**MURAL: An Unsupervised Random Forest-Based Embedding for Electronic Health Record Data**

Michal Gerasimiuk*
Dept. of Comp. Sci.
Yale University
New Haven, CT, USA
michal.gerasimiuk@yale.edu

Dennis Shung*
Sch. of Medicine
Yale University
New Haven, CT, USA
dennis.shung@yale.edu

Alexander Tong
Dept. of CS & Oper. Res.
Univ. de Montréal; Mila
Montreal, QC, Canada
alexander.tong@mila.quebec

Adrian Stanley
Glasgow Royal Infirmary
Glasgow G4 0SF, United Kingdom
adrianstanley99@gmail.com

Michael Schultze
Department of Medicine
Otago Med. Sch. — Dunedin Campus
Dunedin, New Zealand
michael.schultze@otago.ac.nz

Jeffrey Ngu
Christchurch Hospital
Christchurch, New Zealand
jeffrey.ngu@cdhb.health.nz

Loren Laine
Sch. of Medicine
Yale University
New Haven, CT, USA
loren.laine@yale.edu

Guy Wolf
Dept. of Math. and Stat.
Univ. de Montréal; Mila
Montreal, QC, Canada
guy.wolf@umontreal.ca

Smita Krishnaswamy†
Depts. of Genetics & Comp. Sci.
Yale University
New Haven, CT, USA
smita.krishnaswamy@yale.edu

Abstract—A major challenge in embedding or visualizing clinical patient data is the heterogeneity of variable types including continuous lab values, categorical diagnostic codes, as well as missing or incomplete data. In particular, in EHR data, some variables are missing not at random (MNAR) but deliberately not collected and thus are a source of information. For example, lab tests may be deemed necessary for some patients on the basis of suspected diagnosis, but not for others. Here we present the MURAL forest – an unsupervised random forest for representing data with disparate variable types (e.g., categorical, continuous, MNAR). MURAL forests consist of a set of decision trees where node-splitting variables are chosen at random, such that the marginal entropy of all other variables is minimized by the split. This allows us to also split on MNAR variables and discrete variables in a way that is consistent with the continuous variables. The end goal is to learn the MURAL embedding of patients using average tree distances between those patients. These distances can be fed to nonlinear dimensionality reduction method like PHATE to derive visualizable embeddings. While such methods are ubiquitous in continuous-valued datasets (like single cell RNA-sequencing) they have not been used extensively in mixed variable data. We showcase the use of our method on one artificial and two clinical datasets. We show that using our approach, we can visualize and classify data more accurately than competing approaches. Finally, we show that MURAL can also be used to compare cohorts of patients via the recently proposed tree-sliced Wasserstein distances.

Index Terms—data analysis, electronic medical records, random forests, unsupervised learning

I. INTRODUCTION

Unsupervised nonlinear embedding methods have allowed for exploration manifold learning of big high dimensional datasets in many fields ranging from epidemiology, to biology, to physics. However, a major limitation of using unsupervised embeddings in healthcare data is the large amount of missingness in the data as well as the mixed modality of the variables collected. In a typical EHR or patient dataset the range of missing data range from 20% to 80%, varying across broad categories of possible fields such as demographics, laboratory values, and treatment information [1], [2], [3]. Further there is a mix of real-valued, categorical and binary data which can be difficult to normalize or scale. This makes it difficult to compute distances and affinities between datapoints—the first step in nonlinear dimensionality reduction methods such as tSNE [4], UMAP [5], diffusion maps [6] or PHATE [7].

Similar distance/affinity computations are also required for spectral clustering [8], which operates on a graph Laplacian computed from the affinity matrix. Thus data with missing values cannot be used, and if the values are MNAR they cannot be imputed.

To tackle these issues, we propose to use an intermediary representation called the MURAL-forest, an unsupervised random forest in which tree distances between datapoints form

* Equal contribution. † Corresponding author. This research was partially supported by IVADO Professor funds, CIFAR AI Chair, and NSERC Discovery grant 03267 [G.W.]; Chan-Zuckerberg Initiative grants 182702 & CZF2019-002440 [S.K.]; and NIH grants R01GM135929 & R01GM130847 [G.W., S.K.], K23DK125718-01A1 [D.S.]. The content provided here is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

978-1-6654-3902-2/21/$31.00 ©2021 IEEE
an accurate measure of dissimilarity and can be used for data distance/affinity computation, as needed in [4], [7], [3], [6]. MURAL creates a set of trees by splitting on any variable type (categorical, continuous, with or without missingness) using a marginal entropy criterion that is computed on other variables. Further, MURAL ensures that heterogeneity within categorical or MNAR variables is immediately broken down using low dimensional entropy to create 4-way splits at such levels. We test MURAL on ground truth data that the resulting tree distances result in accurate embeddings.

While random forests are normally supervised and trained for prediction, there have been some efforts to learn random forests in an unsupervised manner. [9] describes a method called manifold forests which effectively uses a splitting criterion based on intra-versus-inter split affinity or density. However, these and other methods often presuppose the ability to compute distances or affinities between high dimensional datapoints. By contrast, we use our MURAL unsupervised random forests in order to be able to compute an accurate distance between datapoints with missing and mixed-mode variables.

