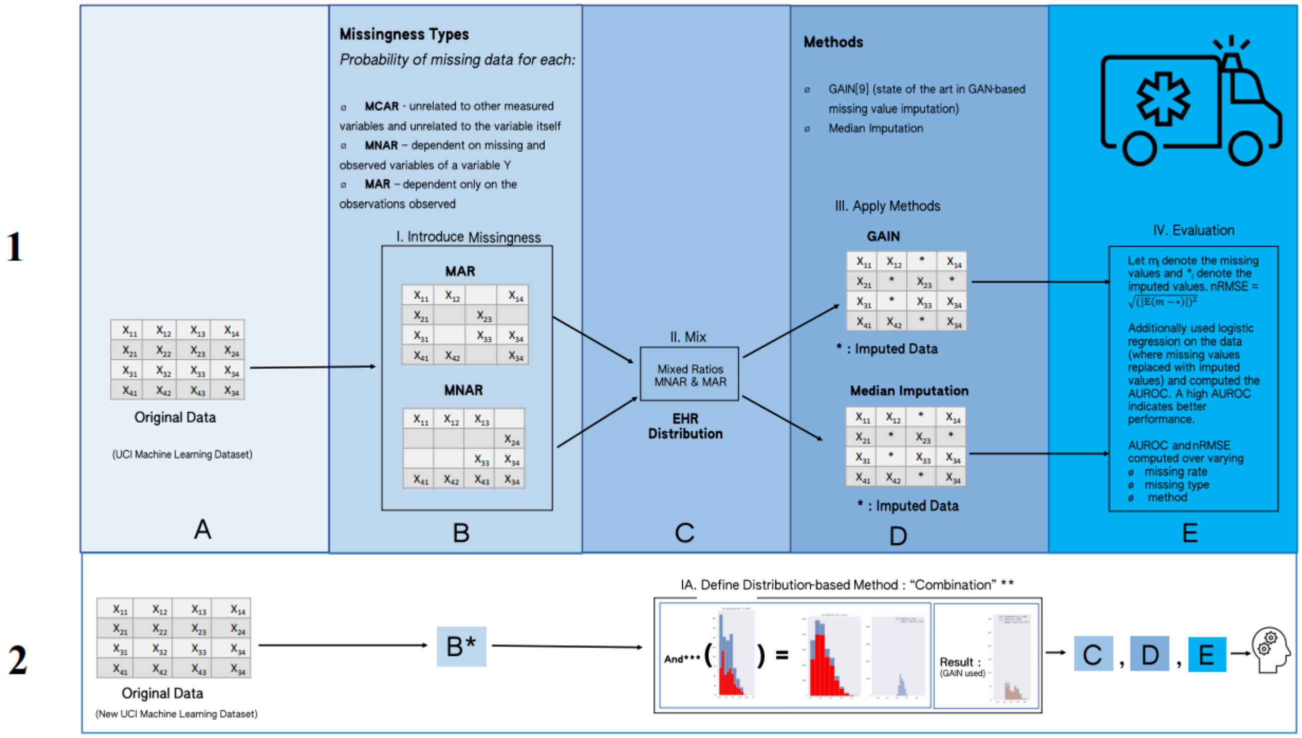


Fig.1. ROC curve, AUC, and F1-score for GAN-based Anomaly Detection in the Cardiotocography Dataset[12]

Fig.1. is showing results for when , in [12], MIM-GAN was used for an anomaly detection task on the Cardiotocography dataset which has 21 features in each sample, such as Fetal Heart Rate (FHR) and Uterine Contraction (UC) features. 9.6% data belongs to the pathologic class, namely the anomalous events class. In Fig.1. we can see that the ROC curve of the MIM method is farther from the dotted line than the other methods, indicating better performance. In addition, in both AUC and F1-score the other methods in Fig.1. were outperformed in terms of all of median, mean and stability (narrow interquartile range (IQR)). MIM not only has theoretically been shown to detect rare events but also has been shown to detect rare events in application.

IV. Project Description



Missingness Types

Probability of missing data for each:

- **MCAR** - unrelated to other measured variables and unrelated to the variable itself
- **MNAR** – dependent on missing and observed variables of a variable Y
- **MAR** – dependent only on the observations observed

Methods

- GAIN[9] (state of the art in GAN-based missing value imputation)
- Median Imputation

Let m_i denote the missing values and $*_i$ denote the imputed values. nRMSE = $\sqrt{(|\mathbb{E}(m-*)|)^2}$

Additionally used logistic regression on the data (where missing values replaced with imputed values) and computed the AUROC. A high AUROC indicates better performance.

