Supervised Manifold Learning via Random Forest Geometry-Preserving Proximities*

Jake S. Rhodes
Department of Statistics
Brigham Young University
Provo, Utah, USA
rhodes@stat.byu.edu

***Abstract*—Manifold learning approaches seek the intrinsic, low-dimensional data structure within a high-dimensional space. Mainstream manifold learning algorithms, such as Isomap, UMAP, t-SNE, Diffusion Map, and Laplacian Eigenmaps do not use data labels and are thus considered unsupervised. Existing supervised extensions of these methods are limited to classification problems and fall short of uncovering meaningful embeddings due to their construction using order non-preserving, class-conditional distances. In this paper, we show the weaknesses of class-conditional manifold learning quantitatively and visually and propose an alternate choice of kernel for supervised dimensionality reduction using a data-geometry-preserving variant of random forest proximities as an initialization for manifold learning methods. We show that local structure preservation using these proximities is near universal across manifold learning approaches and global structure is properly maintained using diffusion-based algorithms.**

***Index Terms*—supervised learning, manifold learning, random forest, data visualization, data geometry**

I. Introduction

Manifold learning algorithms are often used for exploratory data analysis. They are typically applied to noisy data in an attempt to find meaningful patterns or relationships across time, classes, or variables [1]. Most manifold learning approaches are unsupervised in that they do not use auxiliary information (e.g., data labels) in the embedding construction process. In many contexts, only unsupervised models are applicable as auxiliary information can be expensive or inaccessible. However, when available, label information can provide valuable insights into the data's intrinsic structure relative to the labels. Subjecting the embedding process to the use of auxiliary information can help to uncover a data geometry unattainable without labels. In this paper, we discuss weaknesses of current supervised manifold-learning approaches and show improvements on existing methods by applying a new variant of a random forest-based [2] similarity measure in a manifold-learning setting. We use Geometry- and Accuracy-Preserving proximities (RF-GAP [3]) and demonstrate their ability to meaningfully encode a similarity measure and subsequent embedding that naturally incorporates labels. As opposed to distance or similarity measures which condition upon class labels to artificially exaggerate the separation of points of opposing classes, random forest proximities serve as a measure of similarity that uses labels (continuous, categorical, or otherwise) in a manner consistent with the model's learning. Additionally, forest-based proximities appropriately denoise the data, providing a meaningful metric or graph for the embedding process.

II. Supervised Manifold Learning

Manifold-learning algorithms use distance or similarity graphs to encode local data structure. For example, Isomap forms a k-nearest neighbor (k-NN) graph using Euclidean distance and seeks the shortest path between observations to approximate true geodesic distances upon which multidimensional scaling is applied [4]. Diffusion Map (DM) uses a Gaussian kernel applied to a k-NN graph to form local similarities upon which eigendecomposition is applied [5]. T-distributed Stochastic Neighbor Embedding, or t-SNE, estimates probabilities as a normalized Gaussian kernel to define similarities between points. The Kullback-Leibler (KL) divergence between these and a lower-dimensional mapping is estimated via gradient descent to form the target embedding [6].

Each of these methods is unsupervised; data labels are not used in any part of the embedding process. However, supervised variants of these methods have been developed. Most of these supervised extensions of the algorithms adapt the existing algorithm at the distance- or similarity-learning level. In some cases, distances are rescaled [7], additively incremented [8], or otherwise adapted conditionally upon class association [9]. Often, these dissimilarity measures can provide perfect linear separation where such discrimination is not possible using traditional classifiers. See Equation 1, for an example of a class-conditional dissimilarity, where $D(.,.)$ denotes a distance function (e.g., Euclidean), β is usually set as the average distance between points, α lessens separation between similar points of opposing classes, and y_i, y_j are the respective labels of x_i and x_j. This dissimilarity has been used to create supervised variants of t-SNE [10], Isomap [9], Locally-Linear Embedding [11], and Laplacian Eigenmaps [12]. In each of these extensions, the within-class structure is partially maintained, but manifold structures are distorted at a global level as a result of exaggerated class separation. Such dissimilarity measures are order non-preserving bijections [13].