We show the accuracy of our method by comparing the MURAL derived distances to known ground truth and recovering embeddings in a 5-dimensional Swiss roll. We then apply our method to a complete case subset of an intensive care unit dataset and of an international patient registry dataset of patients presenting with symptoms of upper gastrointestinal bleeding. We induce missingness in the complete case subsets in specific ranges of laboratory values and compare imputed values using mean imputation and multiple imputation with chained equations to the original ground truth. We then construct MURAL-embeddings on the full datasets with missingness. We show that MURAL-embeddings consistently display more structure and create separations that are more clinically meaningful than commonly used imputation methods. Finally, we show an application of our method in comparing entire cohorts of patients by computing a tree-based Wasserstein distance on the MURAL-forest, which can be used to quantify similarities or distances between patient cohorts.

II. BACKGROUND

A. Manifold Learning, Dimensionality Reduction, Clustering

Though there are many nonlinear dimensionality reduction and embedding methods, we focus our results on methods that can learn the data manifold or intrinsic low dimensional shape and structure of the data. We believe that this is useful in biomedical settings where many measurements of the patient reflect non-orthogonal aspects of the same underlying entity, essentially indicating the data in fact lies in a lower dimensional space.

High dimensional data can often be modeled as a sampling \( Z = \{z_i\}_{i=1}^N \subset \mathcal{M}^d \) of a \( d \) dimensional manifold \( \mathcal{M}^d \) that is mapped to \( n \) dimensional observations \( X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^n \) via a nonlinear function \( x_i = f(z_i) \). Intuitively, although measurement strategies, modeled here via \( f \), create high dimensional observations, the intrinsic dimensionality, or degrees of freedom within the data, is relatively low. This manifold assumption is at the core of the vast field of manifold learning [10], [6], [11], [12], which leverages the intrinsic geometry of data, as modeled by a manifold, for exploring and understanding patterns, trends, and structure that displays significant nonlinearity.

In [6], diffusion maps were proposed as a robust way to capture intrinsic manifold geometry in data by eigendecomposing a powered diffusion operator. Using \( t \)-step random walks that aggregate local affinity, [6] were able to reveal nonlinear relations in data and allow their embedding in low dimensional coordinates. These local affinities are commonly constructed using a Gaussian kernel:

\[
K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\varepsilon}\right), \quad i, j = 1, \ldots, N \quad (1)
\]

where \( K \) forms an \( N \times N \) Gram matrix whose \((i, j)\) entry is denoted by \( K(x_i, x_j) \). A diffusion operator is defined as the row-stochastic matrix \( P = D^{-1}K \) where \( D \) is a diagonal matrix with \( D(x_i, x_i) = \sum_j K(x_i, x_j) \). The matrix \( P \), or diffusion operator, defines single-step transition probabilities for a time-homogeneous diffusion process, or a Markovian random walk, over the data. Furthermore, as shown in [6], powers of this matrix \( P^t \), for \( t > 0 \), can be used to simulate multi-step random walks over the data, helping understand multiscale organization of \( X \), which can be interpreted geometrically when the manifold assumption is satisfied. \( P \) has been used in many downstream unsupervised learning tasks, eigendecomposition of \( P \) yields the popular diffusion map dimensionality reduction method that can be used as input to clustering. \( P \) is also used by the PHATE [7] method for visualization. PHATE transforms the diffusion operator with a pointwise logarithm \( \log(P) \), derives distances between points \( x_i, x_j \) as \( \|\log(P) - \log(P)\|_2 \), and then embeds the resulting distances, known as potential distances, with metric MDS.

Other methods for visualization, such as [4], [5], use \( K \) rather than \( P \) to focus on near neighbors rather than learning the entire data manifold. The diffusion operator \( P \) is related to the graph Laplacian that, depending on the normalization used, can be written as \( L = I - K \) or \( L = I - P \). Thus the graph Laplacian has the same eigenvectors and eigenvalues that are in the opposite order. Spectral clustering [8] is often described in terms of the graph Laplacian, i.e., \( k \)-means over a graph Laplacian rather than data.

B. Decision Trees

A tree \( T \) is a connected directed acyclic graph \( T = (V, E) \) with vertices (or nodes) \( V = \{t_1, t_2, \ldots, t_n\} \) and \( n - 1 \) edges \( E \) such that every node has at most one incoming edge. A rooted tree has a root node \( t_1 \) with no incoming edges, while \( t_i, i > 1, \) all have exactly one incoming edge. A node \( t_j \in \text{children}(t_i) \) if and only if \([t_i, t_j] \in E\), i.e., there is a directed edge from \( t_i \) to \( t_j \). A descendant(\( t_i \)) is any node \( t_k \) that is connected to \( t_i \) by a directed path \( t_1, \ldots, t_k \) with a directed edge between each consecutive pair of nodes.
Decision trees contain nodes that split on a variable to create partitions of the data such that datapoints on one side of the partition are more similar to each other in terms of the decision variable. Recursive splits create finer granularity branches where data points are similar with respect to all of the variables that have been split on the path to the node. A specific strength of decision tree is the ability to naturally split multiple types of data—binary, ordinal, and missing.