AUROC and nRMSE computed over varying

- missing rate
- missing type
- method

Fig.2. Project Description Summary

The project description for this project can be summarized in Fig.2. Let us start with **1**, the first cycle. In **1**A we load in a dataset from UCI Machine Learning [19], a website that houses free datasets. After **1**A we move to **1**B where we get rid of a set amount if values of a certain missing rate using a missing type. In this case we use both the MNAR and MAR case then mix the two in **1**C since EHR distributions are made up of either or both of MNAR and MAR. Example splits include 80% MNAR, 20% MAR, 50% MNAR, 50% MAR, etc. After we have our dataset with missing and observed values, we apply both GAIN and median imputation in **1**D. In **1**E nRMSE is calculated where nRMSE is defined as the RMSE between the missing values and known values. For this case we will be doing nRMSE feature wise to figure out which method works best for which distribution. Note that by distribution we mean the distributions of the features in the dataset we created in **1**C. In the real world we don't know the missing values. The other evaluation is for AUROC. Since we have the original dataset we know the targets so we compute linear regression post-imputation and assess AUROC to understand how well the dataset can now be used for secondary use. Let us move to **2.**

2A is the same as **1**A however we use a different dataset. **2**B uses the same split and missing rate as **1**B to apply missingness. After **2**B we declare the new method "combination". Using the Anderson test (denoted in Fig.2. as And) we can compare the distributions between that of **2**C and **1**C. If a distribution maps to that of **1**C then whatever algorithm performed best on the distribution in **1**C is used. **2**C,D,E function the same as their counterparts in **1.**

The datasets that I use for testing are the UCI Machine Learning Datasets [19] : credit card and messidor. The reason for this choice is that median imputation outperforms GAIN in credit card and GAIN outperforms median imputation in messidor so there exists some range.