$$D'\left(\mathbf{x}_i, \mathbf{x}_j\right) = \begin{cases} \sqrt{1 - e^{\frac{-D^2(\mathbf{x}_i, \mathbf{x}_j)}{\beta}}} & y_i = y_j \\ \sqrt{e^{\frac{D^2(\mathbf{x}_i, \mathbf{x}_j)}{\beta}}} - \alpha & y_i \neq y_j \end{cases} \tag{1}$$

These approaches are problematic in several ways: (1) The class-conditional distances form an attempt to maintain with-class structure but cause disruption between classes. Artificial class separation diminishes inter-class relationships, thus distorting the global data structure. (2) The manifold disruption reduces the integrity of resulting downstream tasks. For example, classification tasks following dimensionality reduction can have unrealistically low error rates. (3) Class-conditional measures do not provide an avenue for continuously-valued labels. These extensions have not been adapted to regression problems. (4) These approaches are not extendable to new, unlabeled points (e.g., a test set used for subsequent predictions). To overcome each of these weaknesses, we propose the use of random forests [2], [3] to generate supervised similarities to be used in manifold learning.

III. Random Forest Proximities

Random forests [2] provide a number of benefits for prediction problems that are supportive of metric learning. For example, random forests apply to both regression and classification problems, handle mixed variable types, provide an unbiased estimate of the generalization error, are insensitive to monotonic transformations, are relatively robust to outliers, and provide a natural way of assessing variable importance, ignoring noise variables in the presence of meaningful ones.

Random forests form an ensemble in binary-recursive decision trees each of which partitions a bootstrap sample of the training data. Observations within the bootstrap sample are called in-bag, while those in the training data not included in the sample are called out-of-bag, or OOB. Each partition forms a decision space used for classification or regression. The partitions naturally form a channel for generating similarity measures using data labels. These similarities are referred to in the literature as random forest proximities.

Leo Breiman originally defined the proximity between two observations, x_i, and x_j, denoted by $p(x_i, x_j)$, as the number of terminal nodes they share across all trees, divided by the number of trees in the forest [17]. This simple approach applies to all training points regardless of bootstrap status. As a result, proximities constructed on the training set tend to slightly overemphasize class segregation. To overcome this, an alternative formulation was derived to only calculates pair-wise similarities between points x_i and x_j using trees in which both of these observations are OOB [18]. Subsequently, these proximities combat overinflated class separation. However, it has been shown that OOB-only proximities do not fully benefit from the random forest's learning and are a noisier similarity measure [3].

In [19], the authors demonstrated that random forests behave like a nearest-neighbor regressor with an adaptive bandwidth. That is, random forest predictions (in the regression context) can be determined as a weighted sum using a kernel function, as shown in Equation 2.

$$\hat{y}_i(k) = \sum_{j \neq i} k(x_i, x_j) y_j \tag{2}$$

Here, k is a weighted kernel function determined by the number of training examples sharing a terminal node with x_i. This is comparable to other kernel methods, such as the SVM, which uses a kernel to define similarity and ultimately the decision space. Ideally, random forest proximities should serve as a kernel capable of mimicking random forest predictions. Using normalized proximities as weights can serve as a test for the proximities' consistency with the forest's learning. In this regard, existing random forest proximity formulations do not adequately incorporate the forest's learning [3].

Both the original formulation [17], as well as that using only OOB observations [18], are not capable of reconstructing the random forest predictions for two reasons: (1) Random forests train on a set of bootstrap (in-bag) samples and predict on another set (the OOB samples or a test set). The original formulation doesn't discriminate between in-bag or OOB observations, and the OOB proximity definition does not use in-bag samples in their construction. (2) Decision tree voting takes into account the number of in-bag observations within a given terminal node, while the proximities are constructed without regard to the number of "voting points" [19]. To construct proximities that serve as a kernel for random forest prediction, these two points must be accounted for.