\[ P \text{ is given by} \]

\[ \text{and child nodes. Shannon entropy of a probability distribution} \]

\[ \text{the difference in Shannon entropy between the parent node} \]

\[ \text{gain is also an information theoretic measure which measures} \]

\[ \text{parent node is} \]

\[ k \]

\[ | \]

\[ I \]

\[ \text{impurity index is given by} \]

\[ \text{affinities can be defined between datapoints. However, for} \]

\[ \text{the data variables are continuous and distances or Gaussian} \]

\[ \text{direction trees} \]

\[ \text{density forests} \]

\[ \text{PCA trees} \]

\[ \text{approximate principal} \]

\[ \text{in the absence of labels: random projection trees} \]

\[ \text{variants of decision trees have been used to cluster data} \]

\[ \text{Decision trees contain nodes that split on a variable to} \]

\[ \text{to create partitions of the data such that datapoints on one side} \]

\[ \text{the subset is considered pure.} \]

\[ \text{chooses a splitting variable and threshold. These weak learners} \]

\[ \text{minimize an impurity function to ensure that each child node} \]

\[ \text{a threshold so as to optimize a local criterion such as the} \]

\[ \text{Gini impurity index or information gain. The Gini impurity} \]

\[ \text{is an information theoretic measure that is based on} \]

\[ \text{information gain is} \]

\[ P \]

\[ \text{is the class distribution of the parent node, and} \]

\[ P^a \]

\[ \text{the class distribution of the} \]

\[ a \]

\[ \text{data points split by the} \]

\[ \text{total number of data points split by the} \]

\[ a \]

\[ \text{number of data points from} \]

\[ \text{node is} \]

\[ k \]

\[ \text{Note that these criteria are with respect to a} \]

\[ \text{classification label that is given in a supervised setting.} \]

\[ \text{The original random forest classifier used labeled data to} \]

\[ \text{randomly train an ensemble of decision trees with a majority} \]

\[ \text{vote aggregating the classifications. Decision trees are} \]

\[ \text{constructed through recursively partitioning the space occupied} \]

\[ \text{by data as observations travel from the tree’s root to its} \]

\[ \text{leaves, each nonterminal node containing a weak learner that} \]

\[ \text{chooses a splitting variable and threshold. These weak learners} \]

\[ \text{minimize an impurity function to ensure that each child node} \]

\[ \text{receives a “purer” cohort than its parent. Purity is determined} \]

\[ \text{by the proportion of labels; if all examples belong to the same} \]

\[ \text{class, the subset is considered pure.} \]

\[ \text{D. Unsupervised Random Forests} \]

\[ \text{Variants of decision trees have been used to cluster data} \]

\[ \text{in the absence of labels: random projection trees} \]

\[ \text{density forests} \]

\[ \text{PCA trees} \]

\[ \text{approximate principal} \]

\[ \text{in the absence of labels: random projection trees} \]

\[ \text{previous work in demonstrated unsupervised forest construc-} \]

\[ \text{Wasserstein Distance over Trees} \]

\[ \text{The 1-Wasserstein distance (also known as the earth mover’s} \]

\[ \text{distance} \]

\[ \text{measures the total cost of moving shifting the mass from} \]

\[ \text{one probability distribution to another. For discrete} \]

\[ \text{probability distributions over a general metric space this can} \]

\[ \text{be computed exactly in} \]

\[ O(n^3) \]

\[ \text{time using the Hungarian algorithm} \]

\[ O(n^2) \]

\[ \text{time using the Hungarian} \]

\[ \text{approximated using entropy regularization in} \]

\[ O(n^2) \]

\[ \text{However, for discrete probability distributions over a tree metric} \]

\[ \text{space the 1-Wasserstein distance can be computed exactly in} \]

\[ \text{linear time} \]

\[ \text{Given two probability distributions} \]

\[ \mu, \nu \]

\[ \text{over a measurable space} \]

\[ \Omega \]

\[ \mu, \nu \]

\[ \text{have bounded support, then the dual} \]

\[ \Pi(\mu, \nu) \]

\[ \text{is the set of joint probability distributions} \]

\[ \pi \]

\[ \Omega \]

\[ \Omega \times \Omega \]

\[ \text{The 1-Wasserstein} \]

\[ \text{between} \]

\[ \mu \]

\[ \nu \]

\[ \text{is defined as:} \]

\[ W_p(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int_{\Omega \times \Omega} \rho(x, y)\pi(dx, dy). \]

\[ \text{Let} \]

\[ \| \cdot \|_{L_p} \]

\[ \rho \]

\[ \mu, \nu \]

\[ \text{have bounded support, then the dual of equation} \]

\[ \text{known as the Kantorovich-Rubinstein dual, can be} \]

\[ \text{expressed as:} \]