V. Results

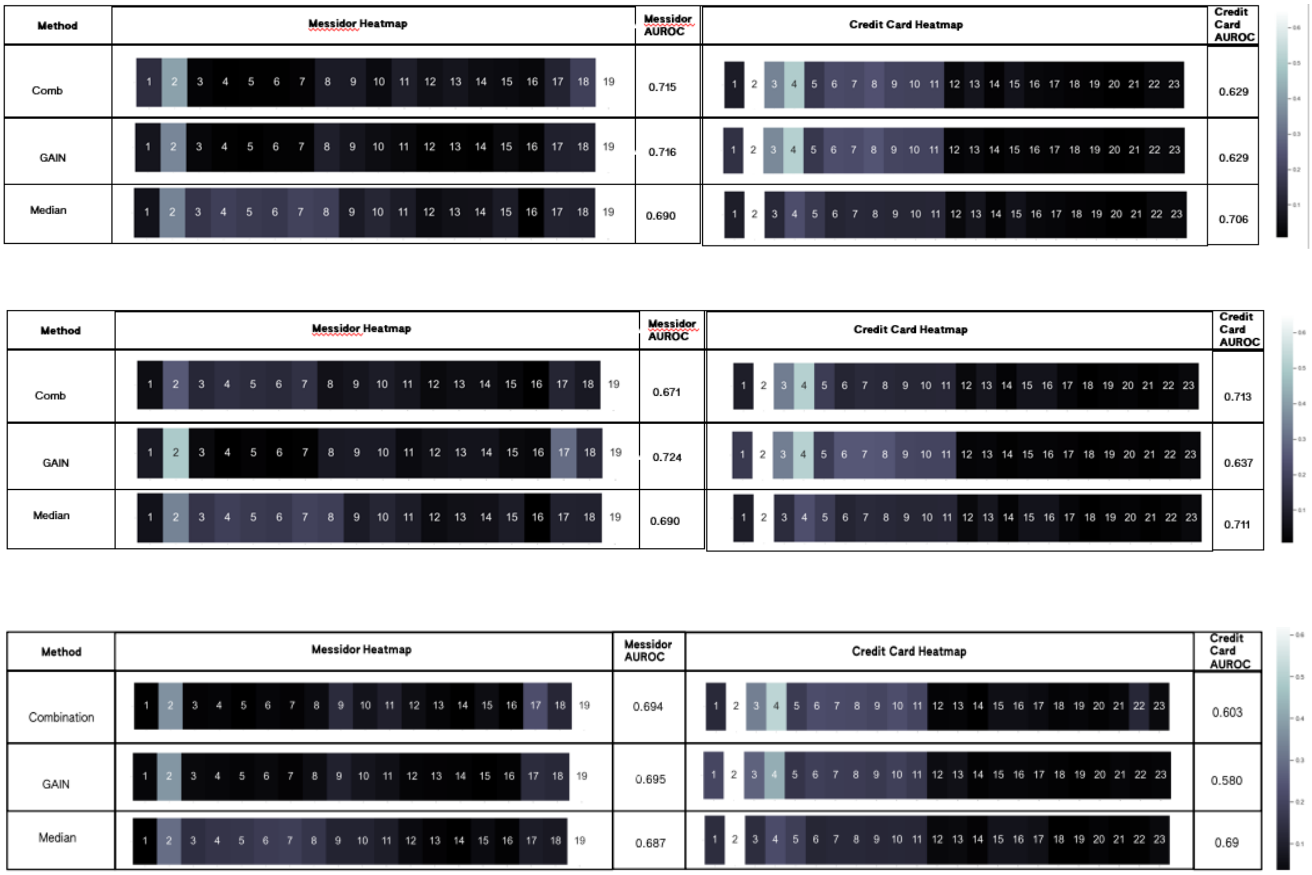


Fig.3. Featurewise Heatmaps of nRMSE and total AUROC over Messidor and Credit Card w/ 80% missing rate and varying missing type: (**Top**) 20% MAR, 80% MNAR, (**Middle**) 50% MAR, 50% MNAR, (**Bottom**) 80% MAR, 20% MNAR

In looking across datasets in Fig.3. 5 out of the 6 of the cases for the combination method was consistently never outperformed by both methods. The idea that the combination method is just taking advantage of the idea that median imputation works better over a skewed distribution could be possible but in looking at the Messidor heatmap combination taking on values from GAIN (which is the outperforming method here) and never getting outperformed by both methods in this dataset that idea is refuted. Skewed distributions could also work better for GAIN. Consider Fig.4.

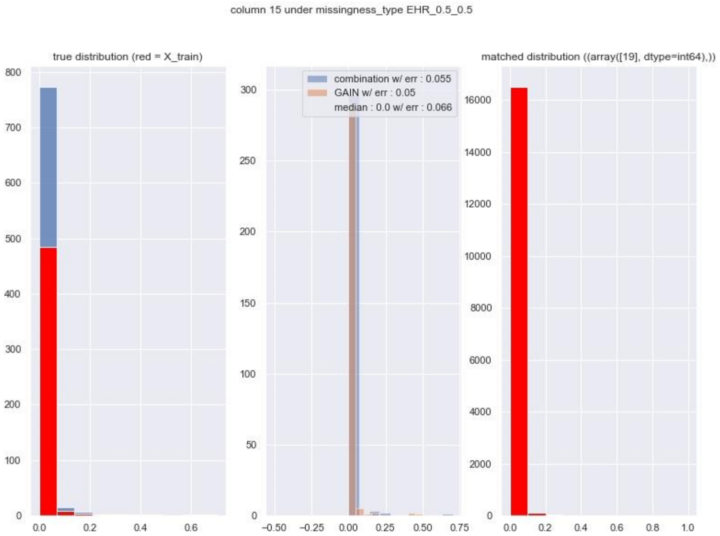


Fig.4. column 15 of Messidor with 80% missing rate, 50% MNAR. 50% MAR : (**Left**) True distribution (**Middle**) Results (**Right**) Mapped Distribution

Hence distribution has been shown to play a role in data imputation and not just for median (in the case of a skewed distribution) or mean (in the case of a normal distribution). GAIN also outperforms dependent on certain distributions. There may exist certain bounds.

VI. Conclusion

The distribution-based method “combination” is consistent across datasets but has its faults when there does not exist two similar distributions. Future work could be to add more methods (such as mean) although not too many. Adding too many methods will increase computational time. Another subject for future work is to implement a distribution search API for UCI Machine Learning [19]. If imputation is dependent on distributions then finding the same distribution online via an API would guarantee success and consistency. The process of searching through an API, however, may be time-consuming so this case would be more so for if the authors know what the important predictors are and only want to target those for distribution-based implementation. Lastly, one more subject of future work could be to extend [1] to improve endpoint prediction on myocardial infarction.

Works Cited

[1] D.A. Kaji, J.R. Zech, J.S. Kim, *et al.An attention based deep learning model of clinical events in the intensive care unit.* PLoS One, 14 (2) (2019),Article e0211057. https://www.ncbi.nlm.nih.gov/pubmed/30759094

[2] Jinsung Yoon and James Jordon and Mihaela van der Schaar. *GAIN: Missing Data Imputation using Generative Adversarial Nets. 2018.* GAIN: Missing Data Imputation using Generative Adversarial Nets (arxiv.org)

[3] Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* **13**, 395–405 (2012).

[4] Weiskopf, N. G., Hripcsak, G., Swaminathan, S. & Weng, C. Defining and measuring completeness of electronic health records for secondary use. *J. Biomed. Inform.* **46**, 830–836 (2013).