In [3], the authors propose a new proximity formulation capable of reconstructing random forest OOB and test predictions as a kernel method. They call these proximities Random Forest-Geometry- and Accuracy-Preserving proximities (RF-GAP) and show improvement across multiple applications using this new definition, including data imputation, outlier detection, and visualization via MDS.

The RF-GAP proximities, however, do not form a proper kernel function as originally defined. They are asymmetric and self-similarity is defined to be 0 to account for the kernel prediction problem. To overcome this, we normalize the similarities to set the maximum similarity to 1, symmetrize the proximities, and define the diagonal entries to be 1. In doing so, the proximities can serve in any capacity which requires a kernel matrix. Using this modified RF-GAP formulation, these proximities serve as a similarity measure that overcomes the weaknesses of the class-conditional supervised distances in the following ways: (1) Rather than conditionally adapting an existing distance measure and thereby distorting the global data structure, random forest proximities provide a measure of local similarity which partially retains global information through the forest's recursive splitting process. Therefore, instead of exaggerating class separation, natural observational relationships are retained, as can be seen in low-dimensional visual representations (see Figure 1). (2) Proximities formed using OOB data points retain the random forest's learning, thus, downstream task integrity is not jeopardized but relevant

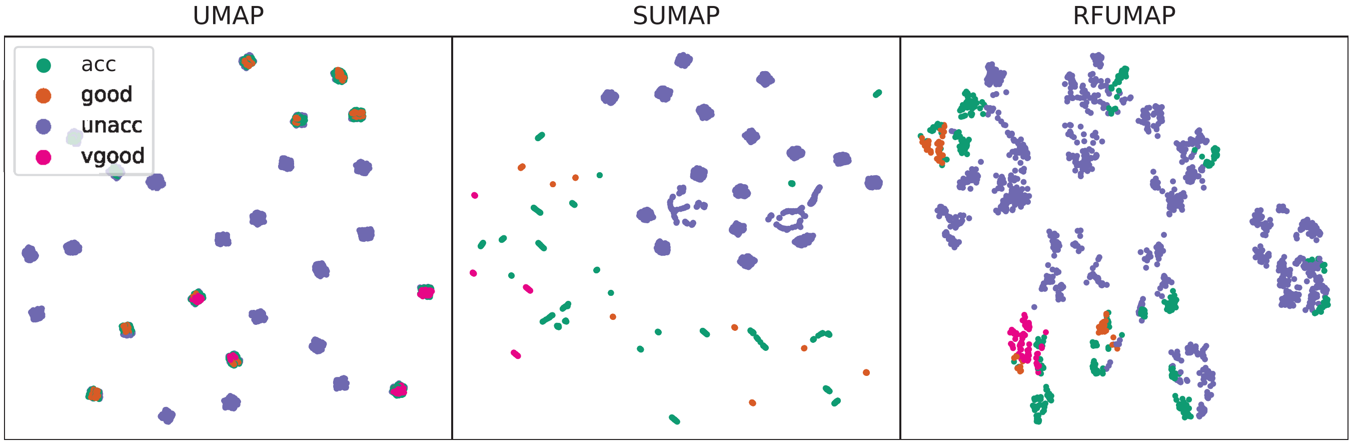

Fig. 1. This figure provides UMAP-based 2-dimensional representations of the cars dataset [14]. The left figure is the original UMAP [15] implementation which shows distinct clusters of overlapping classes. A supervised version of UMAP [16] (center) provides perfect class separation as well as distinctive clusters, which does not coincide with a cross-validated k-NN accuracy of 0.899. The RF-GAP-based UMAP implementation (right) shows observational relationships which correspond to variable interactions within the dataset.

supervised information is retained. (3) Random forests are not limited to classification problems but also work with continuous labels. This provides an avenue for supervised metric learning in a regression context, as shown in Figure 2. (4) A trained random forest model can extend similarity measures to unlabeled or out-of-sample observations, providing a means for semi-supervised metric learning or subsequent prediction. Additionally, noise variables are not likely to be used for splitting unless relevant variables are not included in the random subset of splitting variables. Subsequently, generated proximities naturally account for variable importance and can serve as a means of denoising.