\[ W_p(\mu, \nu) = \sup_{\| f \|_{L_p} \leq 1} \int_{\Omega} f(x)\mu(dx) - \int_{\Omega} f(y)\nu(dy). \]

\[ \text{When} \]

\[ \text{is a tree metric over a rooted tree} \]

\[ T \]

\[ \text{for every pair of points} \]

\[ x, y \]

\[ \mu \]

\[ \nu \]

\[ \text{is the total weight of the (unique) path between nodes} \]

\[ x \]

\[ y \]

\[ T \]

\[ \text{Denote the edge weight associated with each node} \]

\[ t \]

\[ \omega \]

\[ D(t, \mu) \]

\[ D(t, \nu) \]

\[ \text{the Wasserstein distance between two distributions on} \]

\[ T \]

\[ \text{can be expressed as:} \]

\[ W_{yp}(\mu, \nu) = \sum_{t \in T} w_t |D(t, \mu) - D(t, \nu)|. \]
MURAL, we construct an unsupervised random forest over a high dimensional Ω that consists of continuous, categorical and missing variables. These trees subsequently define a distance on Ω, which in turn defines a Wasserstein distance between distributions on Ω, and because of the specific construction of MURAL trees, admits a simple feature importance measure described in [IVC].

III. MURAL

Next we present the MURAL algorithm for building unsupervised random forests from continuous and categorical data with missing values on healthcare data. Our code is available at https://github.com/mgerasimituk/mural.

A. Problem Formulation

Our goal is to build a distance matrix $D$ whose $(i,j)$-th entry contains the distance $d(x_i,x_j)$ between observations $x_i$ and $x_j$. Desirable properties for $D$ are that neighbors found using $D$ have similar clinical manifestations, and moreover that $D$ can be used in a nonlinear dimensionality reduction method to create nonlinear axes corresponding to largest patient variation, and clusters that group patients by overall similarities.

B. Distinguishing Randomly versus Non-Randomly Missing Variables

A key insight in MURAL is that healthcare data consists of variables that are intentionally missing, i.e., missing not at random (MNAR), that are a source of significant information since the data is related to unobserved patient characteristics (e.g. there appears to be higher levels of missing values for reported income in individuals with higher income levels). However, patient data also often contains data that is missing completely at random (MCAR) or missing at random (MAR). MCAR data is when missingness does not depend on the observed or missing values, and MAR data is when missingness does not depend on the missing values but may depend on the observed values. MCAR or MAR data are usually the result of absence of documentation through the extraction, transformation and loading of clinical data [25], [26]. We note that MCAR or MAR variables can be imputed on the basis of informational redundancy with other variables using conditional probability modeling, regression or other techniques, with multiple imputation leading to the most unbiased results [27], [28]. However, MNAR variables cannot be imputed well as we show in Section IV.

MURAL distinguishes between MCAR variables and MNAR features for continuous variables. We may or may not have prior knowledge about which continuous variables are MCAR vs. MNAR. If we do not, then we can distinguish between the two cases based on Little’s test [29], which examines patterns of missingness for correlations with other variables. Little’s test gives each variable a significance value for rejecting the null hypothesis that says it is missing at random. If this significance value $p \leq 0.05$ then we conclude that the variable is MNAR, otherwise we deem the variable to be MCAR.

Next, for variables that are MNAR we use imputation to fill the values in as a preprocessing step using fully conditional specification (FCS) multi-variable imputation [28]. After this step, all variables are either fully imputed or MNAR.

C. Mixed-modality Variable Splitting Scheme

MURAL incorporates multimodal variables into an unsupervised random forest framework by using a nuanced splitting scheme. Key aspects of the MURAL splitting scheme are as follows:

- At each iteration MURAL chooses a variable $v_i$ to split on at random.
- If the $v_i$ is MNAR for some observations then we create a preliminary binary split between observations where it is measured and observations where it is missing.
- For the branch where the variable is measured, we find a single threshold based on the unsupervised information gain criterion we define in the following section to create two child nodes.
- For the branch where the variable is missing, we randomly select another variable $v_j$ with no missingness and create two child nodes based using the residual multidimensional entropy described in Equation 7.
- If $v_i$ has no missingness then it is split into two child nodes again using the same information gain criterion.
- The two-level splits described above are flattened into the same level to create four child nodes if the first split was missing/not-missing (see Figure 1).

This scheme effectively creates four child nodes when a variable has MNAR since there are actually two variables worth of information, first variable describing the missingness, and second variable describing a further split. However the reason for choosing an additional variable on the branch with missingness is to create a branch with controlled heterogeneity (entropy) instead of the forced split that comes with a naturally binary variable.

A similar four-way splitting scheme can avoid fragmentation caused by the presence of binary variables. For a binary variable $v_b$, we create two hidden nodes, one for the value of 0 and one for the value of 1, and choose a second, non-binary, variable $v_c$ for the hidden nodes to optimize binary splits on. The edges are weighted as in the case of missingness, resulting in a four-way split.