[5] Weiskopf, N. G. & Weng, C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J. Am. Med. Inform. Assoc*. **20**, 144–151 (2013).

[6] Giraud T, Dhainaut JF, Vaxelaire JF, Joseph T, Journois D, Bleichner G, Sollet JP, Chevret S, Monsallier JF. Iatrogenic complications in adult intensive care units: a prospective two-center study. *Crit Care Med.* 1993;21:40–51. doi: 10.1097/00003246-199301000-00011.

[7]0 Bracco D, Favre JB, Bissonnette B, Wasserfallen JB, Revelly JP, Ravussin P, Chiolero R. Human errors in a multidisciplinary intensive care unit: a 1-year prospective study. *Intensive Care Med.* 2001;27:137–145. doi: 10.1007/s001340000751.

[8] Abramson NS, Wald KS, Grenvik AN, Robinson D, Snyder JV. Adverse occurrences in intensive care units. *JAMA.* 1980;244:1582–1584. doi: 10.1001/jama.1980.03310140040027.

[9] Osugi J, Owada Y, Yamaura T, Muto S, Okabe N, Matsumura Y, Higuchi M, Suzuki H, Gotoh M. "Successful Management of Crizotinib-Induced Neutropenia in a Patient with Anaplastic Lymphoma Kinase-Positive Non-Small Cell Lung Cancer: A Case Report. "Case Rep Oncol. 2016 Jan 15;9(1):51-5. doi: 10.1159/000443662. PMID: 26933419; PMCID: PMC4748772.

[10] Enders, C. K. (2010), Applied Missing Data Analysis, The Guilford Press.

[11] P. Fan, Y. Dong, J. X. Lu, and S. Y. Liu, "Message importance measure and its application to minority subset detection in big data", in Proc. IEEE Globecom Workshops (GC Wkshps 2016), Washington D.C., USA, Dec. 4–8, 2016, pp. 1–6.

[12] Rui She and Pingyi Fan. *MIM-Based GAN: Information Metric to Amplify Small Probability Events Importance in Generative Adversarial Networks.* 2021. 2003.11285.pdf (arxiv.org)

[13] Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035 10.1038/sdata.2016.35

[14] C. E. Shannon, "A mathematical theory of Communicaiton, " The Bell Syst. Tech. J., Vol.27, pp.379-423, 623-656, July–Oct. 1948.

[15] T. M. Cover and J. A. Thomas, Elements of Information Theory 2nd Edition, Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, 2006.

[16] T. V. Erven, P. Harremoes "Renyi divergence and Kullback Leibler divergence," IEEE

Trans. on Inf. Theory, vol. 60, no. 7, pp. 3797–3820, 2014.

[17] H. Akaike, B. N. Petrov, and F. Caski, “Information theory and an extension of the maximum likelihood principle,” in Proc. IEEE Int. Symp. Inf. Theory,(ISIT) Budapest, Hungary, 276– 281, 1973.

[18] Rubin, D. B. (1976), ‘Inference and Missing Data’, Biometrika 63(3), 581– 592. [19]

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

[20] Van Buuren, S.; and Groothuis-Oudshoorn, K. 2011. mice: Multivariate imputation by chained equations in R. Journal of statistical software 45: 1–67.

[21] Daniel J. Stekhoven, Peter Bühlmann, MissForest—non-parametric missing value imputation for mixed-type data, *Bioinformatics*, Volume 28, Issue 1, 1 January 2012, Pages 112–118,

[22] Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with de noising autoencoders. In Proceedings of the 25th international conference on Machine learning, 1096–1103

[23] Ramon Vinas and Xu Zheng and Jer Hayes. *A Graph-based Imputation Method for Sparse Medical Records.* 2021. [2111.09084] A Graph-based Imputation Method for Sparse Medical Records (arxiv.org)

[24] Beaulieu-Jones, Brett K, and Jason H Moore. “MISSING DATA IMPUTATION IN THE ELECTRONIC HEALTH RECORD USING DEEPLY LEARNED AUTOENCODERS.” *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* vol. 22 (2017): 207-218. doi:10.1142/9789813207813_0021

[25] Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; and Altman, R. B. 2001. Missing value estimation methods for DNA microar rays. Bioinformatics 17(6): 520–52

[26] Nelwamondo FV, Mohamed S, Marwala T. [Accessed September 30, 2016];*Missing Data: A Comparison of Neural Network and Expectation Maximisation Techniques.* 2007 Apr;

[27] Gelman A, Hill J. [Accessed August 10, 2016];*Data Analysis Using Regression and Multilevel/hierarchical Models.* 2006