We show that supervised manifold learning methods generally improve with the use of RF-GAP proximities. We compare the embedding mapping using common manifold learning algorithms including Isomap, t-SNE, Diffusion Map, Laplacian Eigenmaps, UMAP, PHATE, MDS, and Kernelized-PCA. Visualizations for each of these can be found in Appendix A. Additionally, we demonstrate that diffusion-based embeddings generated using RF-GAP proximities better retain the random forest's learning in low dimensions, as shown in Figure A.1.

IV. Random Forest-Based Manifold Learning

Local connectivity is encoded via a distance metric. A kernel function, (e.g., Gaussian kernel) can be applied to the graph distances to provide a local measure of similarity between observations from which global relationships can be learned. For example, in diffusion processes, a stochastic matrix, or diffusion operator, is formed by row-normalizing the pair-wise similarities. The global structure is learned by powering the diffusion operator, simulating a random walk between observations.

The quality of learned embedding is highly dependent on the kernel construction as well as global-structure mapping. Unlike an unsupervised kernel function, random forest proximities form noise-resilient, locally-adaptive neighborhoods ensuring that subsequent embeddings are constructed in a manner relevant to and consistent with the data labels. Similarities between points of different classes are still reflected in the proximity values, whereas this inter-class preservation is lost or diminished in class conditional measures such as the one given by Equation 1. Continuous labels can also be reflected in the embedding using random forest proximities. We provide an example in Figure 2. In this figure, embeddings are colored both by the life expectancy (the target label) as well as the country's economic status (developed vs. developing). It is clear that the unsupervised embeddings create separate clusters for each economic status, while the RF-PHATE embedding shows a continuum consistent with life expectancy in lower dimensions.

V. Quantifying Results

We use two methods to evaluate the low-dimensional embeddings in the supervised context. The first approach determines the extent to which the embedding can be used for the original classification problem. To this end, we use the 2-dimensional embeddings of unsupervised, supervised, and RF-GAP-based models to train a k-NN classifier to predict the original labels and compare the accuracy with that of a model trained on the full dataset. All accuracies were averaged using leave-one-out cross-validation and the overall results were aggregated across all datasets given in Table B.1. Ideally, the difference in accuracies should be minimal, that is, we want to retain useful information without overfitting. In Figure 3, we see that the unsupervised embeddings produce accuracies lower than those using the full datasets, demonstrating a loss of information in low dimensions. The class-conditioned, supervised approaches tend to overfit the labels, generally producing much higher accuracies (near-perfect, in some examples) as a result of artificial class separation. The RF-GAP-based embeddings have accuracies more consistent with models trained on the full dataset, though results vary