D. Splitting with Residual Multidimensional Entropy

When each variable without MNAR is chosen, we choose a threshold $\text{thres}(v_i)$ based on an unsupervised information gain criterion, which we term residual multidimensional entropy. Instead of choosing a split that maximizes the information gain in the class label, we choose a split that maximizes the residual entropy of the remaining variables. Since the remaining variables are not naturally discrete like class labels, we discretize the continuous variables. Thus for each node $i$, splitting on variable $v_i$ we compute the following summation:
fraction of the information available, so we prefer the sum of marginal entropies. As noted before, we cannot directly use within-split affinity or density estimates as that is indeed the end result of MURAL.

E. MURAL-derived Distances and Embeddings

For every tree $T_k$ in a trained forest, we define the tree distance between two nodes to be

$$d_k(t_i, t_j) = \sum_{p \in \mathcal{P}} w_{tp}$$

where $\mathcal{P}$ is the index sequence of the edges in $T_k$ on the unique undirected path $(t_1, \ldots, t_j)$ between $t_i$ and $t_j$ with no node repetition, and $w_{tp}$ is the weight of edge $p$. We opt for all edges having unit weights.

Then we define distances between $x_i, x_j$ by noting that nodes corresponding to leaves of the tree $t_{1,1}, t_{1,2}, \ldots$ each contain sets of datapoints, i.e., $t_{1,1} = \{x_{1,1}, x_{1,2} \ldots x_{1,m}\}$. Thus if $x_i \in t_{1,1}$ and $x_j \in t_{1,1}$ then $d_k(x_i, x_j) = d_k(t_{1,1}, t_{1,1})$

The constructed tree metrics can be averaged over the MURAL-forest, which results in a composite MURAL-distances.

$$D_M = \frac{1}{k} \sum_k D_k.$$  \hspace{1cm} (9)

Here $D_k$ is the matrix of pairwise distances of the $k$th tree, and $D_M$ the the averaged distance matrix over all trees.

This leads directly to a MURAL-based distance matrix, which can be converted into an affinity matrix using any kernel function. For embedding, the distance matrix can be passed as an input to a nonlinear dimensionality reduction algorithm such as PHATE [7], which we choose for its manifold-approximation capabilities. We call the resultant PHATE embedding using $D_M$ the MURAL-embedding.

IV. RESULTS

A. Empirical Validation

In this section we induced missingness, i.e., MNAR, in datasets (or data subsets) where there was no missingness to validate the ability of MURAL to recover meaningful distances. We used 3 datasets for this experiment:

1) A synthetic Swiss roll constructed of 3,000 points embedded in a 5 dimensional space.

2) A dataset of patients requiring intensive care unit stay (ICU) from the publicly available Phillips eICU Collaborative Research Database (eICU-CRD) of critical care units across the United States from 2014 to 2015 [31], [32]. Patients with available data within the first 24 hours of ICU stay (148,532 unique patient encounters) had 10 variables selected: five laboratory values (bilirubin, blood urea nitrogen, creatinine, hematocrit, and albumin), age, 3 ordinal variables from nursing assessment of eye, motor, and verbal responses using the Glasgow Coma Scale, and binary variable of invasive mechanical ventilation within the first 24 hours.
3) An international registry of consecutive, unselected patients presenting with symptoms of upper gastrointestinal bleeding between March 2014 and March 2015 from Yale–New Haven Hospital (United States), Glasgow Royal Infirmary (Scotland), Royal Cornwall Hospital Truro (England), Odense University Hospital (Denmark), Singapore General Hospital (Singapore), and Dunedin Hospital (New Zealand). 7 variables were selected: 4 laboratory variables (hemoglobin, urea, albumin, INR), 1 vital sign (systolic blood pressure), 1 binary demographic variable (gender), and 1 ordinal variable (degree of liver disease).

The eICU and UGIB datasets already have missingness in a significant portion of the entries. Thus, to create an artificial ground truth, we used only entries with all variables present. For the UGIB dataset, 2,761 patients with complete data were selected. For computational efficiency we subsampled the eICU dataset to 10,000 patient encounters with complete data across 10 variables. Using this subset of complete data as ground truth, we artificially induced missingness in order to test the ability of MURAL to recover meaningful distances.

We induced missingness in the Swiss roll dataset in a similar way to what is observed in real clinical data, with one variable that has a pattern of missing values deemed to be missing not at random by pairwise Little’s test [29], and two variables with random values dropped out. In the eICU dataset we induced missingness by dropping the values of bilirubin > 3, the threshold chosen since these are clearly physiologically abnormal values and thus would be missing not at random. In the UGIB dataset we induced missingness by dropping the values of INR > 3, also chosen since they are physiologically abnormal values and would be missing not at random. All datasets were standardized after missingness was induced but before they were used for constructing MURAL-forests.