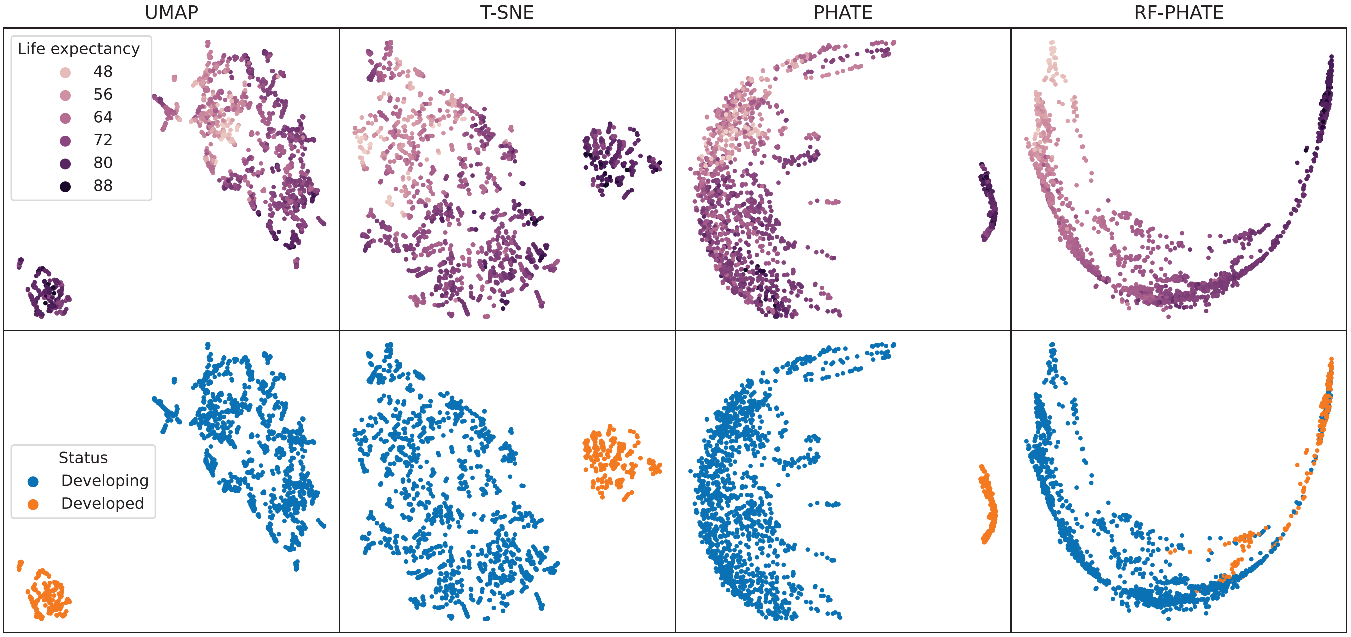

Fig. 2. This figure demonstrates the ability of RF-GAP methods to work with continuous labels. The figures provided here are two-dimensional embeddings of the Life Expectancy dataset [20] using UMAP, t-SNE, PHATE, and PHATE with the RF-GAP proximities, colored by life expectancy (top) and country economic status (bottom). The single most important variable for determining life expectancy is the country's economic status: developed or developing. All three unsupervised models completely separate these labels into two clusters, while RF-PHATE [21] produces a progression incorporating these discrete labels in a continuous manner.

by method. The slightly higher values for the majority of RF-GAP methods can be accounted for by the superior predictive power of random forests over k-NN. We show in Figure A.1 that the k-NN predictive accuracy of the RF-GAP embeddings typically aligns very well with the OOB score of the forest which generated the proximities, demonstrating the embeddings' abilities to retain the forest's learning in low dimensions.

The second evaluation technique provides an assessment of the hierarchy of variable importance retained in the embedding. In a supervised context, variables that provide higher class-discriminatory power are considered to be more important. For this evaluation, we first determine a permutation-based variable importance score using a k-NN classification model on the original supervised task. We then produce a second set of variable importance scores by regressing onto the embedding using a k-NN model trained on the original dataset. We calculate the correlation between the two variable importance scores. In Figure 4, we see that supervised models generally outperform unsupervised models in retaining variable importance, bearing in mind that class-conditional methods inflate class discrimination. Diffusion-based RF-GAP methods tend to perform best with this metric, suggesting they better preserve global, hierarchical variable importance.