To quantitatively compare the preservation of the underlying manifold structure of these datasets in the presence of missing values, we used the accuracy of a kNN graph derived from the embedding compared to the ground truth kNN graph from each of these graphs. We compared the performance to mean imputation, and to another standard tree-based imputation method, multiple imputation with chained equations (MICE) using classification and regression decision trees. We find that near neighbors are recovered in the MURAL-embedding with greater accuracy than baselines that first use imputation in this case where ground truth is known in IV-A. In the artificial Swiss roll dataset, the MURAL-forest with 100 trees outperformed mean imputation at 5, 10, and 100 neighbors by 2%, 3%, and 22%, and outperformed MICE with CART at 5 and 10 neighbors by 2% and 3%. This reflects the specific nature of the Swiss roll dataset, since smaller neighborhoods were more likely to be perturbed due to the intrinsic coiled data structure manifold. For the eICU dataset and the UGIB dataset, the MURAL-forest with 100 trees outperformed mean imputation at 5, 10, and 100 neighbors by 7%, 10%, and 15%; MURAL-forest outperformed MICE with CART at 5, 10, and 100 neighbors by 15%, 15%, and 14%.

### TABLE I

| Dataset       | P@5       | P@10      | P@100     |
|---------------|-----------|-----------|-----------|
| Swiss Roll    | 0.729 ± 0.01 | 0.752 ± 0.01 | 0.743 ± 0.01 |
| Mean Imputation | 0.403 ± 0.00 | 0.429 ± 0.00 | 0.522 ± 0.00 |
| MICE with CART | 0.664 ± 0.01 | 0.729 ± 0.01 | 0.850 ± 0.01 |
| UGIB dataset  | 0.227 ± 0.01 | 0.259 ± 0.02 | 0.387 ± 0.06 |
| Mean Imputation | 0.158 ± 0.00 | 0.162 ± 0.00 | 0.231 ± 0.00 |
| MICE with CART | 0.119 ± 0.01 | 0.135 ± 0.01 | 0.215 ± 0.00 |
| UGIB dataset  | 0.239 ± 0.02 | 0.230 ± 0.03 | 0.307 ± 0.03 |
| Mean Imputation | 0.080 ± 0.00 | 0.080 ± 0.00 | 0.150 ± 0.00 |
| MICE with CART | 0.085 ± 0.01 | 0.084 ± 0.01 | 0.165 ± 0.01 |

Fig. 2. PHATE plots of the Swiss roll embedded in 5 dimensions. (I-1) Mean imputation. (I-2) MICE. (M-1) MURAL-embedding without any restrictions on choosing splits in variables with missing values. (M-2) MURAL-embedding with trees of depth 4. (M-3) MURAL-embedding with each node choosing the best split from among four variables. (M-4) MURAL-embedding with 100 trees of depth 10, each node choosing the best split from only one variable, not splitting on variables with missing values in the first three levels.

### B. Ablation Study

Next, we carried out an ablation study using this Swiss roll dataset to investigate which parameter choices lead to the best embeddings. Results are shown in Table II. For each parameter choice, we trained MURAL-forests with 5 different random initializations. In the run testing 100 trees of depth 10, we rejected one outlier forest that generated pathological distances. Generally we find that low-dimensional (3-d) entropy works for our proposed residual multidimensional entropy from Equation 7. Surprisingly we found that splitting on a single residual variables works best for the discrete and MNAR case. An additional insight was that restricting MNAR variables to low levels in the tree worked best as they would have minimal effect on other distances in this scheme. Full results are seen in Table II.

### C. Applications

1) **Embeddings:** Our initial goal was to derive distances that provide a faithful representation of a dataset that contains
We note that PHATE contains similar ability to preserve data manifold-affinities as quantified by the DeMAP metric in [7].

| Model | DeMAP | P@5 | P@10 | P@100 | P@500 | Distortion | Time (s) |
|-------|-------|-----|------|-------|-------|------------|----------|
| Mean Imputation | 0.495 | 0.637 | 0.635 | 0.581 | 0.675 | 5530 | — |

**Entropy Dimensions**

| ID | 0.504 ± 0.08 | 0.720 ± 0.01 | 0.742 ± 0.01 | 0.739 ± 0.01 | **0.783 ± 0.01** | 1880 ± 270 | 376 ± 240 |
| 2D | 0.618 ± 0.04 | 0.728 ± 0.01 | 0.748 ± 0.01 | 0.746 ± 0.01 | **0.782 ± 0.00** | 1800 ± 180 | **150 ± 5.4** |
| 3D | 0.577 ± 0.08 | 0.733 ± 0.01 | 0.755 ± 0.01 | **0.749 ± 0.01** | 0.779 ± 0.01 | 1860 ± 250 | 240 ± 22 |
| 5D | **0.628 ± 0.03** | 0.704 ± 0.01 | 0.725 ± 0.00 | 0.727 ± 0.02 | 0.763 ± 0.01 | 1.23 ± 2.0 × 10^7 | 1150 ± 260 |
| RME | 0.556 ± 0.06 | **0.738 ± 0.01** | **0.757 ± 0.01** | 0.738 ± 0.01 | 0.773 ± 0.01 | **1540 ± 120** | 448 ± 41 |

**SHOWS SUPERIOR SILHOUETTE SCORES COMPARED TO OTHER IMPUTATION METHODS**

**TABLE II**

**Ablation Study for MURAL-Embedding on the Swiss Roll Embedded in 5 Dimensions.** $(\mu \pm \sigma)$ over 5 runs. **Bold** represents best in each parameter.

**TABLE III**

MURAL-EMBEDDING ON PHATE WITH SPECTRAL CLUSTERS k=4 SHOWS SUPERIOR SILHOUETTE SCORES COMPARED TO OTHER IMPUTATION METHODS

| Experiment | Approach | Silhouette Score |
|------------|----------|------------------|
| UGIB dataset | MURAL | **0.46** |
| Mean Imputation | 0.043 |
| MICE with CART | 0.44 |
| eICU dataset | MURAL | **0.390** |
| Mean Imputation | 0.386 |
| MICE with CART | 0.378 |

By visual inspection we see that the MURAL-embeddings are much more structured than PHATE embeddings of the raw data with imputed values. Furthermore, visualization of several clinical variables on the embeddings show that the separations correspond to clinical groupings that are used in generating the embedding and also meaningful clinical groups not used to generate the MURAL-embeddings. In Figure 3 risk for hospital-based intervention (A) and age (B) were not used to generate the MURAL-embeddings, yet the MURAL-embeddings show separation into high risk and low risk groupings (A) and separation by age, in this case the MURAL-embeddings show separation into high risk and (B) were not used to generate the MURAL-embeddings, yet Figure 3 risk for hospital-based intervention (A) and age groups not used to generate the MURAL-embeddings. In generating the embedding and also meaningful clinical separations correspond to clinical groupings that are used several clinical variables on the embeddings show that the raw data with imputed values. Furthermore, visualization of which we term as MURAL-embeddings are shown in Figure 2, Figure 3 and Figure 4.

By visual inspection we see that the MURAL-embeddings are much more structured than PHATE embeddings of the raw data with imputed values. Furthermore, visualization of several clinical variables on the embeddings show that the separations correspond to clinical groupings that are used in generating the embedding and also meaningful clinical groups not used to generate the MURAL-embeddings. In Figure 3 risk for hospital-based intervention (A) and age (B) were not used to generate the MURAL-embeddings, yet the MURAL-embeddings show separation into high risk and low risk groupings (A) and separation by age, in this case < 30 years old versus ≥ 80 years old (B). For factors used to generate the MURAL embeddings, there are separations into two major structures horizontally based on gender (C) and minor structures based on liver disease status (D). By contrast, mean imputation mixes genders despite having the information, and MICE imputation overlaps the two disease cohorts. In Figure 4 risk for in-hospital mortality (B) and admission diagnosis of acute coronary syndrome (C) were not used to generate the MURAL-embeddings, but the MURAL-
Fig. 3. MURAL-embedding preserves data structure and separation better than mean imputation and multiple imputation with chained equations using classification and regression trees (MICE with CART) after visualization with PHATE. MURAL-embedding separates groups of patients with clinically relevant subgroups: A) high risk (red) and low risk (blue) groups as defined by need for hospital-based intervention B) different age groups <30 years old (blue) versus >80 years old (red) C) gender, male (red) versus female (blue) D) liver disease (yellow to red) versus no liver disease (blue). Spectral clustering (E) of the graph subsets the known groups including the group of men with liver disease (green), female patients (red), young males at low risk for hospital-based intervention (purple), and older males at high risk for hospital-based intervention (blue).

### TABLE IV

| Experiment | Approach         | Risk    | <30 y/o v >80 y/o | <30 y/o v 30-40 y/o |
|------------|------------------|---------|------------------|---------------------|
| UGIB dataset | MURAL-forest with TSWD | **12.0 ± 0.33** | 12.6 ± 0.56 | 5.2 ± 0.22 |
|            | EMD on Mean Imputation | 2.39   | 2.26 | 1.27 |
|            | EMD on MICE with CART | **1.13 ± 0.001** | 1.12 ± 0.001 | 0.94 ± 0.0001 |
| eICU dataset | MURAL-forest with TSWD | **10.6 ± 1.76** | 28.5 ± 5.8 | 1.96 ± 0.44 |
|            | EMD on Mean Imputation | 1.78   | 3.78 | 1.82 |
|            | EMD on MICE with CART | **1.73 ± 0.01** | 3.60 ± 0.01 | 1.68 ± 0.08 |

embeddings show separation into low and high risk groups for in-hospital mortality (B) and subgroup of patients with admission diagnosis of acute coronary syndrome (C) (red) and those with other diagnoses (blue). Need for mechanical ventilation (A) and nursing assessment of verbal response (D) were used to generate the MURAL-embedding; clear separations are found between patients with mechanical ventilation (A, red) and those who did not (A, blue), as well as patients with normal verbal response (D, light yellow), mildly impaired verbal response (D, orange), those with no verbal response (D, purple). These separations are not seen in the embedding methods on imputed data despite also containing these variables. These visual results suggest that these embeddings are useful and amenable to further quantitative analysis. The structures in the MURAL-embedding are largely retained even after including 108 variables derived from the first 24 hours of patient stay in the ICU for the same patient cohort in the eICU dataset (Figure 5).