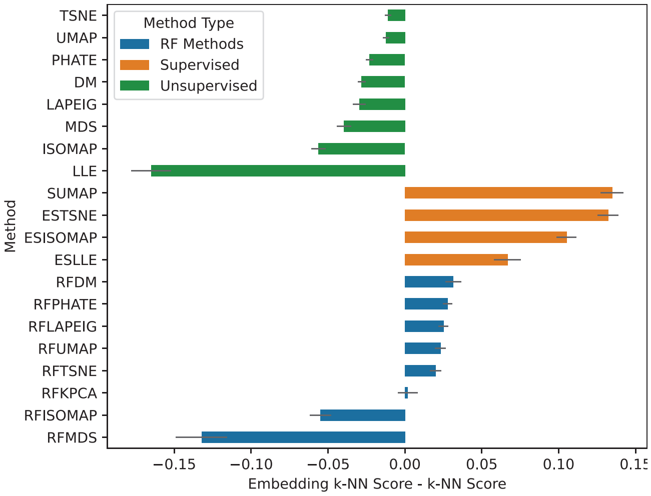

Fig. 3. A k-NN model is trained on each embedding and the accuracies are compared with those of a model trained on the full dataset. Unsupervised embeddings tend to produce lower accuracies, while supervised embedding accuracies are inflated. Generally, RF-GAP-based embeddings produce accuracies more consistent with the full model. All k-NN accuracies were assessed using leave-one-out cross-validation. Score differences are aggregated across 20 random initializations across each dataset provided in Table B.1

VI. Conclusion

In this paper, we discussed the weaknesses of existing supervised manifold learning methods. We showed that RF-GAP-based manifold learning methods preserve local structure while maintaining global structure relative to data classes as can be seen in scatterplots of the embeddings. The visual quality of the RF-GAP embedding depends on the method used, but variable importance is maintained in low dimensions regardless of the method. Diffusion-based RF-GAP embeddings

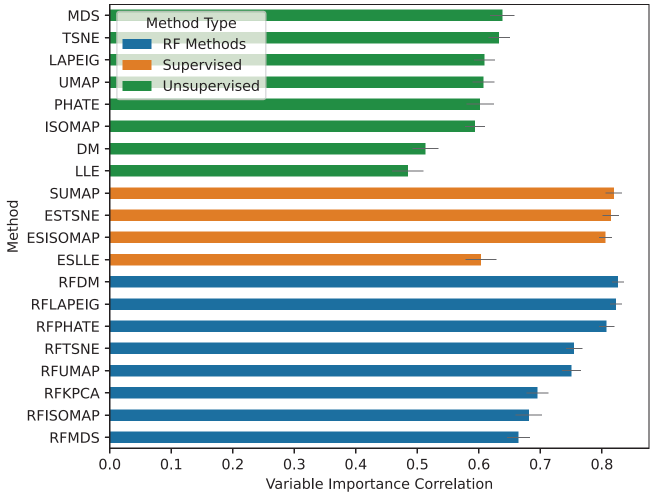

Fig. 4. The correlation between importance scores shows to what extent variable importance is retained in the embedding. All RF-GAP methods outperform each of the unsupervised methods, and most other supervised methods have higher correlations than supervised methods.

tend to retain the random forest's learning in low dimensions, suggesting that such methods better maintain the integrity of the kernel from which the embeddings were derived.

References

[1] K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. v. d. Elzen, M. J. Hirn, R. R. Coifman, N. B. Ivanova, G. Wolf, and S. Krishnaswamy, "Visualizing structure and transitions in high-dimensional biological data," *Nat. Biotechnol.*, vol. 37, no. 12, pp. 1482–1492, Dec 2019. [Online]. Available: https://doi.org/10.1038/s41587-019-0336-3

[2] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: http://doi.org/10.1023/A:1010933404324

[3] J. S. Rhodes, A. Cutler, and K. R. Moon, "Geometry- and accuracy-preserving random forest proximities," 2022. [Online]. Available: https://arxiv.org/abs/2201.12682

[4] J. B. Tenenbaum, V. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000. [Online]. Available: https://doi.org/10.1126/science.290.5500.2319

[5] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 5–30, 2006, special Issue: Diffusion Maps and Wavelets. [Online]. Available: https://doi.org/10.1016/j.acha.2006.04.006

[6] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[7] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios *et al.*, "Non-linear dimensionality reduction techniques for classification and visualization," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '02. New York, NY, USA: Association for Computing Machinery, 2002, p. 645–651. [Online]. Available: https://doi.org/10.1145/775047.775143

[8] D. de Ridder, O. Kouropteva, O. Okun, M. Pietikäinen, and R. P. W. Duin, "Supervised locally linear embedding," in *Artificial Neural Networks and Neural Information Processing — ICANN/ICONIP 2003*, O. Kaynak, E. Alpaydin, E. Oja, and L. Xu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 333–341.