2) Spectral Clustering: Spectral clustering [8] is a key unsupervised clustering method that follows the data manifold by operating on a data affinity matrix. We used k-means with $k = 4$ on the diffusion operator created by PHATE, which is equivalent to spectral clustering and compared it to similar clustering using imputed data. To evaluate cluster quality we
Fig. 4. MURAL-embedding better separates out patients in the intensive care unit at risk for in-hospital mortality compared to mean imputation and MICE with CART on PHATE visualizations. MURAL-embedding separates patients who A) received invasive mechanical ventilation (red) versus not (blue) B) at high risk for in hospital mortality (red) versus not (light blue) C) presenting with admission diagnosis of acute coronary syndrome (red) versus other diagnosis (blue) D) nursing assessment (1 to 5, higher is healthier). Spectral clustering (E) of the graph subsets the groups of patients who required mechanical ventilation (blue), patients with slightly impaired verbal responses on nursing assessment (green), and patients with admission diagnosis of acute coronary syndrome with bilirubin measurement, reflecting concern about liver disease (purple) versus without bilirubin measurement (red), impairment of verbal responses (blue).

| A | B | C | D | E |
|---|---|---|---|---|
| **MURAL-embedding** | | | | |
| Mean Imputation | | | | |
| MICE with CART Imputation | | | | |

use the silhouette score, and found that the clustered MURAL-embedding has highest silhouette score in both datasets.

In addition, the resultant subgroups can be interpreted clinically. For example, the red cluster in Figure 4 corresponds to admission diagnosis of acute coronary syndrome who either have a bilirubin measured or not measured within 24 hours of ICU admission, which may suggest a concern for concurrent hepatobiliary dysfunction or injury. In Figure 4, the green cluster correspond to male patients with severe liver disease, which correspond to a specific type of gastrointestinal bleeding, portal hypertensive bleeding.

3) Wasserstein Distance between cohorts: Seeing that groups of patients who differ in a clinically significant way form distinct clusters on the embeddings, we decided to quantitatively check whether MURAL-forests themselves separate dissimilar cohorts more than similar ones. This is directly relevant to the task of characterizing clusters within the data representations. For example, if clusters are two groups of patients with the same diagnosis, if there is a meaningful difference between their measured laboratory or clinical characteristics that could reflect risk for a poor clinical outcome. To that end, we calculated tree-sliced Wasserstein distances between the low risk and high risk patient cohorts. For the eICU dataset, we defined this as patients who died in the hospital, and in the UGIB dataset we defined this as a need for hospital-based intervention (red blood cell transfusion, hemostatic intervention, or 30-day mortality). We then extract feature importances by aggregating the variables associated with nodes where the Wasserstein distances were most disparate. The feature importances in Figure 6 are consistent with the top two variables predictive of in-hospital mortality from supervised machine learning algorithms (regression models) trained and validated on the eICU dataset [34], and the top three variables predictive of need for hospital based intervention from a high performing supervised machine learning algorithm (gradient boosted decision trees) trained and validated on the UGIB dataset [35]. As a sanity check, we also compared tree-sliced Wasserstein distances between different age groups: first, a very different age group (< 30 years old versus > 80 years old) and a similar age group (< 30 years old versus 30-40 years old). The similar age group had much lower TSWD compared to the group with different ages, and the TSWD from the MURAL-forest for high risk versus low risk was more distinct compared to the other imputation methods. More generally, we believe these types of Wasserstein distances can be used to measure distances or similarities between treatments, diagnostic variants and other differences between sub-cohorts.
Fig. 5. MURAL-embedding better separates out patients in the intensive care unit at risk for in-hospital mortality compared to mean imputation and MICE with CART on PHATE visualizations even with 108 variables. MURAL-embedding separates patients who A) received invasive mechanical ventilation (red) versus not (blue) B) at high risk for in hospital mortality (red) versus not (light blue) C) presenting with admission diagnosis of acute coronary syndrome (red) versus other diagnosis (blue) D) nursing assessment (1 to 5, higher is healthier). Spectral clustering (E) of the graph subsets the group of patients who required mechanical ventilation (blue/green).

Fig. 6. TSWD on the MURAL-forest can be used to generate feature importance graphs that are consistent with feature importances in supervised approaches on the same data. For the eICU dataset, the first two factors were within the top 3 predictive factors in supervised models predicting mortality on the same dataset. For the UGIB dataset, the first 3 factors were identified as the top 3 predictive factors in high-performing supervised models on the same dataset.

V. Conclusion

We present MURAL, a random-forest based framework for deriving distances between patients using mixed-model electronic health record data. We showed that the resultant MURAL-embeddings recapitulate the structure and heterogeneity of patient populations better than alternatives—thus paving the way for unsupervised learning to be used on clinical data. We note that most of the machine learning methods that are currently used for modeling clinical data require supervised training and large sets of annotated and labeled samples. However, by making clinical data amenable to unsupervised approaches, we can diminish this burden and even discover novel, clinical groupings of patients that could be meaningful for diagnosis, prognosis, or treatment.