[9] B. Ribeiro, A. Vieira, and J. Carvalho das Neves, "Supervised isomap with dissimilarity measures in embedding learning," in *Progress in Pattern Recognition, Image Analysis and Applications*, J. Ruiz-Shulcloper and W. G. Kropatsch, Eds. Berlin, Heidelberg: Springer, 2008, pp. 389–396. [Online]. Available: https://doi.org/10.1007/978-3-540-85920-8_48

[10] L. Hajderanj, I. Weheliye, and D. Chen, "A new supervised T-SNE with dissimilarity measure for effective data visualization and classification," in *Proceedings of the 2019 8th International Conference on Software and Information Engineering*, ser. ICSIE '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 232–236. [Online]. Available: https://doi.org/10.1145/3328833.3328853

[11] S. Zhang, "Enhanced supervised locally linear embedding," *Pattern Recognit. Lett*, vol. 30, no. 13, pp. 1208 – 1218, 2009. [Online]. Available: https://doi.org/10.1016/j.patrec.2009.05.011

[12] Q. Jiang and M. Jia, "Supervised laplacian eigenmaps for machinery fault classification," *2009 WRI World Congress on Computer Science and Information Engineering*, vol. 7, pp. 116–120, 2009. [Online]. Available: http://doi.org/10.1109/CSIE.2009.765

[13] L. Hajderanj, D. Chen, and I. Weheliye, "The impact of supervised manifold learning on structure preserving and classification error: A theoretical study," *IEEE Access*, vol. 9, pp. 43 909–43 922, 2021. [Online]. Available: 10.1109/ACCESS.2021.3066259

[14] M. Bohanec and V. Rajkovič, "V.: Knowledge acquisition and explanation for multi-attribute decision," in *Making, 8 th International Workshop "Expert Systems and Their Applications*, 1988.

[15] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv*, vol. abs/1802.03426, 2018. [Online]. Available: https://arxiv.org/abs/1802.03426

[16] L. McInnes, J. Healy, N. Saul, and L. Grossberger, "Umap: Uniform manifold approximation and projection," *The Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[17] L. Breiman and A. Cutler, "Random forests," https://www.stat.berkeley.edu/ breiman/RandomForests/cc_home.htm, accessed: 03/02/2023.

[18] T. Hastie, R. Tibshirani, and J. Friedman, "Random forests," in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY: Springer New York, 2009, pp. 587–604. [Online]. Available: https://doi.org/10.1007/978-0-387-84858-7_15

[19] Y. Lin and Y. Jeon, "Random forests and adaptive nearest neighbors," *J. Am. Stat. Assoc.*, vol. 101, no. 474, pp. 578–590, 2006. [Online]. Available: https://doi.org/10.1198/016214505000001230

[20] N. Mohamed Amine Mairech, "Datac'ept : Life expectancy prediction," 2019. [Online]. Available: https://kaggle.com/competitions/datacept-life-expectancy-prediction

[21] J. S. Rhodes, A. Cutler, G. Wolf, and K. R. Moon, "Random forest-based diffusion information geometry for supervised visualization and data exploration," *2021 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 331–335, 2021. [Online]. Available: https://doi.org/10.1109/SSP49050.2021.9513749

[22] D. Dua and C. Graff, "UCI machine learning repository," 2017, (Accessed on 03/02/2023). [Online]. Available: http://archive.ics.uci.edu/ml

[23] R. Gorman and T. J. Sejnowski, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural Netw*, vol. 1, no. 1, pp. 75–89, 1988. [Online]. Available: https://doi.org/10.1016/0893-6080(88)90023-8